Outage Cause Classification of Power Distribution Systems with Machine Learning and Real-World Data

Haoyuan Sun, Fangxing Li
Deptartment of EECS
University of Tennessee at Knoxville
Knoxville, TN, USA

Christopher Sticht, Srijib Mukherjee
Power Systems Resilience Group
Oak Ridge National Laboratory (ORNL)
Oak Ridge, TN, USA

***Abstract*—Power distribution systems are geographically dispersed by nature. It may be affected by various factors, such as vegetation, weather, animal and human behaviors. Present response procedures to an outage event massively rely on expert experience and thus tend to be time-consuming. Automatic outage event detection and classification will help to reduce the responding and restoration time. However, this issue is less addressed with existing research done in this area. In this applied research, a set of waveform pre-processing techniques are first proposed to prepare the waveform data for being used as inputs to the classification algorithm. Further, a machine learning-based algorithm is proposed to classify the outage events according to their root causes, e.g. tree contact, animal contact, lightning, etc. Available data include three phase current & voltage waveforms and contextual information during the distribution system outages. The proposed machine learning algorithm takes the current and voltage waveforms as direct inputs in search of features that humans are unable to capture. Real data provided by a distribution company in the East Tennessee region is used to test the proposed pre-processing techniques and the classification algorithm.**

***Index Terms*—Waveform pre-processing, outage cause classification, machine learning, neural network, distribution power system.**

I. INTRODUCTION

Power distribution systems tend to be affected by various factors, such as vegetation, weather, animal, and human behaviors, because of its geographically dispersed nature. Therefore, a common task at distribution companies is to identify the root cause of an outage event. This information could provide crucial guidance on how to deal with the outage and thus help clear the outage within a short time.

In present practice, outage data are checked manually, which is time-consuming and causes significant delay between outage detection and system recovery. In addition, the amount of anomalous waveform data is often large such that most of them is left unprocessed. An automated algorithm can save significant manpower, help with timely system recovery, and make better use of the great amount of data.

This work is to develop a machine learning-based algorithm to process outage event recordings and to identify the cause of each outage event based on real-world data provided by a distribution company in the Eastern Tennessee region.

The available data set is a collection of outage event recordings. Each of these recordings has a cause label. Therefore, this problem can be formulated as a classification task in which we try to categorize each outage event into one of the existing cause classes. An event cause classification algorithm can be trained on the past outage event recordings such that it should be able to identify the cause of new outage events.

In the field of power system outage event analysis, there are a large number of works focusing on affected phase identification [1][2][3]. However, there are not many works devoted to the root cause identification. Very few of these existing works directly process the voltage and current waveforms. Instead, they use contextual information [4][5] or extracted features from the waveform data [6][7]. Contextual information includes weather, affected phase(s), season, event time, and interrupting device, etc. Extracted features from waveform data include derivative of current and voltage signal, energy, amplitude, correlation coefficient, etc.

For example, in [8], five features, namely self-recoverability, zero current time, degree of distortion, transition time (time duration from event occurrence to during-event stable stage), and waveform randomness, are extracted from the recorded waveforms and used as the input to a fuzzy inference system for event cause recognition.

Features that are extracted from waveform data are manually and intentionally designed. As a result, these features tend to work well only on purposes that they have been designed for but are less effective in other circumstances. However, with the ever-growing complexity and uncertainty in modern power systems, unseen circumstances are emerging continually. To address these unseen circumstances, a practical solution is to identify some general features. In fact, the waveforms are features themselves and should be more informative than any manually designed features. In this sense, it should be beneficial to include the waveform data as inputs to an event classification algorithm, as a crucial complement alongside the manually designed features.

A significant reason why contextual information and manually extracted features from waveforms were used instead of waveforms themselves is that PMU data cannot offer sufficient waveform details. Its sampling rate is typically around 60Hz, which means only one sample per grid cycle is recorded. Using waveforms as a direct source to analyze the cause of outage events would require much higher resolution event recordings than PMU data. The necessity of high-resolution event recordings has already been realized [9] and has been implemented in some power systems [10]. These recordings with high resolution are called point-on-wave (POW) data, featuring a sample rate of 1 kHz or even higher. Unlike PMU data, POW data offer much more waveform details including

instantaneous phase shift and distortions, and thus can be used as a direct source for outage event cause analysis.

To fill this gap in existing research, this paper works on the identification and classification of root causes of distribution power system outage events. Real POW current and voltage waveforms are directly processed with the proposed algorithm.

