IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 19, 2019, accepted January 17, 2020, date of publication February 4, 2020, date of current version February 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2971582

Stacked Auto-Encoder-Based Fault Location in Distribution Network

GUOMIN LUO, YINGJIE TAN, MENG LI, MENGXIAO CHENG, YANMEI LIU, AND JINGHAN HE

School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Meng Li (mengl@bjtu.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2018JBM056, in part by the National Natural Science Foundation of China under Grant 51507008, in part by the China Postdoctoral Science Foundation under Grant 2019M660436, and in part by the State Grid Corporation Technology Project under Grant 5100-201955018A-0-0-00.

ABSTRACT Accurate fault location is crucial for finding and clearing faults in distribution networks. It can help to reduce the loss of power failure and ensure safe and stable operation. Without efficient observation points and the lack of phase angle information, traditional fault location methods of distribution networks usually have large errors. As the application of μPMUs in the distribution network becomes more and more common, voltage and current waveforms, as well as phase angle can be recorded and provide more information for accurate fault location. In this paper, an intelligent location method is proposed to pinpoint the fault location based on the information of μPMUs which are properly allocated. The fault section is firstly determined by comparing the zero-sequence current waveforms on both sides of the fault section. Then, a Stack Auto-Encoder (SAE) is modeled to provide an end-to-end solution to pinpoint fault point with the voltage and current phasors. Finally, the performance of the proposed method is tested by a simulated distribution network on the platform of PSCAD. The results show that the proposed method is effective in locating faults and can withstand the effects of transition resistance, fault type, and noise. Compared to another popular method, the proposed method shows better location accuracy.

INDEX TERMS Fault location, stacked auto-encoder, distribution network.

I. INTRODUCTION

The distribution network connects power transmission network and power loads, and directly provides the power to users [1]. In some countries such as China, distribution systems are ungrounded or grounded with Petersen coil. When a single-phase grounding fault occurs, the phase voltage of the faulted phase will drop and only a small fault current will be produced. As only current relays are installed in distribution networks, the fault features are not obvious, and the current relay does not operate immediately. The phase-to-ground voltage of non-faulted phases will rise and might lead to insulation breakdown and serious faults. Therefore, quick and reliable determination of the fault point is critical for the maintenance and operation of the distribution network.

Fault location in the distribution network is always a tough job and faces a lot of challenges. The first challenge is the complicate operating condition. The topologies of distribution networks are complicated and change with operation status [2]. Transmission lines are short, and multiple branches are included. It is difficult to discriminate which section or part the fault occurs in. The second challenge is the unobvious fault feature. Since the neutral point of the transformer is ineffectively grounded, the post-fault currents are small. With such a slight difference, to pinpoint the exact fault position is not an easy job. Besides, limited measurement information is also a challenge. Usually, only currents are recorded for feeders. Voltages are only detected on buses. Phase information is missed in distribution networks.

Although precise fault location in the distribution network is a tough job, a lot of research work has been done to solve this problem [3], [4]. Those methods include the injection-based methods, the steady-state based ones, and the transients-based ones. The injection-based methods are usually offline locations. A signal is generally injected into the faulted system. The fault location could be calculated with the reflection or transition of the injected signal. But this kind of method requires additional injection equipment which is uneconomic [5]. Also, if the topology of the faulted system is complicated or transmission lines are long,

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

the injected signals may become too weak to be detected. Furthermore, the location accuracy of the injection-based method is affected by distributed capacitance, grounding resistance and other factors. Steady-state methods adopt voltages and currents to calculate fault locations iteratively based on known topologies and parameters of distribution network, for example, the interconnections, branches, line parameters, and power of loads. But its location performance depends on the accuracy of known parameters [6]. It is ineffective in locating single-phase-to-ground faults. The transients-based methods utilize post-fault transient measurements. As more and more transient waveforms in the distribution networks can be obtained, the application and research of transients based fault location become more popular. Its location performance is reported to be effective in many research works. In reference [7,] [8], the fault is located with wavelet-based frequency characteristics as well as the topology of the distribution network. In reference [9], [10], the instants of sudden variations of post-fault transients are calculated with Hilbert-Huang transform, and the fault location is calculated with these instants and transients propagating velocities.

Electrical measurements, including injected pulse waves, steady-state phasors, and transient waveforms, can provide enough information for mapping the nonlinear relationship between electrical features and fault locations. Machine learning-based models have been adopted in pinpointing faults in distribution networks, for example, artificial neural networks (ANN), support vector machine (SVM), multilayer perceptron neural network (MLP-NN), extreme learning machine (ELM), and other intelligence algorithms [11], [12]. Effective performance was reported. In some research, the wavelet spectrums of voltage and zero-sequence currents are analyzed by SVM to identify possible fault zone and accurate fault distance based on the voltage drop data and the change of zero-sequence current during faults [13], [14]. Also, the wavelet characteristics of three-phase post-fault currents and voltages are used to train an ANN and determine the fault location on the distribution lines under various scenarios [15], [16].

However, traditional machine learning-based methods were not widely discussed and used. On the one hand, the features which are extracted on the basis of experts' experience might not fully represent the complicated, unobvious, but unique fault characteristics; on the other hand, the traditional machine learning models such as ANN and SVM are shallow learning ones. They cannot learn the complicated relationship between electrical measurements and fault locations. Their generalization is poor, so they usually have unsatisfied performance when the system parameters or operating environment changes [17].

Benefiting from the application of the microphase measurement unit (μPMU), the real-time dynamic monitoring of the distribution system can be realized [18]. μPMUs can provide both voltage and current measurements with phase and time information. With properly assigned μPMUs, the dynamic process of distribution network can be known and the fault zone or even fault point can be found out. Some research work has reported fault location in distribution network with μPMUs. For example, the PMU information on both ends of the faulted line is used in reference [19] to estimate fault points. Those research results show a potential application of μPMU-based fault location in distribution networks. But some problems such as error of line parameters and the effect of multiple branches have not been well solved. Further investigations are needed to adopt μPMU-based information in the distribution network fault location.

In order to improve the fault location accuracy in distribution networks, this paper proposes a location method based on μPMU measurements and stack auto-encoder (SAE). SAE is a kind of unsupervised learning algorithm, which can reduce the dimension of raw data and establish a stable representation of the original input. As the depth of the SAE network can be easily increased, SAE can effectively fit complicate functions and represent nonlinear features. It has been widely employed to solve big data problems or issues with strong uncertainties. In this paper, based on a suitable setting of μPMUs, the distribution network is segmented into different zones. Once a fault occurs, the faulted zone is firstly found out according to the similarity of zero-sequence waveforms provided by μPMUs. An SAE is then trained to learn the relationship between voltage-current-phase vectors and fault distances. Finally, the μPMU measurements on both ends of the fault zone are processed by the trained SAE to obtain the fault distance.

