Knowledge Distillation and Multi-task Feature Learning for Partial Discharge Recognition

1st Jinsheng Ji
School of Electrical and Electronic Engineering
Nanyang Technological University
639798, Singapore
jinsheng.ji@ntu.edu.sg

2nd Zhou Shu
School of Electrical and Electronic Engineering
Nanyang Technological University
639798, Singapore
zhou.shu@ntu.edu.sg

3rd Hongqun Li
Grid Digitalisation
SP Group
349277, Singapore
lihq@spgroup.com.sg

4th Kai Xian Lai
Asset Sensing&Analytics
SP Group
349277, Singapore
kaixian@spgroup.com.sg

5th Yuanjin Zheng
School of Electrical and Electronic Engineering
Nanyang Technological University
639798, Singapore
yjzheng@ntu.edu.sg

6th Xudong Jiang
School of Electrical and Electronic Engineering
Nanyang Technological University
639798, Singapore
exdjiang@ntu.edu.sg

***Abstract*—To achieve accurate detection and recognition of partial discharge (PD) in switchgear, developing an intelligent PD diagnosis system has garnered significant attention in recent years. Due to inevitable noise interference and high similarity of different PD signals, detecting and identifying PDs using a portable PD detector poses significant challenges. In this study, we aim to transfer the knowledge acquired by the large-scale network to a lightweight network for precise PD recognition. To achieve this, we employ a k-means clustering model to effectively separate signals originating from different sources, thereby obtaining Phase Resolved Partial Discharge (PRPD) patterns. Then, we introduce knowledge distillation and a multi-task feature learning framework to extract discriminative features from PRPD patterns. We conduct experiments and compare the proposed method against some state-of-the-art methods on our constructed PD recognition dataset to evaluate the superiority of the proposed method.**

***Index Terms*—Knowledge distillation, Partial discharge, Pattern recognition**

I. Introduction

Partial discharge (PD) denotes localized breakdowns in insulation between conductors within electrical systems. Although these discharges may seem insignificant, they can profoundly affect the performance and reliability of high-voltage equipment [1]. PD classification serves as a vital tool for assessing and mitigating risks associated with partial discharges. Accurate PD classification plays a pivotal role in proactive maintenance planning, reducing downtime, and preventing catastrophic failures, thus safeguarding the reliable operation of high-voltage equipment. By categorizing PDs based on measurable attributes, engineers and experts gain a comprehensive understanding of their characteristics and implications. By categorizing PDs according to parameters such as magnitude, frequency, and waveform shape, engineers gain valuable insights into their severity and type. This enables

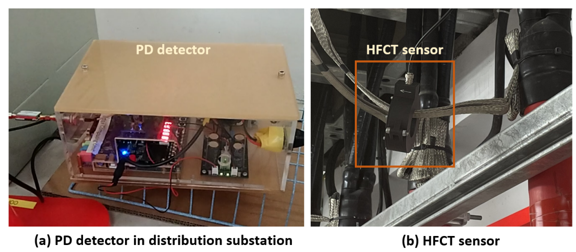

Fig. 1. The designed PD detector and high-frequency current transformer (HFCT) sensor deployed in distribution substation.

them to make informed decisions regarding maintenance or repair [2].

To detect partial discharges that originate from multilevel pulsewidth in power electronic devices, a machine learning (ML) based model is introduced in [3] by concatenating a sequence of extracted features to capture the temporal dependence of consecutive PDs. In [4], a novel approach that combines wavelet kernels with convolutional neural networks (CNNs) is presented to accurately localize partial discharge sources in power apparatus. To handle partial discharge recognition with complex data sources, a deep convolutional neural network (DCNN) model is employed in [5] to automatically extract the features from the PRPD patterns and achieves impressive results. PRPD patterns are divided into multiple phase windows and features of discharge magnitude and number of pulses are extracted from each window to form 1-D histograms for the PD classification in [6].

