Voltage Sag Source Identification Based on Few-Shot Learning

HAOTIAN SUN[1], HAO YI[1], (Member, IEEE), GUANGYU YANG[1],
FANG ZHUO[1], (Member, IEEE), AND ANPING HU[2]
[1]State Key Laboratory of Electrical Insulation and Power Equipment, Xi'an Jiaotong University, Xi'an 710049, China
[2]China Electric Power Research Institute, Nanjing 100192, China

Corresponding author: Hao Yi (yi_hao@xjtu.edu.cn)

ABSTRACT Efficient identification of the voltage sag sources is significant in the power quality studies. This paper presents a novel method for voltage sag source identification which performs automatic feature extraction and shows a superior performance regardless of the insufficient amount of training samples. In the proposed strategy, the input data are preprocessed and fetched into the feature extractor, which is designed based on the convolutional neural network. Then the weighted k-nearest neighbor classifier generates the identification results. In the training period, the few-shot learning technique is harnessed, and the siamese network is constructed such that the proposed model learns efficiently even with a small number of samples. The proposed scheme is implemented in Python and PyTorch framework. Case studies and comparisons with other methods are carried out on 700 samples of voltage sag events in Jiangsu Province, China. Experimental results show the superiority of the proposed method over other identification methods in the tested cases.

INDEX TERMS Voltage sag identification, few-shot learning, siamese network, convolutional neural network, weighted k-nearest neighbors.

I. INTRODUCTION

In the last decades, voltage sag (VS) has been regarded as one of the most significant issues in power quality [1]. In spite of their short duration, such events can cause serious problems in transmission and distribution system, microgrids, industrial or customer facilities [2], [3]. Various sources such as faults, induction motors starting, and energization of transformers may cause the VS events that are propagated throughout the power system [4]. Therefore, monitoring and analyzing the VS events help to mitigate the substantial loss on the industrial utilities, improve the power quality indices, and also clarify the responsible agent for the VS events. Identifying the causes of the VS phenomena is an important strategy to analyze and characterize such events. Nowadays, such work is carried out mostly by the manual investigation of the specialists [5]. It is thus significant to formulate a VS-type identification method that can automatically handle a large scale of recorded events.

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

The developed VS identification methods can be categorized into three main families: 1) expert systems based on the human expert knowledge or rules; 2) methods based on unsupervised learning and clustering techniques, and 3) methods utilizing supervised learning and deep learning techniques.

Empirical rules are usually established to make decisions in the first group of methods. An expert system that makes use of the VS waveform information for VS data segmentation is explored in [6]. An approach based on the wavelet entropy and a dedicated expert system is suggested in [7]. These methods show great simplicity and are usually convenient to implement. However, their performance can be greatly deteriorated if the model parameters such as the threshold values are selected improperly.

The second group of methods does not require the labeling of the input data. Instead, these methods formulate the clusters of the VS samples based on the predefined mapping or feature extracting schemes. A method identifying the VS events by exploring the polarization ellipse representations of these events is presented in [8]. The technique presented in [9] and [10] characterize the voltage-space vectors
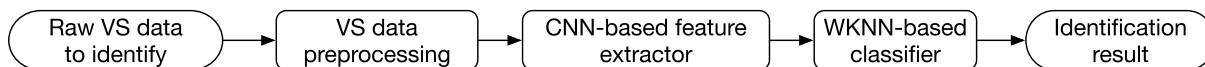
| Raw VS data to identify | → | VS data preprocessing | → | CNN-based feature extractor | → | WKNN-based classifier | → | Identification result |

FIGURE 1. Structure of the proposed model after training period.

and the RMS-voltage transition and make use of the K-means clustering algorithm to distinguish different VS groups. The methods in this group obviate the labor work of labeling input data. However, these methods only split the VS data into several groups. The further recognition of the VS sources still needs to be carried out manually or by other identification schemes. Besides, the clustering performance can also be strongly affected by different selections of the predefined VS features.

The emerging machine learning and deep learning concepts are also employed to identify the VS types. As suggested in [11], a binary decision tree based on several support vector machines (SVMs) is utilized to classify the VS data. Reference [5] formulates the VS identification by implementing a long short-term memory (LSTM) network to avoid the extra step of extracting the VS features. In [12], K-means-singular value decomposition is adopted for feature extraction, followed by the least-square SVM to obtain the predicted VS type. Authors of [13] train an auto-encoder network with unlabeled VS data and develop a back-propagation (BP) network with labeled data for VS event categorization.

From the previous literature reviews, the performance of some methods is greatly affected by the threshold values, the feature selections, and sometimes the accuracy of the inception point of the VS events. Therefore, it is significant to develop a VS identification method that can automatically extract features from the VS data. On the other hand, the methods of the last group generally require the specialists to manually recognize and label a sufficient number of monitored data of VS events for training purpose. The manual labeling of these data usually consumes lots of time and yields large cost on human resources. These methods will suffer performance degradation when the number of correctly labeled data are limited. The amount of VS training data may be augmented with synthetic data such as the simulated VS events. However, the recorded VS data and the simulated ones generally come from different distributions. The mismatching in the data distribution may produce further deterioration in the identification performance [5], [11], [14]. Consequently, an identification method that can correctly identify a large scale of unlabeled VS samples even with an insufficient number of labeled training data is strongly required.

