IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 June 2023, accepted 14 July 2023, date of publication 24 July 2023, date of current version 2 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3298358

RESEARCH ARTICLE

Disturbance Classification Method for Microgrids Based on EEMD-Transformer-SVM

QIAN ZHOU[1], SHILING ZHANG[2], (Member, IEEE), YANJUN LI[3], ZHE LI[2], (Member, IEEE), QIN FANG[4], AND HAN HUANG[3]

[1]State Grid Chongqing Electric Power Company, Chongqing 401123, China
[2]Chongqing Electric Power Company Scientific Research Institute, Yubei, Chongqing 401123, China
[3]Bishan Power Supply Bureau of State Grid Chongqing Electric Power Company, Bishan, Chongqing 402760, China
[4]State Grid Chongqing South Shore Power Supply Bureau, Nanan, Chongqing 400000, China

Corresponding author: Yanjun Li (2140632311@qq.com)

This work was supported by State Grid Chongqing Bishan Company (Research on Safety Risk of Power Equipment in the Construction of New Power System in Chongqing) under Grant 20B331-9003001-0908.

ABSTRACT Accurate real-time power system disturbance classification is beneficial in avoiding system faults. However, in the process of disturbance detection, the quality of data obtained from the synchronous phase measurement unit (PMU) can be problematic, seriously affecting its application in disturbance classification. Moreover, existing methods are unable to accurately classify data with excessive noise. To address this problem, a disturbance classification method based on ensemble empirical mode decomposition, Transformer neural network, and support vector machine (EEMD-Transformer-SVM) was proposed. First, considering the nonlinear and non-stationary characteristics of microgrid disturbance data, using ensemble empirical mode decomposition to extract data features could effectively reduce the difficulty of fitting nonlinear fluctuation patterns in machine learning models, while avoiding interference between local features. Moreover, to capture and amplify the effective information in the data, a Transformer with a multilayer self-attention encoder network was proposed, which could further transform the data features after EEMD. Finally, the features were classified using a support vector machine. Based on the Consortium for Electrical Reliability Technology Solutions (CERTS) microgrid system, the proposed method was tested under different disturbance data to verify its accuracy and efficiency. By testing the data classification performance in different scenarios, the method demonstrated a high level of generalization.

INDEX TERMS Disturbance classification, empirical mode decomposition, Transformer-based, self-attention, support vector machine, CERT, generalizability.

I. INTRODUCTION

A. RELATED WORK

With the enormous use of nonlinear loads—such as high proportional power electronic devices within the power sector—the power-quality problem is becoming increasingly prominent, having a great impact on the stable operation of power systems. Therefore, to manage power quality problems and to improve the quality of power supply, it is important to identify and classify power quality disturbances accurately and efficiently, and to provide sufficient information for further proposed remedial measures.

Most of the power quality disturbances are non-smooth signals, and different methods of feature extraction combined with classifiers are often used to solve the power quality disturbance classification problem [1], [2]. The main methods of feature extraction include the S-transform (ST) [3], discrete wavelet transform (DWT) [4], short-time Fourier transform (STFT) [5], and Hilbert–Huang transform (HHT) [6], [7]. The ST has good noise immunity and performance in time-frequency analysis, but is still limited by Heisenberg's inaccuracy principle and has poor real-time performance. The DWT has variable time-frequency resolution, but its analysis can

The associate editor coordinating the review of this manuscript and approving it for publication was Gab-Su Seo.

be influenced by the number of decomposition layers and wavelet basis selection.

In recent years, scholars have proposed new improved feature extraction methods, such as successive variational mode decomposition (SVMD) [8] and multi-scale fluctuation-based dispersion entropy (MFDE) methods [9]. After feature extraction, the extracted features need to be entered into the classifier model. Most of the existing studies use the direct multi-classification method, but this strategy is only applicable to the identification of single disturbances and a small number of composite disturbances. With increasing disturbance types, the classifier faces dimensional disaster and the classification efficiency and accuracy is greatly reduced. For this reason, using a support vector machine (SVM) [10] for classification, the idea of multi-label classification [11] was introduced into the disturbance classification problem, its implementation methods being specifically divided into two categories—that is, algorithm modification strategies and problem transformation strategies. For example, the artificial neural network (ANN) [12] effectively uses deep learning to solve the classification problem end-to-end. The disturbance signal can be converted into a two-dimensional image [13] through a phase space reconstruction algorithm, with the generated two-dimensional image being used as the convolutional neural network (CNN) training and testing dataset to achieve power quality disturbance classification. The original signal of power quality disturbance is grayed out [14], the generated grayed-out feature map then being used as the input dataset of the CNN model, and the grayed-out feature data being automatically extracted by the CNN to recognize power quality disturbances.

A combination of a self-encoder and CNN [15] can be used to recognize disturbances, mapping the high-dimensional input feature data to the low-dimensional hidden variable features through the self-encoder and using the output hidden variable features as the input features of the CNN to complete the classification of power quality disturbances. The recursive graph method [16] can be used to generate a two-dimensional trajectory map of power quality disturbance signals, building a long short-term memory (LSTM) network as a classification model, and using the classification model to extract deep features of power quality disturbance; simulations proved that the algorithm could effectively realize the classification of power quality disturbance.

The power quality disturbance signal can be constructed as a time-frequency matrix [17] by using an incomplete S-transformation algorithm, from which a variety of effective classification features can be extracted as the sample dataset of a deep feedforward network, a dropout regularization layer being added to optimize the neural network model to improve the classification accuracy; simulations proved that the method could effectively achieve a total of 17 disturbance types—including single and compound disturbances. The power quality (PQ) composite disturbances were identified based on the modified multi-label radial basis function (MLRBF) neural network and three-layer Bayesian Network (TLBN) [18], [19], respectively, which could fully exploit the association between the disturbance labels to improve the classification effect, but normally at an enormous computational cost. Additionally, combining the LSTM and broad learning system (BLS) algorithms [20] or the particle swarm optimization (PSO) and BLS algorithms [21] could further improve the classification accuracy of the model. The binary relevance (BR) method [22] has been used to convert the multi-label problem into multiple binary classification subproblems, the time complexity of the algorithm being greatly reduced, its classification accuracy being largely dependent on the generalization performance of each sub-classifier.

B. CONTRIBUTION

However, there are still many aspects of the disturbance data classification problem that need to be further improved. Traditional microgrid disturbance detection methods rely on manually designed features, which cannot fully cover the complexity of measuring disturbance data. Moreover, the combination of a certain feature extraction technique and a certain classification model is almost arbitrary. Consequently, the process of feature design and feature selection lacks generalizability. In this work, a data-driven deep learning model was established to discriminate disturbance types. To enhance the input features of the model and the classifier model to effectively capture and amplify the valid information in the data for fast and accurate classification and identification of complex disturbances, this study proposed a method that combined ensemble empirical mode decomposition with a Transformer neural network and support vector machine (EEMD-Transformer-SVM) for disturbance data classification—in short, a class of data-driven machine learning techniques used for efficient identification and categorization of disturbance types to enable swift protection and remedial measures.

