Series Arc Fault Detection of Grid-Connected PV System via SVD Denoising and IEWT-TWSVM

Wei Gao, *Member, IEEE*, and Rong-Jong Wai, *Senior Member, IEEE*

***Abstract*—Series arc fault (SAF) is one of the most harmful faults during the operation of photovoltaic (PV) systems. It is a challenging task to find SAFs promptly for avoiding the PV fire. Aiming at the SAFs under different operating conditions, a novel detection algorithm by combining the Hankel singular value decomposition (Hankel-SVD) denoising method and the improved empirical wavelet transform–twin support vector machine (IEWT-TWSVM) is proposed. The Hankel-SVD algorithm is used to denoise the dc-bus current, which alleviates the influence of switching frequency and irrelevant background noise effectively. Then, the denoised current is decomposed by the IEWT, and the composite multiscale permutation entropy of each frequency band is input into the TWSVM classifier of the salp swarm optimization for completing the fault detection. The proposed algorithm not only can detect SAFs at various fault locations, but also resist dynamic shading, inverter startup, strong wind, and other interference phenomena. Moreover, the detection performance due to the arc transient process, a long-line fault, a single string array fault, a multiply string array fault, and different sampling rates of this model is verified in this study, and the corresponding results are relatively ideal. Experimental results show that the detection accuracy of the proposed method is as high as 98.10% for the measured data, which is more superior than other methods, such as wavelet decomposition, empirical mode decomposition, statistics methods including mean, standard deviation, and entropy.**

***Index Terms*—Composite multiscale permutation entropy (CMPE), improved empirical wavelet transform (IEWT), photovoltaic (PV) power system, series arc fault (SAF), singular value decomposition (SVD), twin support vector machine (TWSVM).**

I. Introduction

DUE to the advantages of photovoltaic (PV) power generation of quick installation, convenient operation, environmental friendliness, and low maintenance cost, it has attracted more and more attention all over the world. With the concerted efforts of governments and enterprises in various countries, the PV industry has developed rapidly. According to the data from the 21st century renewable energy policy network [1], the global new installed capacity of PV power generation in 2018 was 100 GW, with a total capacity of 505 GW. The total installed capacity of global PV power generation is expected to reach 1 TW in 2021. In many countries, PV power generation has become an important and ever-increasing power generation way, which accounted for 12.1% of the total power generation in Honduras in 2018, besides, accounted for 8.2%, 7.7%, and 6.5% in Italy, Germany, and Japan, respectively. As the PV power generation system is exposed to the external environment and suffers from the effects of weathering and aging for a long time, it is easy to occur the deterioration of cables, joints, system components, etc. If repairs and maintenance are not provided regularly, the probability of dc arc faults will be greatly increased, and even serious PV fire accidents can happen [2]. With the increase in the number of residences, commercial PV facilities, and large-scale PV power generation, it has been an important and challenging task to avoid PV fire accidents effectively. Therefore, the National Electrical Code (NEC) [3] requires all PV systems with a dc operating voltage of more than 80 V to be equipped with arc fault circuit interrupters.

Comparing with the ac arc fault, the dc arc fault cannot be detected by the zero crossing of the current. There are two main arc fault types in PV systems including the parallel arc fault (PAF) and the series arc fault (SAF). Moreover, the PAF also includes the grounding arc fault, and it is usually accompanied by large current and voltage changes, which can be easily detected by conventional detection equipment [4]. On the contrary, the current and voltage changes of the SAF are not obvious, so that they are not enough to melt the fuse or activate the overcurrent protection device [5]. In addition, the output characteristics of the PV system are nonlinear. Under the control of the maximum power point tracking (MPPT) algorithm for converters, the waveform of the SAF loses the original variation characteristic, which will easily lead to misjudgment by the detection equipment. Therefore, it is necessary to design a novel SAF detection method with the properties of high flexibility, high reliability, and good robustness to improve the identification ability for SAFs in PV systems.

Previous research methods of arc fault diagnoses for PV systems are mainly studied from three directions, namely the time domain, the frequency domain, and the time–frequency domain. The time-domain method mainly uses statistical indicators, such as mean, standard deviation, correlation coefficient, entropy, to detect arc faults. Lu *et al*. [6] collected ripple components

Manuscript received March 21, 2021; revised May 2, 2021; accepted July 12, 2021. Date of publication August 10, 2021; date of current version October 21, 2021. This work was supported by the Ministry of Science and Technology of Taiwan through under Grant MOST 108-2221-E-011-080-MY3. *(Corresponding author: Rong-Jong Wai.)*

Wei Gao is with the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan, and also with the College of Electrical Engineering and Automation, Fuzhou University, Fuzhou 350108, China (e-mail: gaowei0203@fzu.edu.cn).

Rong-Jong Wai is with the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: rjwai@mail.ntust.edu.tw).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JPHOTOV.2021.3098376.

Digital Object Identifier 10.1109/JPHOTOV.2021.3098376

from the line current and voltage of a dc system, and then calculated their rate of change and standard deviation. After that, multiple thresholds were predetermined to judge arc faults. Although the effectiveness is verified on the PV system, it does not consider the influence of the waveform abnormality caused by the meteorological environment and operating state changes on the algorithm. Chen *et al.* [7] calculated the Euler integral for the mean current of the two time windows before and after, found that there would be a large pulse signal when the arc was started, and then define a detection variable. An arc fault was considered to occur when the detection variable was greater than the predetermined threshold value. However, the arc fault condition is only tested under ideal conditions, and the influence of various interference conditions on the application of the algorithm is not analyzed further. Ahmadi *et al*. [8] proposed the signal-to-noise ratio (SNR) of the cross-correlation function to detect arc faults. By calculating the correlation coefficient of the dc-bus voltage under normal and fault conditions of a PV system, the SNR relationship between them was obtained. Then, the arc fault was determined by the judgment of whether the SNR exceeds the limit value. Unfortunately, this method is prone to misjudgment in the PV system with massive noise. Generally speaking, the time-domain-based method is relatively simple and convenient, but it is relatively weak in the anti-interference ability. Especially, when some of the time-domain-based methods are applied for other types or sizes of PV systems, it is inevitable to reselect suitable thresholds again.

The frequency-domain methods mainly use the discrete Fourier transform (DFT) or the fast Fourier transform (FFT) for arc fault detection. Chae *et al*. [9] investigated an SAF detection method based on relative amplitude comparisons. The DFT was adopted to calculate the standard deviation of the frequency content of several consecutive time windows to determine the threshold of each frequency component. When the detected frequency components in several consecutive cycles were all exceed the predetermined threshold, it was judged as an arc fault. The method in [9] has been tested in a dc microgrid and can effectively avoid some misjudgments caused by interference in a normal state. Once it is applied in a PV system, the difference of switching frequency between different inverters will affect the threshold setting. Besides, the adjustment of the MPPT will also interfere with the algorithm and affect its accurate judgment. In fact, there are relatively few studies on arc fault diagnoses with frequency-domain methods alone.

The time–frequency method combines the time domain and frequency domain information of the waveform for fault analyses whose effective information is more sufficient compared with only the time-domain-based or the frequency-domain-based method. Zhu *et al*. [10] proposed an arc fault detection algorithm based on the discrete wavelet transform (DWT). The signal of the dc-bus current was collected at a sampling frequency of 200 kHz, and the ratio of the average power of the first-level DWT coefficient to the reference average power was used as the features. If they exceeded a certain threshold, it was considered as an arc fault. However, it does not further study the influence of meteorological environment and operating state changes on the algorithm. Xia *et al*. [11] applied the wavelet packet decomposition (WPD) to decompose the current signal at a sampling frequency of 250 kHz into three layers and four frequency bands. The energy and wavelet coefficients of each frequency band were extracted as the features of the support vector machine (SVM) classifier. The WPD can divide the spectrogram equally into different frequency bands, but the fault components of the arc are also easy to be segmented by multiple bands at the same time. Thus, it is difficult to choose an appropriate frequency band to analyze the arc fault. Liu *et al*. [12] used the variational mode decomposition (VMD) to extract the information of the current signal, and the Shannon entropy was used to calculate the complexity of the modal components. After that, the mutation of entropy was used to determine the occurrence of arc faults. This method has strong anti-interference so that it can correctly identify the startup of inverters, dynamic shading, etc. However, the arc fault cannot be successfully identified in [12] if the occurrence time of arc fault is too short. To make it worse, the calculation of the VMD is too complex to be used in practical applications. Miao *et al.* [13] used a noninvasive magnetic sensor to obtain the current signal when an arc occurs. Then, an improved empirical mode decomposition (EMD) algorithm was investigated to decompose it, and the Hurst exponent was adopted to select the appropriate component. Moreover, the variance and standard error of the component were extracted and input to an SVM for the fault detection and a higher identification accuracy can be obtained. However, the method in [13] was only verified by PV simulators without practical experiments. Wu *et al.* [14] proposed a combined fault detection method based on the ensemble empirical mode decomposition (EEMD) and the fuzzy C-means clustering (FCM). Numerical simulations and experimental data were used to prove that the proposed method had a good anti-interference ability. However, only simulation data were adopted to test some interference conditions. Besides, the method in [14] also admitted that there may be misjudgments in the case of weak arcs and arc flashes.

The application of artificial intelligence technology has the advantages of effectively combining the methods of time domain, frequency domain, and time–frequency domain to construct an automatic fault classifier for improving the efficiency and accuracy of arc fault diagnoses. Telford *et al*. [15] adopted the mean value of the movement of the 50-ms time window and the approximate coefficients and detail coefficients, obtained by the DWT decomposition as input data. After that, they trained a hidden Markov model to detect arc faults. But this method is only suitable for dc systems with linear outputs, and it cannot work well when it is applied in the PV systems with nonlinear output characteristics. Khamkar and Patil [16] combined the DWT and cascaded fuzzy logic algorithms to detect and locate arc faults. However, only numerical simulations were provided in [16] to verify the model. To address the problem of the lack of effectively measured arc signal samples, Lu *et al*. [17] used the method of domain adaptation based on the deep convolutional generative adversarial network (DC-GAN) to enhance the arc waveform recorded in the laboratory environment to the actually measured arc waveform, and generated a large number of data.