To attain the aforementioned functionalities, this paper introduces a novel VS identification method. The adoption of the convolutional neural network (CNN) obviates manually predefining the VS features or determining the inception point-on-wave. The siamese network and the few-shot learning technique are utilized such that the proposed method learns efficiently even from a small number of

VS training data. Finally, the distance-weighted k-nearest neighbors (WKNN) classifier is adopted to generate the identification results. Besides, the training data, as well as the testing data, are assumed in the same distribution.

This paper is organized as follows. Section II illustrates the proposed methodology of VS identification. Section III shows the case studies, comparison, and discussions on several different identification methods. The conclusion is presented in Section IV.

II. METHODOLOGY OF VS IDENTIFICATION

Fig. 1 depicts the structure of the proposed VS identification method. As shown, the identification of VS sources are carried out in three major stages: 1) the VS data that are captured by the power quality monitors are firstly loaded and transformed into the unified dimension which is accepted by the following feature extractor; 2) the CNN-based network, which is trained in advance using the representation learning technique, produces an embedding vector for each recorded VS event; 3) each embedding vector is then analyzed by the WKNN-based classifier, and finally the identification result is generated.

A. VS DATA PREPROCESSING

Suppose that the examined raw sample is of size $3 \times T_{raw}$ where the three channels correspond to the three phases recording, whereas each channel contains T_{raw} sampling points. Note that T_{raw} of different VS events recorded by the power quality monitors differs from each other since the VS duration varies in a wide range, typically from 0.5 cycle to 1 minute [1]. Therefore, each channel in the raw sample is downsampled and truncated into the appropriate dimension that is compatible with the feature extractor. Suppose that the feature extractor accepts data of size $3 \times T_{ext}$. The downsampling rate R_d is then defined as follows:

$$R_d = \lfloor \frac{T_{raw}}{T_{ext}} \rfloor \tag{1}$$

where $\lfloor \cdot \rfloor$ indicates the floor operator which takes the integer part of the input value. Then for each channel of the downsampled data, the first T_{ext} points are retained, such that the downsampled data are truncated into the size $3 \times T_{ext}$, which is compatible to the following feature extractor.

It should be noted that T_{ext} is determined by considering the features of the power quality monitors. In this paper, the examined data are collected from the system operating at 50 Hz, in the sampling frequency of 10 kHz. In order to avoid the impact of the signal detection delay, the power quality monitors that are installed in the examined system retain one period before the detected VS starting point, as well as
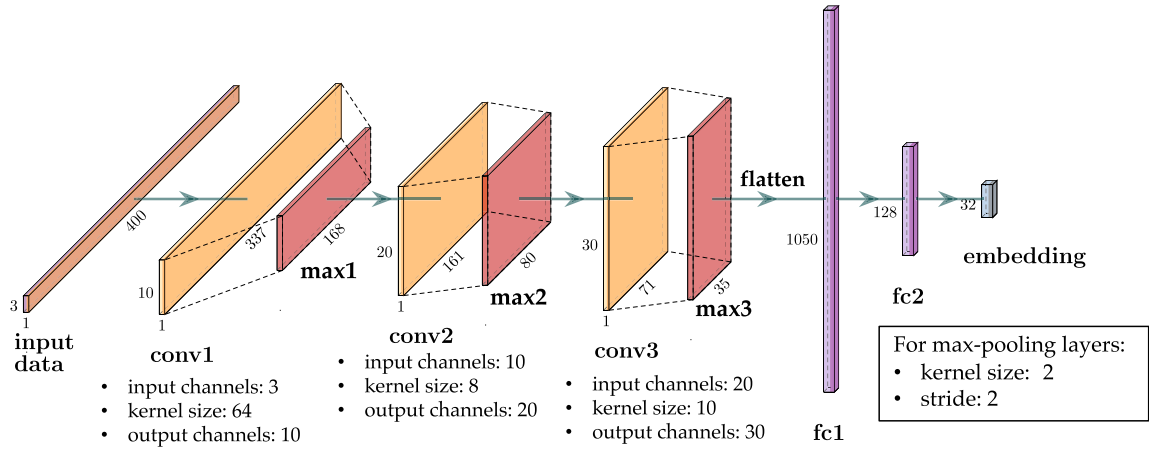
FIGURE 2. Structure of the proposed feature extractor.

one period after the detected VS ending point. Consequently, the recorded signal contains at least two cycles which are equivalent to 400 sampling points in the examined case. Therefore, T_{ext} is set to 400 in the following discussions. It should be noted that the T_{ext} can vary in different scenarios regarding the recording mode of the employed monitors and the value of their sampling rate.

To accelerate the model convergence in the training period, the training data are further separated into several batches. The number of samples in each batch is denoted as N.

B. CNN-BASED FEATURE EXTRACTOR

The convolutional neural network (CNN) usually consists of several convolution layers, pooling layers, and finally fully-connected layers which are usually connected in series [15]. Fig. 2 presents the CNN structure and the layer parameters of the proposed feature extractor, using the network visualization program implemented in [16]. The input data of the feature extractor is transformed into the length of T_{ext} (400 points) by the VS data preprocessing procedure. In the proposed model, the input signal passes through alternately three convolutional layers (*conv1* to *conv3*) and three max-pooling layers (*max1* to *max3*), then two fully-connected layers (*fc1* and *fc2*).