The contributions of this study can be outlined as follows:

1) This paper mined the maximum feature information of disturbance data by combining EEMD with the Transformer model, and addressed the challenge of disturbance classification using a data-centric approach.
2) To further strengthen the classification effect of the model, the tail of the model was connected with an SVM to further classify the results of the Transformer model.
3) The PSCAD simulation was used to produce disturbance data, compared and evaluated the effectiveness of the proposed classification approach and conducted a comparative analysis with existing methods.

C. WORK ORGANIZATION

This work is structured as follows: **Section II** introduces the studied Microgrid(MG) system; **Section III** focuses on the principles involved in the proposed model; **Section IV** describes the overall architecture of the proposed model. In **Section V**, the simulation results are discussed. **Section VI** presents the conclusions.

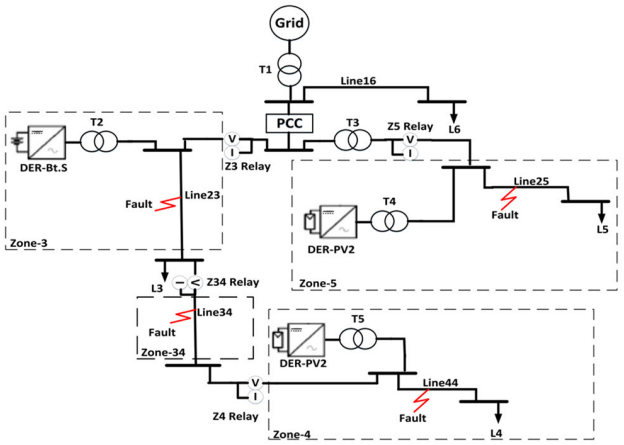

FIGURE 1. CERTS microgrid system structure [24].

TABLE 1. CERTS microgrid system parameters [22].

Transformer Parameter	Value
T1, T2, T3, T4 90 kW, 45 kVAR and T5	13.8/0.48 kV, X/R=6, Z=5% 0.48/0.48 kV, X/R=5, Z=1%
Load Parameter	Value
L3, and L4 L5 L6	90 kW, -40VAR 90 kW, -20VAR
Line Parameter	Value
Line16, Line25, Line23, and Line44 Line34	Size (AWG2), 68.58 m Size (AWG2/0), 22.86 m
DER Parameter	Value
DER-Bt.S, DER-PV2, and DER-PV2	100 kW, unity power factor (3-phase capacitor bank (15 kVA))

II. THEORY: MICROGRID SYSTEM

In this paper, an improved microgrid system based on the CERTS microgrid is the research subject [23], [24], [25]. As shown in Fig. 1 and Fig. 2, the operating voltage of the microgrid adopted is 0.48 kV and 50 Hz, which can support grid-connected or islanded load modes and is controlled by the common coupling point (PCC) switching state. Additionally, loop switches enable the system to operate in either radial or loop topology. Three distributed energy resources (DER) is used in the system—that is, a battery storage system (DER-Bt.S in Fig. 1) and two photovoltaic (PV) power sources (DER-PV). The DER-Bt.S is connected to the system via a current-controlled voltage-source inverter, its control strategy switching to a frequency-controlled inverter in island mode. The DER-PV controls the voltage-source converter interface through current, maintaining its control strategy in grid-connected and islanded modes.

The parameters of the load, transformer, and transmission line is set according to the literature [22] and made simple improvements to some parameters. The specific CERTS microgrid parameters are shown in Table 1. The construction of the model in PSCAD simulation software is completed and generated data, which were used to detect the type and location of disturbances. For further details about the simulation construction, please refer to [26] and [27].

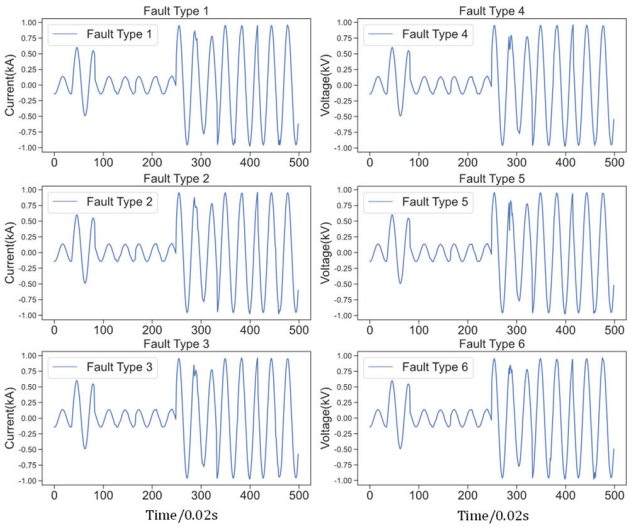

FIGURE 2. Disturbance data generated by microgrid systems.

III. METHODOLOGY

A. THE ENSEMBLE EMPIRICAL MODE DECOMPOSITION (EEMD)

Considering that traditional empirical mode decomposition [28] can suffer from mode mixing defects due to extreme value mutations caused by transient pulses—resulting in the loss of some time or frequency scales—thus causing errors in the decomposed modal components which cannot perfectly restore the characteristics of the original signal, affecting the final disturbance discrimination effect. Consequently, in this study, Ensemble empirical mode decomposition [29], [30] is adopted to decompose the sequence into several intrinsic mode functions (IMFs) and residuals of different frequencies, each of which represents a local feature on a certain time scale. The frequency components contained in each decomposed mode component are different, which effectively suppresses the phenomenon of mode mixing and avoids mutual interference between local features. EEMD can help the model more efficiently learn the data features of different disturbance types and improve the accuracy of disturbance discrimination.

The principle of the EEMD [31] method can be described as follows: the added Gaussian white noise fills the entire time-frequency space uniformly, promoting the natural separation of frequency scales, and reducing the occurrence of mode overlap problems due to IMF discontinuities in the EMD method. The EEMD method can be described as follows:

Step 1: Initialize the EMD method. Set the original signals' processing M times. Add random white noise of different amplitudes to each of these M original signals to form a new series of signals. The amplitude coefficient of the noise is k, and the execution number is $m = 1$.

Step 2: The numerically generated white noise $x_m(t)$ with the given amplitude to the original signal $x(t)$ can be generated, as follows:

$$x_m(t) = x(t) + k^* N_m(t) \tag{1}$$

where $x(t)$ denotes the original signal, $N_m(t)$ denotes randomly added white noise, and k is the amplitude coefficient.

Step 3: The EMD method can be used to decompose $x_m(t)$ into a series of IMF components $C_{j,m}.C_{j,m}$ is the j^{th} IMF of the m-th decomposition:

$$x_m(t) = \sum_{j=1}^{S} C_{j,m}(t) + r_{j,m}(t) \tag{2}$$

where S denotes the number of IMFs, $C_{j,m}(t)$ denotes the IMFs, which include different frequency bands ranging from high frequency to low frequency, and $r_{j,m}(t)$ denotes the final residue, which is the mean trend of the signal.