The rest of this paper is organized as follows. In Section II, the settings of μPMUs in the distribution network and the fundamentals of SAE are introduced. The SAE based location method is proposed on the basis of fault features analysis. The proposed method is implemented in Section III. Both data processing and training of SAE are discussed. A series of simulations are conducted to prove the effectiveness of the proposed method in Section IV. Finally, this paper is concluded in Section V.

II. PROPOSED METHOD

In order to find the location of the fault point, the fault zone should be determined first. Based on the analysis of fault features, fault location method will be proposed in this section.

A. RELATED ANALYSIS

1) DIVISION OF FAULT ZONE

Since unobvious current features will be generated, it is crucial to find the faulted zone in the distribution network in order to reduce the computation complexity. The assignment of μPMUs at all nodes in the distribution network is a good solution but an uneconomic one. Besides, it is impossible to process such a large amount of real-time and synchronous μPMU measurements at the same time. The communication capacity in the distribution network is limited. Therefore, the assignment of μPMUs should be balanced. More μPMUs

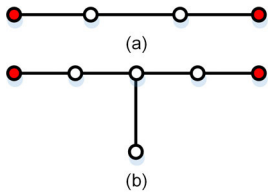

FIGURE 1. Topology of fault zones, (a) zone I, (b) zone II.

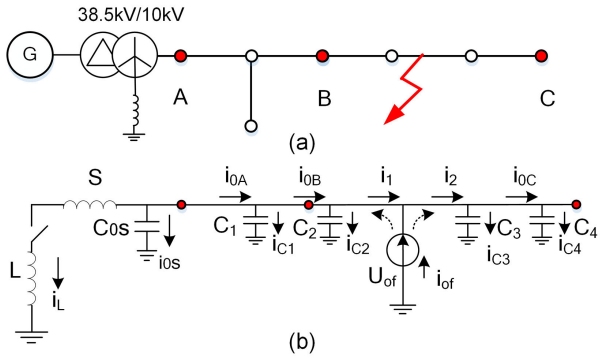

FIGURE 2. Characteristics of zero-sequence currents in distribution network, (a) equivalent circuit of distribution network, (b) zero-sequence currents in the system.

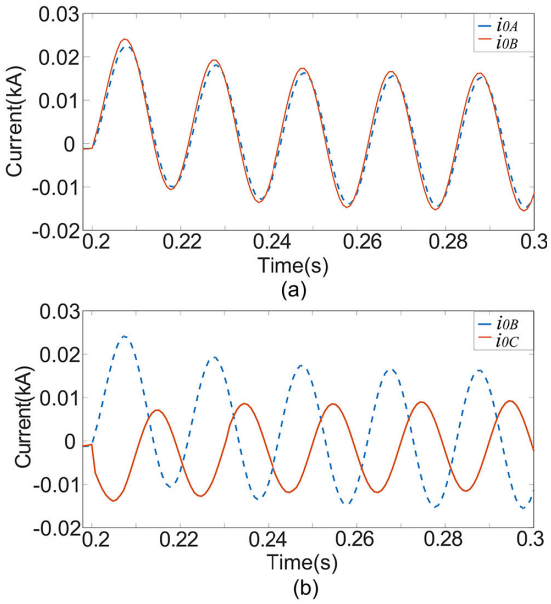

FIGURE 3. Comparisons between zero-sequence currents, (a) non-faulted zone, (b) Faulted zone.

are preferred in the aspect of obtaining useful information, while fewer ones are needed when considering communication bandwidth and economical operation. In this paper, the distribution network is divided to be different zones according to the following principles: 1) only two μPMUs are included in each zone at the boundaries. 2) less than 3 nodes (except boundary nodes) are contained in each zone. Therefore, the zone topologies of distribution networks can be divided into two main types: 1) zone I, as shown in Fig. 1(a), which has only one branch; and 2) zone II, as shown in Fig. 1(b), which includes "T" shape interconnection. In both kinds of zones, two μPMUs are installed at zone terminals, which are denoted by red dots in Fig. 1.

After the division of distribution network, fault location can be simplified by finding fault zone and pinpointing fault points within the faulted zone.

2) ZERO-SEQUENCE ANALYSIS OF FAULT CURRENT

After the occurrence of a single-phase grounding fault, the first thing is to find the faulted zone. As shown in Fig.2, the zero-sequence equivalent of a single-phase-to-ground fault in the distribution network is to add a zero-sequence fault component U_{0f} at the faulted point. Such component U_{0f} provides zero-sequence currents I_{0f} in this faulted system. According to the μPMU assignment principle in the previous section, the electrical measurements at both boundaries of each zone can be collected.

In the non-fault section AB and faulted section BC, the current relationships can be described as in equation (1) and (2).

$$i_{0A} - i_{0B} = i_{c1} \tag{1}$$

$$i_{0B} - i_{0C} = i_{0f} - i_{c2} - i_{c3} - i_{c4} \tag{2}$$

As the capacitance to ground, for instance, C_1, C_2, C_3 and C_4 in Fig. 2(b), in the distribution system, is very small, the current in capacitance i_{c1} should be small, and the zero-sequence currents at both boundaries should be similar if no fault components is included. But due to the existence of the zero-sequence component, such as i_{0f} in (2), the waveform and polarity of zero-sequence currents at both boundaries of the faulted zone will be different.

Generally, the zero-sequence impedances of ineffective grounded networks are at least 4 or 5 times larger than those of directed grounded ones. Therefore, in Fig.2, the transformer is delta-wye connected and grounded with a Petersen-coil, which equivalent inductance is set to be 0.45H. Fig. 3 compares the zero-sequence currents of μPMUs installed at nodes A, B and C in Fig. 2(a). Fault is added between point B and C. The similarity between zero-sequence currents i_{0A} measured at μPMU A and i_{0B} at μPMU B is much higher than that between i_{0B} and i_{0C} at μPMU C due to the existence of the zero-sequence component U_{of}. Therefore, it is possible to adopt the differences between the zero-sequence currents of all zones to find the faulted one.

3) ANALYSIS OF POST-FAULT PHASE ANGLE

Once the faulted zone is determined, finding the precise fault point is the most important job. Fig.4 shows the equivalent post-fault circuit of the faulted zone when ignoring the effect of capacitance to ground. Here, V_M, V_N, I_M and I_N are the measured voltages and currents of μPMUs at both boundaries of the faulted zone. The total length of the faulted zone is L, and its unit impedance is z_{MN}. Voltage V_f and current I_f is the voltage and current at fault point, and R_f denotes the grounding resistance. The distance between μPMU M and fault point f is x.