In the convolutional layers, the input signal that contains three channels can be regarded as a one-dimensional image where these three channels are recognized as the RGB color representations, as described in Fig. 3. Note that the length of the VS signals is already transformed into 400 points and the height of the one-dimensional image is enlarged for the visualization propose.

The input signal is scanned and characterized by the one-dimensional kernels with learnable parameters. Let C_i denotes the number of channels and T_i indicates the number of sampling points in each channel. For the i-th convolution layer, the input to this layer x_i is of size $N \times C_i^{in} \times T_i^{in}$ whereas the output y_i is with size $N \times C_i^{out} \times T_i^{out}$. The output y_i of
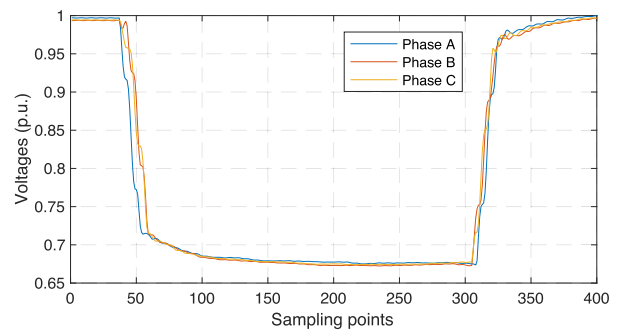
(a) transformed VS waveforms (VS due to three-phase fault), with $T_{ext} = 400$
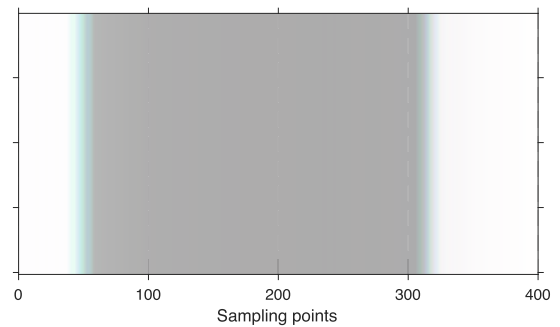
(b) corresponding recognization in the CNN

FIGURE 3. Example of one-dimension image representation of VS signal.

the *conv-i* is expressed as follows:

$$y_i = \sigma\left(\sum_{c=0}^{C_{i-1}-1} \omega_{i,c} \star x_{i,c} + b_i\right) \quad (2)$$

where $\star$ signifies the cross-validation operator whose definition is given by (3); $\omega_{i,c}$ and b_i are the learnable parameter of this layer; $\sigma(\cdot)$ represents the activation function of the parametric rectified linear unit (RReLU) which is expressed as in (4) [17].

$$\omega_{i,c} \star x_{i,c}(t) = \sum_{m=-\infty}^{+\infty} \omega_{i,c}(m) \cdot x_{i,c}(m+t) \quad (3)$$

$$\sigma(x) = \max(0, x) + 0.25 \cdot \min(0, x) \quad (4)$$

The intermediate features are extracted in the outputs y_i of the convolutional layer *conv-i*. The most significant components of these features are filtered and retained by the following max-pooling layer *max-i*. Note that despite the possible information loss, the max pooling layers generally improve the model performance and contribute to the time-shift invariance of the proposed model [18], [19].

In the proposed model, the *max-i* splits the input with the stride of 2 and takes the maximum value of every input slice. The output p_i of *max-i* is presented as follows:

$$p_{i,c}(t) = \max(y_{i,c}(2t), y_{i,c}(2t + 1)) \tag{5}$$

The two fully connected layers(*fc1, fc2*) reduce the length of the output $p_{i,c}$ and generate the 32×1 embedding vector z that can be regarded as the feature representation of the input signal [20]. Note that the time translation invariance of the one-dimensional CNN structure makes the proposed method immune to the inception time of the VS events. Also, the parameters such as the dimensions of the proposed neural network can be changed regarding different sizes of the input signals, in order that the performance of the proposed feature extractor can be optimized, and that the signal information can be fully extracted.

C. WKNN-BASED CLASSIFIER

In the trained model, the embedding outputs that correspond to the inputs of the same VS types are clustered altogether in the sense of L-2 norm [21]. Denote the examined testing input as x_{test}. The L-2 distance $d(x_i, x_j)$ between x_i and x_j is defined in (6). Let d_j represent the L-2 distance between x_{test} and the element x_j in training set.

$$d(x_i, x_j) = [(x_i - x_j)^T \cdot (x_i - x_j)]^{\frac{1}{2}} \tag{6}$$

Data clustering can be carried out by the classifier such as the ensemble learning classifier and the k-means algorithms [22]. These algorithms usually require the assembling of sufficient number of un-clustered samples to perform the unsupervised training. In the proposed method, the simple technique that do not require the training period is needed. Therefore, the distance-weighted k-nearest neighbor (WKNN) technique can be applied to classify the testing inputs. For a testing sample, the WKNN algorithm calculates directly several training samples that are nearest to the testing one. The prediction is then carried out by analyzing the VS types of these selected training samples.

Denote $x_1, \ldots, x_k$ and $l_1, \ldots, l_k$ the first k training inputs and their labels with the minimum L-2 distance to x_{test}. Note that $x_1, \ldots, x_k$ are sorted in increasing order regarding their distances to x_{test}. The weight M_n for the n-th nearest neighbor x_n is expressed as in (7).

$$M_n = \begin{cases} \dfrac{d_k - d_n}{d_k - d_1}, & \text{if} \quad d_k \neq d_1, \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

Then the VS type prediction of the testing input x_{test} can be generated:

$$l_{test} = \arg\max_l \sum_{n=1}^{k} M_n \cdot \delta(l = l_n) \tag{8}$$

where the $\delta(\cdot)$ equals 1 if $l = l_n$ and equals 0 otherwise.