Although real data are realistic and informative, they usually come with noise. Therefore, pre-processing is necessary before any analysis could be carried out with these data. First, real data may contain redundant segments which does not help with event cause identification. To locate the useful segments in an event recording, a normalized RMS envelope method is adopted in [11] to detect the start and end points of an event. This paper presents a new approach to deal with this issue. Second, there are repetitive recordings caused by a single event being recorded simultaneously by different devices. This could become a contamination to the training set and test set if not handled properly. This paper presents a method to remove the repetitive recordings. Third, the data may not be sufficient in quantity to train a machine learning algorithm. This paper proposes a waveform data augmentation method to enlarge the dataset.

Convolution neural network (CNN) has achieved huge success in image processing. It can be also used to process time-series data, which can be thought of as 1-D images. The approach of 1-D CNN is used in [12] on waveform data for fault location. Artificial neural network (ANN) is suitable for processing categorical and numerical features. ANN is used to deal with contextual information and thus to differentiate between tree-caused events and animal-caused events [13]. This paper uses CNN and ANN as the backbone of the proposed event cause classification algorithm, in which CNN processes the current and voltage waveform data and ANN processes a group of contextual features.

The main contributions of this paper are: (1) a set of waveform data pre-processing techniques to deal with redundant waveform segments, repetitive recordings, and data insufficiency; (2) an outage cause classification algorithm based on CNN and ANN that takes three phase current and voltage waveforms as direct inputs.

This paper is organized as follows: Section II provides an overview of the dataset used in this work; Section III introduces a set of waveform data pre-processing procedures; Section IV presents the proposed outage cause classification algorithm; Section V tests the performance of the proposed algorithm on a dataset of real event recordings; Finally, Section VI concludes the paper.

II. Overview of the Dataset

The data used in this work are real event recordings of year 2020 provided by a local distribution company in the Eastern Tennessee region [10]. The dataset has over three thousand recordings in total. Each recording comes with the exact date and time of the event, and three phase current and voltage waveforms. All waveforms are POW data with a sample rate of 3840 Hz, which means 64 data points per grid cycle are recorded. The current and voltage waveforms of each recording are around 0.5s in length , i.e. 30 system cycles, and contain pre-event, during-event, and post-event stages. Note that the start and end points of these stages are not marked explicitly and need to be identified. Each record also has a label indicating the cause of this event, although it is inferred by some linemen examining the scene of the event and thus could be inaccurate. Possible causes include tree contact, animal contact, lightning, equipment failure, and weather. Some causes only have very few occurrences, and thus do not have sufficient data to train a machine learning algorithm. In this work, causes that have more than a hundred occurrences in the dataset are retained for further analysis.

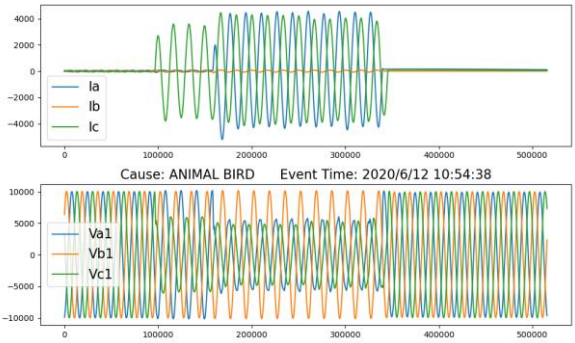

Fig. 1. Example of an event recording.

Fig. 1 gives an example of an event recording. The current and voltage data are plotted to clearly show the event. In this example, first, a single-phase fault occurred on phase C and then it developed into a double-phase fault involving phase A-C. This event occurred on 6/12/2020 at 10:54:38, and was caused by an animal, a bird to be exact, as indicated by the cause label.

III. Waveform Pre-Processing

The purpose of the waveform pre-processing procedures introduced in the following subsections is not to extract any manually designed features, but rather to technically prepare the waveform data for being used as direct inputs to the classification algorithm.

A. Waveform Truncation based on Anomaly Detection

A typical recording contains pre-event, during-event, and post-event stages. The pre-event and post-event stages correspond to normal states or operation after the isolation of faulted grid sections, and thus usually have steady sinusoidal waveforms. In contrast, the during-event stage corresponds to on-going disturbances and faults, and thus have unsteady or non-sinusoidal waveforms.