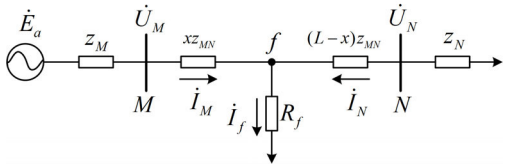

FIGURE 4. Equivalent circuit of faulted line.

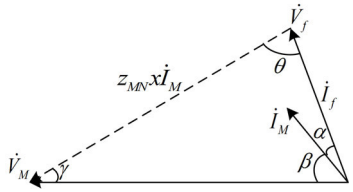

FIGURE 5. Phasor diagram of faulted line.

According to the equivalent circuit, the voltages at μPMU M and faulted point f have a relationship, as described in equation (3):

$$\dot{V}_M = \dot{V}_f + z_{MN}x\dot{I}_M = R_f\dot{I}_f + z_{MN}x\dot{I}_M \tag{3}$$

Their phasor diagram is displayed in Fig. 5.

Here, angle α is the phase angle between $\dot{I}_f$ and $\dot{I}_M$, β is the phase angle between $\dot{I}_M$ and $\dot{V}_M$, θ is the phase angle between $\dot{I}_f$ and the line voltage drop $z_{MN}x\dot{I}_M$, and γ is the phase angle between $\dot{V}_M$ and the line voltage drop $z_{MN}x\dot{I}_M$. The three voltage phasors construct a triangle. According to the Law of Sines, equations in (4) can be obtained, and the fault distance can be written as in (5).

$$\frac{|\dot{V}_M|}{\sin\theta} = \frac{|\dot{V}_M|}{\sin(\alpha+\beta+\gamma)} = \frac{|z_{MN}\dot{I}_M|\,x}{\sin(\alpha+\beta)} \tag{4}$$

$$x = \frac{|\dot{V}_M|\sin(\alpha+\beta)}{|\dot{I}_M z_{MN}|\sin(\alpha+\beta+\gamma)} \tag{5}$$

When the grounding resistance R_f is zero, the angle γ is zero, the phase difference between $\dot{V}_M$ and $\dot{I}_M$ is β which can be measured by μPMUs. But with the increase of R_f, the angle α and γ change accordingly, their effect cannot be ignored. But those two angles are difficult to be measured as the grounding resistance R_f and grounding voltage $\dot{V}_f$ are unknown. In some traditional methods [20], angle α is ignored and assumed to be 0°. Due to the change of load, fault location and system impedance in distribution networks, α usually varies between 0 to 15° [21]. The value of angle γ depends on the modulus of $\dot{V}_f$. It can be an acute angle less than 60° since the modulus of both $\dot{V}_f$ and $z_{MN}x\dot{I}_M$ cannot be greater than that of $\dot{V}_M$. As two angles α and γ are unknown in equation (5), it is difficult to calculate the value of x directivity.

Fig.6 illustrates the variations of μPMU-based measurements, including voltages, currents and phase angles with fault distance x. They change with the increase of fault distance. The waveform variations of μPMU-based measurements still suggest the mapping relationship between fault distance x and phasors $\dot{V}_M$ and $\dot{I}_M$, which would be learned with machine learning models.

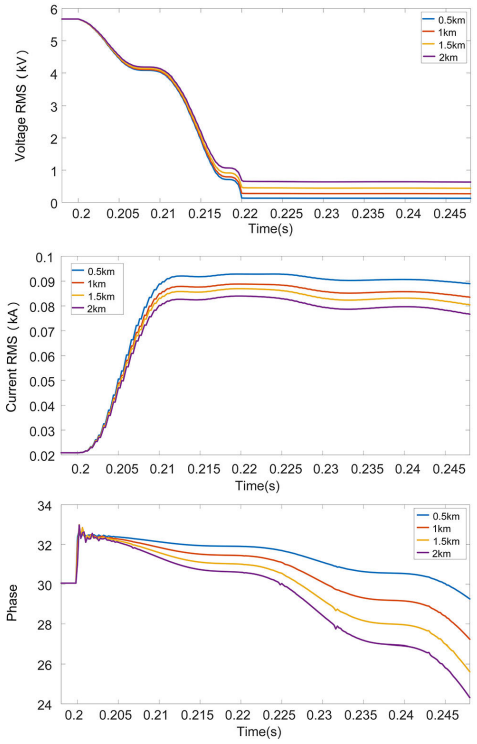

FIGURE 6. Waveform variations with fault distance, (a) voltage waveforms, (b) current waveforms, (c) phase angle waveforms.

B. FUNDAMENTALS OF SAE

A stacked autoencoder (SAE) becomes more widely used for learning generative models of data, and shows great potential in representing electrical measurements with reduced dimensions.

1) AUTOENCODER (AE)

As the basic unit of SAE, an AE usually contains an encoder and a decoder. When used as a coding tool, only the encoder part is used. The typical structure of an AE is shown in Fig. 7.

The dimensions of the input layer and the output layer are the same. The input vector $x = [x^{(1)}x^{(2)}\cdots x^{(n)}]$ is encoded at the hidden layer and reconstructed at the output layer. When the reconstruction error between original x and reconstructed $\hat{x} = [\hat{x}^{(1)}\hat{x}^{(2)}\cdots\hat{x}^{(n)}]$ is small enough, the output of the hidden layer, or the primary code $h = [h^{(1)}h^{(2)}\cdots h^{m}]$ can be regarded to be an extracted feature of input x. The encoding and decoding mapping are as shown in equation (6) and equation (7), respectively. Here, n is the dimension of the input layer, m is the dimension of the hidden layer, W is weight matrix, b is the bias vector, and S represents the activation function.

$$h = f(x) = S(Wx + b) \tag{6}$$

$$\hat{x} = g(h) = S(W'h + b') \tag{7}$$

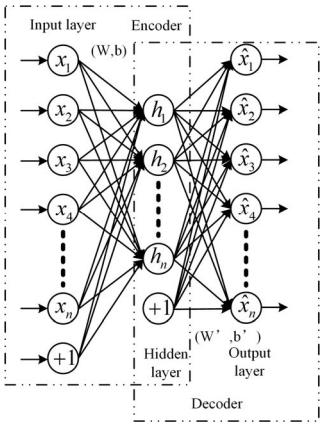

FIGURE 7. A typical AE structure.