TABLE I
SUMMARY OF RESEARCH BACKGROUND AND MERITS/LIMITATIONS OF PREVIOUS METHODS

Type	Ref.	Research background or merits	Limitations
Time-domain-based methods	[6]	It detected DC SAFs by detecting the line current drops, the change rates of the line current average value, and the standard deviations of the line current and the ac component of the supply voltage.	It does not consider the influence of the waveform abnormality caused by the meteorological environment and operating state changes on the proposed algorithm.
	[7]	A relatively simple detection algorithm without the complicated calculation was realized in the hardware level, which could be used in various DC resistive fields.	The arc fault condition is only tested under ideal conditions, and the influence of various interference conditions on the application of the proposed algorithm is not analyzed further.
	[8]	It investigated the SNR of the cross-correlation function to detect arc faults.	This method is prone to misjudgment in a PV system with the massive noise.
Frequency-domain-based methods	[9]	By calculating the standard deviation of frequency components, the corresponding detection threshold was established.	The difference of switching frequency between different inverters will affect the threshold setting. Besides, the adjustment of MPPT will also interfere with the algorithm and affect its accurate judgment.
Time-frequency-domain-based methods	[10]	The ratio of the average power of the first-level DWT coefficient to the reference average power was used to detect arc faults.	It does not further study the influence of meteorological environment and operating state changes on the algorithm.
	[11]	It adopted the WPD algorithm to divide the frequency band into several segments and calculated the corresponding energy to judge whether the fault occurred or not.	It is difficult to choose an appropriate frequency band for analyzing the arc fault.
	[12]	It applied the VMD and the Shannon entropy to calculate the complexity of the modal components, and it had higher resistance interference.	The arc fault will not be successfully detected if the occurrence time of an arc fault is too short.
	[13]	An algorithm based on EMD combined with SVM is proposed, which uses the variance and standard deviation of the components as features for the fault detection.	It was only verified by PV simulators without practical experiments.
	[14]	A combined fault detection method based on EEMD and FCM is proposed.	Only simulation data are adopted to test some interference conditions, and there may be misjudgments in the case of weak arcs and arc flashes.
Artificial-intelligence-based methods	[15]	It used the DWT and the HHM to realize the high-precision fault detection.	This method is only suitable for DC systems with linear outputs, and it may not work well when it is applied in PV systems with nonlinear output characteristics.
	[16]	It combined the DWT and cascaded fuzzy logic algorithms to detect and locate arc faults.	Only numerical simulations were provided to examine the model without experimental verifications.
	[17]	It used the DCGAN algorithm for data enhancement and realized reliable detection.	The proposed algorithm needs to obtain a lot of data through numerical simulations, and the corresponding execution time is longer than the requirement of the threshold method.
Other methods	[18]	It proposed a DC arc-fault detection method based on electromagnetic radiation signals.	Advanced sensors and measurement equipment are required to accurately capture information, which makes the application field to be subject to certain constraints.
	[19]	It designed a detection technology for PV arc faults via the method of spread spectrum time-domain reflectometry.	An expensive detection equipment with rather high sampling frequency is indispensable.

Then, a convolutional neural network was applied to realize the fault identification of SAFs. However, the algorithm in [17] needs to obtain a lot of data through numerical simulations, and the corresponding execution time is longer than the requirement of the threshold method.

In addition to some of the aforementioned conventional detection methods, some researchers also use other characteristics to identify and detect arc faults, such as the increment of heat energy, ultraviolet, acoustic signals, and electromagnetic radiation detection [5]. Xiong *et al.* [18] proposed a dc arc fault detection method based on electromagnetic radiation signals. The fourth-order Hilbert curve fractal antenna was used to detect the electromagnetic radiation signal of the dc arc, and the amplitude and frequency spectrum of the electromagnetic radiation signal measured under different circuit currents were analyzed. It was found that the characteristic frequency of the electromagnetic radiation of the dc arc generated by a PV system is about 39 MHz, and the rule that the duration of the electromagnetic radiation pulse of the dc arc fault is much longer than the switching operation, which can effectively distinguish the fault from normal interference. The disadvantage in [18] is that advanced sensors and measurement equipment are required to accurately capture information, which makes the application field to be subject to certain constraints. Alam *et al.* [19] designed a detection technology for PV arc faults via the method of spread spectrum time-domain reflectometry (SSTDR). The occurrence of the fault was confirmed by detecting the impedance change and autocorrelation before and after the failure of the PV array. The advantage of the method in [19] is that it is unnecessary to measure the voltage and current, and the potential arcs in the PV array can also be predicted. However, an expensive detection equipment with rather high sampling frequency is indispensable.

As for the review and summary of previous researches as shown in Table I, it can be seen that traditional time-domain-based or frequency-domain-based methods have poor expression of arc features, and are vulnerable to interference from the system or the surrounding environment, resulting in poor

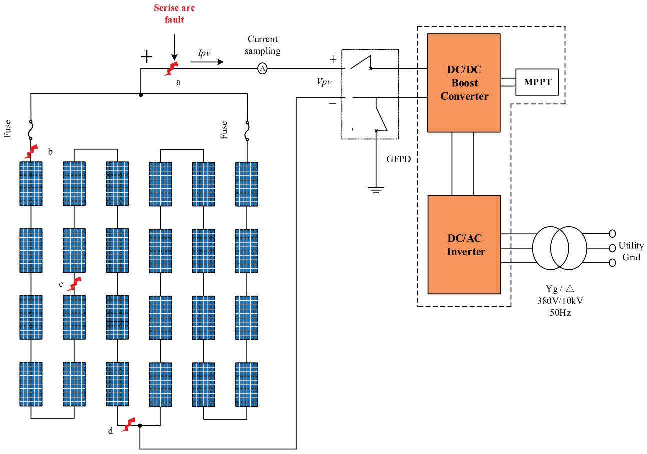


Fig. 1. Composition of a PV grid-connected system.

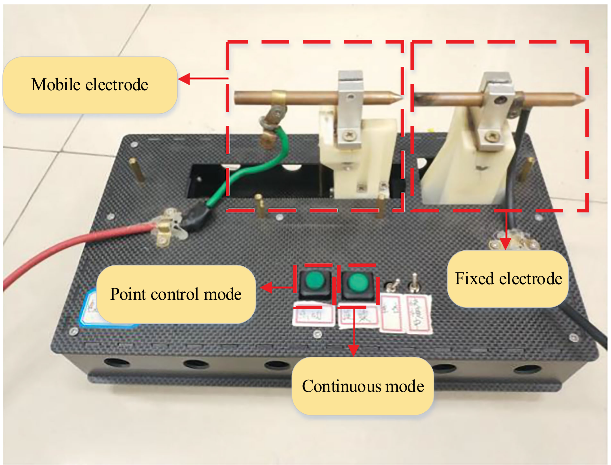


Fig. 2. Arc fault generator (AFG).

detection effect. The time–frequency domain method can make full use of the features differences in the time domain and the frequency domain before and after the occurrence of arc faults for capturing more effective characteristic expressions to greatly improve the detection accuracy. However, typical time–frequency decomposition methods, such as the WT and the WPD, cannot divide frequency bands adaptively such that they are unsuitable for various high-efficiency inverters with different switching frequencies. The fixed partition may divide valid features into different regions to form fragmentation features, which bring difficulties to the fault detection. Therefore, on the basis of removing the switching frequency and the background noise, an improved empirical wavelet transform (IEWT) is used in this study to segment relatively complete feature bands adaptively; a composite multiscale permutation entropy (CMPE) and a twin support vector machine (TWSVM) are combined into the proposed strategy to achieve the objectives of feature extraction and classification. Some previous methods in [11], [12], and [17] via the same current data of PV arrays to realize the arc fault detection will be compared with the proposed strategy in this study to support the aforementioned discussions.

The rest of this article is organized as follows. Section II introduces the composition of a PV system, and the states of PV currents in the time–frequency domain when a SAF occurs. Section III explains the proposed algorithm and the diagnostic process. The effectiveness of the proposed method is verified and analyzed by the measured data in Section IV. In Section V, the performance of the proposed method is compared with other methods to highlight superior advantages. Finally, Section VI concludes this article.

II. Composition of PV System and Waveform of Arc Fault Time–Frequency

A. *Composition of a PV System*

A PV grid-connected generation system generally consists of PV arrays, dc combiner boxes, inverters, protection devices, and transformer, as shown in Fig. 1. Among them, the PV array is composed of several PV modules. The output power of the array will be increased when they are connected in series or parallel. The output current and voltage of PV modules are nonlinear and varied due to the irradiance and temperature. Thus, the MPPT algorithm is universally used to extract the maximum power from a PV system.

B. *SAF Characteristics of a PV System*

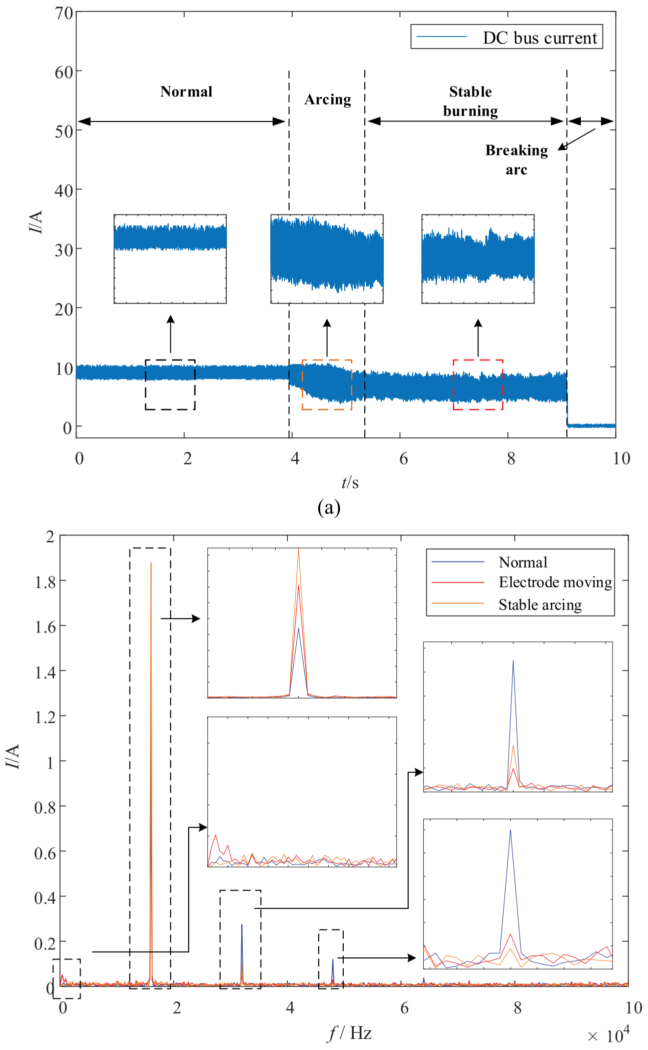
Various types of arc faults, including SAF, PAF, and grounding arc fault, due to hardware aging and loose joints may be caused by PV generation systems in daily use. Among them, the SAF is difficult to detect due to the weak changes of currents, which is the major issue to be solved in this study. Fig. 1 shows a PV system with a structure of 2 × 12, that is, 12 of the PV modules are connected in series, and 2 series modules are connected in parallel for power supply. Under standard test conditions, the maximum power, the open-circuit voltage, and the short-circuit current of individual modules are 270 W, 38.5 V, and 9.09 A, respectively. That is to say, the maximum power, the open-circuit voltage, and the short-circuit current of the entire array are 6.48 kW, 462 V, and 18.18 A, respectively. In Fig. 1, the positions labeled from *a* to *d* indicate the SAFs occurred at the bus rod, and the front, the middle and the end of the PV substring. An arc fault generator (AFG) shown in Fig. 2 is a device used to simulate the arc generation. It includes two electrodes, which are fixed, and a mobile one whose material is a bar copper. In the AFG, the structure of the discharge electrode includes tip-tip and tip-post. The movement of the electrode is driven by a stepper motor and its controller. After the control pulse is sent out by the controller, the electrode will move at a certain speed. By considering the randomness of actual arc length variation mentioned in [20], the fault waveforms of different arc lengths can be obtained by separating the electrodes during the speed range from 3 to 9 mm/s randomly, and the signals captured by short time windows are taken as the research object. Through these measures, the negative influence of artificial simulation operation on the actual fault detection can be avoided. The detailed experimental conditions are listed in Table II.