Step 4: If the decomposition number is $m < M$, then $m = m + 1$. Then, return to Step 2.

Step 5: The corresponding average can be calculated for the M-th decomposition. The result of the calculation can be obtained as follows:

$$\overline{C}_j(t) = \frac{1}{M}\sum_{i=1}^{M} C_{j,m}(t) \tag{3}$$

where $\overline{C}_j(t)$ denotes the jth IMF decomposed by EEMD when $m = 1, 2, 3, \ldots, M$.

The total operation number M and the noise amplitude α are the two parameters that need to be specified in the EEMD method. The final average value of the corresponding IMF can then be offset with the added white noise to improve the signal-to-noise ratio. So, the two parameters need to be chosen carefully. It should be noted that the number of sets should increase as the amplitude of the noise increases to reduce the contribution of additional noise to the decomposition result.

B. COMPOSITE MULTISCALE ENTROPY (CMSE)

For each decomposed IMF column, the composite multiscale entropy (CMSE) is computed and reorganized the IMFs with close entropy values. The flowchart of the CMSE algorithm is shown in Fig. 3. At a given scale factor τ, The SampEns of all coarse-grained time series can be computed at that scale factor, and the multiscale entropy value can be defined as the average of all entropy τ values [32], as follows:

$$\mathrm{CMSE}(x, \tau, \mathrm{m}, \mathrm{r}) = \frac{1}{\tau}\sum_{k=1}^{\tau} \mathrm{SampEn}\left(y_k^{(\tau)}, m, r\right) \tag{4}$$

where $x = \{x_1, x_2, \cdots, x_N\}$ denotes a one-dimensional time series of length N, m denotes the dimension of each pair in the matched vector pair for the SampEn calculation, and r denotes a predefined tolerance threshold.

SampEn quantifies the probability that two sequences of m consecutive data points that are similar to each other remain similar when an additional consecutive point is included.

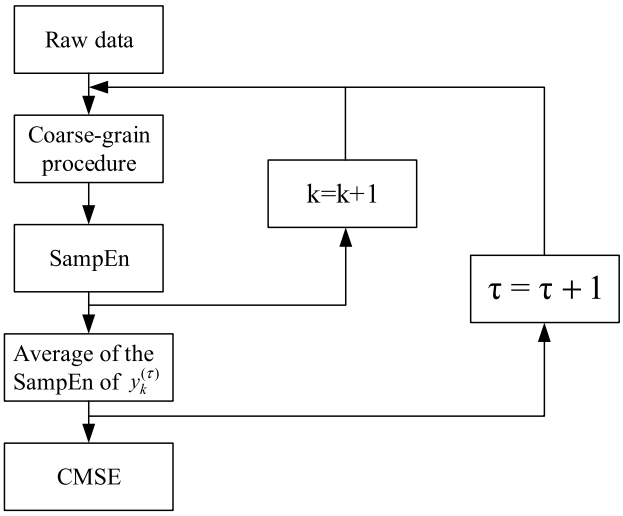

FIGURE 3. Flow chart of the CMSE algorithm.

Being "similar" means that the distance between two vectors is less than a tolerance threshold value r, the distance being defined as the maximum of the absolute differences between their components in the two vectors. Here, the $m = 2$ and $r = 0.15\sigma$ can be used to calculate SampEn, where σ denotes the standard deviation of the time series (raw THz time domain data). The cases of $m = 2$ and r values between 0.1σ and 0.25σ have been widely adopted in previous SampEn calculations, because (in the case of $m = 2$), SampEn is not related to the time series length. The r value used here is also within the reported range of SampEn applications and has been adopted in previous CMSE studies.

$y_{\mathrm{k}}^{(\tau)} = \left\{y_{k,1}^{(\tau)}, y_{k,2}^{(\tau)}, \cdots, y_{k,p}^{(\tau)}\right\}$ is the k_{th} coarse-grained time series at a scale factor of τ and can be defined as follows:

$$\mathrm{y}_{\mathrm{k,j}}^{(\tau)} = \frac{1}{\tau}\sum_{i=(j-1)\tau+k}^{j\tau+k-1} x_i,\ 1 \le j \le \frac{\mathrm{N}}{\tau},\ 1 \le k \le \tau \tag{5}$$

C. SELF-ATTENTION

The self-attentive layer can be split into h attention heads for input. In the output section, the results of each header can then be concatenated to form the output of the layer, and a parametric linear transformation can be applied. Each sequence is entered separately with h input headers, $x = (x_1, \ldots, x_n)$ of n elements where $x_i \in \mathbb{R}^{d_x}$, and a new sequence $z = (z_1, \ldots, z_n)$ of the same length can be computed, where $z_i \in \mathbb{R}^{d_z}$. Each output element z_i can be calculated as a weighted sum of the input elements of the linear transformation, as follows:

$$z_i = \sum_{j=1}^{n} \alpha_{ij}\left(x_j W^V\right) \tag{6}$$

Each weight coefficient, α_{ij}, can be computed using a *softmax* function, as follows:

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}} \tag{7}$$

where $j = 1, 2, 3, \ldots, S$. Moreover, e_{ij} can be computed using a compatibility function that compares two input elements, as follows:

$$e_{ij} = \frac{\left(x_i W^Q\right)\left(x_j W^K\right)^T}{\sqrt{d_z}} \tag{8}$$

The scaled dot product can be used as a compatibility function to optimize the computational efficiency, the linear transformation of the input adding sufficient expressiveness. Finally, $W^Q, W^K, W^V \in \mathbb{R}^{d_x \times d_z}$ are parameter matrices and for each layer and attention header, these parameter matrices are unique.

D. SUPPORT VECTOR MACHINES (SVM)

The support vector machine (SVM) can be considered to be a supervised learning method used to accomplish classification tasks or regression tasks, the purpose of which is to find a hyperplane in a high-dimensional space that maximizes the distance between classes. The SVM for classification tasks can be as follows:

$$\begin{aligned} &\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \\ &\text{subject to } \left| y_i - w^T \phi\left(x_i\right) - b \right| \le \epsilon + \zeta_i \\ &\qquad \zeta_i, \zeta_i^* \ge 0, i = 1, 2, \cdots n \end{aligned} \tag{9}$$

where $x_i \in R^p$, $i = 1, 2, \ldots, n$, and $y \in R^n$, y_i denotes the true label for x_i input and $y_i = w^T \phi\left(x_i\right) + b$ can be referred to as the predicted results for x_i, ζ_i denotes the penalty applied if the difference between the prediction $w^T \phi\left(x_i\right) + b$ and the label y_i is greater than the tolerance limit ϵ, and $\phi\left(\cdot\right)$ denotes a function that can be applied when a linear separation of observations is not possible.

In this paper, the low error rate of SVM generalization, its fast classification speed, and the easy interpretation of results are taken advantage of to replace the final output layer of the model for the final classification task of the model.

E. LONG SHORT-TERM MEMORY NETWORK (LSTM)

The LSTM network [33]—a modification of the recurrent neural network (RNN) network structure [34]—emerged to address the challenge of long-range dependencies in long-sequence that beyond the capability of recurrent neural networks (RNNs). Its network structure is as shown in Fig. 4.