Different from previous methods, a knowledge distillation and multi-task feature learning framework is proposed to extract informative features from the PRPD patterns. Specifically, we adopt a teacher network to learn useful information from a large amount of PD data collected in the power substation, and then soft-labels are generated by the teacher network to guide a light-weight student network to obtain accurate PD type predictions. Besides, a multi-task learning scheme is further

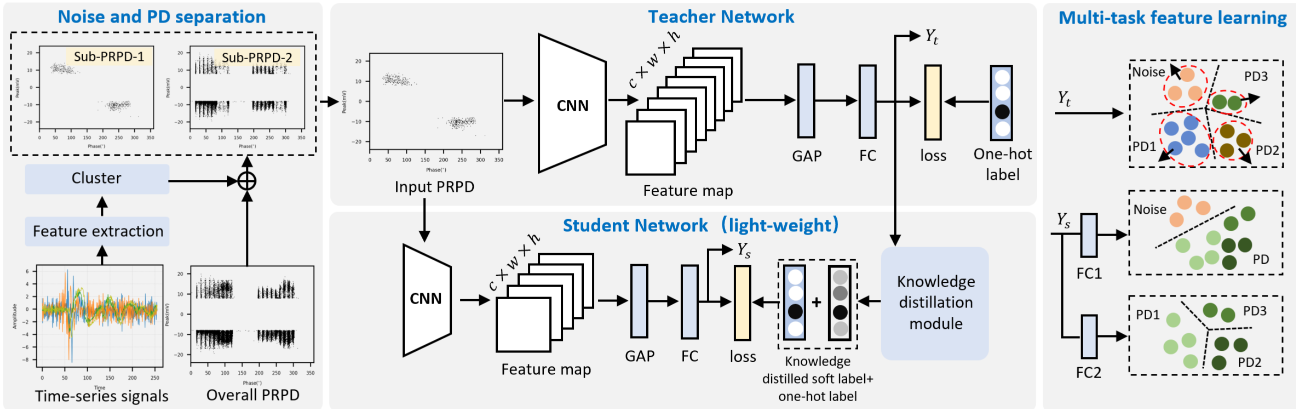

Fig. 2. Framework of knowledge distillation and multi-task feature learning for partial discharge recognition. Multiple sub-PRPD patterns are obtained according to the signal clustering results. Light-weight student network is guided by the huge teacher network to learn more effective knowledge. Multi-task classifiers are trained to distinguish between noise and PDs as well as classify different types of PDs simultaneously. FC and GAP represent global average pooling and fully connected layers respectively.

proposed to learn more robust and generalized features from the PRPD patterns.

II. Methodology

A. PRPD Pattern Preparation

The overall framework is demonstrated in Fig. 2. As the power substation is a complex environment with diverse electrical interferences, it is challenging to recognize partial discharge accurately due to the presence of various noises within the signals captured by the PD monitoring system. Consequently, eliminating the interference of noise on PD identification is the primary task. Based on this fact, frequency transform using Eq.(1) is used to represent the captured signals.

$$X(k) = \sum_{n=0}^{N-1} x_n e^{-\frac{i2\pi}{N}kn}, \tag{1}$$

where $X(k)$ is the kth element of the DFT features vector, N is the length of pulse, x_n is the nth element of the time-series signal. Then, a k-means based clustering algorithm is applied to separate signals into different groups according to their characteristics. With the help of frequency features X, each clustered signal group can be regarded as single-source signals with similar frequency distribution.

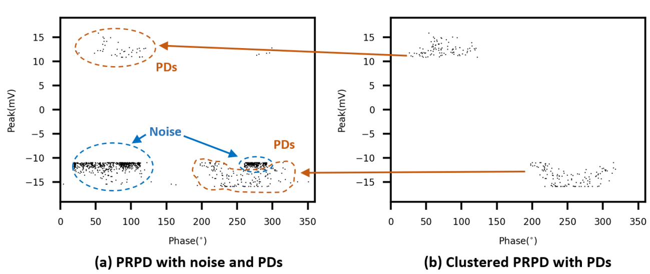

Fig. 3. Frequency feature based clustering for noise and PD signals separation.

As shown in Fig.3, PD signals and noise are separated and clean PRPD patterns can be generated for further PD recognition. In total, we collect 2924 PRPD patterns containing noise(2104), corona(177), internal(485) and surface(158). These PRPD patterns are manually checked and labeled for PD recognition.