TABLE II
EXPERIMENTAL CONDITIONS

Item	Content
Size of array	2×12
Location of fault	Bus rod; Front, middle, and end of substring
MPPT algorithm	Perturbation and observation
Electrode structure of AFG	Tip-tip, tip-post
Electrode material of AFG	Copper
Others	Equipped with blocking diode and bypass diode

Here, one takes the time–frequency domain waveform when an SAF occurs at position as an example to illustrate the characteristics of the SAF. First, the electrode is controlled to be closed via the AFG; at this time, it is in unfaulty condition. Then, electrodes are controlled to separate and enter the stage of arcing. The movement of the electrode is stopped after 5.5 s and it is the stage of stable burning. After 9.2 s, the arc disappears naturally. The time–frequency domain waveform of the current in the entire arc simulation stage is depicted in Fig. 3. It shows the development of the waveform in the time domain when an SAF occurs; four stages are included: normal, arcing, stable burning, and breaking arc. The frequency-domain waveforms after the FFT transformation (dc components have been removed), including three stages of normal, arcing, and stable burning, are illustrated in Fig. 3(b). In [5], it is recorded that the frequency-domain characteristics of arc faults mainly exist within 100 kHz. According to the Shannon–Nyquist theorem, this study collects current signals at a sampling rate of 200 kHz, and analyzes the corresponding results. There are three obvious high-energy frequency bands in Fig. 3(b), where the component at 16 kHz is the switching frequency of the inverter, and the components at 32 and 48 kHz are two and three times the switching frequency. As can be seen from Fig. 3(a), after an SAF occurs, the dc-bus current will drop to a certain extent, and then immediately rise; it is mainly caused by the operation of the MPPT controller. Due to the increase in the arc gap, the dc-bus current continues to drop. When the arc gap is fixed, the dc-bus current fluctuates up and down. With the ablation of the electrode, the voltage drop of the arc itself increases with the increase in the gap, and the arc is extinguished due to the insufficient applied voltage. As can be seen from Fig. 3(b), there are three main characteristics of arc faults. First, the noise in a low frequency of the current spectrogram of normal and stable burning is smaller, but the one in the stage of arcing is larger than the former. The second characteristic is that the components of the normal current near the frequency doubling of 32 and 48 kHz are much higher than those in the case of a fault. Finally, the frequency spectrum of the normal current is more stable than that of the fault current, namely the noise of the normal current is slightly smaller than the fault current.

From the abovementioned analyses, it can be found that when an SAF occurs in a PV system, the dc-bus current has a certain degree of fluctuation in the time domain, but the difference between the states of stable burning and normal operation is not obvious. Moreover, the sudden changes in illuminance or the adjustments of the MPPT algorithm have the same effect, which makes the traditional time-domain analysis unreliable. In addition, from the current spectrogram, it can be observed that the holistic amplitude of the fault current and the normal current in the frequency domain are not much different. Unfortunately, traditional methods, such as the wavelet transform, are generally only suitable for the case of the obviously different frequency spectrum, otherwise, the misjudgments are prone to occur. Besides, due to the existence of the switching frequency with its multiple frequency components, the algorithm of the frequency domain is easily affected by the switching action signal so that the recognition accuracy decreases. Therefore, the bandpass filtering algorithm will be used in this study to reduce the adverse effects of switching frequency. The series time–frequency domain signals and machine learning algorithms are combined to achieve a simple, fast, safe, and reliable fault diagnosis.



(a)

(b)

Fig. 3. Time–frequency characteristics of SAF. (a) Time-domain waveform and (b) frequency-domain waveforms at different stages.

III. Algorithm Principle and Diagnosis Process

A. Hankel Singular Value Decomposition (Hankel-SVD)

1) Principle of Singular Value Decomposition (SVD): For a matrix ***A*** with a size of *m* × *n*, its SVD can be expressed as

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T \tag{1}$$

where $\mathbf{U} \in R^{m\times m}$ and $\mathbf{V} \in R^{n\times n}$ are both orthogonal matrices, and $\sum = (\mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r), \mathbf{O}) \in R^{m\times n}$ is the singular value matrix, in which ***O*** is the zero matrix, $r = \min(m, n)$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. The nonzero diagonal elements in $\sum$ are the singular values of the matrix ***A***. If the SVD is applied to matrices with different construction methods, the processing effects also differ. Yue *et al.* [21] showed that if a Hankel matrix was formed by a 1-D signal, various frequency components of the original signal could be obtained by the SVD. The singular values $(\sigma_1, \sigma_2, \ldots, \sigma_r)$, respectively, correspond to the components of the original signal frequency spectrum, whose amplitudes are arranged from large to small. As can be seen from Fig. 3(b), the amplitude of the switching frequency is the largest, whereas the amplitude of the background noise is minimal. In this way, after the SVD decomposition, the front-ranked singular values denote the amplitude of the switching frequency, and the last-ranked singular values represent the amplitude of the background noise. Once these singular values are removed, the switching frequency and the background noise component disappear.

2) Filtering and Signal Reconstruction: The Hankel matrix is a type of matrix that has been widely applied in many fields, such as numerical analysis, optimization theory, and system identification. There is a set of a digital signal vector $X = [x(1), x(2), \ldots, x(N)]$, and the Hankel matrix can be constructed by the phase space reconstruction as

$$\mathbf{A} = \begin{bmatrix} x(1) & x(2) & \cdots & x(n) \\ x(2) & x(3) & \cdots & x(n+1) \\ \vdots & \vdots & \vdots & \vdots \\ x(m) & x(m+1) & \cdots & x(N) \end{bmatrix} \tag{2}$$

where *m, n,* and *N* have a relation of $m + n - 1 = N$. The Hankel matrix is filtered by the SVD to remove the corresponding singular values, and then a new singular value matrix $\widetilde{\Sigma}$can be obtained. After that, a new Hankel matrix $\widetilde{\mathbf{A}}$ can be calculated by the formula of $\widetilde{\mathbf{A}} = \mathbf{U}\widetilde{\Sigma}\mathbf{V}^T$. It can be found from the construction formula of the Hankel matrix that the first row and last column (or the first column and last row) of the matrix $\widetilde{\mathbf{A}}$ can be directly restructured to acquire a new digital signal, *f*(*t*).

Yue *et al.* [21] illustrated that if the row (*m*) satisfies the relation of $2e < m \leq N/2$ (*e* is the number of the main component of the signal in frequency domain), the singular values processed by the SVD will appear in pairs. To obtain the best effect of the noise reduction, the product of *m* and *n* of should be maximized. Thus, *m* is equal to *N*/2, and

$$P_{f_n} \approx \sigma_{2n-1} + \sigma_{2n} \tag{3}$$

where P_{f_n} represents the amplitude of the *n*th frequency component, which is in descending order of the frequency spectrum complex amplitude in the signal X, and σ_{2n-1} and σ_{2n} stand for a pair of singular values corresponding to the *n*th frequency.

B. Improved Empirical Wavelet Transform

1) Empirical Wavelet Decomposition: Gilles [22] proposed the concept of EWT in 2013, which is adaptive for analyzing nonstationary time-varying signals. That is, the amplitude modulation and frequency modulation (AM–FM) components at different frequency bands can be extracted by a series of the wavelet filter bank, which is constructed based on the frequency-domain distribution characteristics of the signal. Specifically, the steps can be expressed as follows:

a) Extract the signal spectrum by the FFT. The Fourier transform of the signal *f*(*t*) can be performed by the following formula:

$$\hat{f}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} f(t)e^{-iwt}dt \tag{4}$$

where the frequency spectrum of *f*(*t*) is denoted as $f(\omega) = |\hat{f}(\omega)|$, in which $\omega \in [0,\ \pi]$.

b) Segment the signal bands. First, the number of local maximum values labeled as Num can be found from the obtained frequency spectrum, and they are arranged in the descending order. Then, local maximum values are further arranged according to the corresponding frequencies (ω) in the ascending order ($\omega_1,\ \omega_2,\ \ldots,\ \omega_{\mathrm{Num}}$). After that, the normalized frequency band $[0, \pi]$ is divided into subbands *K*(*K*≤Num), and the subbands of the AM–FM component with the center of $\omega_k (k = 2, 3, \ldots, K)$ is set as $[\Omega_{k-1},\ \Omega_k]$, where $\Omega_{k-1} = (\omega_{k-1} + \omega_k)/2$, $\Omega_0 = 0$, and $\Omega_K = \pi$.