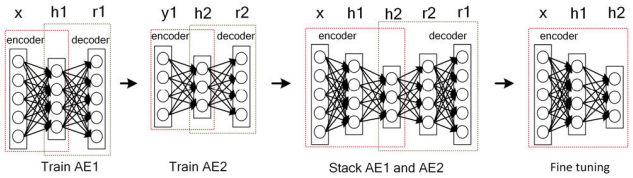

FIGURE 8. Formation of a two-layer SAE.

2) SAE

A stacked autoencoder is a neural network consisting of multiple layers of sparse autoencoders in which the outputs of each layer are wired to the inputs of the successive layer [5]. The construction process of a two-layer SAE is illustrated in Fig. 8. To train a SAE, four steps are usually included.

(1) Train the AE_1 with input x, and produce primary code h_1.

(2) Train the succeeded encoder AE_2 with primary code h_1 of the previous autoencoder.

(3) Combine the encoders of both AE_1 and AE_2 to construct an SAE. An output layer should be added to form a regression model.

(4) Fine-tune the entire network. Labeled samples such as fault measurements and fault distances are used to train the network in a supervised way.

C. PROPOSED METHOD

According to the aforementioned analysis, a μPMU-based fault location method is proposed in this paper. It includes two main steps: waveform similarity-based fault zone determination and SAE based fault distance calculation.

1) FAULT ZONE DETERMINATION

As discussed in Section II part A, the zero-sequence current are similar when a fault is not included. The similarity which is characterized by Pearson correlation coefficient $C(x, y)$ is used to discriminate the faulted zone from non-faulted ones.

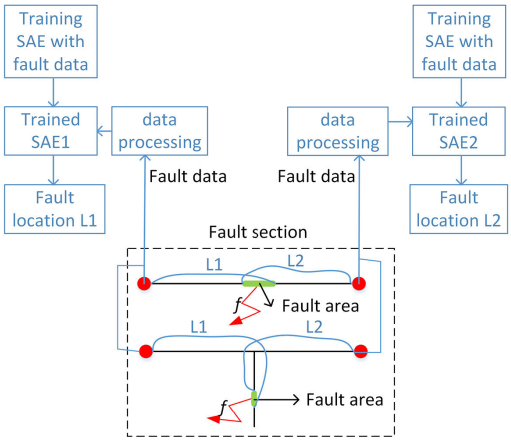

FIGURE 9. Flow chart of fault distance calculation.

The definition of $C(x, y)$ is shown in equation (8):

$$C(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{8}$$

where, x_i and y_i refer to the i^{th} element in vector x and vector y, respectively. $\bar{x}$ and $\bar{y}$ are the mean values of x and y. The value of $|C(x, y)|$ is between 0 and 1. Larger $|C(x, y)|$ suggests more similar waveforms.

When a fault occurs in the distribution network, the similarity between zero-sequence currents measured by μPMUs is calculated zone by zone, from the source to the load. A threshold θ is adopted to identify the faulted zone. The first zone with a $|C(x, y)|$ smaller than θ is regarded to be faulted.

2) FAULT DISTANCE CALCULATION

With the voltage, current and phase information, the fault distance can be deduced as shown in equation (5). The voltage and current vectors measured by μPMUs on both sides of the faulted section are used as the inputs of SAE. SAE model is used in this paper to learn the mapping relationship between μPMU measurements and fault distance. The fault distance calculation procedure is displayed in Fig. 9. The mapping relationship will be learned separately by SAE, and the fault distance will also be calculated separately with the μPMU measurements at each boundary. Since two main topologies: single line and T shape, are considered, the middle point of two calculated distance is selected to avoid the influence of calculation error and grounding resistance.

III. IMPLEMENTATION

When a single-phase to ground fault occurs in the distribution network, the μPMU measurements in this system will vary. Fig. 10 shows the flowchart of the proposed location method. As the most important step, the training of the SAE model will affect the location performance. Therefore, the parameters

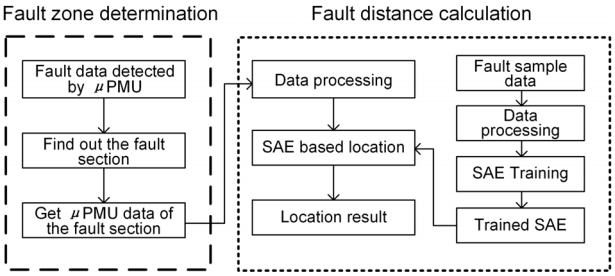

FIGURE 10. Flowchart of proposed fault location method.

TABLE 1. Measurements used for fault distance calculation.

Fault type		V	I
Single phase to ground fault	Ag	V_A	I_A
	Bg	V_B	I_B
	Cg	V_C	I_C

and structure of the proposed SAE model are discussed and analyzed in the following content.

A. FAULT ZONE DETERMINATION

When single-phase to ground fault occurs in distribution networks, zero-sequence measurements can be detected. If continuous zero-sequence sampling points are detected, for instance, three periods, the system is regarded to be faulted. The waveform similarity coefficient $|\mathrm{C}(x,\ y)|$ of zero-sequence currents will be calculated zone by zone from source to load. The threshold value θ is selected according to experimental results and field experiences. As the line length is only a few kilometers, the similarity of zero-sequence currents of non-faulted zone usually follows $|\mathrm{C}(x, y)| > 0.85$. When a single-phase-to-ground fault with zero grounding resistance occurs, such coefficient would be $|\mathrm{C}(x, y)| < 0.3$. The threshold value θ should be within the range from 0.5 to 0.8. The median value of this range, 0.65 is thus selected to be the value of similarity threshold θ.

B. DATA PROCESSING

The raw data collected by μPMUs should be processed and normalized to eliminate the influences from magnitude, initial state, and other factors. According to the practical application of μPMU, the sampling frequency is selected to be 5kHz and the post-fault data of 0.05s will be recorded [22]. Different from fault zone determination, phase voltage and phase current is adopted in fault distance calculations. Table 1 lists the measurements used in SAE-based fault distance calculation when a single-phase-to-ground fault occurs. For example, the phase voltage V_A and phase current I_A will be used when phase A is grounded.