D. MODEL TRAINING USING FEW-SHOT LEARNING

In the proposed model, the embedding outputs of different VS types are supposed to be sufficiently distinguishable, even in the cases where only a few labeled data are available for training. One efficient solution to train the model with very small-scaled training set is the few-shot learning technique [23]. Since all trainable parameters of the model are centered in the CNN-based feature extractor, the few-shot learning technique is applied in this part. More specifically, the siamese structure of the feature extractor is designed in the training period.

Let us define the positive pairs as the pairs of the same labeled samples and the negative pairs as the dissimilarly labeled samples. Denote $\{Z_P^{(1)}, Z_P^{(2)}\}$ the set of positive pairs and $\{Z_N^{(1)}, Z_N^{(2)}\}$ the set of negative pairs.

As presented in Fig. 4, a certain pair of training samples are selected and separately fed into two CNN-based feature extractors. These two feature extractors have an identical structure and share the same parameter values. Then the contrastive loss is calculated based on the embedding outputs of the two feature extractors. The loss function is formulated such that the positive pairs produce small contrastive loss compared with the negative pairs, which result in a relatively large loss. The parameters, such as the weights and the bias that are shared by both feature extractors, are updated according to the contrastive loss. In this paper, the optimization process for the shared parameters is not illustrated in detail, since this process can be handled automatically by the deep learning framework such as PyTorch and TensorFlow.

Fig. 5 presents the strategy of hard-negative pairs selection. All possible positive pairs are generated by the pair-selection algorithm, whereas only the same number of pairs with the smallest distance is selected as the negative pairs, so as to simulate the most challenging cases for the model to fit.

The positive loss measures the similarity between the embeddings of the samples in the same VS types [24]:

$$\mathcal{L}_P = \sum_{i=1}^{N_p} d(z_i^{(P1)}, z_i^{(P2)})^2 \tag{9}$$

where $z_i^{(P1)}, z_i^{(P2)}$ represent the i-th pair in the set $\{Z_P^{(1)}, Z_P^{(2)}\}$; N_p is the number of pairs in $\{Z_P^{(1)}, Z_P^{(2)}\}$.

Besides, the negative loss estimates the similarity of the embeddings from the different VS types:

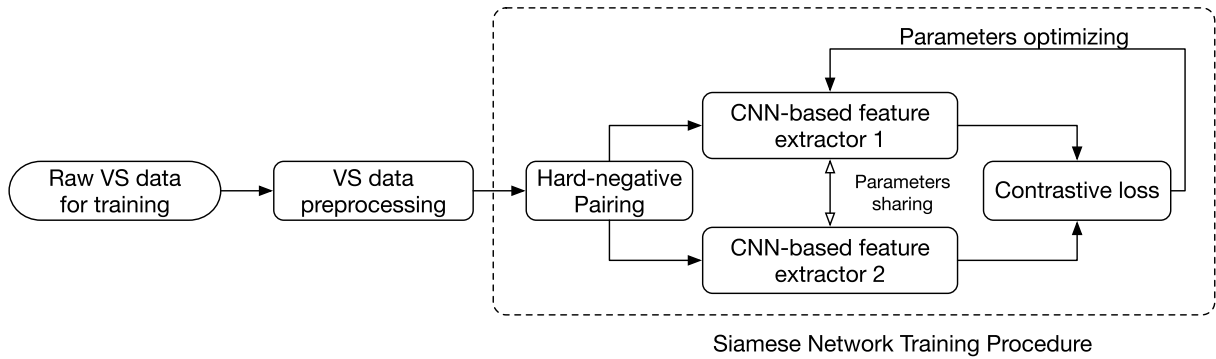$$\mathcal{L}_N = [\max(0, \delta - \sum_{i=1}^{N_p} d(z_i^{(N1)}, z_i^{(N2)}))]^2 \tag{10}$$

FIGURE 4. Structure of siamese network in training period.

Algorithm 1: Hard-negative pairs selection

Get full embedding set $\{Z\}$ and full label set $\{L\}$;
begin data preperation
    generate pair combination list $\{Z_i, Z_j\}$;
    foreach (z_i, z_j) *in* $\{Z_i, Z_j\}$ **do**
        sort $\{Z_i, Z_j\}$ in increasing order of $d(z_i, z_j)$;
    end
    generate corresponding label list $\{L_i, L_j\}$;
end
begin positive/negative pairs selection
    foreach $(z_i, z_j), (l_i, l_j)$ *in* $\{Z_i, Z_j\}, \{L_i, L_j\}$ **do**
        if $l_i = l_j$ **then**
            add the pair (z_i, z_j) to $\{Z_P^{(1)}, Z_P^{(2)}\}$;
        else
            add the pair (z_i, z_j) to $\{Z_N^{(1)}, Z_N^{(2)}\}$;
        end
    end
end
begin negative pairs truncation
    set N_p the number of pairs in $\{Z_P^{(1)}, Z_P^{(2)}\}$;
    truncate $\{Z_N^{(1)}, Z_N^{(2)}\}$ by reserving the first N_p
    pairs;
end
Result:
- set of positive pairs: $\{Z_P^{(1)}, Z_P^{(2)}\}$
- set of negative pairs: $\{Z_N^{(1)}, Z_N^{(2)}\}$
- number of positive (negative) pairs: N_p.

