







Loop Optimization Noise-Reduced LSTM Based Classifier for PD Detection

Ning Xu , Hoay Beng Gooi , *Life Senior Member, IEEE*, Lipo Wang , *Senior Member, IEEE*,
Yuanjin Zheng , *Senior Member, IEEE*, Wensong Wang , *Senior Member, IEEE*,
and Jiawei Yang , *Student Member, IEEE*

Abstract—Partial discharge (PD) is an important inducement of power failures such as insulation degradation. Effective, timely, and economical detection of the PD is the basis for maintaining power system stability. This paper proposes a loop optimization noise-reduced LSTM based classifier for PD detection of large-scale data disturbed with measurement noise. It establishes the loop optimization to greatly coordinate the data processing and data-driven based machine learning algorithm. The noise reduction, feature extraction, and the long short-term memory (LSTM) neural network are combined as a whole with internal linkages. The proposed method combines these three necessary steps to achieve a better solution for the PD detection, while other methods are in general single ones, and not dedicated designed for the solution. The combination of the different methods has combined advantages and can overcome each individual's drawback. This comprehensive solution is verified using large real trial data measured from medium voltage (MV) overhead power lines. It exhibits superior performance and great practicality for PD detection of MV overhead power lines with an accuracy of 95.4%. It outperforms other techniques in the literature based on the experimental results. This paper also summarizes the loop optimization rules for PD detection solutions based on extensive experiments and analysis. The loop optimization rules are competent for the optimization of wider engineering applications of PD detection.

Index Terms—Detection solution, feature extraction, loop optimization, neural network applications, noise reduction, partial discharges, signal processing.

I. INTRODUCTION

PARTIAL discharge (PD) is a kind of localized electrical discharge that only partially bridges the insulation between conductors [1], [2]. PD is usually the result of localized electrical

stress concentrations in an insulator or on the surface of an insulator. Typically, such discharges appear as pulses of much less than 1-microsecond duration [3]. The PD is detrimental to the insulation with cumulative effect [4]. It is known as one of the most conclusive indicators of the deterioration of the electric insulation [5]. Particularly, the PD monitoring is a significant part of insulation maintenance for the covered conductors (CC) of the medium voltage (MV) overhead power lines. The insulation of CC prevents interphase short circuits when faults occur on it. As the value of the abnormal leakage current is very small, the standard digital relay protection is unable to react with such a low current. An efficient and effective solution of PD detection is a valuable tool to detect abnormal events and anomalies in MV overhead lines and set up preventive maintenance plans before the failure of electrical components [6].

The two common PD detection and recognition principles are the PD pulse shape [7], [8] and the phase-related energy exchange [9], [10]. The development of artificial intelligence (AI) technologies provides machines with more advanced capabilities to use data and knowledge [11], [12]. The AI technologies create more possibilities for various automatic fault detection and pattern recognition tasks. As for the interpretation of the microsecond-level data, the data volume is usually less than several thousand. Such lengths are suitable for the AI algorithms to display their intelligence functions [13], [14]. While for the measurement data at the level of milliseconds, the number of measurements in one signal can be one hundred million. Such a large scale of data cannot be analyzed directly due to the limitation of computation. Moreover, the PD detection of large-scale data confronts more challenges of noise corruption and irrelevant information interference. The noise reduction and data refining are more important for such PD detection tasks [15], [16]. In the literature, various noise reduction methods, features, and AI techniques are employed to different PD detection tasks [15], [16], [17], [18], [19]. However, the procedure of noise reduction, feature extraction, and AI processing are selected and optimized independently. There is a lack of the internal coordination and linkage optimization. And there are often deficiencies in the analysis for the rules of determining scheme details, which are as important as the general methodologies for practical engineering applications.

This paper dedicates to the PD detection method that coordinates the data processing and data-driven based machine learning algorithm with internal linkages. It aims to explore

Manuscript received 2 January 2022; revised 26 March 2022 and 31 May 2022; accepted 2 October 2022. Date of publication 19 October 2022; date of current version 19 January 2023. Paper 2021-PSEC-1654.R2, presented at the 2021 IEEE 12th Energy Conversion Congress & Exposition - Asia, Singapore, May 24–27, 2021, and approved for publication in the IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS by the Power Systems Engineering Committee of the IEEE Industry Applications Society. This work was supported in part by SP Group and the National Research Foundation, Singapore, in part by the Energy Market Authority under its Energy Programme under EP Award EMA-EP010-SNJL-004, and in part by Nanyang Technological University. (*Corresponding author: Ning Xu.*)

The authors are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: xuni0001@e.ntu.edu.sg; ehbgooi@ntu.edu.sg; elpwang@ntu.edu.sg; yjzheng@ntu.edu.sg; wangws@ntu.edu.sg; jiawei003@e.ntu.edu.sg).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TIA.2022.3215642>.

Digital Object Identifier 10.1109/TIA.2022.3215642

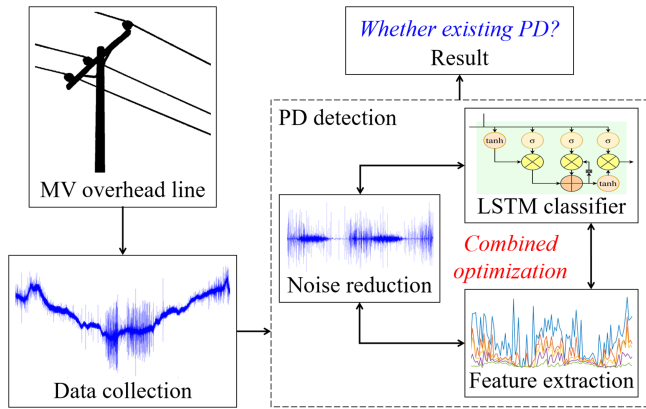


Fig. 1. The proposed method establishes the loop optimization to greatly coordinate the noise reduction, feature extraction, and the LSTM classifier as a whole. It is applied to detect PD from the large-scale high-frequency voltage signals measured from the MV overhead power lines.

an optimal solution for the automatic online PD detection of MV overhead power lines. A loop optimization noise-reduced LSTM based classifier is proposed based on large real trial data measured from the MV overhead power lines. Fig. 1 presents the overall application procedure of the proposed method from the data collection to the PD detection. The proposed method employs the loop optimization to greatly coordinate the noise reduction, feature extraction, and the long short-term memory (LSTM) neural network as a whole. The individual components in the whole are selected based on the most essential characteristics of PD. And they have great practicality with low requirements for computation and hardware. The proposed method maximizes their effectiveness through the optimal combination to achieve a better solution for the PD detection, while other methods in the literature are in general single ones, and not dedicated designed for the solution. This comprehensive solution is with advantage than the literature ones. Moreover, this paper proposes the loop optimization rules of the adaptive framework based on the systematic analysis. The loop optimization rules have great potential for the promotion of wider engineering applications.

The structure of this paper is as follows: Section II introduces the architecture and methodologies of the proposed method. Section III presents the experiments and results with analysis and discussions. The common loop optimization rules are also established. Section IV concludes the paper.