c) Construct a wavelet filter bank. For subbands $[\Omega_{k-1},\ \Omega_k]$, the scale function $\hat{\phi}_k(\omega)$ and the wavelet function $\hat{\varphi}_k(\omega)$ can be, respectively, expressed as follows:

$$\hat{\phi}_k(\omega) = \begin{cases} 1 & |\omega| \leq (1-\gamma)\Omega_k \\ \cos\left(\frac{\pi}{2}\beta(\frac{|\omega|-(1-\gamma)\Omega_k}{2\gamma\omega_k})\right) & (1-\gamma)\Omega_k \leq |\omega| \leq (1+\gamma)\Omega_k \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$\hat{\varphi}_k(\omega) = \begin{cases} 1 & (1+\gamma)\Omega_k < |\omega| < (1-\gamma)\Omega_k \\ \cos\left(\frac{\pi}{2}\beta\left(\frac{|\omega|-(1-\gamma)\Omega_{k+1}}{2\gamma\omega_k}\right)\right) & (1-\gamma)\Omega_{k+1} \leq |\omega| \leq (1+\gamma)\Omega_{k+1} \\ \sin\left(\frac{\pi}{2}\beta\left(\frac{|\omega|-(1-\gamma)\Omega_k}{2\gamma\omega_k}\right)\right) & (1-\gamma)\Omega_k \leq |\omega| \leq (1+\gamma)\Omega_k \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $\beta(x)$ is a random function, and the same definition in [22] is used as follows:

$$\beta(x) = x^4(35 - 84x + 70x^2 - 20x^3). \tag{7}$$

d) Signal reconstruction. The inverse Fourier transform is used to calculate $f(\omega) \times \hat{\phi}_k(\omega)$ and $f(\omega) \times \hat{\varphi}_k(\omega)$, and then the time-domain representations of each frequency band can be obtained, namely the modal component ewt(k).

2) Algorithm Improvement Based on Mathematical Morphology: The EWT has the defect of too dense segmented frequency spectrum. To make it worse, the switching frequency components and noise in the dc-bus current of a PV system make it difficult to reasonably segment the frequency spectrum, which results in a lack of characteristics. Therefore, the method of mathematical morphology is used in this study to improve the defects of the EWT. The basic idea of mathematical morphology is to use the structure element as the probe to search for information, and the relationship between various parts of the data can be obtained by the continuous translation of the structure element. With the purpose of achieving different processing targets, the morphological operations consist of four forms, namely erosion, dilation, open operation, and close operation [23]. The close operation is mainly adopted to process the frequency spectrum for reducing the adverse effects of the original spectrum noise on the EWT segmentation in this study.

Suppose that $F(o)$ is the sampled signal to be the structural elements, and the domain definition of them can be expressed as $D_F = \{0, 1, 2, \ldots, N-1\}$ and $D_S = \{0, 1, 2, \ldots, M-1\}$, respectively, where N and M are both natural numbers ($N > M$). The formulas for the morphological dilation and erosion can be, respectively, represented as

$$(F \oplus s)(o) = \max[F(o-p) + s(o)] \tag{8}$$

$$(F \Theta s)(o) = \min[F(o+p) - s(o)]. \tag{9}$$

Moreover, the formula for the close operation is defined as

$$(F \centerdot s)(o) = (F \oplus s \Theta s)(o) \tag{10}$$

where $\oplus$, Θ, and $\centerdot$ are the symbols of dilation, erosion, and close operation, respectively.

To retain the characteristics of the original signal frequency spectrum to a maximum extent, the linear structure elements $s(p) = 0$ are used for the close operation. The size of it is the width of the straight line, but usually, it is required to adjust according to different types of signals. The improvement of the EWT is to add an operation after the first step of the calculation, that is, after $f(\omega)$ is calculated, and suppose that $f(\omega) = F(o)$, then $f(\omega)$ is processed by the close operation of the formulas in (8)–(10), so that a new $\hat{f}(\omega)$ can be obtained. Subsequently, the decomposition quality of the EWT can be improved by the segmentation of the frequency spectrum on $\hat{f}(\omega)$. The dc-bus current in the normal state of a PV system is taken as an example to explain the improvement effect of the EWT. The segmentation region of the frequency band of the EWT before and after the algorithm improvement is illustrated in Fig. 4. It can be obviously observed that, before the improvement, the cutting lines of the frequency spectrum are mainly concentrated in the front half part and they are so dense. It implies that there is a great adverse effect on the extraction of subsequent frequency spectrum characteristics. On the contrast, the entire spectrum can be divided orderly by the segmentation of the frequency spectrum after the algorithm improvement.

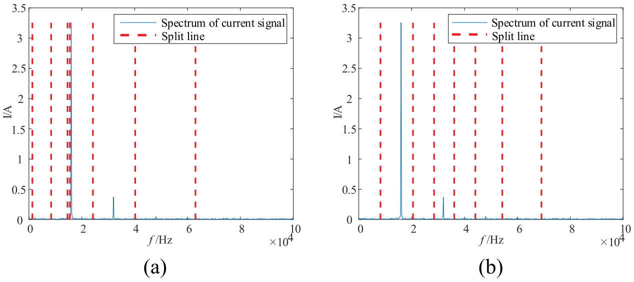


Fig. 4. Spectrum segmentation of (a) EWT and (b) IEWT.

C. Composite Multiscale Permutation Entropy (CMPE)

1) Permutation Entropy (PE): The PE is used for measuring the complexity of time series, and it has the advantages of high computational efficiency, good robustness, and invariance [24]. The complexity of each sequence obtained after the EWT decomposition is calculated by the PE, and it is adopted as the classification feature in this study. The principle of the PE is assuming that there is a discrete sequence vector $\text{ewt} = [w(1),\ w(2),\ \ldots,\ w(N)]$, and the phase space reconstruction can be performed on the sequence vector of ewt by setting the delay time (t) and the embedding dimension (d). The reconstructed sequence vector of ewt can be expressed as follows:

$$\hat{\text{ewt}} = \begin{bmatrix} w(1) & w(1+t) & \cdots & w(1+(d-1)t) \\ \vdots & \vdots & & \vdots \\ w(i) & w(i+t) & \cdots & w(i+(d-1)t) \\ \vdots & \vdots & & \vdots \\ w(v) & w(v+t) & \cdots & w(v+(d-1)t) \end{bmatrix} \tag{11}$$

where $i = 1, 2, \ldots, v$, in which $v = N - (d-1)t$ represents the number of components in the reconstruction matrix, and each reconstruction component is composed of a row of data in the matrix. Each reconstruction component is rearranged in the ascending order of numerical value, and the column of each element of the reconstruction component is denoted as $j_1, j_2, \ldots, j_d$. By this way, each reconstructed sequence $\hat{\text{ewt}}$ can find the corresponding symbol order as

$$s(l) = (j_1, j_2, \ldots, j_d) \tag{12}$$

where $l = 1, 2, \ldots, e$ and $e \leq d!$. Similar to this, the way of an arbitrary symbol order can be represented by $s(l)$. If $P_1, P_2, \ldots, P_e$ stand for the probability of each symbol sequence, respectively, the normalized PE [25] can be represented as

$$\text{PE} = -\ln{(d!)}^{-1} \sum_{j=1}^{d!} P_j \ln(P_j). \tag{13}$$

The entropy of the PE is between 0 and 1, and it is believed that the smaller the entropy is, the more structured the original time series will be.

2) Composite Multiscale Permutation Entropy (CMPE): The PE can only reflect the spatial structure on the overall time scale of the data, while the structure on different time scales of the data will be ignored. Fortunately, the abovementioned problem can be solved by the multiscale permutation entropy (MPE) [26]. The MPE adds a new scale factor (τ) on the basis of the PE, and then the coarse-graining processed time series $y^{(\tau)}$ can be obtained by the following formula:

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x(i), 1 \leq j \leq \frac{N}{\tau}, \tau \leq N. \quad (14)$$

Thus, the MPE can be expresses as

$$\mathrm{MPE}(\mathrm{ewt}, \tau, d, t) = \mathrm{PE}(y^{(\tau)}, d, t). \quad (15)$$

Although the MPE remedies the defect of single-scale in the PE, the length of the coarse-graining sequence is equal to the length of the original time sequence divided by the scale factor, which makes the deviation of entropy increase with the decrement in the length of the coarse-graining sequence. The proposed CMPE solves the above-mentioned problems [27]. The time series $y_g^{(\tau)} = \{y_{g,1}^{(\tau)}, y_{g,2}^{(\tau)}, \ldots, y_{g,(N-g)/\tau}^{(\tau)}\}$ of the gth dimension of the scale (τ) is calculated by the CMPE to be on the basis of the MPE, in which $y_{g,j}^{\tau}$ can be calculated by the following formula:

$$y_{g,j}^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+g}^{j\tau+g-1} x(i), 1 \leq g \leq \tau. \quad (16)$$

The CMPE is calculated by taking the mean of the sum of the MPE under the corresponding scale, and can be represented as

$$\mathrm{CMPE}(\mathrm{ewt}, \tau, d, t) = \frac{1}{\tau} \sum_{k=1}^{\tau} \mathrm{PE}(y_k^{(\tau)}, d, t). \quad (17)$$

D. TSVM Based on Salp Swarm Optimization

1) Twin Support Vector Machine (TSVM): Jayadeva and Chandra [28] proposed a TWSVM to construct two nonparallel hyperplanes by computing two quadratic programming problems. Since the number of constraint conditions of each quadratic programming problem is half than that of the classic SVM, the training speed of the TWSVM is theoretically about four times the speed of the traditional SVM. However, the TWSVM is not perfect. For instance, the issue of parameter selection cannot be handled well by the TWSVM, which affects the results of classification. The nonlinear TWSVM with Gaussian kernel functions is mainly used in this study. First, suppose that there are m_{train} training samples in the q-dimensional real space R^q, m_1 samples belong to the positive class, and $m_2 = m_{\mathrm{train}} - m_1$ samples belong to the negative ones. The matrices $\mathbf{A} \in R^{m_1 \times q}$ and $\mathbf{B} \in R^{m_2 \times q}$ represent the positive and negative class training samples, respectively, and each row of them represents a sample. The training process of the TWSVM is to find two nonparallel hyperplanes, that is,

$$\begin{cases} K_f(H^T, C^T) w_1 + b_1 = 0 \\ K_f(H^T, C^T) w_2 + b_2 = 0 \end{cases} \quad (18)$$

where K_f is the undetermined kernel function, $C^T = [A,\ B]^T$, w_1 and w_2 are normal vectors, and b_1 and b_2 are deviating variables. These two hyperplanes can be obtained by solving the following two quadratic programming problems:

$$\begin{cases} \min \frac{1}{2} \left\| K_f(A, C^T) w_1 + e_1 b_1 \right\|^2 + c_1 e_2^T \xi \\ s.t. - (K_f(B, C^T) w_1 + e_2 b_1) + \xi \geq e_2, \xi \geq 0 \end{cases} \quad (19)$$

$$\begin{cases} \min \frac{1}{2} \left\| K_f(B, C^T) w_2 + e_2 b_2 \right\|^2 + c_2 e_1^T \eta \\ s.t. (K_f(A, C^T) w_2 + e_1 b_2) + \eta \geq e_1, \eta \geq 0 \end{cases} \quad (20)$$

where e_1 and e_2 are both unit column vectors, whose numbers of columns are the same to the kernel function $K_f(A, C^T)$ and $K_f(B, C^T)$, respectively. In addition, c_1 and c_2 are penalty factors, and ξ and η are slack variables. Therefore, each class in the binary classification problem corresponds to a hyperplane. If the distance between the sample and the hyperplane of the positive class is less than that to the negative class, the sample is classified as a positive class; otherwise, it is regarded as a negative class. The kernel function used in this study is a Gaussian kernel function to be expressed as

$$K_f(h_{i_v}, h_{j_v}) = \exp\left(\frac{-\left\| h_{i_v} - h_{j_v} \right\|^2}{2\sigma^2} \right) \quad (21)$$

where σ is the width of the Gaussian kernel function, and h_{i_v} and h_{j_v} are the *i*th and *j*th input samples, respectively.