FIGURE 5. Pseudo-code describing the hard-negative pairing algorithm.

where D_m denotes the predefined distance margin between the positive and the negative sets; $z_i^{(N1)}$, $z_i^{(N2)}$ represent the i-th pair in the set $\{Z_N^{(1)}, Z_N^{(2)}\}$.

The proposed model adopts the contrastive loss which averages the positive loss as well as the negative loss with the tuning parameter μ [25]:

$$\mathcal{L}_{\text{contrastive}} = (1 - \mu)\mathcal{L}_P + \mu\mathcal{L}_N \qquad (11)$$

By optimizing the parameters of the siamese network, the contrastive loss $\mathcal{L}_{\text{contrastive}}$ are attenuated. In the optimized

model, both $\mathcal{L}_P$ and $\mathcal{L}_N$ are reduced to the range that is determined by the margin D_m, which indicates that the examples from the same VS types are clustered together and the example from the different VS types are mutually separated. Finally, the aforementioned WKNN-based classifier can be adopted to identify the VS type of testing data.

III. CASE STUDIES

To validate the proposed identification method, several tests are carried out with 700 labeled VS samples that are monitored in 2018, in Jiangsu Province, China. Each sample contains a three-phase voltage signal that is captured at the sampling rate of 10 kHz (200 samples per cycle). As mentioned above, the data preprocessing adjusts the signal size to be $3 \times T_{ext}$, with $T_{ext} = 400$. The labeling correctness, as well as the signal integrity of these data, are also manually verified. According to the cause of the VS, all labeled data are classified into five categories: 1) VS due to three-phase faults (3P); 2) VS due to two-phase faults (2P); 3) VS due to single-phase faults (1P); 4) VS due to motor starting (MS); 5) VS due to transformer energizing (TE). Note that both grounded and ungrounded cases are included in the 3P and 2P events and that each of the three phases can be the faulty phase for 2P and 1P events.

The proposed method has been implemented in Python, using the deep learning framework PyTorch. The hyperparameters of the feature extractor, as well as other parameters of the model, are listed in Table 1.

In order to simulate the scenarios where different amount of labeled data are available, the identification performance in five cases where the examined methods are trained with 20, 50, 100, 200, and 500 samples are investigated. In addition, 200 samples are extracted from the rest of the data and utilized as the test data.

For the set of 200 testing samples, the performances of the examined methods on the test set are assessed by the following indexes:

- identification accuracy:

$$A\% = \frac{1}{200}\sum_{i=1}^{200}\delta(l_i^{(\text{predicted})} = l_i^{(\text{true})}) \times 100\% \quad (12)$$

TABLE 1. Hyper-parameters of the feature extractor and model parameters.

Model parameters/hyper-parameters	values
margin in negative pairs D_m	1.0
averaging weight μ	0.5
learning rate lr	5×10^{-3}
input signal length T_{ext}	400
number of epochs n_e	50
optimizer	Adam
number of nearest neighbors	5

FIGURE 6. Convergence of testing loss from the proposed model with different number of training samples.

TABLE 2. Performance of the proposed method on the testing set with different number of training samples.

No. of samples	Accuracy on different types ($A\%$)					Average
	3P	2P	1P	MS	TE	
20	60.0	40.0	62.5	100	100	72.5
50	52.5	67.5	72.5	100	100	78.5
100	97.5	62.5	67.5	95.0	97.5	84.0
200	100	87.5	80.0	100	100	93.5
500	95.0	90.0	87.5	100	100	94.5

where $l_i^{(predicted)}$ and $l_i^{(true)}$ are the i-th predicted label and the real label, respectively;

- identification error:

$$E\% = \frac{1}{200} \sum_{i=1}^{200} \delta(l_i^{(predicted)} \neq l_i^{(true)}) \times 100\% \quad (13)$$

- macro-precision:

$$P\% = \frac{1}{5} \sum_{c \in C} \frac{TP_c}{TP_c + FP_c} \times 100\% \quad (14)$$

where $C = \{3P, 2P, 1P, MS, TE\}$ denotes the set of categories; TP_c, TN_c, FP_c are the number of samples in category c that are true positive, true negative, false positive, respectively.

- macro-F score:

$$F\% = \frac{2 \cdot P \cdot A}{P + A} \times 100\% \quad (15)$$

The accuracy, the macro-precision, and the macro-F score are the typical indexes for the performance evaluation of the examined model. The values of these three indexes are positively correlated to the identification ability.

A. IMPACT OF THE NUMBER OF VS DATA

The amount of training data can affect the training convergence and the testing accuracy of the proposed model. The insufficient number of training samples may result in the diverging model and loss oscillations [26]. Therefore, five testing cases, involving 20, 50, 100, 200, and 500 training samples, are investigated to show the impact of the number of training samples.