As aforementioned in equation (5), the fault distance is related to the phasor $\dot{V}_M$ and $\dot{I}_M$, whose relationship can be modeled with machine learning algorithms. SAE which is effective in finding features in an unsupervised way is adopted. As complex numbers, the phasor measurements $\dot{V}_M$ and $\dot{I}_M$ cannot be used as the inputs of SAE directly, the values of their real parts (V^r, I^r) and imaginary parts $\left(V^i, I^i\right)$ are employed, as shown in equation (9).

$$\begin{cases} V^r = V\cos\theta, & I^r = I\cos\theta \\ V^i = V\sin\theta, & I^i = I\sin\theta \end{cases} \tag{9}$$

In order to eliminate the influence of the initial state, only the fault increments are analyzed. The increments are obtained according to equation (10), where $V_0^r, V_0^i, I_0^r, I_0^i$ are the initial vectors before fault occurrence.

$$\begin{cases} V_t^{r^*} = V_t^r - V_0^r & I_t^{r^*} = I_t^r - I_0^r \\ V_t^{i^*} = V_t^i - V_0^i & I_t^{i^*} = I_t^i - I_0^i \end{cases} \tag{10}$$

Min-max normalization is adopted here to avoid the influence of magnitude. The value range of normalized vector is between [0], [1]. The normalized voltage and current vectors are combined together to form a fault transient vector g, as shown in equation (11).

$$g = \left[V_A^{r^*}, V_A^{i^*}, I_A^{r^*}, I_A^{i^*}\right] \tag{11}$$

Fig. 11 illustrates the original voltage measurements V_t, the current measurements I_t, and the normalized fault transient vectors g. It can be shown from Fig. 11(a) and Fig. 11(b), the voltage and current vectors vary with fault distance. After normalization, both voltage and current vectors are in the same range and only the variation of their waveforms is considered.

C. SAE NETWORK TRAINING

The aim to train an SAE model is to reduce the error between calculated distance and actual fault distance. Training error defined in equation (12) is a common criterion to measure the performance of a trained network.

$$E_{error} = \frac{1}{N}\sum_{i=1}^{N}|p_i - t_i| \tag{12}$$

where, N is the number of training samples, p_i is the fault distance calculated by the i^{th} input, and t_i is the i^{th} actual fault distance. Training error is used in the selection of network architecture and functions to ensure better performance of fault distance calculation.

1) SELECTION OF NETWORK ARCHITECTURE

The selection of SAE network architecture refers to two aspects: the depth of network and the width of each layer. To provide an end-to-end fault distance calculation model, the size of the input and output layers should be the same as the input vector, fault transient vector, and the output, fault distance. So, the size of input layer is 1000, and the size of output layer is 1. Only the number of hidden layers and the size of each hidden layer should be determined here. A deeper network could learn more complex mapping relationships and larger amounts of data. Fewer units will be contained in hidden layers and better generalization will be obtained

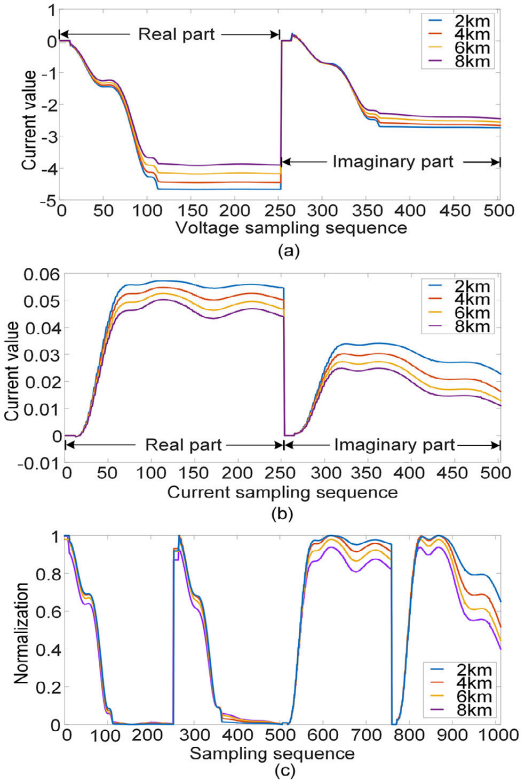

FIGURE 11. Formation of fault transient vectors, (a) normalized currents, (b) normalized voltages, (c) fault transient vectors.

TABLE 2. Training time and training errors with different depths.

Structure of net	Training time	Training errors
1000-500-1	192.52s	0.0873
1000-500-200-1	242.20s	0.0624
1000-500-200-100-1	296.65s	0.7366
1000-500-200-100-50-1	351.62s	1.7888

in test sets. But it is not the deeper the better. A too large or too deep network may result in time-consuming training or overfitting. Properly selection of architecture is needed to satisfy its application.

a: NUMBER OF HIDDEN LAYERS

In order to choose the number of hidden layers, four kinds of network structures are roughly selected to test their performances. The size of each hidden layer is only half of their previous layer. The training error in equation (12) is used to evaluate the performance. A mean error of 5 training results are used to avoid the effect of random initial states, and are listed in Table 2. Also, the mean training time of SAEs with the same structure is considered.

As discussed before, the training time of the SAE network increases with the number of layers. But the training error does not decrease with the number of layers. It reaches a minimum when the number of the hidden layers is 2, and rises rapidly when the number of hidden layers continues increasing. The SAE network with 2 hidden layers works better when its fitting capability matches the complexity of the mapping task. So, the number of hidden layers is selected to be two to balance the training time and the training error.

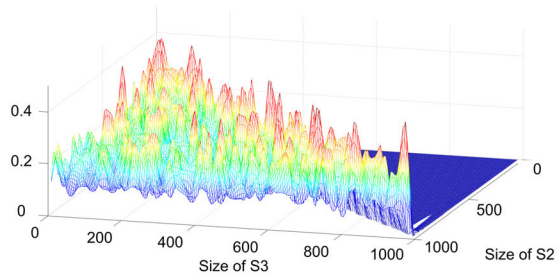

FIGURE 12. Training errors with different sizes of hidden layers.

b: SIZE OF HIDDEN LAYERS

As 2 hidden layers are chosen, it is necessary to determine the number of nodes S2 and S3 in the second and third layers in the SAE network, respectively. Comparing training errors is still a popular method in determining the size of hidden layers [23]. The value of S3 should be smaller than S2 to realize dimensional reduction. The changing step of the hidden layer is set to be 5 to balance the training time and performance. So, the range of S3 is from 10 to 1000, and the range of S2 is from 5 to S3. All the possible numbers of S2 and S3 are trained and the mean training error of 5 networks are recorded. Fig. 12 shows the distribution of training errors of all possible combinations of S2 and S3. The size of hidden layers are finally determined to be S2 = 435 and S3 = 145 for SAE1 of μPMU at the source side. Similarly, the hidden layer sizes of SAE2 of μPMU at the load side are selected to be S2 = 485 and S3 = 165.