To help the algorithm better focus on the during-event stage, we identify the start and end points of an event with the residual component method [14] and the Circular Trajectory Approach [16]. The residual component method is to superimpose a cycle of waveform onto its previous cycle, and the difference is the residual component. The Circular Trajectory Approach is to form a circular trajectory in an x-y plane with a sinusoidal signal itself and its derivative.

The start point of an event is set as the first sample point in a recording that has a residual greater than a threshold, indicating that an anomaly arises. The end point of an event is

set as the sample point that meets one of the following conditions, whichever comes earlier:

(1) It is the last sample point that has a residual greater than a small threshold, indicating that the system is back to normal operation from there on; or

(2) It is the last sample point that has a circular trajectory with a radius greater than a small threshold, indicating that this section of the grid is isolated from the main system from there on.

This approach of determining the end point ensures that the opening and reclosing processes of the breaker are not included in any of the samples. Two cycles of the waveform before the start point and after the end point are also included in the truncated samples, to represent normal system states and in turn to assist the algorithm with capturing the transition between normal and abnormal states.

B. Removal of Repetitive Waveforms

A single event could be recorded simultaneously by several devices from different locations. These event recordings could be very similar to each other. If some of them are in the training set while others are in the test set, the algorithm will be able to correctly classify the test samples simply because it has seen similar ones during the training phase, instead of truly extracting critical features out of the event recordings. This situation is obviously undesired, and can be avoided by only selecting one sample out of a group of recordings that are very similar to each other.

To find similar recordings, a measure of waveform similarity is needed. Fig. 2 gives an example of two highly similar recordings. Measuring the similarity of waveforms like these directly with, for example, Euclidean distance-based methods will not work well, because these recordings are similar but not exactly the same. They are different in amplitude because the two recording devices have different electrical distance from the event site.

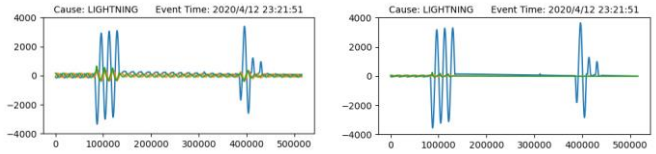

Fig. 2. Example of two similar recordings.

In this work, normalization on a per half cycle basis is performed to elminate the amplitude difference among recordings. Every data point is divided by the maximum absolute value of the half cycle it belongs to. This approach of normalization is inspired by the scaling transformation used in [15]. Note that the normalization procedure described above is only for waveform similarity measurement purpose.

Waveform similarity check is then performed on the normalized data with a three-second moving time window. For waveforms that are recorded no more than three seconds apart, Euclidean distance is calculated between each pair of them. Two waveforms are determined as similar if their Euclidean distance is less than a threshold.

A practical issue with recordings of the same event from different devices is that they may not be perfectly aligned along the time axis. For example, the recording shown in Fig. 2 on the left is two sample points ahead of the one on the right, in terms of event start point. If they are compaired directly, they would be marked as dissimilar, which is incorrect. After examining hundreds of event recordings, it is observed that this lead or lag along the time axis is typically only a few sample points. Therefore, to address this waveform alignment issue, a waveform shifting operation is designed to assist the waveform similarity check, as shown in Fig. 3. That is, while we compare two waveforms A and B, A is shifted by x sample points, where $x > 0$ means A is shifted to the left and vice versa. For every integer x in [-16, 16], an Euclidean distance between A and B is calculated. The x value that leads to the minimum distance between A and B is thus the point of alignment. Therefore, this minimum distance is used as the similarity score between waveforms A and B.

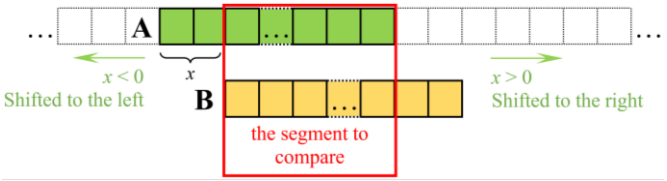

Fig. 3. Waveform shifting operation.

In a group of similar recordings, the ideal one to retain is the one from the device that has the least electrical distance from the event site, which is the most informative recording. Although it is usually hard to access the network topology due to security concerns of distribution companies, the waveform itself can also reveal the electrical distance between the device and the event site. The recording of a device that is closer to the event site should have larger voltage drop during an event. Consequently, in a group of similar recordings, the one with the largest voltage drop is retained, while others are excluded.