The LSTM block contains a memory cell instead of neurons, which can be considered to be a memory unit with state c_t at time t. Apart from the memory cell, three adaptive and multiplicative units control the flow of information in the block. These units include the input gate i_t, output gate o_t, and forget gate f_t. Access to the block or other parts of the network is controlled by the input and output gates, while forgetting gates are used to reset the state of the memory cell. The LSTM block can be described and the hidden state h_t calculated as

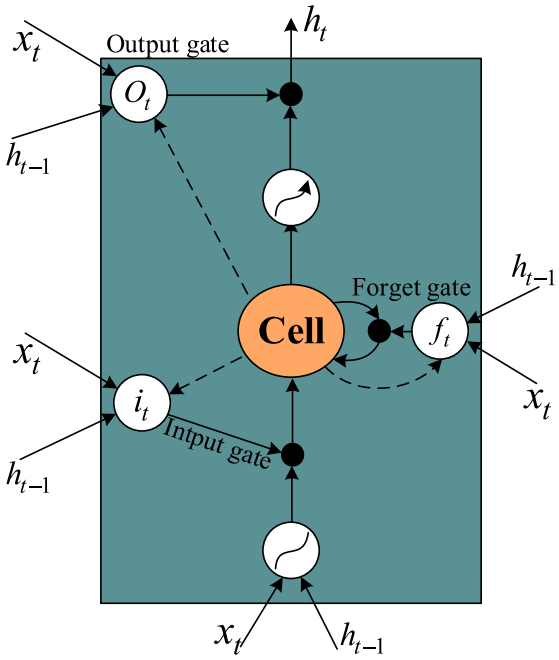

FIGURE 4. The LSTM structure.

follows:

$$f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{10}$$

$$i_t = \sigma_g\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{11}$$

$$o_t = \sigma_g\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{12}$$

$$c_t = f_t \odot c_{(t-1)} + i_t \odot tanh(W_c x_t + U_c h_{(t-1)} + b_c) \tag{13}$$

$$h_t = o_t \odot tanh(c_t) \tag{14}$$

where f_t, i_t and o_t denote the forget gate, input gate, and output gate, respectively; W_f, U_f, W_i, U_i, W_o, U_o, W_c and U_c denote weight matrices; b_f, $b_i b_o$, and b_c denote bias vectors; x_t denotes the current input; h_{t-1} denotes the output of the LSTM at previous time $t - 1$; $\sigma()$ denotes the Sigmoid activation function; and $\odot$ denotes the Hadamard production.

F. TRANSFORMER

The Transformer forms an encoder-decoder network structure by stacking encoder and decoder layers. As shown in Fig. 5, the encoder layer comprises two sub-layers—that is, a self-attention layer followed by a position-wise feed-forward layer. The decoder layer is similar to the encoder and comprises three sub-layers—that is, the self-attention layer followed by the encoder-decoder attention layer, and then the position-wise feedforward layer. It uses residual connections around each of the sublayers, followed by layer normalization. To prevent future information leakage during the training process, masking can be added to the decoder self-attention layer.

Self-attention extracts feature to obtain global relationships between each word, but does not explicitly preserve temporal information, so additional encoding of position is needed to introduce sequence information. In this paper, the position

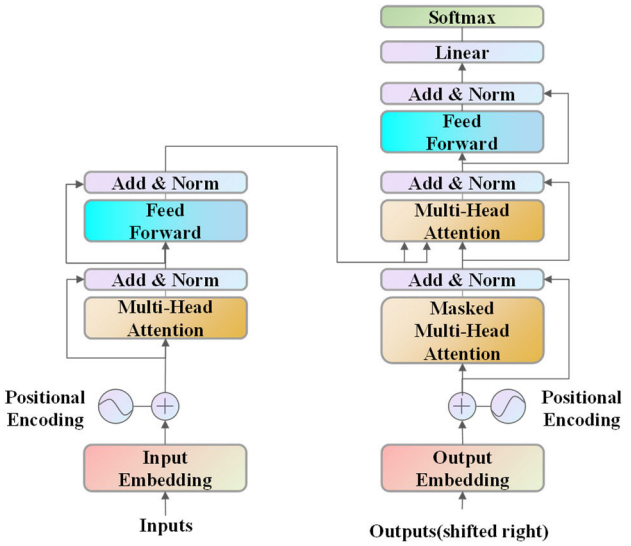

FIGURE 5. The overall structure of the transformer.

TABLE 2. Configurations for disturbance cases simulation [35].

Parameter	Possible Configuration	Count
Topology	Radial or loop	2
Operating mode	Grid-connected or islanded	2
Disturbance type	(A/B/C) G, AB, AC, BC, (AB/AC/BC) G, or ABCG	10
Disturbance resistance (Ω)	0.01, 1, 10, or 100	4
Disturbance line	Line 12, 23, 34, or 56	4
Disturbance location	10%, 20%, or 90% on disturbance line	9

encoding uses the learnable position encoding in Bert. Similar to general word embedding encoding, the learnable position encoding proposed by Bert is randomly generated and trainable with the dimension [*seq_length*, *width*], where *seq_length* denotes the sequence length and *width* denotes the vector length corresponding to each token. Residual connections help propagate position information to higher layers.

IV. RESULTS: CASE STUDY

In this research, the CERTS microgrid model was constructed utilizing PSCAD simulation, and a set of ten predefined disturbance types were incorporated. Various line disturbances were introduced at different locations within the microgrid model to generate corresponding data waveform sequences, as shown in Table 2.

A. SIMULATION ENVIRONMENT

In the simulations conducted, disturbance data for the microgrid were generated through PSCAD simulations. The hardware platform used in this work included a laptop equipped with a GeForce RTX 1080 GPU and Intel Core i7-7700HQ. The software components consisted of PSCAD 4.6.2, Python 3.8, and the following packages with their respective versions used in data processing and deep learning modeling—namely, Keras 2.6.0 with Tensorflow 2.6.0 as the backend, Numpy 1.19.5, Pandas 1.2.4, pyemd 0.5.1, and scipy 1.9.3.