II. ARCHITECTURE AND METHODOLOGY

The overall flowchart of the proposed method is shown in Fig. 2. The voltage signal refers to the large-scale high-frequency voltage signal from the large real trial data set shared by Technical University of Ostrava (VSB). The voltage signals are measured online from 23 MV overhead power lines. Reference [20] introduces the measurement principle: An inductor is wound on the surface of the CC as a voltage sensor. The inductor has connections to the capacitor divider. The capacitive divider is loaded by a properly selected resistor. The output of this divider is connected to the DAQ oscilloscopic card. Voltage signals are measured from the sensor through its coupling capacity

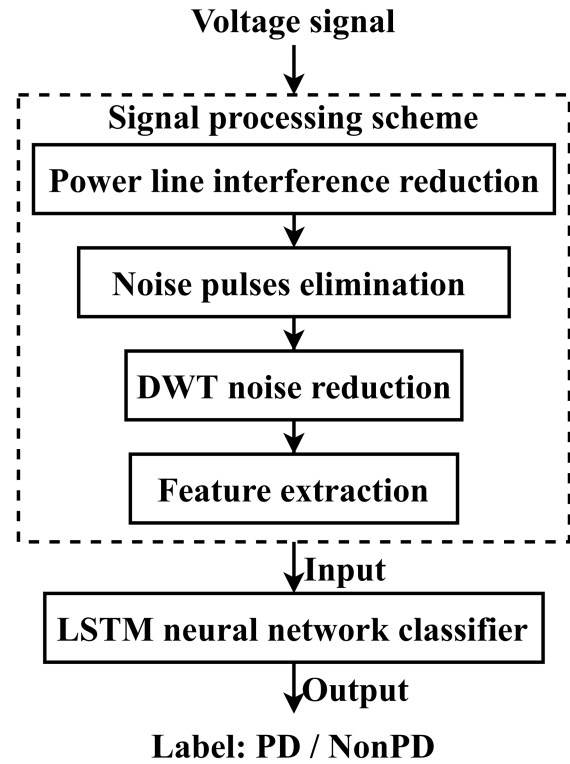


Fig. 2. The flowchart of the proposed method.

to the CC. Each signal contains 800000 measurements of the voltage over 0.02 seconds, i.e., a single complete grid cycle. Such voltage signals capture the impulse components generated by PD activities in an insulation system of the CC.

To effectively process such large-scale voltage signals, series of noise reductions are conducted to reduce the complicated background noises in the complex real environment. Multiple sequence features are employed to capture the PD characteristics that are closely related to phase-dependent variation rather than independent values. The LSTM neural network classifier classifies the input features into the PD or NonPD class. Being labeled as PD means that the voltage signal contains the impulse components of PD. Being labeled as NonPD means that there is no PD activity. The proposed method aims to achieve the effective distinction of PD components from other ambient background noises through the comprehensive contribution of the optimal noise reduction methods, feature extraction methods, and the LSTM neural network classifier. This section introduces the methodologies for each step.

A. Noise Reduction

Firstly, the digital trap filter is used to reduce the power line interference. Secondly, three kinds of noise pulses: abnormal negative pulses; symmetrical pulses; and abnormal high amplitude pulses, which are irrelevant to PD, are removed. The experts who measure the data set determine these noise pulse patterns according to the background noises of the actual measurement environment. In the proposed method, the specific noise pulses elimination method is to focus on the measurements that have abnormally large negative values, or abnormally large negative

derivatives. Local signals near such measurements are considered disturbed by noise pulses. They will be removed from the voltage signals.

In addition to the aforementioned targeted noise pulses elimination, the wavelet transform (WT) noise reduction is carried out on the voltage signal. WT is one of the most popular tools for signal noise reduction [21]. The WT decomposes a signal as a linear combination of the mother wavelet and the summation of the product of the wavelet coefficients [22], [23]. It is commonly known to have the following categories: discrete wavelet transform (DWT); discrete wavelet packet transform (DWPT); and continuous wavelet transform (CWT). Among them, the DWT is more computationally efficient. Its expression is as follows:

$$y_n = \sum_{i=1}^{N/2^j} a_j(i) \varphi_{j,i}(n) + \sum_{j=1}^J \sum_{i=1}^{N/2^j} d_j(i) \psi_{j,i}(n) \quad (1)$$

where $\varphi_{j,i}(n)$ represents the scaling function; $\psi_{j,i}(n)$ represents the wavelet function; J is the series of wavelet decomposition; N is the total number of coefficients of wavelet decomposition; $a_j(i)$ is the approximate coefficient; and $d_j(i)$ is the detailed coefficient. $a_j(i)$ and $d_j(i)$ are expressed as follows:

$$a_j(i) = \langle y(n), \varphi_{j,i}(n) \rangle \quad (2)$$

$$d_j(i) = \langle y(n), \psi_{j,i}(n) \rangle \quad (3)$$

The proposed method adopts the DWT noise reduction. The noise is reduced by applying the soft threshold and heuristic variant of Stein's unbiased risk threshold selection rules to the wavelet coefficients. And the denoised signal is reconstructed with manipulated detail signals.

B. Feature Extraction

According to the limitation of the practical conditions, it is impossible to directly establish the mapping relationship between nearly millions of voltage measurements and the PD. The amount of the data must be reduced prior to the classification. Hence, it is necessary to extract features from the voltage signal. The feature can be a value or a sequence. This paper adopts sequences as the features. Compared with a single-element feature, the sequence feature makes the variation and correlation between voltage amplitude and phase more involved in the analysis process of the PD detection, while reducing the degree of manual intervention of the proposed method. The feature length, which means the length of each sequence, must be consistent. To construct a sequence feature with N elements, the voltage signal is divided into N segments. The length of each segment is $800,000/N$. One value is extracted from each segment. There are various extraction methods, such as taking the maximum value; taking the standard deviation; taking the average value, and so on. When a method is selected, all the N segments use this method to extract values. All the N values form a sequence with length N as an N -element feature for the voltage signal. To reduce the confusion, the sequences are always rearranged in phase order of 0° – 360° . All the features in this paper are extracted in this manner.

The proposed method adopts 13 features. Feature 1 is the maximum value, i.e., the first largest value. Feature 2 is the 3rd largest value. Feature 3 is the 5th largest value. Feature 4 is the 10th largest value. Feature 5 is the 25th largest value. By introducing Features 2 to 5, it is able to reduce the effect of the individual large amplitude noise pulses. The numbers 3rd, 5th, 10th, 25th are randomly selected. Besides the amplitude, PD also changes the data distribution of the voltage signal. The data distribution is also considered as a feature. The extraction method is explained as follows: In order to describe the distribution of data belonging to interval M , the number of measurements belonging to interval M is firstly counted. The ratio of this number to the total number of measurements is the desired feature. Feature 6 is the distribution of data less than 1 mV. Feature 7 is the distribution of data less than 0.5 mV. Feature 8 is the distribution of data belonging to interval $[-1 \text{ mV}, 1 \text{ mV}]$. Feature 9 is the distribution of data belonging to interval $[-0.5 \text{ mV}, 0.5 \text{ mV}]$. Feature 10 is the distribution of data less than 0.1 mV. Feature 11 is the distribution of data belonging to interval $[-0.1 \text{ mV}, 0.1 \text{ mV}]$. The distribution of the large amplitude measurements fluctuates greatly due to different measuring environments. Therefore, the above small values are selected to reflect the commonness of PD. The specific values of these intervals are randomly selected. Feature 12 is the average value. It reflects the energy of the voltage signal. Feature 13 is the standard deviation value. It reflects the degree of the dispersion of the voltage signal.