2) SELECTION OF FUNCTIONS

a: ACTIVATION FUNCTION

In SAE networks, the activation function maps the input of the neutron to its output. A non-linear activation function can help the network to fit non-linear mapping relationships which are common in practical applications. Three non-linear activation functions: sigmoid function, hyperbolic function and rectified linear function are discussed here.

A sigmoid function is a mathematical function having a sigmoid curve. A standard choice for a sigmoid function is the logistic function defined by equation (13). The value of the function $f(z)$ is between 0 and 1. It is sensitive to the input only when z is close to zero, and saturates quickly when z is too large or too small. Gradient-based learning with sigmoid function can be very difficult due to its wide saturation area.

$$f(z) = \frac{1}{1 + e^{-z}} \tag{13}$$

Hyperbolic functions (Than) are defined as combinations of the exponential functions. A typical form of the hyperbolic function is shown in equation (14), and its value also ranges from 0 to 1. The derivation of the hyperbolic function is very

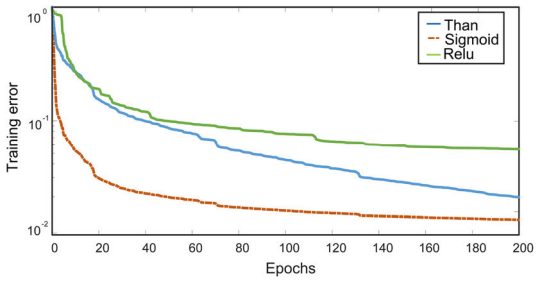

FIGURE 13. Training errors of SAE with different activation functions.

large. In the back-propagation algorithms, the gradient update with hyperbolic function is fast, but the vanishing gradient problem still exists.

$$f(z) = \frac{1 - e^{-2z}}{1 + e^{-2z}} \tag{14}$$

Rectified linear function (Relu), which is defined in equation (15), overcomes the vanishing gradient problem, allowing models to learn faster and perform better. But it cannot activate inputs with zeros when gradient-based algorithms are adopted. Also, the local minimum of cost functions cannot be achieved.

$$f(z) = \max\{0, z\} \tag{15}$$

The training error curves of different activation functions with the same iterations are shown in Fig. 13. For the application of fault distance calculation, the sigmoid function performs best. Its errors are smaller than those of the others, and converge rapidly. The sigmoid function is thus used as the activation function in this research.

b: LOSS FUNCTION

The cost function $J(\theta)$ is used to measure the difference between the predicted value and the actual value. Reducing the value of cost function is the primary goal of SAE training. Two typical definitions of cost functions are discussed here: mean squared error and cross-entropy.

The mean squared error (MSE) measures the average of the squares of the errors. As defined in equation (16), y is the actual value, x is the input value, θ is the network parameter, and $f(x;\theta)$ is the prediction value.

$$J(\theta) = \frac{1}{2} E_{x,y\sim\hat{p}data} \|y - f(x;\theta)\|^2 \tag{16}$$

Cross-entropy is used to quantify the difference between two probability distributions: the true distribution of outputs y calculated with the input x, and the estimated probability distribution y calculated with the maximum likelihood theorem. It is defined as in equation (17).

$$J(\theta) = -E_{x,y\sim\hat{p}data} \log p_{\text{mod } el}(y|x;\theta) \tag{17}$$

The training error of SAE networks with the same structure but different activation functions are listed in Table 3. The network performances better when cross-entropy is used as its cost function. The training error of cross-entropy is smaller.

TABLE 3. Training errors of SAE with different cost functions.

Cost Function	Cross entropy	MSE
Error	0.05197	0.05526

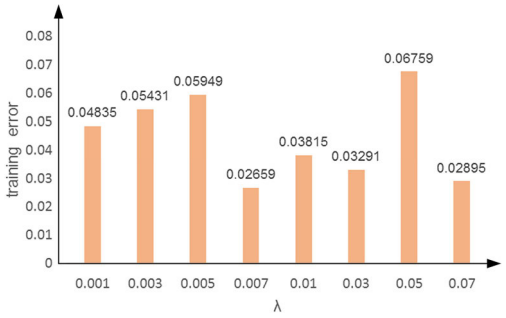

FIGURE 14. Training error of SAE with different Regular coefficient.

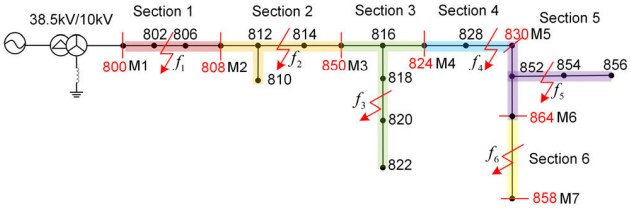

FIGURE 15. Simulated distribution network.

Therefore, cross-entropy is used as the cost function for the SAE model in this research.

c: REGULARIZATION COEFFICIENT

L^2 regularization method is usually employed during SAE training to avoid overfitting and improve generalization capability. A regularization item is added to the loss function, as shown in equation (18):

$$J = J_0 + L^2 = J_0 + \frac{\lambda}{2n} \sum_{\omega} \omega^2 \tag{18}$$

Here, the original cost function is denoted by J_0, L^2 is the regularization item. It equals to the sum of network parameter squares ω^2 times regulation coefficient λ and divided by the sample size n. The relationship between regulation coefficient and training errors is shown in Fig. 14. Therefore, $\lambda = 0.007$ is selected as the regulation coefficient, according to Fig. 14.

IV. SIMULATION RESULTS AND COMPARISONS

To demonstrate the effectiveness of proposed method, a typical 10kV distribution network is modeled on the platform of PSCAD. The simulation system is shown in Fig. 15. The source voltage is 38.5kV, and the rated ratio of the transformer is 38.5kV/10.5kV. Bergeron model is adopted as line models. This system is divided into 6 zones according to the μPMU assignment principles in Section II part A. The measuring points M1-M7 stand for μPMUs, where voltage and current waveforms can be collected. Six single-phase to

TABLE 4. Similarity coefficients of different faults.

Fault	f_1	f_2	f_3	f_4	f_5	f_6
C(M1,M2)	0.104	0.914	0.927	0.924	0.971	0.924
C(M2,M3)	\	0.074	0.968	0.959	0.983	0.972
C(M3,M4)	\	\	0.106	0.986	0.954	0.968
C(M4,M5)	\	\	\	0.077	0.963	0.974
C(M5,M6)	\	\	\	\	0.145	0.978
C(M6,M7)	\	\	\	\	\	0.112

TABLE 5. Simulation scenarios for training and test samples.