2) Salp Swarm Algorithm (SSA): According to the above-mentioned analyses, there are three hyperparameters that remain to be optimized in the TWSVM, namely c_1, c_2 , and σ. It is time consuming for parameter tuning by human, making it worse, and the optimal parameters are difficult to be found. Mirjalili *et al.* [29] proposed a novel metaheuristic intelligent algorithm, in which the salp swarm is searched in the form of a chain. This algorithm has a simple structure, and there is almost no parameter setting involved. Additionally, it has an excellent performance of calculation. The corresponding optimization process is explained as follows.

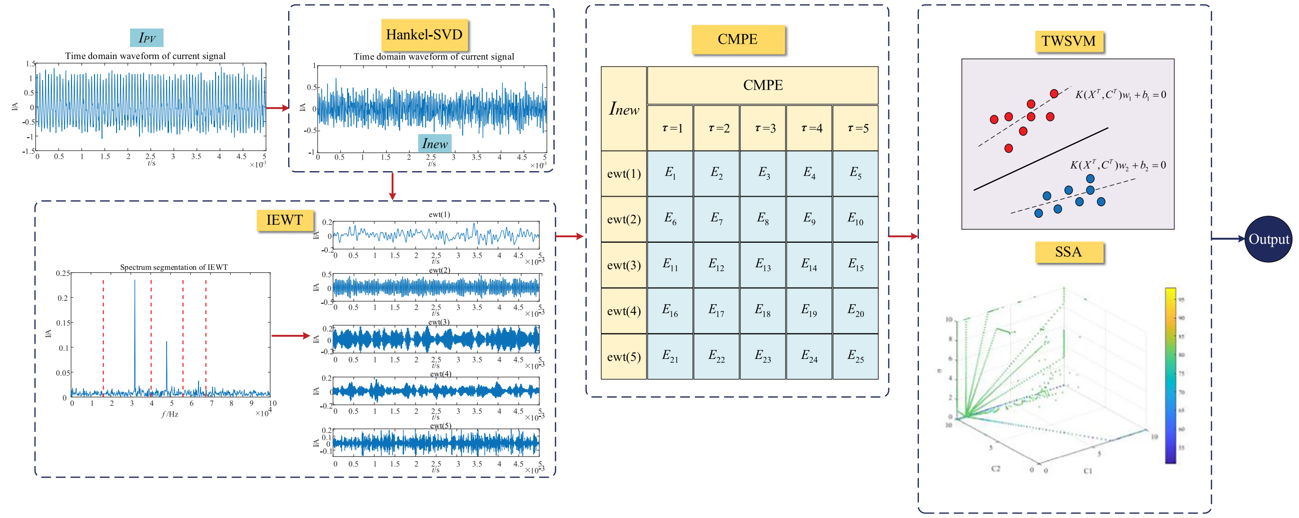
S1: Initialization population. According to the upper and lower limit values of each dimension of the search space, the position $\mathrm{Pos}_{j_{\mathrm{SSA}}}^{i_{\mathrm{SSA}}}(0)$ of the salpa is initialized, where $i_{\mathrm{SSA}} = 1, 2, \ldots, N_{\mathrm{SSA}}$, and $j_{\mathrm{SSA}} = 1, 2, \ldots, D_{\mathrm{SSA}}$, in which N_{SSA} is the population size of the salp swarm, and D_{SSA} is the spatial dimension

$$\begin{aligned} Pos_{j_{\mathrm{SSA}}}^{i_{\mathrm{SSA}}}(0) &= rand(N_{SSA}, D_{SSA}) \\ &\quad \times (ubj_{SSA} - lbj_{SSA}) + lbj_{SSA} \end{aligned} \quad (22)$$

where rand () is the random operator; $ub_{j_{SSA}}$ and $lb_{j_{SSA}}$ are the upper and lower limits of the search space, respectively.

S2: Calculate the adaptation degree of each salpa with respect to the objective function, and determine the initial position of the food source. Then, the position with the smallest adaptation degree is chosen as the target one.

S3: Determine the leader and the follower. Those salpas located in the first half of the salpa chain are the leaders, and the rest are the followers.


I_{PV}
Time domain waveform of current signal
Hankel-SVD
Time domain waveform of current signal
Inew
IEWT
Spectrum segmentation of IEWT
f/Hz
ewt(1)
ewt(2)
ewt(3)
ewt(4)
ewt(5)
t/s
I/A
CMPE
TWSVM
$K(X^T, C^T)w_1 + b_1 = 0$
$K(X^T, C^T)w_2 + b_2 = 0$
SSA
C1
C2
Output

Inew	CMPE				
	$\tau=1$	$\tau=2$	$\tau=3$	$\tau=4$	$\tau=5$
ewt(1)	E_1	E_2	E_3	E_4	E_5
ewt(2)	E_6	E_7	E_8	E_9	E_{10}
ewt(3)	E_{11}	E_{12}	E_{13}	E_{14}	E_{15}
ewt(4)	E_{16}	E_{17}	E_{18}	E_{19}	E_{20}
ewt(5)	E_{21}	E_{22}	E_{23}	E_{24}	E_{25}

Fig. 5. Processing framework of the proposed method.

S4: Update the position of the leader of the salpa

$$\mathrm{Pos}_{j_{\mathrm{SSA}}}^{i_{\mathrm{SSA}}}(t_{\mathrm{SSA}}) = \begin{cases} \mathrm{PF}_{j_{\mathrm{SSA}}} + z_1((\mathrm{ub}_{j_{\mathrm{SSA}}} - \mathrm{lb}_{j_{\mathrm{SSA}}})z_2 + \mathrm{lb}_{j_{\mathrm{SSA}}}) & z_3 \geq 0.5 \\ \mathrm{PF}_{j_{\mathrm{SSA}}} - z_1((\mathrm{ub}_{j_{\mathrm{SSA}}} - \mathrm{lb}_{j_{\mathrm{SSA}}})z_2 + \mathrm{lb}_{j_{\mathrm{SSA}}}) & z_3 < 0.5 \end{cases} \tag{23}$$

where t_{SSA} is the current iteration times, $\mathrm{Pos}_{j_{\mathrm{SSA}}}^{i_{\mathrm{SSA}}}(t_{\mathrm{SSA}})$ is the position of the leader (i_{SSA}) of the current generation salpas in the j_{SSA}th dimension space, $\mathrm{PF}_{j_{\mathrm{SSA}}}$ is the position of the current food source in the j_{SSA}th dimension space, z_2 and z_3 are the random number uniformly distributed between [0,1], c_1 adaptively decreases as the time of iteration increases, and the value of z_1 is selected as

$$z_1 = 2e^{-(4t_{\mathrm{SSA}}/\mathrm{Max\_iter})^2}. \tag{24}$$

In (24), Max_iter is the maximum iteration. Moreover, the values of z_1, z_2 , and z_3 are the same as those used in [29].

S5: Update the position of the followers of salpas

$$\mathrm{Pos}_{j_{\mathrm{SSA}}}^{i_{\mathrm{SSA}}}(t_{\mathrm{SSA}}) = \frac{1}{2}(\mathrm{Pos}_{j_{\mathrm{SSA}}}^{i_{\mathrm{SSA}}}(t_{\mathrm{SSA}}-1) + \mathrm{Pos}_{j_{\mathrm{SSA}}}^{i_{\mathrm{SSA}}-1}(t_{\mathrm{SSA}}-1)) \tag{25}$$

where $\mathrm{Pos}_{j_{\mathrm{SSA}}}^{i_{\mathrm{SSA}}}(t_{\mathrm{SSA}})$ is the position of the followers (i_{SSA}) of the current generation salpas in the ${j_{\mathrm{SSA}}}^{\mathrm{th}}$ dimension space.

S6: Perform boundary treatment on each dimension of the updated individual, and update the position of the food source based on the updated global optimization position of salpa.

S7: Judge whether the maximum iteration are achieved, and if so, output the current optimal hyperparameters (c_1, c_2, σ) and the objective function value; otherwise, jump to S3 to continue iterative evolution.

E. Diagnosis Process

As shown in Fig. 5, the aforementioned algorithms are combined in this study to construct a novel model of SAF detection of a PV system. The processing procedure of the proposed scheme is expressed as follows.

S1: Collect the dc-bus current (I_{PV}) of a PV system with a sampling frequency of 200 kHz, and select the data during a certain time window for analyses.

S2: Perform bandpass filtering on I_{PV}. First, the dc component in the current will be filtered out. Then, the Hankel matrix is constructed, and the SVD is performed to filter the high-frequency switching frequency and noise components. Finally, a new value of the SVD is used to reconstruct the new signal waveform, I_{new}.

S3: Decompose I_{new} by the IEWT and the multiple sets of modal components. It implies that information from different frequency bands can be obtained, and the CMPE value of each modal component can be calculated as the input data of the classifier.

S4: Divide the input data into training and testing sets. The training data are used to train the TWSVM classifier, meanwhile, the SSA is used to find the optimal hyperparameters of the model.