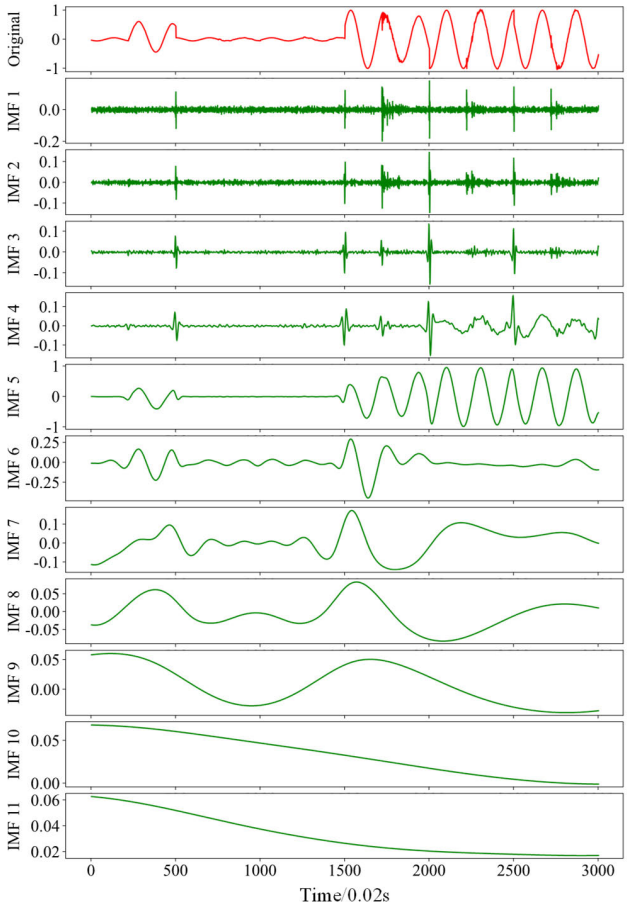

FIGURE 6. Data decomposition results in the time and frequency domains.

B. DATA TIME-FREQUENCY DOMAIN TRANSFORMATION

In this paper, the EEMD is used to decompose the data in the time-frequency domain, and to decompose the perturbed raw data into *s* modal components and one residual component. The composite multiscale entropy values of each component can be calculated, and the sequences with similar CMSE entropy values superimposed to form the recombined component sequences. The recombinant component sequences can then be spliced and input to the model for training. The decomposition results are shown in Fig. 6.

The CMSE of each component is calculated, as shown in the Table 3, and merged the IMFs according to their entropy. Consequently, The IMF1 with IMF2, IMF3 and IMF4, IMF5 with IMF6, IMF7 and IMF8, and IMF9 with IMF10 and IMF 11 were merged to form four subsequences including Res.

C. OVERALL STRUCTURE OF THE MODEL

To maximize the disturbance of data information mining and knowledge learning [36], during the data preprocessing stage, the original data is subjected to time-frequency transformation, and the implicit feature information contained in the data is further calculated by collecting EEMD data to obtain the CMSE of each decomposed mode component. The sequences with similar entropy are merged, and the

TABLE 3. Composite multiscale entropy of each component.

IMFs	CMSE
IMF1	1.00782738993894239
IMF2	0.93782738273827383
IMF3	0.53784823948237498
IMF4	0.37384792838434850
IMF5	0.14783485394085039
IMF6	0.088934898058274059
IMF7	0.05897096770042723
IMF8	0.02186979009867003
IMF9	0.0049745657345837728
IMF10	0.0019697940492394804
IMF11	0.0005085962195798039
Res	1.6434961961271035

decomposed sequence is used as prior knowledge injected into the model, which can help it to learn the characteristics of different disturbance data and converge faster, thereby improving the accuracy of the model in discriminating different disturbance types. The representation learning for the multi-dimensional data which has been transformed from the time domain to the frequency domain and mapped to the high-dimensional space is achieved through the embedding layer. After positional embedding, the data at each position is then tagged. Moreover, after the multi-layer Transformer encoder, data mapping is performed multiple times for feature conversion, after which it is processed through the bidirectional LSTM network. The obtained data is decoded and output by the fully connected layer and SVM. The established model structure is as shown in Fig. 7.

Since most deep learning models embed high-dimensional representations of the input data at the first stage, the same architectural setup was followed. As shown in Fig. 7, the first layer is set as the embedding layer, and the encoder part of the Transformer structure is used, which has advantages in extracting features from disturbance data. The original linearly inseparable disturbance data is then mapped through multiple feature-mapping layers to become linearly separable in high-dimensional space.

Parameter setting of the model is shown in Fig. 8, the number of attention mechanism heads being determined based on the data feature dimension; the design principle being to ensure that the number of heads is equal to the feature dimension. The decomposed data and original data are jointly used as the input data, the input feature number being six. However, to further expand the parallelism of computation, the number of attention mechanism heads were slightly increased to eight, which was determined through simulation. Since the length of the perturbation measurement data is long (n=3000), to improve the computational efficiency, the query to access and interact with the neighborhood must be limited within a small window size (m=100). The feature mapping process of this encoder is repeated five times to complete the perturbed data feature extraction. As the bidirectional LSTM has good decoding abilities for the perturbed data after feature extraction, the model selects the LSTM to complete the decoding work after data encoding. The decoded data features are then constrained to ten types of disturbance using

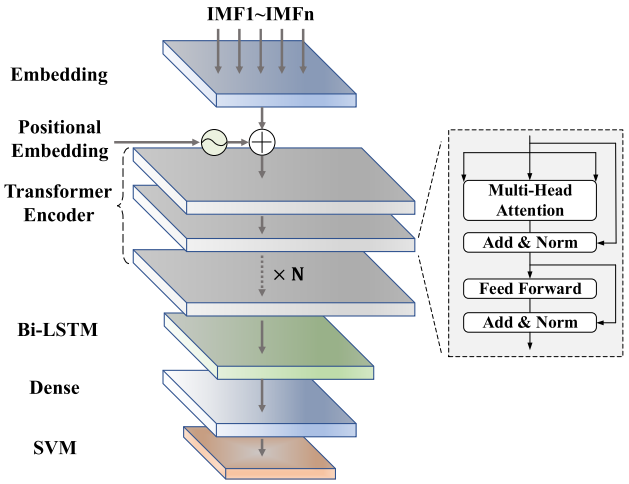

FIGURE 7. The structure of the proposed model.

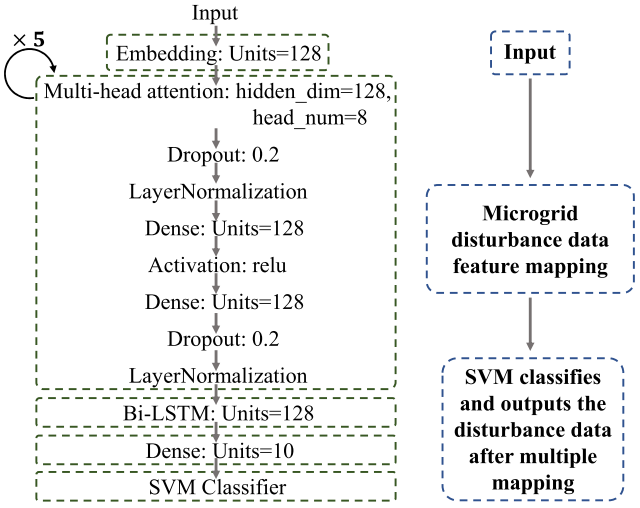

FIGURE 8. Model parameter setting and brief architecture of the model.

TABLE 4. Parameter setting of the model.

Layer	Parameter
Embedding	Hidden_dim: 128
Self-Attention	Hiden_dim: 128; head_num: 8; dropout_rate: 0.2
Transformer Encoder	Block_num: 5
LSTM	Units:128; Return_sequences: True
Loss	Categorical_crossentropy
Optimizer	Adam
Batch_size	45
Epochs	100

SVM classifiers to achieve the perturbation type discrimination output. The configuration of model layers, along with their corresponding parameter values, is provided in Table 4.