C. LSTM Neural Network Classifier

The proposed method utilizes an LSTM neural network classifier to classify the features. The LSTM neural network classifier, as shown in Fig. 3, is consist of a sequence input layer that receives the input features, an LSTM layer, a feed forward layer with fully connected feedforward neural network (FNN), a softmax layer, and an output layer. The LSTM neural network classifier uses the advantage of the optimal recurrence mechanism of the LSTM neural network. As an optimal type of recurrent neural networks (RNNs), the LSTM neural network is able to not only store information as an internal state, but also precisely control the time scale variables of the hidden state [24], [25]. It is applicable to sequence analysis, such as the voltage signal classification. The LSTM cell, as shown in Fig. 4, introduces forget gate, input gate, and output gate. Such a structure is able to flexibly control the influence of the previous information on the present state. The LSTM cell also introduces the addition operation between different internal state information. This could effectively reduce the gradient disappearance problem caused by the multiplicative effect of small gradients.

In Fig. 4, x_t is the input of the present time step, and h_{t-1} is the output of the previous time step. x_t and h_{t-1} are combined into the four internal variables: g , i , f , and k . Among them, g is squashed by the hyperbolic tangent activation function. Variable i of the input gate, f of the forget gate, and k of the output gate utilize the sigmoid activation function. The calculation formulas are as follows:

$$g = \tanh(x_t W_x^g + h_{t-1} W_h^g + b^g) \quad (4)$$

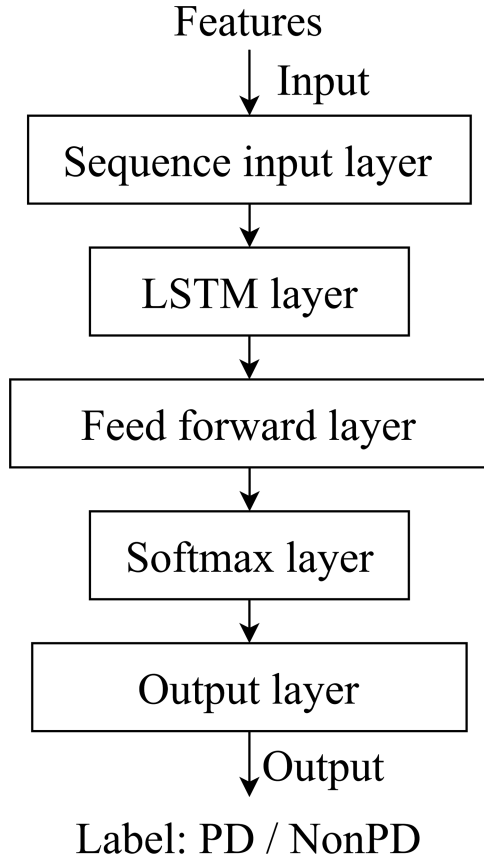


Fig. 3. The architecture of the LSTM neural network classifier.

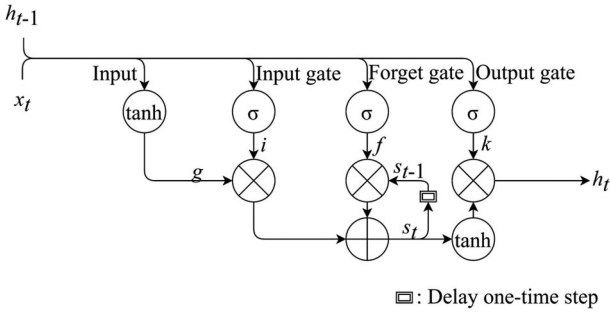


Fig. 4. Diagram of the LSTM cell.

$$i = \sigma(x_t W_x^i + h_{t-1} W_h^i + b^i) \quad (5)$$

$$f = \sigma(x_t W_x^f + h_{t-1} W_h^f + b^f) \quad (6)$$

$$k = \sigma(x_t W_x^k + h_{t-1} W_h^k + b^k) \quad (7)$$

where W_x^g, W_x^i, W_x^f , and W_x^k are the weights for the current input; W_h^g, W_h^i, W_h^f , and W_h^k are the weights for the previous output; and b^g, b^i, b^f and b^k are the input biases. s_t is the inner state of the LSTM cell. The current inner state s_t is related to s_{t-1} , f , g , and i :

$$s_t = s_{t-1} f + gi \quad (8)$$

Therefore, the cell constructs an internal recurrence loop which establishes a connection between inputs at different times. The inner state is through the hyperbolic tangent activation function. The result element-wise multiplies the output of the output gate to be the final output of the LSTM cell:

$$h_t = \tanh(s_t) k \quad (9)$$

With the above logic unit, the LSTM layer learns long-term dependencies between time steps in sequence features of the voltage signal. Its output is used as the input of the fully connected FNN. The softmax function is used as the last activation function to normalize the output of the FNN to a probability distribution over the PD class and the NonPD class. Finally, the output classification layer computes the cross-entropy loss for classification tasks with the PD and NonPD classes and decides the predicted result, PD/NonPD, to which the sequence features belong.

The training-validation data and testing data for the LSTM neural network classifier are the sequence features extracted from the voltage signals randomly selected from the VSB data set. In the training step, both the sequence features and the PD/NonPD labels are given to the LSTM neural network classifier. The classifier is improved by establishing the mapping relationship between the sequence features and the corresponding label. After the training, the LSTM neural network classifier classifies the testing data that are sequence features not used during training. The predicted results are PD/NonPD labels to which the testing data belong. They are predicted according to the mapping relationship established in the training process. There are 29049 voltage signals in the VSB data set. 2104 signals are labeled by experts as PD whereas 26945 signals as NonPD. As for the PD class, the ratio of the signals used for training-validation and testing is 7:3. This ratio is determined after considering the efficiency, effect, and confidence level of training, validation, and testing. The training set is expected to be sufficient to avoid under-fitting and improve the quality of the model. At the same time, the testing set is expected to be as large as possible to increase the reliability of the evaluation. The ratio of 7:3 is considered a balanced choice for the two conditions. The oversampling method is used in the training-validation data set to minimize the impact of the class imbalance problem. As for the NonPD class, the data are allocated to match the training-validation data set of the PD class. The remaining data are adopted for testing. In order to improve credibility, the training-validation set and testing set are always divided randomly. Multiple groups of repeated experiments are conducted. All the evaluation parameter values in this paper are the average of the results of the repeated experiments.

D. Evaluation Parameters

The quality of noise reduction methods and feature extraction methods is reflected in the final classification performance. Classification performance refers to whether the LSTM neural network classifier with such features as inputs can make correct classification decisions for the testing samples after the training. This paper introduces multiple evaluation parameters to evaluate

the classification performance. The terms PD Target, PD Prediction, True PD Prediction, NonPD Target, NonPD Prediction, and True NonPD Prediction are used to explain the meaning of the evaluation parameters. The descriptions for the PD Target, PD Prediction, and True PD Prediction are presented below. The NonPD Target, NonPD Prediction, and True NonPD Prediction can be deduced similarly.

PD Target: The PD Target refers to the voltage signal labeled by experts as PD.

PD Prediction: If a signal is determined by a classifier as PD, this signal is a PD Prediction for this classifier. The classifier may misjudge and predict a NonPD Target as a PD Prediction.