S5: The test data are used to evaluate the diagnostic accuracy of the trained model.

The distinct objectives for each step throughout the process in Fig. 5 are explained as follows. The Hankel-SVD method is investigated to eliminate the interference of the background noise and the switching frequency. Moreover, the EWT method is adopted for the time–frequency decomposition of signals, and the IEWT method is proposed to ensure that the decomposition is self-adaptive, so as to avoid the signal being forced to be decomposed into a fixed frequency band and the decomposition is too intensive. After relatively complete decomposition of the corresponding frequency band features, the CMPE is then used to further mine the multiscale fault information in the sequence waveform. Finally, the TWSVM

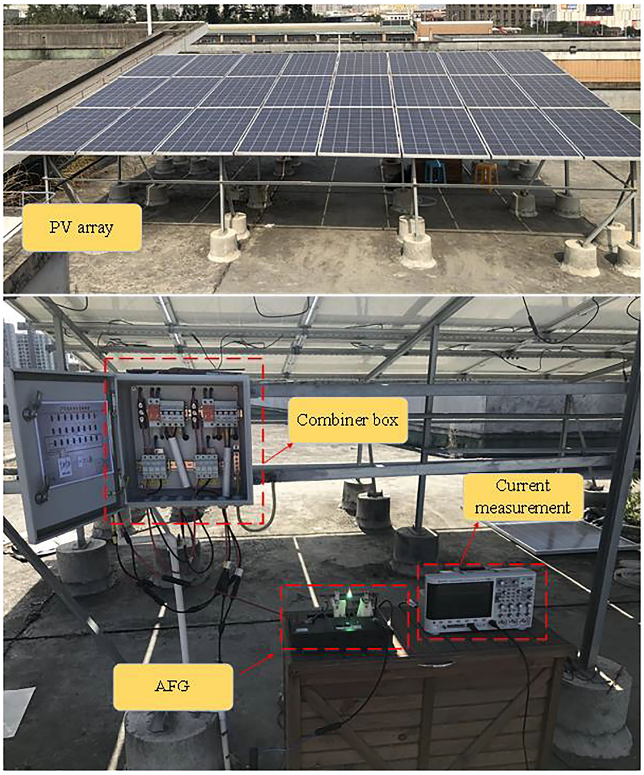


Fig. 6. Photograph of a 2 × 12 scale PV system.

TABLE III
PARAMETERS OF 2 × 12 SCALE PV SYSTEM UNDER STANDARD TEST CONDITION

Equipment	Parameter				
PV module	P_{mpp}	V_{mpp}	I_{mpp}	V_{OC}	I_{sc}
	270W	31.3V	8.63A	38.5V	9.09 A
PV array	2×12 modules (Parallel and Series)				
Inverter	Model	Sungrow SG6KTL-MT			
	Start voltage	250V			
	MPP Voltage range	200-1000V			
	Input	220-1100V			
	Output	220V			
	Switching frequency	16kHz			

is applied to classify the multiscale fault information to improve the training speed.

IV. Experimental Verification and Analyses

A. Experimental Platform

A PV roof grid-connected system with a capacity of 6.48 kWp is used in this study to verify the proposed method. The scale of the PV array is 2 × 12, i.e., 12 of the PV modules are connected in series, and 2 series modules are connected in parallel. The photograph of the experimental setup in this study is depicted in Fig. 6, and the parameters of the laboratory PV system under standard test conditions are summarized in Table III. In Fig. 6, the oscilloscope (DSOX2024A) is adopted to collect the dc-bus current at a sampling frequency of 200 kHz, and connected to the location of the current acquisition in Fig. 1. The AFGs are connected in series at the four positions of *a*, *b*, *c*, and *d* (as shown in Fig. 1) to, respectively, simulate the arc fault occurs at the bus rod, and the front, the middle, and the end of the substring. The recorded current waveforms include the states of normal, arcing and stable burning. The collected data covers various experimental samples under an irradiance range of 150–900 W/m^2, an atmospheric temperature range of 10–40 °C, and an arc gap of 0–15 mm. The distribution of experimental samples is illustrated in Table IV, where the labels of samples are of two classes. The positive category is 1, and the negative one is -1, which represent the states of arc fault and normal situations, respectively. "SAF in *a*" means that an arc fault occurs at position *a* in Fig. 1, and other marks can be deduced by analogy. The ratio of training and test samples is 4:1. The types of the adopted server are XEON W-2123 CPU, 2$*$GTX1080Ti GPU, and 32G RAM.

TABLE IV
SUMMARY OF MULTIPLE DATA SETS

	Fault Type Description	Category Label	Sample Number
1	Normal	-1	100
2	SAF in *a*	1	100
3	SAF in *b*	1	100
4	SAF in *c*	1	120
5	SAF in *d*	1	100

B. Determination of Model Parameters

1) Time Window: The selection of a time window is inconsistent in different literatures. For example, a time window of 50 ms is used in [12], whereas the time-window selection in [17] is 20 ms. For arc faults, the sampling rate is usually higher due to the requirement of high-frequency signals as the features. The bigger the time window is, the more data points will be; obviously, the longer it takes for the signal to perform the time–frequency decomposition. To select an appropriate time window, data for four time windows including 5, 10, 15, and 20 ms are selected, respectively, and then the corresponding execution times are recorded as 0.24, 0.6, 1.2, and 2.4 s, respectively. Owing to the great harm caused by arc faults, it is essential to detect and cutoff the pathway maintained by the current as soon as possible. A complete ripple signal above 200 Hz can be decomposed by a time window of 5 ms. The frequency of the separable waveform signal gradually increases as the time window shortens. In addition, both the arc at steady state and the early arc are required to be accurately identified in this study. The information component of the steady-state arc at the early fault waveform will reduce with the reduction of the time window, which will lead to the reduction in the recognition ability of the proposed algorithm. By considering the timeliness and accuracy of arc fault detection, a time window of 5 ms is used as the detection window in this study.

2) SVD Filtering Parameters: The switching frequency of the inverter in the experimental system is 16 kHz. Through the

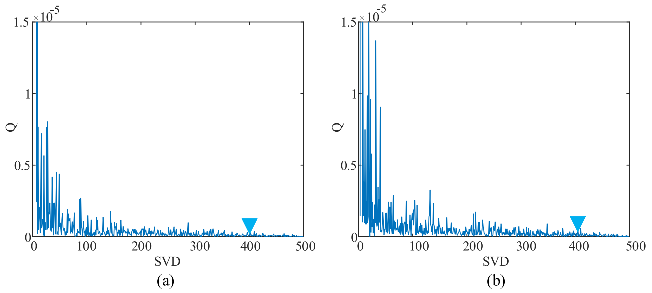


Fig. 7. Energy differential spectrum of current signal. (a) Arc fault. (b) Normal.

analysis of the signal spectrum, it is found that there are obvious high-energy signals at 16, 32, and 48 kHz, where 32 and 48 kHz are two and three times the switching frequency. According to the analytic results noted in Section II-B, it is appropriate to remove the switching frequency and retain the frequency doubling component of the switching frequency. Therefore, for the obtained SVD, the largest two eigenvalues require to be placed to zero. Besides, to remove background noise and retain the characteristics of the original signal as possible, the energy difference spectrum of the singular value is adopted to calculate the most appropriate degree of denoising, and the Pearson correlation coefficient is used to calculate the signal correlation before and after denoising. The energy difference spectrum of the singular value reflects the energy contribution of the effective signal and noise to the singular value, and it is defined as

$$Q(i_\sigma) = \left({\sigma_{i_\sigma}}^2 - {\sigma_{i_\sigma+1}}^2\right) \Big/ \sum_{j_\sigma=1}^{r} {\sigma_{j_\sigma}}^2 \qquad (26)$$

where $i_\sigma = 1, 2, \ldots, r$, and σ_{i_σ} represents a sequence of singular values. Since the background noise has a small proportion of energy and no mutation occurs, the order of the singular value of the noise can be determined and removed. Fig. 7(a) and (b) shows the singular value energy spectrum of the current signal under the fault state and the normal state, respectively. As can be seen from Fig. 7, the distribution of the energy spectrum is relatively uniform if the singular value exceeds 400. Therefore, the effective characteristic information of the first 400 singular values is determined to retain in this study.

Pearson correlation coefficient [30] can be expressed as

$$r_s = \frac{\sum_{i_\sigma=1}^{r}(X_{i_\sigma} - \overline{X})(Y_{i_\sigma} - \overline{Y})}{\left(\sqrt{\sum_{i_\sigma=1}^{r}(X_{i_\sigma} - \overline{X})^2}\right)\left(\sqrt{\sum_{i_\sigma=1}^{r}(Y_{i_\sigma} - \overline{Y})^2}\right)} \qquad (27)$$

where $X_{i\sigma}, Y_{i_\sigma} (i_\sigma = 1, 2, \ldots, r)$ are the sample of two groups of sequences, and $\overline{X}, \overline{Y}$ are the mean of two groups of sequence. By taking the normal signal sample as an example, correlation coefficients of the SVD before and after removing the last 100 values are compared on the basis of removing the switching frequency, and it is calculated that r_s is equal to 99.6%. It shows that the data characteristics before and after filtering are well preserved. The waveform of this normal signal sample after removing the dc component and noise is depicted in Fig. 8.

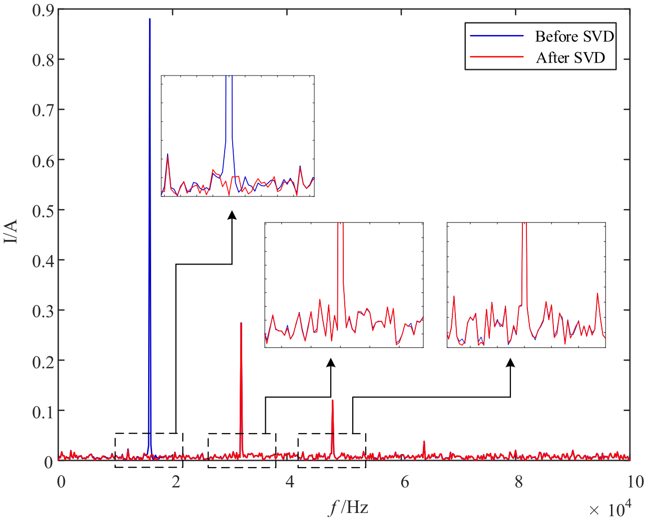


Fig. 8. Current spectrum before and after SVD.