A layer in the model is selected to visualize the distribution of its learnable parameters after initialization, as shown in Fig. 9.

The parameters are initialized by using a uniform distribution with 0 mean.

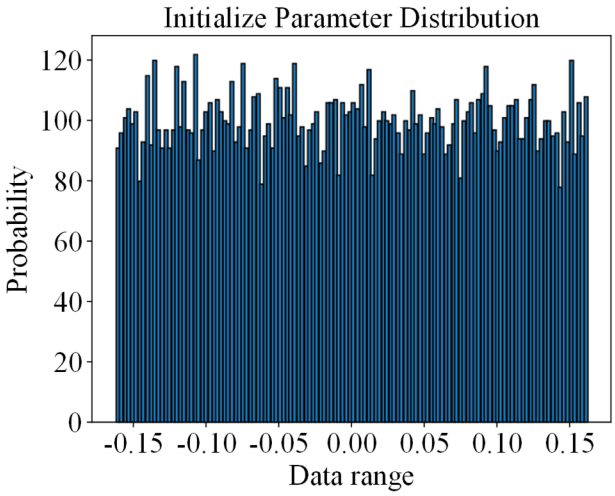

FIGURE 9. Initialized parameter distribution.

FIGURE 10. Training and testing loss of the model.

FIGURE 11. Training and testing accuracy of the model.

V. DISCUSSION: ANALYSIS OF RESULTS

A. SIMULATION RESULTS

After performing the EEMD and reconstruction on the data, it was fed into our proposed model. The model parameters were iteratively updated, and convergence to the optimal state was achieved after 100 training epochs. The training process of the model is depicted in Fig. 10.

As shown in Fig. 11, after training and undergoing multiple layers of Transformer encoders and Bi-LSTM feature mapping, the model is able to effectively extract different feature information from various types of power grid disturbances, causing the originally linearly inseparable power grid disturbance samples to become linearly separable. The advantage of the SVM lies in its principle of only using a portion of support vector samples for classification. By selecting an SVM as the classifier at the end, the model takes on a stacking structure where the next layer of the model further corrects any misclassifications made by the previous layer. This further improves the accuracy of the model's power grid disturbance classification.

The trained model demonstrated excellent performance on both the training and testing datasets, achieving a classification accuracy of nearly 100% on the training dataset and 99% on the testing dataset. The model's consistent performance on both datasets indicates its effectiveness in accurately classifying line disturbances. Additionally, the absence of overfitting or underfitting further strengthens the model's reliability. The confusion matrices of the proposed model can be found in Fig. 12. Using the binary classification problem as an example, the confusion matrix can be expressed as follows:

$$\text{ConfusionMatrix} = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} \tag{15}$$

where TP (true positive) and TN (true negative) denote the number of positive and negative instances that are correctly classified, respectively, and FN (false negative) and FP (false positive) denote the number of positive instances and negative instances that are misclassified, respectively [37]. Based on the confusion matrix, the accuracy can be defined to assess model performances [38], as follows:

$$\text{Accuracy (A)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

In order to minimize the training parameter count of the model, the Bayesian optimization is employed to optimize the hyperparameters and tested with various activation functions during the model debugging phase. After thorough evaluation, the ReLu function was determined to be the most suitable activation function for the residual module.

B. PERFORMANCE COMPARISON

The ablation study is supplemented and tested the effects of other representative time-frequency transform feature mining methods, comparing their classification results. It is evident that the proposed model outperforms all the comparison models, as shown in Table 5.

In Table 5, EMD stands for empirical mode decomposition, WT stands for wavelet transform, WPT stands for wavelet packet transform, and "+" indicates the combination relationship between algorithms. Fig. 13 visualizes the classification accuracy of each model.

The backbone of the proposed model architecture belongs to the deep learning model, which updates the weight parameters through backpropagation and learns an effective disturbance type discrimination model. To control variables and

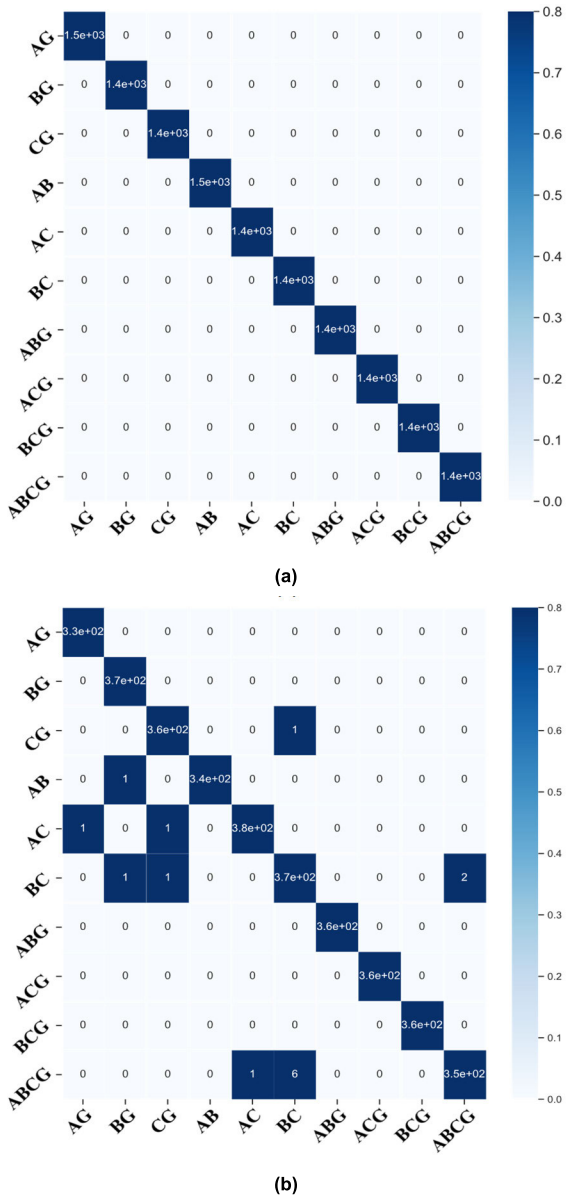

FIGURE 12. Confusion matrices for model disturbance classification performance of the training dataset (a) and testing dataset (b).

TABLE 5. The results of model comparison.