True PD Prediction: If a PD Target is predicted by a classifier as a PD Prediction, this sample is a True PD Prediction for this classifier. In other words, the classifier classifies this sample correctly. The True PD Prediction is the intersection of the PD Target and the PD Prediction.

The following evaluation parameters are adopted to evaluate the classification performance.

Accuracy: In binary classification, the terminology Accuracy is the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined. Accuracy is related to the classification performance of both classes of samples. The higher the Accuracy, the better the overall classification effect. In this paper, Accuracy $Acc \in [0, 1]$ is adopted to evaluate the classification performance of both PD and NonPD classes. It is calculated as follows:

$$Acc = \frac{TP_{PD} + TP_{NPD}}{T_{PD} + T_{NPD}} \quad (10)$$

where TP_{PD} is the number of the True PD Prediction; TP_{NPD} is the number of the True NonPD Prediction; T_{PD} is the number of the PD Target; and T_{NPD} is the number of the NonPD Target.

Recall: The terminology Recall is the proportion of the retrieved relevant instances in all relevant instances. The higher the Recall, the better the classification effect. This paper adopts both $R_{PD} \in [0, 1]$, the Recall for the PD class, and $R_{NPD} \in [0, 1]$, the Recall for the NonPD class, to evaluate the classification performance. The calculation formulas are as follows:

$$R_{PD} = \frac{TP_{PD}}{T_{PD}} \quad (11)$$

$$R_{NPD} = \frac{TP_{NPD}}{T_{NPD}} \quad (12)$$

F1-score: The F1-score is the harmonic mean of the Recall and the Precision. The terminology Precision is the fraction of relevant instances among the retrieved instances. $Pr_{PD} \in [0, 1]$, the Precision for the PD class, and $Pr_{NPD} \in [0, 1]$, the Precision for the NonPD class are calculated as follows:

$$Pr_{PD} = \frac{TP_{PD}}{P_{PD}} \quad (13)$$

$$Pr_{NPD} = \frac{TP_{NPD}}{P_{NPD}} \quad (14)$$

where P_{PD} is the number of the PD Prediction; and P_{NPD} is the number of the NonPD Prediction.

TABLE I
ANALYSIS OF DIFFERENT NOISE REDUCTION METHODS

	Wavelet	Level	Acc	R_{PD}	F_{PD}	R_{NPD}	F_{NPD}
1	-	-	0.901	0.870	0.898	0.933	0.904
2	DB4	2	0.933	0.907	0.931	0.959	0.935
3	DB4	8	0.930	0.898	0.928	0.962	0.933
4	DB1	2	0.914	0.882	0.911	0.946	0.916
5	DB1	8	0.917	0.890	0.915	0.944	0.919
6	Sym8	2	0.936	0.906	0.934	0.965	0.937
7	Sym8	5	0.915	0.884	0.912	0.945	0.917
8	Sym8	8	0.939	0.911	0.937	0.967	0.941
9	Sym8	11	0.947	0.923	0.945	0.970	0.948

$F_{PD} \in [0, 1]$, the F1-score for the PD class, and $F_{NPD} \in [0, 1]$, the F1-score for the NonPD class, are calculated as follows:

$$F_{PD} = \frac{2R_{PD}Pr_{PD}}{R_{PD} + Pr_{PD}} \quad (15)$$

$$F_{NPD} = \frac{2R_{NPD}Pr_{NPD}}{R_{NPD} + Pr_{NPD}} \quad (16)$$

They are adopted by this paper to evaluate the PD detection performance. The higher the F1-score, the better the classification effect.

Training time: The training time Tr of the LSTM neural network classifier is also adopted as an evaluation parameter for the feature length analysis. Under the same hardware conditions, the less the Tr , the faster the computation.

III. RESULT AND ANALYSIS

A. Analysis of Noise Reduction Methods

Different mother wavelets and decomposition levels for the DWT noise reduction are analyzed for the optimization of the noise reduction method. Specifically, three choices of the mother wavelets, the Daubechies 1 (DB1), Daubechies 4 (DB4), and Symlets 8 (Sym8) wavelets, and four choices of decomposition levels for the DWT noise reduction, the levels 2, 5, 8, and 11, are evaluated. The wavelets are randomly selected from the Daubechies series and Symlets series that are recommended to perform noise reduction for voltage signals [26]. The Daubechies wavelets, based on the work of Ingrid Daubechies, are a family of orthogonal wavelets, whereas the Symlets wavelets are a kind of nearly symmetrical wavelets. The Sym8 wavelet has eight vanishing moments; the DB1 wavelet has a maximal one vanishing moment; and the DB4 wavelet has maximal four vanishing moments. The four different decomposition levels are also randomly selected from the natural numbers. Table I lists the evaluation parameters of the different noise reduction methods. The second column of Table I are the mother wavelets used for the DWT noise reduction of the respective noise reduction methods, whereas the third column is the respective decomposition levels. The first method is the control group without noise reduction. The values of the parameters in Table I are the average of the results of ten groups of repeated experiments. Fig. 5 intuitively shows the change of Acc

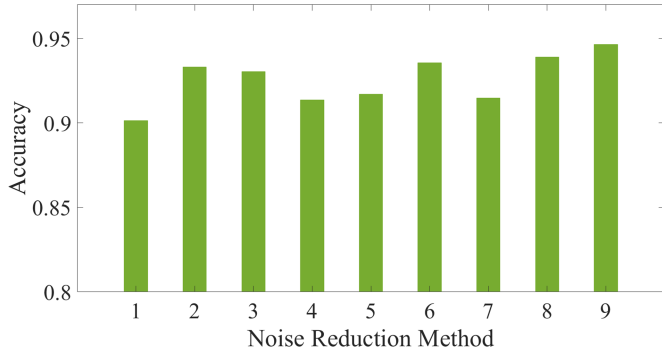
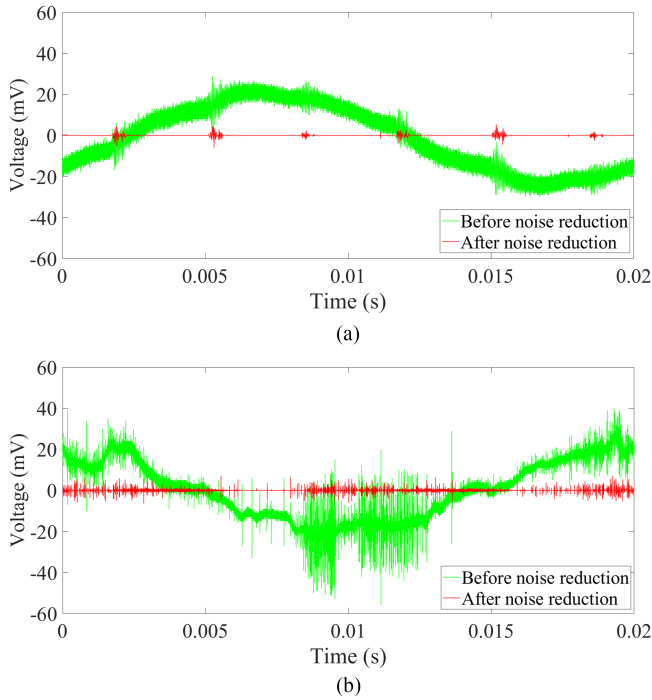
Fig. 5. Accuracy Acc of different noise reduction methods.