Fig. 6 depicts the loss convergence on the testing set with the five cases of the number of training data. As shown, the testing losses with different number of training samples begin to converge after $10 - 15$ training epochs. After 25 epochs, the testing losses for all five cases attain stable values without any oscillations. These observations imply that the proposed method has reliable learning ability. On the other hand, the converged value of losses attenuates when more training samples are provided. The smaller loss indicates
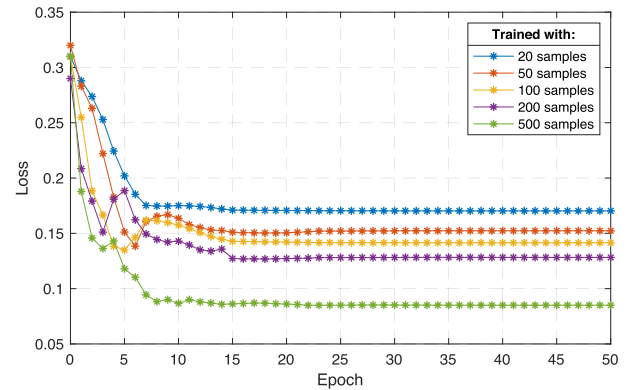
that the samples from different VS categories are more distinguishable in metric of the L-2 distance. Consequently, the increasing number of training samples leads to the rising accuracy of identification. This statement can be verified in Table 2, which shows the accuracies of the proposed method on different VS types with different training data scales.

As presented in Table 2, the proposed method can identify 72.5% of the testing samples when only 20 training samples are available. When the given number of training data increases to 500, the identification accuracy rises consistently at 94.5%, which implies that the proposed method has satisfactory performance in both few-sample and enough-sample cases.

B. IMPACT OF THE VS TYPE

The identification accuracy of the proposed method can vary on the different VS categories. Fig. 7 presents the confusion matrix on the test set, based on the proposed model trained with 500 samples. In Fig. 7, each column presents the distribution of the predicted class, whereas each row describes the distribution of the actual class. According to the confusion matrix given in Fig. 7, 5% of the three-phase faults (3P) samples and 10% of the two-phase faults (2P) samples are misidentified as the single-phase faults, whereas 12.5% of
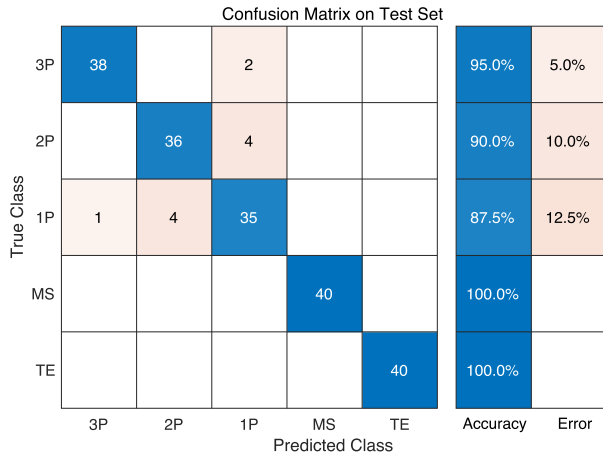
FIGURE 7. Confusion matrix of the proposed model trained with 500 samples.

(a) t-SNE results of 500 training samples

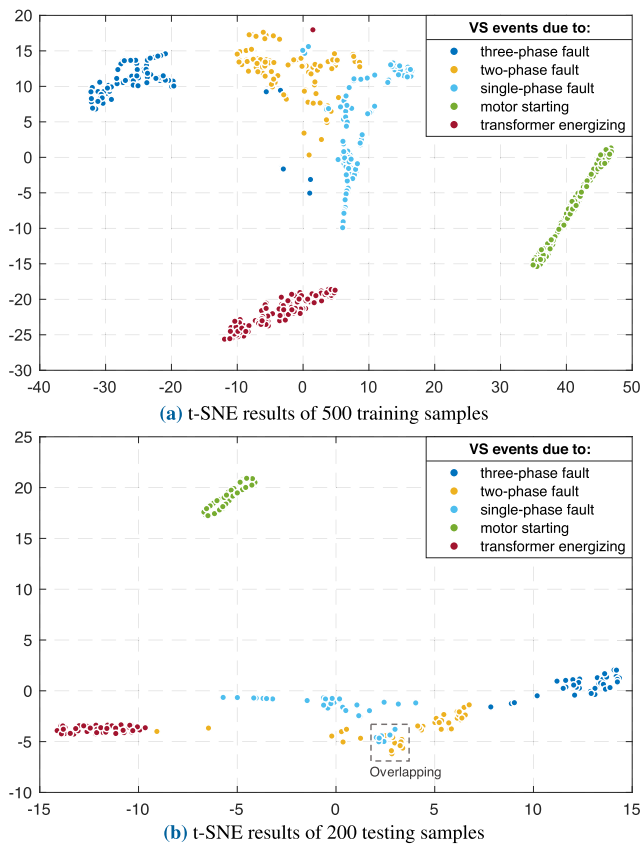**(b)** t-SNE results of 200 testing samples

FIGURE 8. Visualization of training and testing samples using t-SNE.

the single-phase faults (1P) samples are mistaken as the other two faults. On the other hand, the proposed method shows good performance on the non-fault cases, including the motor starting (MS) and the transformer energizing (TE).