Models	Accuracy
EMD+CNN	95.04%
EMD+LSTM	97.02%
EMD+CNN+LSTM	96.56%
EEMD+CNN	95.43%
EEMD+LSTM	97.29%
EEMD+CNN+LSTM	96.89%
EEMD+RF	98.14%
EEMD+XGBoost	98.39%
WT+CNN	94.02%
WT+LSTM	96.33%
WT+CNN+LSTM	95.39%
WPT+CNN	95.12%
WPT+CNN+LSTM	96.41%
Our proposed model	**99.88%**

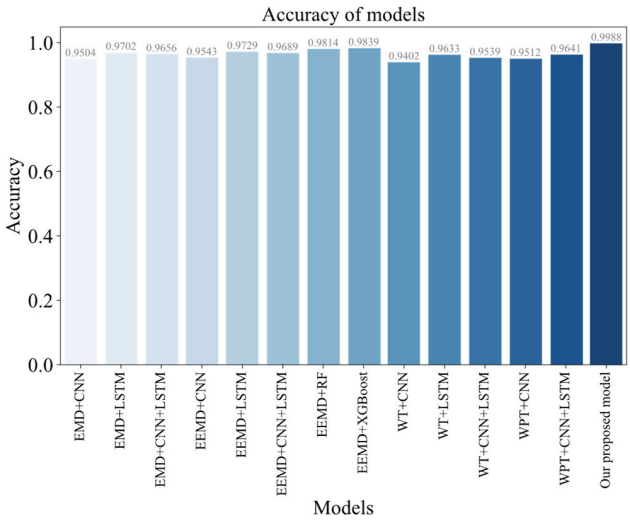

FIGURE 13. Comparison of model accuracy.

TABLE 6. Comparison of computation consumption.

Methods	Complexity per Layer	Params	GPU
Transformer	$O(n^2 \cdot d)$	3.96 M	131 ms
Recurrent Neural Network	$O(n \cdot d^2)$	2.72 M	246 ms
Convolutional Neural Network	$O(k \cdot n \cdot d^2)$	2.41 M	96 ms
Our Proposed Methods	$O(m \cdot n \cdot d)$	3.05 M	112 ms

facilitate the comparison of model computation times, the computational time complexity is compared with multiple typical deep learning model architectures and the actual computation time of each iteration during the training process, as shown in the Table 6.

In Table 6, complexity per Layer denotes the time complexity of each layer in the backbone network of the model, Params denotes the number of parameters in the model (M: million), and GPU denotes the time consumed by each iteration of the model (ms: milliseconds). Moreover, n denotes the length of microgrid disturbance data, d denotes the embedding dimension, k denotes the kernel size of convolutions, and m denotes the size of the neighborhood in restricted self-attention. m is much smaller than the data length n and embedding dimension size d.

As shown in Fig. 14, through visualizing the computation time complexity, computation time, and model parameter volume of various typical networks for microgrid disturbance classification., it is evident that the proposed model has a computationally parallel performance that greatly reduces the model training iteration time while controlling the model parameter volume size. The CNN in the figure is faster through its convolution kernel computation, but its disturbance classification accuracy is 4% lower than that of the proposed model. In summary, by analyzing and comparing the computational consumption of the models and balancing the accuracy of disturbance classification and computation time complexity, it is evident that the proposed model achieved better performance.

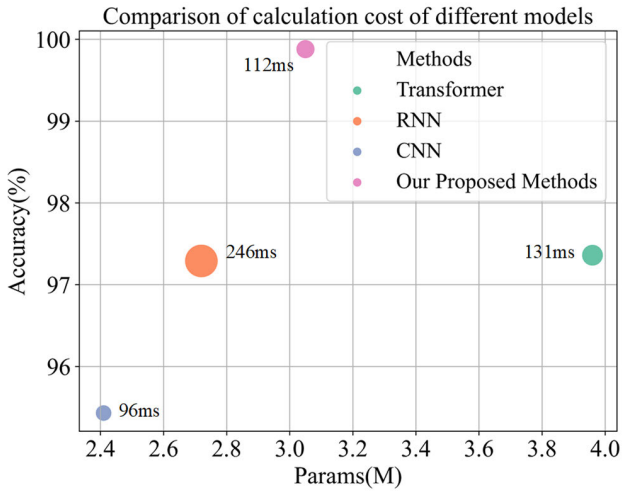

FIGURE 14. Comparison of the calculation cost of different models.

VI. CONCLUSION

The proposed model was superior to the other machine learning models. It could solve the classification of disturbance types end-to-end, without more complex data preprocessing, making it more convenient and effective. Additionally, the model could also consider a variety of factors and maximize the disturbance of the data information mining and knowledge learning. Through a comparative analysis with alternative machine learning models, it was demonstrated that the proposed model exhibits superior robustness and faster convergence rate.

REFERENCES

[1] J. Wang, Z. Xu, and Y. Che, "Power quality disturbance classification based on compressed sensing and deep convolution neural networks," *IEEE Access*, vol. 7, pp. 78336–78346, 2019.

[2] K. Mahapatra, S. Lu, A. Sengupta, and N. R. Chaudhuri, "Power system disturbance classification with online event-driven neuromorphic computing," *IEEE Trans. Smart Grid*, vol. 12, no. 3, pp. 2343–2354, May 2021.

[3] L. Jianmin, L. Haijun, L. Chengbin, T. Zhaosheng, and C. Da, "Detection method of power quality disturbances based on double resolution S transform and learning vector quantization neural network," *Trans. China Electrotech. Soc.*, vol. 34, no. 16, pp. 3453–3463, 2019.

[4] X. Xiangui, L. Kaicheng, C. Delong, W. Menghao, and W. Wei, "A combined de-noising method for power quality disturbances events," *Trans. China Electrotech. Soc.*, vol. 36, no. 21, pp. 4418–4428, 2021.

[5] H. Jianming and L. Xiaoming, "Detection of harmonic in power system based on short-time Fourier transform and spectral kurtosis," *Power System Protection Control*, vol. 45, no. 7, pp. 43–50, 2017.

[6] G. Sharma and R. Sharma, "Cluster-based distributed cooperative spectrum sensing over Nakagami fading using diversity reception," *IET Netw.*, vol. 8, no. 3, pp. 211–217, May 2019.

[7] C. Zhang, S. Zhao, and Y. He, "An integrated method of the future capacity and RUL prediction for lithium-ion battery pack," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 2601–2613, Mar. 2022, doi: 10.1109/TVT.2021.3138959.

[8] Y. Li, B. Tang, and S. Jiao, "SO-slope entropy coupled with SVMD: A novel adaptive feature extraction method for ship-radiated noise," *Ocean Eng.*, vol. 280, Jul. 2023, Art. no. 114677.

[9] Y. Li, Y. Lou, L. Liang, and S. Zhang, "Research on feature extraction of ship-radiated noise based on multiscale fuzzy dispersion entropy," *J. Mar. Sci. Eng.*, vol. 11, no. 5, p. 997, May 2023.

[10] G. Sharma and R. Sharma, "Energy efficient collaborative spectrum sensing with clustering of secondary users in cognitive radio networks," *IET Commun.*, vol. 13, no. 8, pp. 1101–1109, May 2019.

[11] Z. Luowei, G. Chun, and L. Weiguo, "Application of multilabel classification method to categorization of multiple power quality disturbances," *Proc. CSEE*, vol. 31, no. 4, pp. 45–50, 2011.

[12] A. Ali and W. Hamouda, "Advances on spectrum sensing for cognitive radio networks: Theory and applications," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 1277–1304, 2nd Quart., 2017.

[13] Y. Zhu, C. Liu, and K. Sun, "Image embedding of PMU data for deep learning towards transient disturbance classification," in *Proc. IEEE Int. Conf. Energy Internet (ICEI)*, May 2018, pp. 169–174.