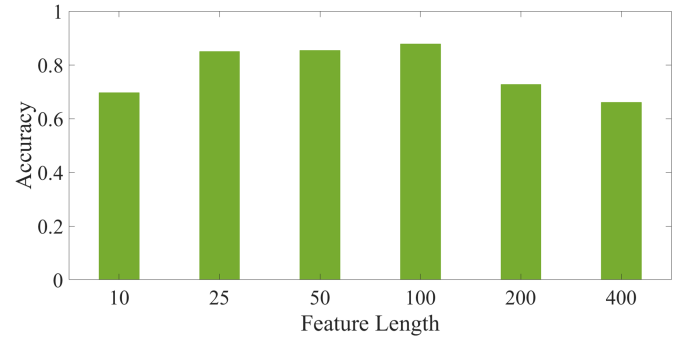
Fig. 6. The examples of the voltage signals before and after noise reduction. (a) is labeled as NonPD while (b) as PD.

caused by different noise reduction methods. The nine numbers of abscissa correspond to the nine different methods described in Table I.

The variation of the values of the evaluation parameters in Table I and Fig. 5 prove the significance and necessity of noise reduction. This preliminary analysis shows that among the DB4, DB1, and Sym8 wavelets, the Sym8 wavelet is more suitable for the proposed PD detection task. With the same decomposition level, the values of Acc , R_{NPD} , F_{PD} , and F_{NPD} are higher when the Sym8 wavelet is used. Among the four different decomposition levels of 2, 5, 8, and 11, the values of Acc , R_{PD} , R_{NPD} , F_{PD} , and F_{NPD} are the highest when level 11 is adopted. As an example, Fig. 6 shows the voltage signals before and after the noise reduction. Specifically, Fig. 6(a) presents a NonPD signal before and after the noise reduction. Fig. 6(b) presents a PD signal before and after the noise reduction. Both

TABLE II
ANALYSIS OF DIFFERENT FEATURE LENGTHS

Length	Acc	R_{PD}	F_{PD}	R_{NPD}	F_{NPD}	Tr (s)
10	0.697	0.643	0.680	0.750	0.712	308.368
25	0.850	0.836	0.848	0.864	0.852	326.344
50	0.854	0.816	0.848	0.891	0.859	369.475
100	0.878	0.894	0.880	0.862	0.876	511.941
200	0.727	0.678	0.713	0.777	0.740	924.404
400	0.661	0.537	0.613	0.785	0.698	2030.125

Fig. 7. Accuracy Acc of different feature lengths.

the voltage signals are processed through the same noise reduction scheme: Firstly, the digital trap filter is used on the signals to reduce the power line interference. Secondly, three kinds of noise pulses: abnormal negative pulses; symmetrical pulses; and abnormal high amplitude pulses, which are irrelevant to PD, are removed from the signals. Thirdly, the 11-level DWT noise reduction with Sym8 wavelet is performed on the signals. It can be seen from Fig. 6 that most irrelevant information is filtered out. This also leads to the reduction of signal amplitude. The intensive pulses caused by PD are much clearer than before.

B. Analysis of the Feature Length

The feature length significantly affects the classification performance and the amount of computation. Six different lengths: 10, 25, 50, 100, 200, and 400 are compared. They are randomly selected from the numbers by which the length of the voltage signal is divisible. The features with over 400 elements lead to the requirement of high-performance computing resources. This paper controls the sequence features within 400 elements to ensure the economic practicability of the proposed method.

Table II lists the evaluation parameters of the different feature lengths. The first column of Table II lists the feature lengths. The values of the parameters in Table II are the average of the results of seven groups of repeated experiments. Fig. 7 shows the change of Acc caused by different feature lengths. On the one hand, If the feature length is too short, the feature will contain insufficient valid information. As the feature length increases from 10 to 100, Acc increases from 0.697 to 0.878. On the other hand, if the feature length is too long, it would add an additional burden to the LSTM neural network classifier. As the feature length increases from 100 to 400, Acc decreases from 0.878 to 0.661. The length of 100 outperforms other lengths in this

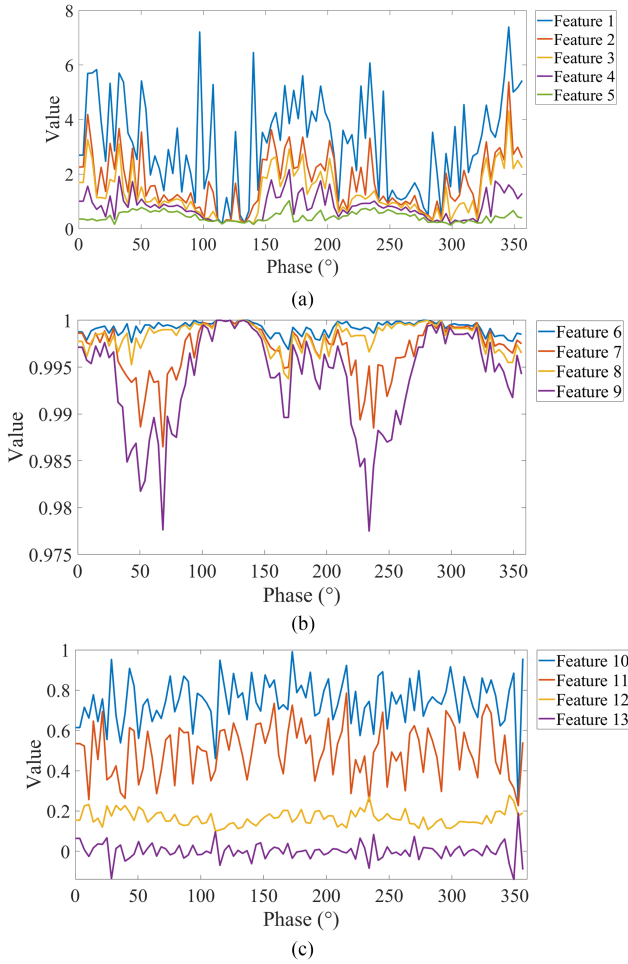


Fig. 8. The selected features of the denoised signal labeled as PD in Fig. 6. (a) is with Features 1–5. (b) is with Features 6–9. (c) is with Features 10–13.

preliminary analysis. Tr for the 100-element feature is under the level acceptable. And it corresponds to the highest values of Acc , R_{PD} , F_{PD} , and F_{NPD} . It is considered to have greater potential with the optimization of noise reduction and feature extraction methods.

C. Analysis of Different Features

The performance of the 13 sequence features described in Section II-B is analyzed as the inputs of the LSTM neural network classifier. As an example, Fig. 8 shows the 13 features extracted from the denoised signal shown in Fig. 6(b). Table III lists the evaluation parameters of the different features. The values of the parameters in Table III are the average of the results of five groups of repeated experiments. It can be seen from Table III that the features of extremum values have nice performance, while the features of the data distribution are less capable of handling the PD class.