Besides, the 500 training samples, as well as the 200 testing samples, are visualized in two-dimensional spaces by using the t-distributed stochastic neighbor embedding (t-SNE). The t-SNE technique performs nonlinear dimensionality reduction which maps the high-dimensional vectors (the 32×1 embeddings) into two components, such that the distribution

of these high-dimensional embeddings can be presented and visualized in the 2-D maps [27]. Fig. 8 presents the t-SNE mapping results of the training data and the testing data. As shown, the points corresponding to the motor starting (MS) and the transformer energizing (TE) are tightly clustered and located distantly to the majority of the points corresponding to the three fault cases (3P, 2P, and 1P), which coincides well with the fact that accuracy on MS and TE are relatively higher than the accuracy on the other three cases. On the other hand, the partial overlapping of several training samples and testing samples can be observed. In Fig. 8b, around 10 yellow points (2P) and 10 blue points (1P) are overlapped, as surrounded by the dash-line box. As shown in Table 2, the partial overlapping between these two types of VS samples yields the relatively lower testing accuracies on these two types (90.0% and 87.5% for 2P and 1P cases, respectively).

Besides, even though the identification accuracy varies among the cases of different VS types, the accuracy on each VS category goes beyond 87.5%, which is still in the acceptable range.

C. COMPARISON AND DISCUSSION

To show the superiority of the proposed method over the other VS identification methods, the proposed method has been compared with the baseline method and the method presented in [12] (Sha's method). The baseline model has a similar CNN structure whose embedding layer is connected directly to a fully-connected layer to generate the outputs. The softmax activation function, as well as the cross-entropy loss function, are selected in the baseline model such that the baseline model can be regarded as the simple six-layer (*conv1-3*, *fc1-2*, and the additional *fc3*) CNN. Note that neither the siamese network nor the few-shot learning technique is adopted in the baseline model. In Sha's method, the K-means-singular value decomposition (K-SVD) technique is utilized to extract the features of the input signal; then the least squares support vector machine (LS-SVM) classifier is adopted to identify the VS type.

Table 3 presents the performance on the 200 testing samples of the three aforementioned methods are examined in five cases where the number of training samples varies among 20, 50, 100, 200, and 500. As shown, both the baseline model and Sha's method suffer degradation on the performance of VS identification when only 20 training samples are available. In the case of 20 training samples, however, the proposed method maintains at least 72.3% precision, 72.5% accuracy, and 72.4% F-score, compared with the baseline model and Sha's method which require 200 training samples to attain the similar values of these three indexes. On the other hand, when the number of training data is sufficient (such as 500 training samples), both the baseline model and Sha's method generate satisfactory precision, accuracy, and F-score which are over 80%. In this case of 500 training samples, all of the three indexes for the proposed method are beyond 94%. Compared to Sha's method, the proposed method enhances at least 8.2% in the precision, 8.5% in the accuracy, and 8.3%

TABLE 3. Identification Precision, Accuracy, and F-score on the testing set of the Baseline Method, the Method introduced in [12], and the Proposed Method on the testing set.

No. of samples	Baseline model			Sha's method			Proposed method		
	P%	A%	F%	P%	A%	F%	P%	A%	F%
20	84.2	25.0	38.6	31.8	37.0	34.2	72.3	72.5	72.4
50	60.8	26.0	36.4	52.2	54.0	53.1	81.9	78.5	80.2
100	65.3	45.3	53.5	56.8	56.0	56.4	84.1	84.0	84.0
200	75.9	64.0	69.4	72.2	71.0	71.6	93.5	93.5	93.5
500	85.4	82.0	83.7	86.4	86.0	86.2	94.6	94.5	94.5

in the F-score. Consequently, the proposed method shows advantages over the other methods in different testing cases.

Other training-based methods such as [5], [11], [13] also require a large scale of training data to optimize the built-in parameters. Although these methods show good performance with a sufficient number of input data, their performance in the few-data case may still be degraded.

In general, these comparisons evinces that the proposed method shows good performance with sufficient number of training data and outperforms other methods in the case where only a few training samples are available.

IV. CONCLUSION

An efficient method for the identification of VS events is introduced in this paper. The proposed method constructs the siamese network and learns from a few available training data. After the training period, the features of the preprocessed data are extracted by the CNN-based feature extractor. Then the identification results are generated from the WKNN-based classifier. The proposed method, as well as the other two identification methods, are implemented and tested with 20 - 500 training samples. The testing results suggest that the proposed method is promising in identifying the VS categories when the number of available training samples is insufficient.

It should be noted that in the training period of the proposed method, the problem of overfitting may occur, especially in the cases where the number of training samples is insufficient. Future work will try to resolve this problem by adding dropout layers. Besides, other disturbance sources such as the capacitor switching may also cause the VS events. Future research will include these sources to make the proposed method more complete.
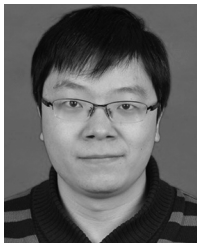
REFERENCES

[1] M. H. J. Bollen and I. Y. H. Gu, *Signal Processing of Power Quality Disturbances*. Hoboken, NJ, USA: Wiley, 2006.

[2] Y. Mohammadi, M. H. Moradi, and R. C. Leborgne, "Locating the source of voltage sags: Full review, introduction of generalized methods and numerical simulations," *Renew. Sustain. Energy Rev.*, vol. 77, pp. 821–844, Sep. 2017.

[3] M. H. J. Bollen, "Voltage sags in three-phase systems," *IEEE Power Eng. Rev.*, vol. 21, no. 9, pp. 8–15, Sep. 2001.

[4] Y. Mohammadi, M. H. Moradi, and R. C. Leborgne, "A novel method for voltage-sag source location using a robust machine learning approach," *Electr. Power Syst. Res.*, vol. 145, pp. 122–136, Apr. 2017.