As can be seen from Fig. 8, the switching frequency at 16 kHz has been basically eliminated, whereas the amplitude of other frequency points does not change before and after processing. Therefore, the influence of the switching frequency and background noise can be removed without affecting the original characteristics of the signal by the Hankel-SVD model, which provides a guarantee for obtaining the effective expression of the fault characteristics.

3) Quantity of Frequency Bands and Scale Factor: There are two ways of signal division by the EWT; one is fully adaptive division. If it is adaptively decomposed into several segments in accordance with signal spectrum characteristics, it will lead to an unstable extracted feature. For example, for multiple groups of signals, the number of segmented frequency bands for each group of signals may differ from each other. The other way of segmentation is to set the number of frequency bands manually, whereas the width of each frequency band is determined adaptively. By this way, the stability of the number of features can be guaranteed due to a certain number of frequency bands of signal decompositions. Thus, the latter treatment is adopted in this study. But the influence of the number of frequency bands on the detection effect must be further investigated.

There is the same problem for the CMPE. Although the CMPE can avoid the problem of incomplete feature extraction of the PE, the selection of the number of scales will also affect the final detection effect.

To find the best combinations, the relationship between the number (N) of frequency bands and the diagnostic accuracy and that between the scale (τ) of the PE and the diagnostic accuracy are analyzed by experiments. The corresponding relationship illustrations are depicted in Fig. 9. If the values of N and τ are selected too large, it will lead to longer computing times for the proposed algorithm, and practical applications will be limited. Because there are at least two high-frequency components in the signal spectrum (i.e., the frequency doubling component), the value of N is suggested to be not less than 3. In summary, the value of N from 3 to 6 and the value of τ from 1 to 10 are

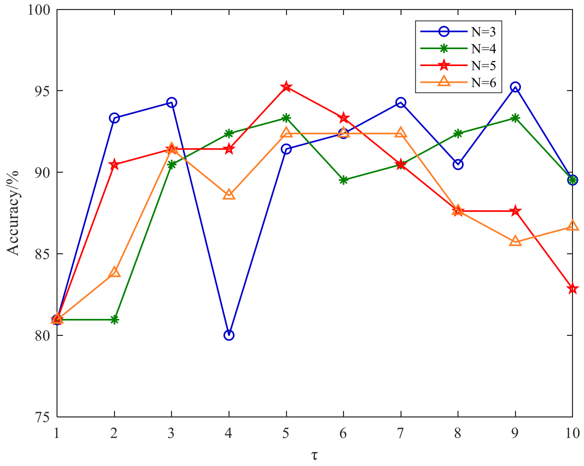


Fig. 9. Detection effect of different frequency bands and scales.

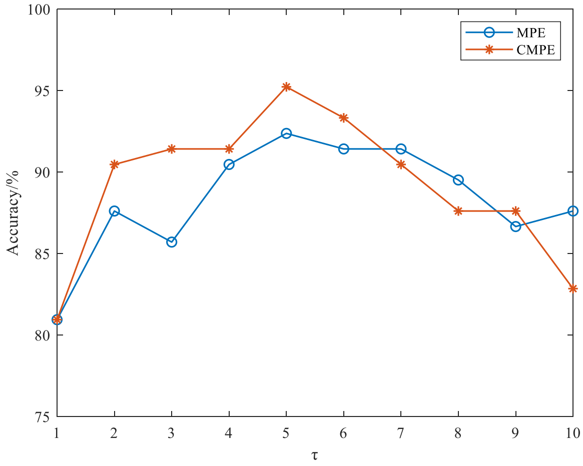


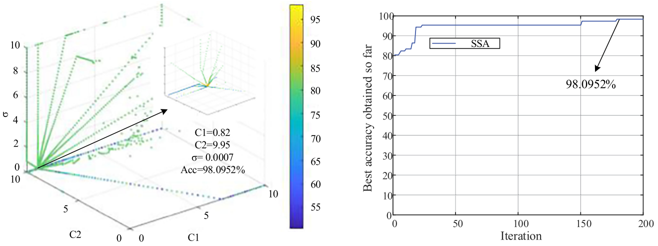
Fig. 10. Accuracy of MPE and CMPE in different scales.

considered in this study. The TWSVM with Gaussian kernel function is used as the classifier, the hyperparameters are set to default values ($c_1 = 4$, $c_2 = 5$, and $\sigma = 7.2 \times 10^{-4}$), and no optimization is performed.

In Fig. 9, blue, green, red, and orange lines represent the detection accuracy when the number of frequency bands (N) are 3, 4, 5, and 6 at different scales (τ), respectively. With the increase of the scale (τ), the accuracy shows different trends as the number of frequency bands (N) are different. When the value of N is equal to 3, the accuracy fluctuating rises during the range of 81–95%, and the highest accuracy reaches 95.24%. When the value of N is equal to 4, the accuracy is basically in the range of 81–93%. When the value of N is 5 or 6, the accuracy first increases and then decreases, and the change is relatively gentle. Among them, the highest detection accuracy about 95.24% is obtained as ($N = 5$, $\tau = 5$) or ($N = 3$, $\tau = 9$). The larger the scale (τ), the longer the computation time. Therefore, $N = 5$ and $\tau = 5$ are selected as the parameters of the model.

To verify the effectiveness of the model, the detection effects of PE, MPE, CMPE, and other methods under the number of the frequency bands equal to five ($N = 5$) are tested in this study, and the classifier used is the same as the foregoing one. For the single-scale PE, the effect is not good and the detection accuracy is only 80.95%. The detection accuracy of the MPE and the CMPE under different scale factors is depicted in Fig. 10. It is clear that the accuracy change trend of the MPE and the CMPE is similar as the scale increases. However, the overall performance of the CMPE is more superior to the MPE, indicating that the defects of the PE and the MPE can be improved by the CMPE, and the excellent arc fault characteristics can be extracted.

4) Optimization of TWSVM Kernel Function: To determine appropriate hyperparameters for the TWSVM model, the SSA is adopted for optimization in this study. The changes of penalty factors (c_1 and c_2) and the kernel function width (σ) in the iteration processes are depicted in Fig. 11(a), where the axes X, Y, and Z represent c_1, c_2, and σ, respectively. The color of the scattered points stands for the accuracy, in which yellow means the highest. The accuracy change of each iteration process of the SSA is illustrated in Fig. 11(b). The maximum iteration times are set to be 200 in this study. As can be seen from Fig. 11(b), the accuracy keeps ascending as the iteration times increase, and finally, it reaches 98.10%, which is improved by about 3% compared with the proposed algorithm without the SSA.



(a) (b)

Fig. 11. Quantitative results of SSA. (a) Parameter space. (b) Objective space.

TABLE V
COMPARISON OF PSO, GA, AND SSA OPTIMIZATION

Type	PSO	GA	SSA
Time	64min	55min	54min
Accuracy	97.14%	96.19%	98.10%

Particle swarm optimization (PSO) and genetic algorithm (GA) are two most widely used algorithms in the field of optimization. With the continuous innovation in this field, other new optimization algorithms, such as the SSA, have received more attention in recent years. To verify the superiority of the SSA used in this study, the optimization performance of the SSA is compared with the ones of the PSO and the GA, and the corresponding results are summarized in Table V. As can be seen from Table V, under the same iteration times of 200, the accuracies of the PSO and the GA are 97.14% and 96.19%, respectively, whereas the SSA found relatively better ones, and increasing the accuracy of the model to be 98.10%. Moreover, the SSA spends less computing time in the optimization process than the PSO and the GA, which improves the efficiency of optimization.

C. *Analysis of Anti-Interference Ability*

In this study, the waveform was collected by a higher sampling rate and a minor time window. In general, the current waveform in the normal state is stable in a minor time window. However, it is inevitable that there will be interference, which will cause the signal to fluctuate, and the phenomenon of a suspected fault will appear. Similarly, arc faults may occur under any circumstances, or the collected time window appears when the changes of the arc are in a transient state. In such conditions, whether the proposed method can effectively identify arc faults is also worthy to discuss. The following situations, which are likely to be misjudged, are recited and discussed.

1) *Dynamic fast shading*: If there are cloud covers or larger flying objects move quickly over a PV array, it will result in dynamic shading for leading to the mutation in the waveform. In this study, large pieces of cardboard are used to shade the PV component quickly for simulating this situation. In normal and arc fault states, a total of 30 sets of dynamic shading samples are collected.

2) *Inverter startup*: At this stage, the semiconductor switch is turned ON, and the current gradually increases from 0. The noise level changes due to the switching action, and it will produce a certain amount of interference. Thirty sets of samples in this case are collected for testing.

3) *MPPT adjustment*: The controller with built-in MPPT functions executes the perturbation and observation algorithm at regular intervals to ensure better power output. In this process, the current signal will appear relatively strong noise, which may lead to the proposed algorithm misjudgment. Thirty sets of samples in this case are collected for testing.

4) *Blowing interference*: As is mentioned in [5] that when the dc arc is interfered by strong wind, the low frequency part of the signal will change, which will affect the detection of the arc fault. When this arc fault is simulated, an industrial fan is used to blow the arc and meanwhile, the arc must keep not extinguished, and 30 sets of samples are collected.

The frequency spectrums under different working conditions are depicted in Fig. 12. As mentioned in Section II-B, when an arc fault occurs, the amplitude of the frequency-doubling component of the switching frequency will decrease, and the noise of the low-frequency part will become stronger. In addition, pink noise may occur [31]. As can be seen from Fig. 12(b) and (c), in the process of the inverter start-up and the MPPT adjustment, the frequency-doubling component also decreases due to the action of the semiconductor switch, and the noise also becomes stronger. That is, its performance forms a domain arc fault with similar characteristics to cause interference. According to the proposed feature extraction method, a total of 25 dimensions of CMPE feature quantities can be obtained for each sample, to better display feature information. Ten samples are taken from each category, and the t-SNE [32] is used for the visualization. The t-SNE can realize the mapping of high-dimensional data to the low-dimensional space, and the distance between two points represents the similarity between them. The closer the distance is, the higher the similarity will be. The corresponding results are depicted in Fig. 13, where "o" represents the failure sample and "×" denotes the normal sample. Although similar samples are distributed in different regions of space, the boundary between different samples is relatively obvious, i.e., the proposed feature extraction method can effectively express the differences. By considering the dispersion of the feature distribution, the TWSVM with kernel functions is used to map the feature quantity to high dimension for achieving the effective classification. As can be seen from the classification results shown in Table VI, the adjustment of MPPT has a little impact on the detection results, and the measurement accuracy can reach 96.67%; that is, there is a misjudgment in one sample. The other categories of samples are 100% identifiable. By observing this misjudged sample, it is caused by the adjustment of MPPT under the occurrence of a large-area cloud shading. Because the corresponding current drop is very large, resulting in a large noise change, it may be misjudged as an arc fault.