D. Analysis of Loop Optimization

Section III-A, III-B, and III-C have discussed the performance of the different noise reduction methods, feature lengths, and

TABLE III
ANALYSIS OF 13 DIFFERENT FEATURES

	Acc	R_{PD}	F_{PD}	R_{NPD}	F_{NPD}
Feature 1	0.905	0.893	0.904	0.918	0.907
Feature 2	0.871	0.856	0.869	0.885	0.873
Feature 3	0.873	0.862	0.872	0.885	0.875
Feature 4	0.854	0.808	0.847	0.900	0.861
Feature 5	0.779	0.696	0.758	0.864	0.796
Feature 6	0.532	0.103	0.180	0.960	0.672
Feature 7	0.563	0.210	0.325	0.917	0.677
Feature 8	0.646	0.586	0.623	0.706	0.666
Feature 9	0.585	0.293	0.414	0.877	0.679
Feature 10	0.562	0.472	0.519	0.651	0.598
Feature 11	0.601	0.363	0.477	0.839	0.678
Feature 12	0.602	0.368	0.481	0.836	0.678
Feature 13	0.644	0.575	0.618	0.713	0.667

TABLE IV
DETAILS OF 7 COMBINATION SCHEMES (SET NO. 1–7)

Set No.	Wavelet	Level	Feature length
1	DB1	8	100
2	DB4	8	100
3	Sym8	5	100
4	Sym8	11	100
5	Sym8	11	50
6	Sym8	11	25
7	Sym8	11	10

feature extraction methods, respectively. The above discussion has preliminarily reflected the general influence of different items on the PD detection performance of the whole proposed method. However, as the internal comparison of the single items controls other items for the single variable, the impact of other items is deficient in the analysis. In this section, the influence of the different combination schemes for the noise reduction, feature lengths, and different feature extraction methods is analyzed.

Table IV presents the analyzed seven combination schemes of the noise reduction methods and feature lengths. The results in IIIA indicate that the DWT noise reduction with the Sym8 wavelet is more likely to obtain a better performance than the wavelets of the DB series with the same decomposition level. Among all the analyzed noise reduction methods, the 11-level DWT noise reduction with the Sym8 wavelet obtains the best performance in the preliminary analysis. Therefore, in this section, the 11-level DWT noise reduction with the Sym8 wavelet is employed to combine with all the feature lengths. As the results in IIIB show that the feature length of 100 outperforms the other lengths in the preliminary analysis, the feature length of 100 is employed in more comparisons of different noise reduction methods.

Table V presents the analyzed five feature groups (Groups A–E) which combine different feature extraction methods (Features 1–13). The results in IIIC indicate that the features of extremum values have nice performance in the preliminary analysis, while the features of the data distribution are less capable of handling the PD class. Groups A, B, and C combine the features of

TABLE V
DETAILS OF 5 DIFFERENT FEATURE GROUPS (GROUPS A-E)

	A	B	C	D	E
1	✓			✓	✓
2	✓			✓	✓
3	✓			✓	✓
4	✓				✓
5	✓				✓
6		✓			✓
7		✓			✓
8		✓		✓	✓
9		✓			✓
10		✓			✓
11		✓			✓
12			✓		✓
13			✓	✓	✓

TABLE VI
ANALYSIS OF DIFFERENT COMBINATION SCHEMES AND FEATURE GROUPS

Set No.	Group	Acc	R_{PD}	F_{PD}	R_{NPD}	F_{NPD}
1	A	0.888	0.877	0.887	0.899	0.889
	B	0.735	0.660	0.714	0.810	0.753
	C	0.558	0.163	0.270	0.954	0.683
	D	0.803	0.767	0.796	0.839	0.810
	E	0.911	0.885	0.908	0.936	0.913
2	A	0.914	0.892	0.912	0.936	0.916
	B	0.773	0.821	0.783	0.725	0.762
	C	0.589	0.786	0.656	0.393	0.489
	D	0.919	0.887	0.917	0.951	0.921
	E	0.942	0.922	0.941	0.961	0.943
3	A	0.913	0.873	0.909	0.952	0.916
	B	0.813	0.674	0.783	0.952	0.836
	C	0.779	0.624	0.739	0.933	0.809
	D	0.917	0.895	0.915	0.938	0.919
	E	0.926	0.886	0.922	0.965	0.928
4	A	0.939	0.921	0.938	0.957	0.940
	B	0.758	0.883	0.785	0.632	0.723
	C	0.660	0.525	0.607	0.796	0.701
	D	0.947	0.926	0.946	0.968	0.948
	E	0.954	0.937	0.953	0.971	0.955
5	A	0.937	0.913	0.936	0.961	0.939
	B	0.781	0.763	0.777	0.798	0.785
	C	0.817	0.797	0.814	0.838	0.821
	D	0.923	0.897	0.921	0.948	0.925
	E	0.935	0.909	0.934	0.962	0.937
6	A	0.912	0.881	0.909	0.944	0.915
	B	0.826	0.844	0.829	0.808	0.823
	C	0.744	0.743	0.743	0.744	0.744
	D	0.899	0.876	0.896	0.922	0.901
	E	0.916	0.887	0.913	0.944	0.918
7	A	0.726	0.606	0.689	0.846	0.755
	B	0.714	0.614	0.683	0.814	0.740
	C	0.669	0.503	0.603	0.835	0.716
	D	0.764	0.733	0.757	0.795	0.771
	E	0.798	0.743	0.786	0.854	0.809

extremum values, data distributions, and statistics, respectively. They are constructed to verify if the combination of features of the same category outperforms the single feature of that category. Group D selects the features with the best preliminary performance from each category. Group E combines all the 13 features.

Table VI lists the evaluation parameters of the different combination schemes with different feature groups. The values of

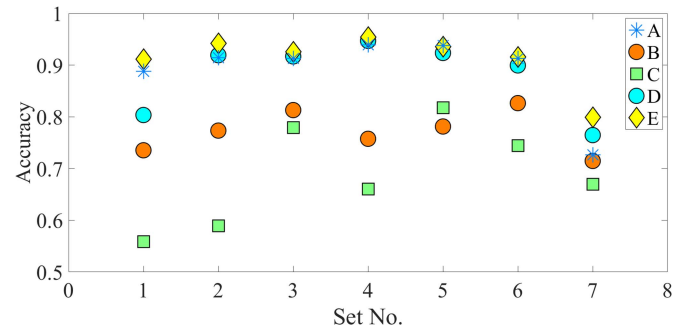


Fig. 9. Accuracy Acc of different combination schemes (Set No. 1–7) With different feature groups (Groups A-E).

the parameters in Table VI are the average of the results of five groups of repeated experiments. Fig. 9 intuitively shows the change of Acc caused by different combinations. It can be seen from Table VI and Fig. 9 that with the same combination scheme, Group A always performs better than Group B and Group C. This is consistent with the performance of the single feature of the three categories. Compared the values of the parameters in Tables III and VI, Group A outperforms the single Features 1-5. Group B outperforms the single Features 6-11. And Group C outperforms the single Features 12-13.

As for Group D and Group E that combine the features from different categories, all the seven combination schemes obtain nice results with these two groups of features. Except Set No. 1, all the other six combination schemes obtain better performance when the features of different categories are combined. And Group E always obtains better performance than Group D no matter the change of noise reduction methods and feature length. Comparing Groups A, D, and E, Group E adds some features that are relatively weak in the preliminary analysis. However, the complete combination outperforms the groups that only combine the strong features. The combination of the different features has combined advantages and can overcome each individual's drawback.

Set No. 1–4 compare the different noise reduction methods with different combination schemes of the 100-elements features. The results of their Groups A, D, and E are consistent with the results of Table I. Set No. 4-7 compare the different feature lengths with the same noise reduction method and different feature combination schemes. With the combination of different features, the advantages of the advanced feature length are further expanded. With the analysis of Section III-A, III-B, III-C, and III-D, this paper determines the comprehensive solution for the PD detection of MV overhead power lines as the LSTM neural network classifier with the Sym8-11-level DWT noise reduction and 13 combined 100-element features.