[14] S. Wang and H. Chen, "A novel deep learning method for the classification of power quality disturbances using deep convolutional neural network," *Appl. Energy*, vol. 235, pp. 1126–1140, Feb. 2019.

[15] E. Yiğit, U. Özkaya, Ş. Öztürk, D. Singh, and H. Gritli, "Automatic detection of power quality disturbance using convolutional neural network structure with gated recurrent unit," *Mobile Inf. Syst.*, vol. 2021, pp. 1–11, Jul. 2021.

[16] H. Chang, W.-Q. Ge, H.-C. Wang, H. Yuan, and Z.-W. Fan, "Laser beam pointing stabilization control through disturbance classification," *Sensors*, vol. 21, no. 6, p. 1946, Mar. 2021.

[17] S. T. Suganthi, A. Vinayagam, V. Veerasamy, A. Deepa, M. Abouhawwash, and M. Thirumeni, "Detection and classification of multiple power quality disturbances in microgrid network using probabilistic based intelligent classifier," *Sustain. Energy Technol. Assessments*, vol. 47, Oct. 2021, Art. no. 101470.

[18] G. Chun, Z. Luowei, and L. Weiguo, "Recognition of multiple power quality disturbances using multi-label RBF neural networks," *Trans. China Electrotech. Soc.*, vol. 26, no. 8, pp. 198–204, 2011.

[19] Y. Luo, K. Li, Y. Li, D. Cai, C. Zhao, and Q. Meng, "Three-layer Bayesian network for classification of complex power quality disturbances," *IEEE Trans. Ind. Informat.*, vol. 14, no. 9, pp. 3997–4006, Sep. 2018.

[20] S. Zhao, C. Zhang, and Y. Wang, "Lithium-ion battery capacity and remaining useful life prediction using board learning system and long short-term memory neural network," *J. Energy Storage*, vol. 52, Aug. 2022, Art. no. 104901.

[21] C. Zhang, S. Zhao, Z. Yang, and Y. Chen, "A reliable data-driven state-of-health estimation model for lithium-ion batteries in electric vehicles," *Frontiers Energy Res.*, vol. 10, Sep. 2022, Art. no. 1013800.

[22] W.-M. Lin, C.-H. Wu, C.-H. Lin, and F.-S. Cheng, "Detection and classification of multiple power-quality disturbances with wavelet multiclass SVM," *IEEE Trans. Power Del.*, vol. 23, no. 4, pp. 2575–2582, Oct. 2008.

[23] R. H. Lasseter, J. H. Eto, B. Schenkman, J. Stevens, H. Vollkommer, D. Klapp, E. Linton, H. Hurtado, and J. Roy, "CERTS microgrid laboratory test bed," *IEEE Trans. Power Del.*, vol. 26, no. 1, pp. 325–332, Jan. 2011.

[24] T. S. Abdelgayed, W. G. Morsi, and T. S. Sidhu, "A new approach for fault classification in microgrids using optimal wavelet functions matching pursuit," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4838–4846, Sep. 2018.

[25] A. Hooshyar, E. F. El-Saadany, and M. Sanaye-Pasand, "Fault type classification in microgrids including photovoltaic DGs," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2218–2229, Sep. 2016.

[26] A. Yazdani and R. Iravani, *Voltage-Sourced Converters in Power Systems: Modeling, Control, and Applications*. Hoboken, NJ, USA: Wiley, Mar. 2010.

[27] A. Yazdani and P. P. Dash, "A control methodology and characterization of dynamics for a photovoltaic (PV) system interfaced with a distribution network," *IEEE Trans. Power Del.*, vol. 24, no. 3, pp. 1538–1551, Jul. 2009.

[28] C. Junsheng, Y. Dejie, and Y. Yu, "A fault diagnosis approach for roller bearings based on EMD method and AR model," *Mech. Syst. Signal Process.*, vol. 20, no. 2, pp. 350–362, Feb. 2006.

[29] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.

[30] H. Wang, J. Chen, and G. Dong, "Feature extraction of rolling bearing's early weak fault based on EEMD and tunable Q-factor wavelet transform," *Mech. Syst. Signal Process.*, vol. 48, nos. 1–2, pp. 103–119, Oct. 2014.

[31] H. Zhao, M. Sun, W. Deng, and X. Yang, "A new feature extraction method based on EEMD and multi-scale fuzzy entropy for motor bearing," *Entropy*, vol. 19, no. 1, p. 14, Dec. 2016.

[32] S.-D. Wu, C.-W. Wu, S.-G. Lin, C.-C. Wang, and K.-Y. Lee, "Time series analysis using composite multiscale entropy," *Entropy*, vol. 15, no. 3, pp. 1069–1084, Mar. 2013.

[33] W. Sima, H. Zhang, M. Yang, and X. Li, "Diagnosis of small-sample measured electromagnetic transients in power system using DRN-LSTM and data augmentation," *Int. J. Elect. Power Energy Syst.*, vol. 137, 2022, Art. no. 107820.
[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[35] J. J. Q. Yu, Y. Hou, A. Y. S. Lam, and V. O. K. Li, "Intelligent fault detection scheme for microgrids with wavelet-based deep neural networks," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1694–1703, Mar. 2019.
[36] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5927–5935.
[37] Y. Wang, I. L. Bennani, X. Liu, M. Sun, and Y. Zhou, "Electricity consumer characteristics identification: A federated learning approach," *IEEE Trans. Smart Grid*, vol. 12, no. 4, pp. 3637–3647, Jul. 2021.
[38] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, "Confusion-matrix-based kernel logistic regression for imbalanced data classification," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 9, pp. 1806–1819, Sep. 2017.

QIAN ZHOU was born in Kaifeng, China, in 1993. She received the B.Sc., M.Sc., and Ph.D. degrees from Chongqing University, Chongqing, China, in 2003 and 2007. Her research interest includes the fault diagnosis of SF6-based high-voltage electrical insulation equipment.

SHILING ZHANG (Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2009 and 2015, respectively.

His research interests include high voltage and insulation technology, UHV insulation structure optimization design, and intelligent operation and maintenance technology.

YANJUN LI received the B.S. and M.S. degrees in electrical engineering from Chongqing University, Xi'an, China, in 2009 and 2012, respectively.

He has been engaged in field practice and business management of power equipment operation and maintenance for ten years. His research interest includes high voltage and systems.

ZHE LI (Member, IEEE) received the B.E. and B.S. degrees in electrical engineering from Wuhan University, Wuhan, China, in 2012 and 2014, respectively.

His research interests include energy storage, distributed energy resources, and demand-side responses that impact the sustainability and resilience of smart grids.

QIN FANG received the B.S. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2014.

His research interests include the application of power equipment operation inspection, power equipment test technology, and intelligent operation inspection technology.

HAN HUANG received the B.S. and M.A.Eng. degrees in electrical engineering from Chongqing University, China, in 2007 and 2010, respectively.

His research interests include power grid technology, energy storage technology, and power data analysis.