C. Waveform Normalization

The recordings in the dataset have various current levels due to different loading conditions. Thus, the absolute amplitude of the waveforms does not offer much information regarding the cause of the outage event. Useful information for event cause classification is rather contained in the shape and the fluctuations of the waveforms. In addition, data with unified amplitude work better for machine learning algorithms. Therefore, normalization is performed to remove the amplitude differences while preserving the shape and fluctuations of the waveforms.

In this work, each truncated event recording is normalized by dividing the entire sequence with the amplitude of the first cycle, i.e. difference of its largest value and smallest value in the first cycle. As such, all normalized samples start with a unit amplitude. This makes it easier for the algorithm to compare different samples. Note that this normalization procedure is for preparing the waveform data for the classification algorithm and therefore is different from the normalization procedure described in subsection III-B.

D. Waveform Data Augmentation

Outage event data are hard to obtain because there are not many outage events. Especially when compared with the large

number of images available for training an image classification algorithm, the dataset used in this work is quite small for a classification task. To augment the dataset and obtain more training data, the following two techniques are developed: (1) Flipping the waveform, as shown in Fig. 4. This technique will double the data. Note that this operation does not change the phase sequence. (2) Alternating the phase sequence of an event recording. Specifically, if the original phase sequence is ABC, then two other permutations of ABC, that is BCA and CAB, can be used to generate new data, as shown in Fig. 5. The other three permutations, ACB, BAC, and CBA, are not used because their phase rotation is opposite to ABC. This technique will triple the data. With these two techniques, the original dataset is enlarged by a factor of six.

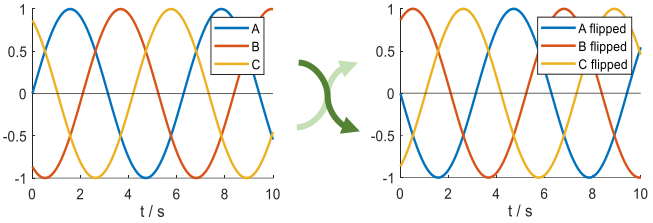

Fig. 4. Waveform data augmentation technique 1 – flipping the waveform.

These two techniques have physical interpretations when applied to three-phase current and voltage waveforms. An anomaly may occur during either the positive half cycle or the negative half cycle of a sinusoidal waveform. Also, it may occur on any of the phase(s). Therefore, by flipping a waveform or alternating the phase sequence, the generated waveforms are still realistic.

Note that these two techniques for data augmentation are performed after the partition of the training set and the test set, and are only performed on the training set. This is to ensure that no test data is seen by the algorithm in the training process either in its original or transformed version, for the validity of the training and test results.

E. Event Duration and Affected Phase(s) Identification

The duration of an event is obtained by taking the time interval between the start point and the end point acquired in subsection III-*A*. All event duration data are normalized by dividing the largest length of all recordings.

Then, any of the A, B, and C phases is marked as affected if the residual component of its current or voltage waveforms is greater than a threshold within the duration of an event.

IV. Waveform Classification

A. Feature Selection

In this work, current and voltage waveforms are used as direct inputs to the algorithm after being pre-processed as introduced in Section III. At the same time, event duration, affected phase(s), season, and daytime / nighttime are selected as contextual features. Season, daytime / nighttime, and cause labels are textual categorical data, and thus are converted to numerical data, which can be processed by a machine learning algorithm, with one-hot encoding. The event durations and affected phase(s) are identified, as introduced in subsection III-*E*. Event cause labels are used as the ground truth for supervised learning.

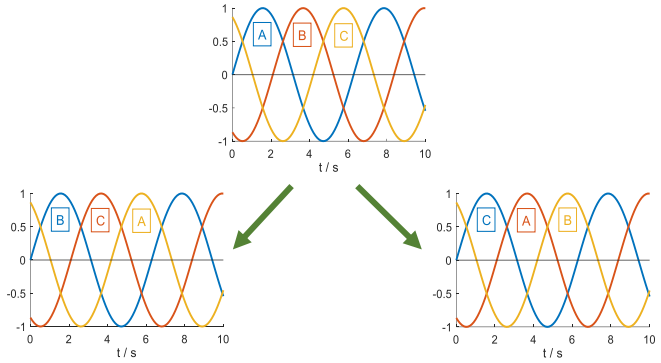

Fig. 5. Waveform data augmentation technique 2 – alternating phase sequence.