B. Knowledge Distillation for PD recognition

The overall framework is presented in Fig. 2. We binary the PRPD patterns and normalize them into image format with a size of 64×64. Besides, we also design PRPD pattern data augmentation strategy as demonstrated in Fig. 4 (a). These augmented PRPD patterns can be used to train the deep neural network for PD recognition task.

$$\begin{aligned}\mathcal{L}_t = & -\sum_{i=1}^{m} \log p_{y_i}^t \\ & + \lambda \sum_{i=1}^{k} max\{d(\boldsymbol{a}_i, \boldsymbol{p}_i) - d(\boldsymbol{a}_i, \boldsymbol{n}_i) + mg, 0\},\end{aligned} \tag{2}$$

where m is the number of one batch samples, $p_{y_i}^t$ is the predicted results corresponding to the ground truth class y_i. where k is number of triplet pairs, $d(\boldsymbol{a}, \boldsymbol{p})$ is the ℓ^2-norm distance between the anchor $\boldsymbol{a}$ and a positive sample $\boldsymbol{p}$, $\boldsymbol{n}$ is a negative sample. λ is a parameter to balance the two losses.

$$\mathcal{L}_s^{k1} = -\frac{1}{c}\sum_{i=1}^{c}(y_i + \gamma \widehat{y}_i^t) \log \widehat{y}_i^s, \tag{3}$$

where $\widehat{y}_i^t$ is the output of the teacher network which acts as a soft label to guide the student network. γ is a hyperparameter to balance the soft label $\widehat{y}_i^t$ and one-hot ground label y_i.

C. Multi-task Feature Learning

In addition to the joint learning with the teacher network, our proposed student network are responsible for another two classification tasks, that is, to judge whether it is PD or not, and the other is to distinguish different types of PDs. The loss functions for the two tasks are defined as in Eq. 4.

$$\mathcal{L} = \mathcal{L}_t + \mathcal{L}_s^{k1} + \frac{-\alpha}{\vec{c}}\sum_{i=1}^{\vec{c}} \vec{y}_i \log \vec{y}_i^s + \frac{-\beta}{\widetilde{c}}\sum_{i=1}^{\widetilde{c}} \widetilde{y}_i \log \widetilde{y}_i^s, \tag{4}$$

where $\mathcal{L}_t$ and $\mathcal{L}_s^{t1}$ are the cross-entropy based classification loss for the teacher and student networks. $\vec{c}$ is 2 indicating two classes of PD and noise, $\vec{y}_i^s$ is the predicted results of FC1 in Fig. 2 for class i. $\widetilde{c}$ is the class number of PDs, $\widetilde{y}_i^s$ is the predicted results of FC2 in Fig. 2 for class i. α and β are two hyperparameters to balance the loss of task $t2$ and $t3$.

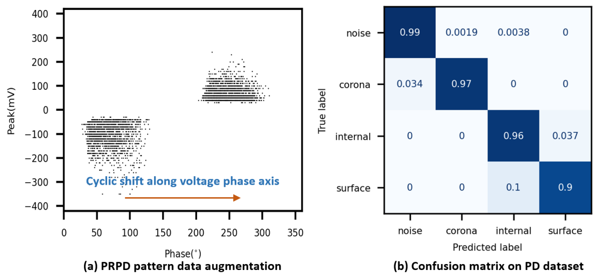

Fig. 4. (a) Demonstration of the data augmentation strategy of the PRPD pattern for deep neural network training. (b) Confusion matrix on the testing set achieved by the proposed method with knowledge distillation and multi-task feature learning.

III. Experiments

A. Implementation Details

The InceptionV3 [10] and light-weight MobileNetV3 [11] are adopted as the backbones of our teacher and student networks respectively. The dataset is split randomly into the training/validating/testing set (30%/20%/50%). During the testing stage, student network will output the prediction without relying on the teacher network. We padding the input image to meet the size requirements of the network input. We set the hyperparameters λ, γ, α, and β to 0.3, 0.06, 0.6, and 0.8 respectively. The initial learning rate is configured to 0.01. In addition to comparing with other methods as the listed methods in Table I and perform experiments, the effectiveness of data augmentation(AUG), knowledge distillation(KD) and multi-task feature learning(MTL) are investigated on our constructed PD dataset. All experiments are implemented using the deep learning framework of PyTorch on server computer with a Nvidia A6000 GPU.