The analysis of Table VI and Fig. 9 also leads to more common discussions of the optimization of PD detection methods. The single features and noise reduction methods discussed in this paper are based on the essential characteristics of the PD. They have low hardware implementation and computation requirements. Hence, they have advantages of adapting to wider PD detection tasks. And the improvement of combination and loop

TABLE VII
COMPARISON OF THE PROPOSED METHOD AND OTHER METHODS IN THE LITERATURE

Method	Acc	R_{PD}	R_{NPD}
STL	0.805	0.838	0.772
DWT	0.871	0.867	0.882
SVM	0.701	0.665	0.737
CNN	0.775	0.742	0.809
Proposed method	0.954	0.937	0.971

optimization analysis of such items are beneficial to practical applications. As combining the noise reduction, feature extraction, and the LSTM neural network classifier as a whole with internal linkages, the comprehensive optimization becomes a problem of permutation and combination. The exhaustive method does not meet the actual economic requirements, while the one-way selection method lacks the internal linkages. Therefore, a loop optimization mechanism is proposed based on the aforementioned experimental results and analysis.

- 1) The principle is that it can continuously conduct the cyclic comparison of different items until the overall PD detection performance satisfies the task requirements. Optimization can stop at any step of a loop. The loop optimization has great improvement potential.
- 2) It is suggested to first determine the feature length preliminarily based on the computation requirements and the detection performance of the single feature. The advanced feature length is likely to have more improvement potential with the further optimization.
- 3) It is suggested to appropriately increase the number of features.
- 4) It is suggested to appropriately increase the decomposition level of the noise reduction methods. The noise reduction methods are suggested to be optimize after selecting the strong features in each loop.

E. Comparison With Other Techniques

The proposed method has achieved a comprehensive and complete solution for PD detection for MV overhead power lines. It is compared with the noise reduction methods, feature extraction methods, and AI classifiers in the literature as shown in Table VII [17], [18]. The “STL” method employs the seasonal and trend decomposition using locally weighted regression (STL). STL transforms the voltage signal into the trend component, seasonal component, and residual component. The peak value, standard deviation, and accumulated peak height of the residual component are used as features. The LSTM neural network is adopted as the classifier. The “DWT” method also employs the LSTM neural network. It employs the DWT noise reduction to the voltage signals. The decomposition components of the DWT are then obtained as features. The “SVM” method uses six features. The features are the energy of the voltage signal, the energy of the voltage signal after the noise reduction, the energy of the abnormal parts of the signal, the average of the segmental maximum values, the amount of the abnormal

measurements, and the standard deviation. The Support Vector Machine (SVM) is adopted for the task of classification. As for the “CNN” method, the voltage signal is firstly denoised. Then the maximum value is extracted as a feature. The convolutional neural network (CNN) is adopted for the task of classification.

It can be seen from Table VII that R_{PD} and R_{NPD} of the “STL” method differ greatly. The “STL” method has obvious shortcomings for the NonPD class. And the STL processing costs more calculation than other noise reduction and feature extraction methods. As for the “DWT” method, its classification results for the PD and NonPD classes are relatively balanced. With the same LSTM neural network, the proposed method greatly improves the PD detection performance with the coordination of the noise reduction, feature combination, and the LSTM based classifier. As for AI techniques, the proposed method is compared with the SVM and CNN. Both CNN and LSTM are considered to have the excellent sequence classification ability. However, with different feature extraction methods, the performance of the two neural networks differs a lot for the same PD detection task. This also demonstrates the significance of the coordination of the feature extraction and the machine learning techniques. The proposed method experimentally demonstrates its superior performance.

IV. CONCLUSION

This paper has presented a loop optimization noise-reduced LSTM based classifier for PD detection of MV overhead power lines. The proposed method greatly coordinates the multiple processing and interpretation procedures for the large-scale data disturbed with measurement noise. The loop optimization mechanism is established to combine the noise reduction, feature extraction, and the LSTM neural network as a whole with internal linkages. The optimal combination has combined advantages and can overcome each individual’s drawback. Extensive experiments have been conducted based on the huge real trial data set of large-scale high-frequency voltage signals measured from the MV overhead power lines. The proposed method outperforms other techniques in the literature based on the experimental results. It has achieved a comprehensive and complete solution for automatic PD detection of MV overhead power lines. It is a valuable tool for insulation monitoring and power maintenance. Moreover, this paper summarizes the common loop optimization rules for PD detection solutions based on extensive experiments and analysis. The loop optimization rules are competent for the optimization of wider engineering applications of PD detection.

REFERENCES

- [1] N. Xu, H. B. Gooi, L. Wang, Y. Zheng, and J. Yang, “Partial discharge detection based on long short-term memory neural network classifier with efficient feature extraction methods,” in *Proc. IEEE 12th Energy Convers. Congr. Expo. - Asia*, 2021, pp. 2328–2333.
- [2] H. D. Ilkhechi and M. H. Samimi, “Applications of the acoustic method in partial discharge measurement: A review,” *IEEE Trans. Dielectrics Elect. Insul.*, vol. 28, no. 1, pp. 42–51, Feb. 2021.
- [3] S. Lu, H. Chai, A. Sahoo, and B. T. Phung, “Condition monitoring based on partial discharge diagnostics using machine learning methods: A comprehensive state-of-the-art review,” *IEEE Trans. Dielectrics Elect. Insul.*, vol. 27, no. 6, pp. 1861–1888, Dec. 2020.