[5] E. Balouji, I. Y. H. Gu, M. H. J. Bollen, A. Bagheri, and M. Nazari, "A LSTM-based deep learning method with application to voltage dip classification," in *Proc. 18th Int. Conf. Harmon. Qual. Power (ICHQP)*, May 2018, pp. 1–5.

[6] E. Styvaktakis, M. H. J. Bollen, and I. Y. H. Gu, "Expert system for classification and analysis of power system events," *IEEE Trans. Power Del.*, vol. 17, no. 2, pp. 423–428, Apr. 2002.

[7] A. M. El-Zonkoly and H. Desouki, "Wavelet entropy based algorithm for fault detection and classification in FACTS compensated transmission line," *Int. J. Elect. Power Energy Syst.*, vol. 33, no. 8, pp. 1368–1374, 2011.

[8] M. R. Alam, K. M. Muttaqi, and A. Bouzerdoum, "A new approach for classification and characterization of voltage dips and swells using 3-D polarization ellipse parameters," *IEEE Trans. Power Del.*, vol. 30, no. 3, pp. 1344–1353, Jun. 2015.

[9] T. García-Sánchez, E. Gómez-Lázaro, E. Muljadi, M. Kessler, and A. Molina-García, "Approach to fitting parameters and clustering for characterising measured voltage dips based on two-dimensional polarisation ellipses," *IET Renew. Power Gener.*, vol. 11, no. 10, pp. 1335–1343, Aug. 2017.

[10] T. García-Sánchez, E. Gómez-Lázaro, E. Muljadi, M. Kessler, I. Muñoz-Benavente, and A. Molina-García, "Identification of linearised RMS-voltage dip patterns based on clustering in renewable plants," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 6, pp. 1256–1262, Mar. 2018.

[11] P. G. V. Axelberg, I. Y. H. Gu, and M. H. J. Bollen, "Support vector machine for classification of voltage disturbances," *IEEE Trans. Power Del.*, vol. 22, no. 3, pp. 1297–1303, Jul. 2007.

[12] H. Sha, F. Mei, C. Zhang, Y. Pan, and J. Zheng, "Identification method for voltage sags based on k-means-singular value decomposition and least squares support vector machine," *Energies*, vol. 12, no. 6, pp. 1115–1137, Mar. 2019.

[13] D. Li, F. Mei, C. Zhang, H. Sha, and J. Zheng, "Self-supervised voltage sag source identification method based on CNN," *Energies*, vol. 12, no. 6, pp. 1014–1059, Jan. 2019.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). Berlin, Germany: Springer-Verlag, 2006.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[16] H. Iqbal, "Harisiqbal88/plotneuralnet v1.0.0," Zenodo, Tech. Rep. 2526396, Dec. 2018. [Online]. Available: https://zenodo.org/, doi: 10.5281/zenodo.2526396.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2015, pp. 1026–1034.

[18] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Oct. 2009, pp. 2146–2153.

[19] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2559–2566.

[20] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1735–1742.

[21] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, no. 4, pp. 325–327, Apr. 1976.

[22] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

[23] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. ICLR*, 2017, pp. 1–11.

[24] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[25] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: https://arxiv.org/abs/1703.07737

[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[27] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

GUANGYU YANG was born in Anhui, China. He received the B.S. degree in electrical engineering from the Hefei University of Technology, Hefei, China, in 2017.
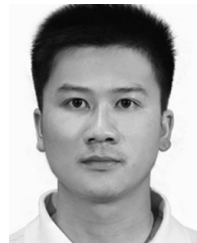
His current research interests include power quality improvement and micro-grid management.

HAOTIAN SUN received the B.S. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2017, where he is currently pursuing the M.S. degree with the Department of Electrical Engineering.

His research interests include power quality improvement and fault location in distribution systems.

FANG ZHUO (M'00) was born in Shanghai, China, in 1962. He received the B.S. degree in automatic control and the M.S. and Ph.D. degrees in automation and electrical engineering from Xi'an Jiaotong University (XJTU), Xi'an, China, in 1984, 1989, and 2001, respectively.

He was an Associate Professor with XJTU, in 1996, and a Full Professor in power electronics and drives, in 2004. He was a Supervisor of Ph.D. students. His research interests include power electronics, power quality, active power filter, reactive power compensation, and inverters for distributed power generation.

Dr. Zhuo is a member of the China Electro Technical Society, Automation Society, and Power Supply Society. He is also the Power Quality Professional Chairman of the Power Supply Society in China.

HAO YI (S'10–M'14) received the M.S. and Ph.D. degrees in electrical engineering from Xi'an Jiaotong University (XJTU), Xi'an, China, in 2010 and 2013, respectively.

Since 2013, he has been a member of the School of Electrical Engineering, XJTU. His current research interests include modeling and control of high-power converters, control and power management of micro-grids, and power quality improvement.

ANPING HU received the B.Sc. degree from the Zhengzhou University of Light Industry, Zhengzhou, China, in 2006, the M.A.Sc. degree from Guizhou University, Guizhou, China, in 2009, both in electrical engineering, and the Ph.D. degree in electrical engineering from the Hefei University of Technology.

He is currently working with the China Electric Power Research Institute. His research interests include power converter systems, electric motor drives, and renewable energy conversion systems.

• • •