B. Proposed Classification Algorithm

A convolutional neural network (CNN) and an artificial neural network (ANN) are used to construct a classification algorithm and to process the aforementioned waveforms and contextual features. The goal here is to identify signature patterns for specific event causes, which may appear at any position in a continuous waveform recording. Theoretically, a signature pattern should always correspond to the same event cause, independent of its position in a waveform. Consequently, CNN is considered suitable for processing the current and voltage waveform data because CNN has the unique properties of local connectivity and spatial invariance. In addition, ANN is a series of fully connected layers (FC) and is thus suitable for processing contextual features.

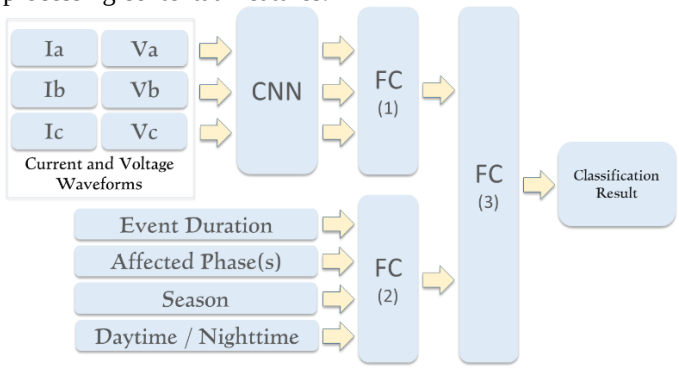

Fig. 6. Overall structure of the proposed neural network.

The overall structure of the proposed neural network is shown in Fig. 6. Its detailed structure is given in Table I. The three-phase current and voltage waveforms, after pre-processing, are first passed to a group of parallel 1-D convolutional layers, where the three-phase current waveforms share the same kernels and weights, and the three-phase voltage waveforms share another group of kernels and weights. The outcomes are concatenated and then passed through a larger convolutional layer and two fully connected layers to correlate features found in the current and voltage waveforms and all three phases. In addition, four contextual features, namely event duration, affected phases(s), season, and daytime / nighttime, are passed through two fully connected layers. Finally, a fully connected layer synthesizes waveform and contextual information to obtain the classification result.

Table I. Detailed Structure of the Proposed Neural Network

Block Name	Layer Name	Structure
CNN	Conv1	4×1, 8, stride 1
		4×1, 8, stride 1
		4×1, 8, stride 1
		4×1, 8, stride 1
		4×1, 8, stride 1
		4×1, 8, stride 1
	Conv2	9×4, 16, stride 2
FC (1)	FC1_1	40
	FC1_2	20
FC (2)	FC2_1	20
	FC2_2	10
FC (3)	FC3_1	10
	FC3_2	Number of cause classes

The neural network is trained for 25 epochs with a batch size of 10. The optimizer is Adam (learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$). The loss function is categorical cross entropy.

V. Performance of the Proposed Algorithm

In this paper, we work on a binary classification task as an initial test of the proposed pre-processing techniques and the classification algorithm. Event recordings with cause labels of 'lightning' and 'high wind' are used. All tests are conducted with 5-fold cross validation. For each fold, 80% data are for training and 20% are for testing. This percentage refers to the dataset before augmentation.

A. Classification Accuracy

The performance of the proposed algorithm is compared with k-nearest neighbors (k-NN) and ANN. All experiments in this subsection are conducted with the augmented dataset. The results are shown in Table II.

Table II. Performance of the Proposed Algorithm

Features	Algorithm	Classification Accuracy		Variance
		Best	Average	
Contextual	k-NN (k=3)	90.24%	83.90%	10.35
	ANN	87.80%	83.17%	9.16
Waveform + Contextual	proposed CNN + ANN	**91.46%**	**86.82%**	**6.19**

By using current and voltage waveforms as direct inputs alongside contextual features, the proposed algorithm outperforms k-NN and ANN and achieves higher classification accuracy in event cause identification. The results of the proposed algorithm have a smaller variance, which means it achieves better consistency.

B. Contribution of the Augmented Dataset

Table III. Contribution of the Augmented Dataset

Training Dataset	Classification Accuracy		Variance
	Best	Average	
Original dataset	87.80%	83.90%	7.97
Augmented dataset	**91.46%**	**86.82%**	**6.19**

To test the effectiveness of the proposed data augmentation techniques, the classification algorithm is trained with only the original dataset or only the augmented dataset. The result, as shown in Table III, indicates that the classification accuracy is improved by the augmented dataset. Algorithm consistency is also improved, as indicated by a smaller variance.