	Fault distance (%)	Fault types	Ground resistance (Ω)	Fault Section	Total number
Network Training	5% 15% 25% 35% 45% 55% 65% 75% 85% 95%	Ag, Bg	1Ω 10Ω 30Ω 50Ω	Sections 1 to 6	480
Network Testing	10% 20% 50% 60% 80%	Ag	5Ω 15Ω 40Ω	Sections 1 to 6	90

ground faults, F1 to F6, are simulated in each zone to test the proposed method.

A. FAULT ZONE DETERMINATION

When a fault occurs, the waveform similarity between two μPMU-based zero-sequence current measurements is calculated. The similarity is compared with an empirical threshold which is chosen to be 0.75. The similarity coefficients $|C(x,y)|$ of different faults are listed in Table 4. Clearly, the waveform similarity is higher when the fault is not included in the analyzed zone, and it will drop greatly when the analyzed zone is a faulted one. The proposed waveform similarity-based fault zone determination is effective to find the faulted zone.

B. FAULT DISTANCE CALCULATION

Once the faulted zone is determined, fault distance is needed to be found out. The trained SAEs analyze the μPMU-based fault transient vectors of both zone boundaries. To train and test the proposed SAE-based fault distance calculation method, 480 training samples and 90 test samples are generated. The simulation parameters are listed in Table 5.

The location results are evaluated with the fault location error defined in equation (19) [24]. Here, L_{MN} is the line length between two μPMUs, L_{esti} represents the calculated fault distance, and L_{act} stands for the actual fault distance.

$$Error(\%) = \frac{|L_{esti} - L_{act}|}{L_{MN}} \times 100 \tag{19}$$

TABLE 6. Location results based on SAE.

Faulted zone	No. of samples	Mean location error	Maximum location error
1	15	1.31%	2.27%
2	15	1.39%	2.31%
3	15	1.36%	2.24%
4	15	1.19%	2.17%
5	15	1.40%	2.43%
6	15	1.23%	2.38%

TABLE 7. Location results with different number of branches.

Number of branches	Mean location error	Maximum location error
1	1.24%	2.27%
3	1.38%	2.32%
5	2.95%	3.78%
7	7.40%	9.35%

The location results are shown in Table 6. The average error of all test samples is 1.41% and the maximum error is 2.43%. That means the maximum location error distance would be around 2 meters, if the total length of faulted zone is 1 kilometer. This error is acceptable in practical applications.

C. EFFECT OF BRANCHES

As described in Fig.9, the proposed method is designed to pinpoint phase-to-ground faults in the faulted section with single branch and "T" shape interconnection. Since load branches are usually included in distribution networks, the effect of multiple branches or missing of μPMU-based measurements should also be discussed. Here, the location performance of proposed method up to 4 branches are studied, including single branch (1 branch), "T" shape (3 braches), "Π" shape (5 branches), and "TΠ" shape (7 branches). Samples in Table 6 are used to estimate the mean and maximum errors of location method with no more than three branches. Zones 1, 4, 6 are regarded to be one-branch sections and zones 2, 3, 5 are regarded to be three-branch sections. Section 2 and section 3 are combined together to form a five-branch section, while sections 2, 3, 4 and 5 form a seven-branch section. The proposed SAE-based location method is implemented for the five-branch section and the seven-branch section. The location results are listed in Table 7. With the increase of the number of branches, the location performance of the proposed method decreases. When 7 branches are included in a faulted section, the mean location error reaches 7.4%, and the maximum error increases to 9.35%. These errors are very large. Additional methods should be applied to improve the location accuracy.

D. EFFECT OF NOISES

In practical applications, the μPMU-based measurements are often polluted by noises, for example, the operating surges from breakers or switches, harmonics from electronic devices, and background Gaussian noises. As the most common noise in measured signals, the effect of Gaussian white noise on the performance of the proposed fault distance calculation method is discussed under different signal-to-noise

TABLE 8. Location results with different SNRs.

SNR	Mean location error	Maximum location error
60dB	1.34%	2.25%
50dB	1.38%	2.33%
40dB	1.69%	2.64%
30dB	2.40%	4.38%
20dB	5.19%	9.17%

TABLE 9. Comparisons between proposed method and existing ones.

Location Methods	Mean location error	Maximum location error
Proposed SAE based method	1.31%	2.3%
BP network based method	2.89%	4.96%
Impedance-based method	4.97%	7.84%

ratios (SNRs). The location results of the SAE-based method are shown in Table 8. The location errors, both mean error and maximum error, increase with the drop of SNRs. The proposed method is still effective when SNR is no less than 40dB. But when SNR continues dropping, the location errors become worse. When the SNR drops to 20 dB, the maximum location error reaches 9.17%. It means the location error distance will be around 90 meters if the total line length is 1 kilometer. In this case, denoising algorithms should be added to improve the location accuracy if SNR is lower than 40dB.

E. COMPARISONS

To demonstrate the performance of proposed method, the location results of the proposed method are compared with two existing popular methods: impedance-based method and BP (back-propagation) neural network-based method.

A single-ended impedance-based method is a traditional fault location method [25]. The capacitance to ground of distribution transmission lines is ignored. Only line resistance and line inductance are kept. This method is usually used when grounding resistance is zero, as in equation (5) when γ equals to zero. Otherwise, the location error is large.

As the most popular machine learning model, BP neural network was often used in solving regression problems such as fault location. In this research, a BP neural network is used. Its architecture is the same as that of the proposed SAE. The same activation function and training algorithm are used.

The location results of these three methods are listed in Table 9.

The location results of the proposed method are better than traditional popular ones. Both the mean error and the maximum error of the proposed method are smaller. The impedance-based method is effective when grounding resistance is zero. But when the effect of grounding resistance is considered, the location error increases. Also, the accuracy of the line parameter will affect the location result. Especially, when the line parameter is inaccurate in practical applications, the location error cannot be ignored. Although BP neural network is good at learning non-linear relationships, its supervised learning algorithm cannot fit complex problems effectively.

V. CONCLUSION

In this paper, a single-phase-to-ground fault location method for distribution networks has been proposed. This method includes waveform similarity-based fault zone determination and SAE based fault distance calculation. With the properly assignment of μPMUs, the distribution network can be divided into separate zones. Once a single-phase-to-ground fault is detected, the faulted zone can be found out by comparing the similarity of zero-sequence current waveforms at the zone boundaries. SAE network is then trained and used to calculate the fault distance and pinpoint the fault with μPMU measurements at both ends. Different from traditional location methods, the phase information provided by μPMUs is used to improve location accuracy. The end-to-end unsupervised learning model can extract features from raw data and avoid the empirical misjudgments from experts. By properly dividing distribution networks, fully exploring μPMU measurements, and extracting more effective features, the proposed method has better performance in pinpointing single-phase-to-ground faults in distribution networks and provides a way for more effective use of end-to-end unsupervised learning models in power systems.