TABLE I
Comparison with some state-of-the-art methods on PD recognition dataset.

Method	Backbone	size	Acc
Ref [7]	Customized	64×64	88.9%
Ref [8]	Customized	64×64	89.6%
MobileNetV3 [11]	MobileNetV3	64×64	90.4%
Ref [9]	InceptionV3	64×64	92.1%
Teacher	InceptionV3	64×64	93.1%
Teacher+AUG	InceptionV3	64×64	97.4%
Student+AUG	MobileNetV3	64×64	93.9%
Student+AUG+KD	MobileNetV3	64×64	97.1%
Student+AUG+KD+MTL	MobileNetV3	64×64	**98.2**%

B. Experimental Results Analysis

Experimental results are demonstrated in Table I. In [7] and [8], a self-defined network is used to extract features from the PRPD patterns. Work [9] also uses InceptionV3 [10] as backbone network to handle the PD recognition task. Compared with MobileNetV3 [11], our student model also adopts the same backbone network and achieves the highest accuracy of 98.2% with the help of knowledge distillation and multi-task feature learning modules. As depicted in Fig. 4 (b), most of the noise and PD patterns are correctly classified and only some PRPD patterns belonging to internal and surface PDs are confused. That means our knowledge distillation framework can learns effective knowledge from the teacher network and help the student network achieve more accurate predictions.

IV. Conclusion

The objective of this study is to transfer knowledge obtained from a complex teacher network to a light-weight student network for accurate PD recognition. Initially, a k-means clustering model is employed to differentiate signals originating from distinct sources, resulting in the extraction of Phase Resolved Partial Discharge patterns. Subsequently, a knowledge distillation technique and a multi-task feature learning framework are employed to extract discriminative features from the PRPD patterns, enabling precise PD type prediction. Experimental results and comparison with some state-of-the-art methods on our PD dataset evaluate the effectiveness of the proposed method.

Acknowledgment

This research is supported by SP Group and the National Research Foundation, Singapore, and the Energy Market Authority under its Energy Programme (EP Award EMA-EP010-SNJL-004) and Nanyang Technological University.

References

[1] Karimi, Masoud, et al. "A novel application of deep belief networks in learning partial discharge patterns for classifying corona, surface, and internal discharges." IEEE Transactions on Industrial Electronics 67.4 (2019): 3277-3287.

[2] Lin, Qingcheng, et al. "Optimized Denoising Method for Weak Acoustic Emission Signal in Partial Discharge Detection." IEEE Transactions on Dielectrics and Electrical Insulation 29.4 (2022): 1409-1416.

[3] Balouji, Ebrahim, Thomas Hammarström, and Tomas McKelvey. "Classification of Partial Discharges Originating From Multilevel PWM Using Machine Learning." IEEE Transactions on Dielectrics and Electrical Insulation 29.1 (2022): 287-294.

[4] Ganguly, Biswarup, et al. "Wavelet kernel-based convolutional neural network for localization of partial discharge sources within a power apparatus." IEEE Transactions on Industrial Informatics 17.3 (2020): 1831-1841.

[5] Song, Hui, et al. "GIS partial discharge pattern recognition via deep convolutional neural network under complex data source." IEEE Transactions on Dielectrics and Electrical Insulation 25.2 (2018): 678-685.

[6] Tharamal, Lakshmi, P. Preetha, and T. K. Sindhu. "Assessing the Degradation of Transformer Oil From Partial Discharge Measurement Data Using Histogram Similarity Measures." IEEE Transactions on Instrumentation and Measurement 71 (2022): 1-10.

[7] Kitani, Ryota, and Shinya Iwata. "Brute-force analysis of insight of phase-resolved partial discharge using a CNN." Electrical Engineering (2023): 1-10.

[8] Florkowski, Marek. "Anomaly Detection, Trend Evolution, and Feature Extraction in Partial Discharge Patterns." Energies 14.13 (2021): 3886.

[9] Raymond, Wong Jee Keen, et al. "Noise invariant partial discharge classification based on convolutional neural network." Measurement 177 (2021): 109220.

[10] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE CVPR. 2015.

[11] Howard, Andrew, et al. "Searching for mobilenetv3." Proceedings of the IEEE/CVF ICCV. 2019.