VI. Conclusion and Future Work

This paper proposed a set of waveform data pre-processing techniques to deal with redundant waveform segments, repetitive recordings, and data insufficiency. This paper also proposed a machine learning-based outage event cause classification algorithm. based on CNN and ANN. CNN takes the current and voltage waveforms as direct inputs while ANN processes the contextual features and synthesizes the waveform and contextual information to obtain the final classification result. The proposed pre-processing techniques and the classification algorithm are tested on real data provided by a distribution company in the Eastern Tennessee region.

Event recordings with other cause labels could be investigated in future works.

References

[1] M. F. Guo, N. C. Yang, and W. F. Chen, "Deep-Learning-Based Fault Classification Using Hilbert-Huang Transform and Convolutional Neural Network in Power Distribution Systems," *IEEE Sens. J.*, vol. 19, no. 16, pp. 6905–6913, 2019.

[2] K. Zhu and P. W. T. Pong, "Fault Classification of Power Distribution Cables by Detecting Decaying DC Components with Magnetic Sensing," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2016–2027, 2020.

[3] V. Ashok and A. Yadav, "Fault Diagnosis Scheme for Cross-Country Faults in Dual-Circuit Line with Emphasis on High-Impedance Fault Syndrome," *IEEE Syst. J.*, vol. 15, no. 2, pp. 2087–2097, 2021.

[4] X. Jiang, S. M. Ieee, B. Stephen, S. M. Ieee, and S. Mcarthur, "Automated Distribution Network Fault Cause Identification with Advanced Similarity Metrics," vol. 8977, no. c, pp. 1–8, 2020.

[5] M. S. Bashkari, A. Sami, and M. Rastegar, "Outage Cause Detection in Power Distribution Systems Based on Data Mining," *IEEE Trans. Ind. Informatics*, vol. 17, no. 1, pp. 640–649, 2021.

[6] M. A. Jarrahi, H. Samet, and T. Ghanbari, "Novel Change Detection and Fault Classification Scheme for AC Microgrids," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3987–3998, 2020.

[7] A. Hooshyar, E. F. El-Saadany, and M. Sanaye-Pasand, "Fault Type Classification in Microgrids Including Photovoltaic DGs," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2218–2229, 2016.

[8] Y. L. Liang, K. J. Li, Z. Ma, and W. J. Lee, "Typical Fault Cause Recognition of Single-Phase-to-Ground Fault for Overhead Lines in Nonsolidly Earthed Distribution Networks," *IEEE Trans. Ind. Appl.*, vol. 56, no. 6, pp. 6298–6306, 2020.

[9] A. Silverstein, A. S. Consulting, J. Follum, A. Silverstein, A. S. Consulting, and J. Follum, "High resolution, time synchronized, grid monitoring devices," *Naspi*, 2020.

[10] EPB Chattanooga, *System Wide S&C IntelliRupter® Waveform Captures, Shared under NDA*. 2020.

[11] A. J. Wilson, D. R. Reising, R. W. Hay, R. C. Johnson, A. A. Karrar, and T. D. Loveless, "Automated Identification of Electrical Disturbance Waveforms within an Operational Smart Power Grid," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 4380–4389, 2020.

[12] M. F. Guo, J. H. Gao, X. Shao, and D. Y. Chen, "Location of Single-Line-to-Ground Fault Using 1-D Convolutional Neural Network and Waveform Concatenation in Resonant Grounding Distribution Systems," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021.

[13] L. Xu and M. Y. Chow, "A classification approach for power distribution systems fault cause identification," *IEEE Trans. Power Syst.*, vol. 21, no. 1, pp. 53–60, 2006.

[14] B. Li, Y. Jing, and W. Xu, "A Generic Waveform Abnormality Detection Method for Utility Equipment Condition Monitoring," *IEEE Trans. Power Deliv.*, vol. 32, no. 1, pp. 162–171, 2017.

[15] A. F. Bastos and S. Santoso, "Universal Waveshape-Based Disturbance Detection in Power Quality Data Using Similarity Metrics," *IEEE Trans. Power Deliv.*, vol. 35, no. 4, pp. 1779–1787, 2020.

[16] H. Sun, F. Li, C. Sticht, S. Mukherjee, "Circular Trajectory Approach for Online Sinusoidal Signal Distortion Monitoring and Visualization," *IEEE PES Letters*, under review.