- [4] L. Zhou et al., "Partial discharge characteristic analysis of distribution network overhead line based on remote detection," in *Proc. IEEE Int. Conf. High Voltage Eng. Appl.*, 2020, pp. 1–4.
- [5] R. Ghosh, B. Chatterjee, and S. Dalai, "A method for the localization of partial discharge sources using partial discharge pulse information from acoustic emissions," *IEEE Trans. Dielectrics Elect. Insul.*, vol. 24, no. 1, pp. 237–245, Feb. 2017.
- [6] M. R. Hussain, S. S. Refaat, and H. Abu-Rub, "Overview and partial discharge analysis of power transformers: A literature review," *IEEE Access*, vol. 9, pp. 64587–64605, 2021.
- [7] R. Ghosh, P. Seri, R. E. Hebner, and G. C. Montanari, "Noise rejection and detection of partial discharges under repetitive impulse supply voltage," *IEEE Trans. Ind. Electron.*, vol. 67, no. 5, pp. 4144–4151, May 2020.
- [8] W. R. SI, C. Z. FU, L. CHEN, S. J. Wang, and P. YUAN, "Study on time-frequency entropy method to make feature extraction for DC PD pulse waveshapes," in *Proc. IEEE Int. Conf. High Voltage Eng. Appl.*, 2018, pp. 1–4.
- [9] R. Albarracín-Sánchez, F. Álvarez-Gómez, C. A. Vera-Romero, and J. M. Rodríguez-Serna, "Separation of partial discharge sources measured in the high-frequency range with HFCT sensors using PRPD-teff patterns," *Sensors*, vol. 20, no. 2, Jan. 2020, Art. no. 3267.
- [10] R. C. Araujo, R. M. de Oliveira, F. S. Brasil, and F. J. Barros, "Novel features and prpd image denoising method for improved single-source partial discharges classification in on-line hydro-generators," *Energies*, vol. 14, no. 11, Jun. 2021, Art. no. 3267.
- [11] I. V. Pustokhina, D. A. Pustokhin, D. Gupta, A. Khanna, K. Shankar, and G. N. Nguyen, "An effective training scheme for deep neural network in edge computing enabled internet of medical things (iomt) systems," *IEEE Access*, vol. 8, pp. 107112–107123, 2020.
- [12] Z. Xu, J. Tang, C. Yin, Y. Wang, and G. Xue, "Experience-driven congestion control: When multi-path tcp meets deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1325–1336, Jun. 2019.
- [13] X. Zhou et al., "Research on transformer partial discharge uhf pattern recognition based on CNN-LSTM," *Energies*, vol. 13, no. 1, Dec. 2019, Art. no. 61.
- [14] 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. Dielectrics Elect. Insul.*, vol. 25, no. 2, pp. 678–685, Apr. 2018.
- [15] A. Gao, Y. Zhu, W. Cai, and Y. Zhang, "Pattern recognition of partial discharge based on VMD-CWD spectrum and optimized cnn with crosslayer feature fusion," *IEEE Access*, vol. 8, pp. 151296–151306, 2020.
- [16] 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.
- [17] 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.
- [18] M. Dong and J. Sun, "Partial discharge detection on aerial covered conductors using time-series decomposition and long short-term memory network," *Elect. Power Syst. Res.*, vol. 184, Jul. 2020, Art. no. 106318.
- [19] 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.
- [20] S. Mišák, Š. Hamacek, P. Bilík, M. Hofinec, and P. Petvaldsky, "Problems associated with covered conductor fault detection," in *Proc. IEEE 11th Int. Conf. Elect. Power Qual. Utilisation*, 2011, pp. 1–5.
- [21] A. A. Soltani and S. M. Shahrtash, "Decision tree-based method for optimum decomposition level determination in wavelet transform for noise reduction of partial discharge signals," *IET Sci., Meas. Technol.*, vol. 14, no. 1, pp. 9–16, Nov. 2019.
- [22] I. Enesi and L. Lic'o, "Optimizing parameters of discrete wavelet transform for noise reduction in fingerprint images," in *Proc. IEEE 10th Mediterranean Conf. Embedded Comput.*, 2021, pp. 1–5.
- [23] Y.-T. Chen, J.-Y. Lin, K.-Y. Liu, and J.-W. Hung, "Smoothing the acoustic spectral time series of speech signals for noise reduction," in *Proc. IEEE Int. Conf. Consum. Electron. - Taiwan*, 2019, pp. 1–2.
- [24] V. Khomenko, O. Shyshkov, O. Radyvonenko, and K. Bokhan, "Accelerating recurrent neural network training using sequence bucketing and multi-GPU data parallelization," in *Proc. IEEE 1st Int. Conf. Data Stream Mining Process.*, 2016, pp. 100–103.
- [25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [26] W. K. Ngui, M. S. Leong, L. M. Hee, and A. M. Abdelrhman, "Wavelet analysis: Mother wavelet selection methods," *Appl. Mechanics Mater.*, vol. 393, pp. 953–958, Sep. 2013.



Ning Xu received the B.E. degree in electrical engineering and automation from Xi'an Jiaotong University, Xi'an, China in 2020. She is currently working toward the M.E. degree in electrical engineering with Nanyang Technological University, Singapore. Her research interests include artificial intelligence technology and related engineering applications.



Hoay Beng Gooi (Life Senior Member, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taiwan, in 1978, the M.S. degree in electrical engineering from the University of New Brunswick, Fredericton, NB, Canada, in 1980, and the Ph.D. degree in electrical engineering from The Ohio State University, Columbus, OH, USA, in 1983. He is an Associate Professor and Co-Director of SP Group-NTU Joint Lab with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include microgrid energy management systems dealing with storage, electricity market, and spinning reserve.



Lipo Wang (Senior Member, IEEE) received the bachelor's degree from the National University of Defense Technology, Changsha, China and the Ph.D. from Louisiana State University, Baton Rouge, LA, USA. He is currently the faculty of the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He has coauthored two monographs and (co-)edited 15 books. He has more than 350 publications, a U.S. patent in neural networks and a patent in systems. He has more than 11000 Google Scholar citations, with H-index 48.

He was keynote speaker for more than 40 international conferences. His research interests include artificial intelligence with applications to power/energy, image/video processing, communications, biomedical engineering, and data mining. He is/was an Associate Editor/Editorial Board Member of more than 30 international journals, including four IEEE Transactions, and Guest Editor for more than 10 journal special issues. He was a Member of the Board of Governors of The International Neural Network Society, IEEE Computational Intelligence Society, and IEEE Biometrics Council. He was the CIS Vice President of Technical Activities and Chair of Emergent Technologies Technical Committee, and Chair of Education Committee of the IEEE Engineering in Medicine and Biology Society. He was the President of the Asia-Pacific Neural Network Assembly (APNNA, recently renamed as APNNS – "Society"). He was the recipient of the APNNA Excellent Service Award. He was the Founding Chair of both the EMBS Singapore Chapter and CIS Singapore Chapter.



Yuanjin Zheng (Senior Member, IEEE) received the B.Eng. and M. Eng. degrees from Xian Jiaotong University, Xi'an, China, in 1993 and 1996, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2001. From July 1996 to April 1998, he was with the National Key Lab of Optical Communication Technology, University of Electronic Science and Technology of China, Chengdu, China. He joined Institute of Microelectronics, A*STAR, Singapore, in 2001 and developed as a Group Technical Manager. Since then, he has

led in developing various wireless systems and CMOS integrated circuits, such as Bluetooth, WLAN, WCDMA, UWB, RF SAW/MEMS, Radar, and wireless implant sensor, and wearable interface circuits. Since July 2009, he has been with the Nanyang Technological University, Singapore, and currently the Director for VIRTUS, IC design center of excellence, working on various radar system development and hybrid circuit and device, such as GaN, SAW, and MEMS designs, and flexible noninvasive sensor circuits, and system for the applications. He has authored or coauthored more than 450 international journal and conference papers, 26 patents filed, and five book chapters. He is currently an Associate Editor for three journals and has been organizing several IEEE conferences as the TPC Chairs and Session chairs. He was also a Lead Guest Editor of a special Issue TBioCAS2019, and Best Paper Award of CAS Life Science and Biomedical Circuits track in 2018 ISCAS.



Wensong Wang (Senior Member, IEEE) received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016. From 2013 to 2015, he was a Visiting Scholar with the University of South Carolina, Columbia, SC, USA. In 2017, he joined Nanyang Technological University, Singapore, as a Research Fellow, and currently a Senior Research Fellow. His research interests include MIMO antenna, antenna array, and advanced sensors. He is also an Associate Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE SENSORS JOURNAL, and *IET Microwaves, Antennas & Propagation*.



Jiawei Yang (Student Member, IEEE) received the B.E. degree in electrical engineering and automation from Wuhan University, Wuhan, China, in 2018, and the M.E. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2020, where he is currently working toward the Ph.D. degree in electrical engineering. His research interests include peer-to-peer energy trading and blockchain application for energy trading.