REFERENCES

[1] H. Zhan, C. Wang, Y. Wang, X. Yang, X. Zhang, C. Wu, and Y. Chen, "Relay protection coordination integrated optimal placement and sizing of distributed generation sources in distribution networks," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 55–65, Jan. 2016.

[2] J. He, Y. Luo, M. Li, Y. Zhang, Y. Xu, Q. Zhang, and G. Luo, "A high-performance and economical multi-port hybrid DC circuit breaker," *IEEE Trans. Ind. Electron.*, to be published.

[3] F. Deng, X. Zeng, X. Tang, Z. Li, Y. Zu, and L. Mei, "Travelling-wave-based fault location algorithm for hybrid transmission lines using three-dimensional absolute grey incidence degree," *Int. J. Electr. Power Energy Syst.*, vol. 114, Jan. 2020, Art. no. 105306.

[4] B. Li, J. He, Y. Li, and B. Li, "A review of the protection for the multi-terminal VSC-HVDC grid," *Protection Control Mod. Power Syst.*, vol. 4, no. 1, p. 21, Nov. 2019.

[5] G. Luo, C. Yao, Y. Liu, Y. Tan, J. He, and K. Wang, "Stacked auto-encoder based fault location in VSC-HVDC," *IEEE Access*, vol. 6, pp. 33216–33224, 2018.

[6] G. Luo, Q. Lin, L. Zhou, and J. He, "Recognition of traveling surges in HVDC with wavelet entropy," *Entropy*, vol. 19, no. 5, p. 184, Apr. 2017.

[7] S. Shi, B. Zhu, A. Lei, and X. Dong, "Fault location for radial distribution network via topology and reclosure-generating traveling waves," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6404–6413, Nov. 2019.

[8] I. M. Karmacharya and R. Gokaraju, "Fault location in ungrounded photovoltaic system using wavelets and ANN," *IEEE Trans. Power Del.*, vol. 33, no. 2, pp. 549–559, Apr. 2018.

[9] M. Guo, N. Yang, and W. Chen, "Deep-learning-based fault classification using Hilbert–Huang transform and convolutional neural network in power distribution systems," *IEEE Sensors J.*, vol. 19, no. 16, pp. 6905–6913, Apr. 2019.

[10] S. Lan, M.-J. Chen, and D.-Y. Chen, "A novel HVDC double-terminal non-synchronous fault location method based on convolutional neural network," *IEEE Trans. Power Del.*, vol. 34, no. 3, pp. 848–857, Jun. 2019.

[11] M. Shafiullah, M. Abido, and T. Abdel-Fattah, "Distribution grids fault location employing ST based optimized machine learning approach," *Energies*, vol. 11, no. 9, p. 2328, Sep. 2018.
[12] M. Shafiullah, M. A. Abido, and Z. Al-Hamouz, "Wavelet-based extreme learning machine for distribution grid fault location," *IET Gener., Transmiss. Distrib.*, vol. 11, no. 17, pp. 4256–4263, Nov. 2017.
[13] S. S. Gururajapathy, H. Mokhlis, H. A. B. Illias, and L. J. Awalin, "Support vector classification and regression for fault location in distribution system using voltage sag profile," *IEEJ Trans. Elect. Electron. Eng.*, vol. 12, no. 4, pp. 519–526, Jul. 2017.
[14] X. Deng, R. Yuan, Z. Xiao, T. Li, and K. L. L. Wang, "Fault location in loop distribution network using SVM technology," *Int. J. Electr. Power Energy Syst.*, vol. 65, pp. 254–261, Feb. 2015.
[15] Y. Aslan, "An alternative approach to fault location on power distribution feeders with embedded remote-end power generation using artificial neural networks," *Electr. Eng.*, vol. 94, no. 3, pp. 125–134, Sep. 2012.
[16] A. Swetapadma and A. Yadav, "Improved fault location algorithm for multi-location faults, transforming faults and shunt faults in thyristor controlled series capacitor compensated transmission line," *IET Gener., Transmiss. Distrib.*, vol. 9, no. 13, pp. 1597–1607, Oct. 2015.
[17] T. Wu and W. U. Bajwa, "Learning the nonlinear geometry of high-dimensional data: Models and algorithms," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6229–6244, Dec. 2015.
[18] M. K. Neyestanaki and A. M. Ranjbar, "An adaptive PMU-based wide area backup protection scheme for power transmission lines," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1550–1559, May 2015.
[19] M. U. Usman and M. O. Faruque, "Validation of a PMU-based fault location identification method for smart distribution network with photovoltaics using real-time data," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 21, pp. 5824–5833, Nov. 2018.
[20] Y.-J. Lee, C.-H. Chao, T.-C. Lin, and C.-W. Liu, "A synchrophasor-based fault location method for three-terminal hybrid transmission lines with one off-service line branch," *IEEE Trans. Power Del.*, vol. 33, no. 6, pp. 3249–3251, Dec. 2018.
[21] Y. Liao and M. Kezunovic, "Optimal estimate of transmission line fault location considering measurement errors," *IEEE Trans. Power Del.*, vol. 22, no. 3, pp. 1335–1341, Jul. 2007.
[22] M. Majidi and M. Etezadi-Amoli, "A new fault location technique in smart distribution networks using synchronized/nonsynchronized measurements," *IEEE Trans. Power Del.*, vol. 33, no. 3, pp. 1358–1368, Dec. 2018.
[23] R. Salim, K. De Oliveira, A. Filomena, M. Resener, and A. Bretas, "Hybrid fault diagnosis scheme implementation for power distribution systems automation," *IEEE Trans. Power Del.*, vol. 23, no. 4, pp. 1846–1856, Oct. 2008.
[24] P. Nan, C. Menghan, L. Rui, and F. Zare, "Asynchronous fault location scheme for half-wavelength transmission lines based on propagation characteristics of voltage travelling waves," *IET Gener., Transmiss. Distrib.*, vol. 13, no. 4, pp. 502–510, Feb. 2019.
[25] R. Krishnathevar and E. E. Ngu, "Generalized impedance-based fault location for distribution systems," *IEEE Trans. Power Del.*, vol. 27, no. 1, pp. 449–451, Jan. 2012.