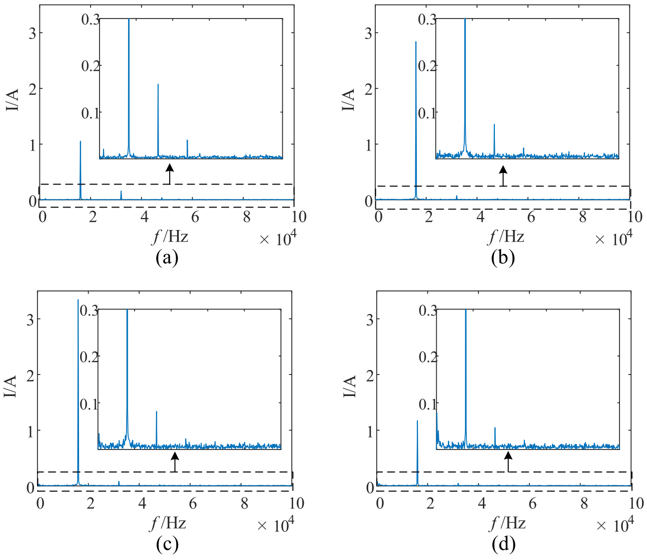


Fig. 12. Frequency spectrum under different working conditions. (a) Shading in normal. (b) Inverter startup. (c) MPPT adjustment. (d) Blowing interference.

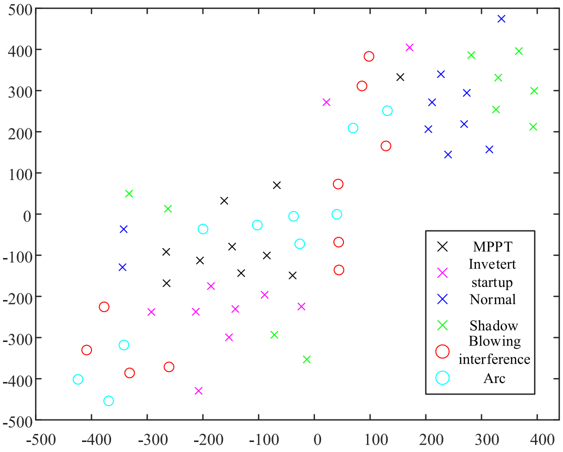


Fig. 13. Features visualization of various samples.

TABLE VI
SUMMARY OF MULTIPLE DATA SETS TO TEST ANTI-INTERFERENCE ABILITY

Fault Type Description	Sample Number	Accuracy
Shadow occlusion	30	100%
Inverter startup	30	100%
MPPT adjustment	30	96.67%
Blowing interference	30	100%

TABLE VII
IDENTIFICATION RESULTS OF EARLY FAULT

Normal time : Arc time	Sample Number	Correct Number
4:1	10	10
3:2	10	9
2:3	10	10
1:4	10	10
Total	40	39

5) *Early fault identification*: The goal of the arc fault detection is to detect the fault as soon as possible when an arc occurs and make a correct judgment. Nevertheless, in the early stage of the fault, there are both normal and fault states in the sampling time window. To test the effectiveness of the proposed method, the situation of the arc faults identification in different stages of the transition from the normal state to the fault state is also evaluated. The sampling time window is selected as 5 ms, and the time of the normal stage are set as 4, 3, 2 and 1 ms, respectively. In each time of the normal stage, ten samples are collected and put into the model for testing. The final test results are displayed in Table VII. All samples can be accurately identified as the occurrence of arc faults, except for one misjudgment. Thus, the proposed algorithm not only has an excellent recognition ability for the steady-state process, but also it can accurately recognize the fault in the early stage.

D. Analysis of Adaptability

The diagnostic effect of the proposed algorithm in long-line and single serial systems is first discussed here to analyze the adaptability of the algorithm. At the same time, the application prospect of the algorithm is verified by studying the relationship between the accuracy and execution time of the algorithm under different sampling frequencies.

1) *Long-line fault*: Generally speaking, the higher the capacity of the electric field is, the more the numbers of components connected in series will be. The content of the background noise is easily affected by the distance between the collection point and the fault point, which will reduce arc characteristics. Consequently, the line is lengthened by 10–20 m in this study, and 30 sets of fault samples are collected under this condition.

2) *Single string system*: The single series system is generally adopted in a household roof PV system. Once a fault occurs on the single series system, the performance of arc is stronger than the inside fault of a double series, which is more harmful. There are 30 sets of fault samples to be collected for verification in this study.

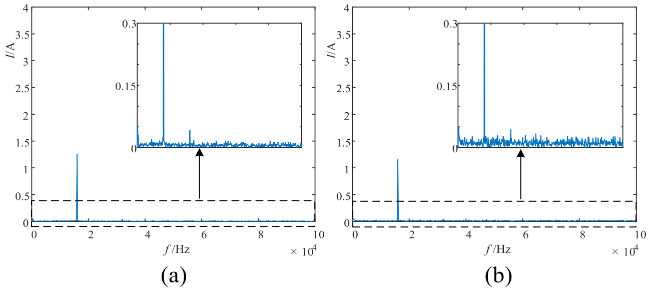


Fig. 14. Frequency spectrum under different systems. (a) Long line. (b) Single string.

TABLE VIII
SUMMARY OF MULTIPLE DATA SETS TO TEST ADAPTABILITY

Fault Type	Sample Number	Accuracy
Long-line & SAF	30	100%
1×12 PV system & SAF	30	100%

TABLE IX
TEST RESULTS OF MULTIPLE STRINGS FAULT SAMPLE

Fault Type	Sample Number	Accuracy
Bus rod	60	96.67%
Front of substring	30	93.33%
Middle of substring	30	100%
End of substring	30	93.33%
Total	150	96%

The FFT spectra of the above-mentioned two fault types, whose dc component has been removed, are depicted in Fig. 14. It is worthy to mention that the waveform amplitude in Fig. 14 is smaller than that in Fig. 12(b) due to a lower irradiance. The diagnostic results from Table VIII show that, whether it is a long-line fault or a single string system, the background noise, which mainly affects the signal, is significantly different from the characteristics produced by the fault state. Thus, judging from the diagnosis result, there is no misjudgment incident.

3) *Multiple strings systems of different sizes*: In some large power plants, multiple-string PV arrays will be constructed. Therefore, existing PV modules are assembled into a 3×8 scale array for verification. Arc locations are selected at the bus rod, the front of substring, the middle of substring, and the end of the substring. The number of total samples is 150. The testing results of the multiple-string systems with different sizes are summarized in Table IX. As can be seen from Table IX, the proposed algorithm still performs well with a total accuracy of 96%. When the fault occurred in the bus rod, the front of string, and the end of string positions, two samples are misjudged, respectively. By observing these misjudged samples, they are caused by the low irradiance of 150–200 W/m^2 when

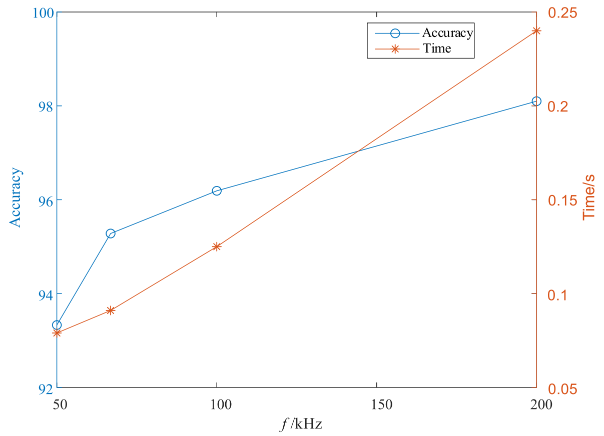


Fig. 15. Accuracy and time of fault detection under different sampling frequencies.

the characteristic information is relatively weak. It is worthy to continuously improve the identification accuracy under the occurrence of a very low irradiance or a huge irradiance drop in the future research.

4) *Change of sample frequency*: In the same time window, the higher the sampling rate, the more data points and the longer the processing time of data will be. To evaluate the relationship among the sampling frequency, the processing speed, and the accuracy, the data in Table IV are downsampled to 100, 66.8, and 50 kHz, respectively. The detection model is retrained, and the ratio of sample classification remains unchanged. The average test results of the samples are depicted in Fig. 15. It can be clearly found that if the sampling frequency is 200 kHz, the accuracy is the highest, but the execution time is also the longest.

With the reduction of the sampling frequency, the accuracy is gradually reduced, and the execution time is also gradually shortened. If the sampling frequency is 50 kHz, the accuracy is 93.33%, and the execution time is only 0.079 s. Therefore, the sampling frequency can be controlled according to application requirements to achieve a balance between the execution time and the accuracy.

5) *Verification of different PV systems*: To study the applicability and robustness of the proposed algorithm, a 3 × 6 scale PV system, which is depicted in Fig. 16, is used for verification. The power plant's PV modules have a capacity of only 1.8 kW, and the aging situation is inevitable after six-year operation. Moreover, the brand and operation parameters of inverters in Fig. 16 are summarized in Table X, which are different from the experimental system of a 2 × 12 scale PV system in Fig. 6. In this system, 230 sets of samples are obtained, half normal and half faulty. Sixty sets are randomly selected from normal and fault samples for testing, and the accuracy rate is only 62%. The major reason is that the switching frequency of two inverters is quite different, and the number of frequency doubling components and noise level are also different, which leads to the fact that the original learning features are not applicable. To this end, 200 sets of samples are selected to retrain the diagnostic model, and the remaining 30 sets of samples are used for testing. The test accuracy rate can reach 96.7%. From the verification results, it can be found that the proposed method is based on the changes in switching frequency and noise level before and after the fault to dig out the fault characteristics, which is dependent on the inverter used. Thus, the model should be retrained when it is applied for different types of inverters. Considering that each inverter has a fixed switching frequency, it will also have the value and significance of promotion if a diagnostic model can be trained for each specific type of inverter. In the future research, the further mining of characteristic information can be investigated to solve the algorithm adaptability under different switching frequencies of inverters.

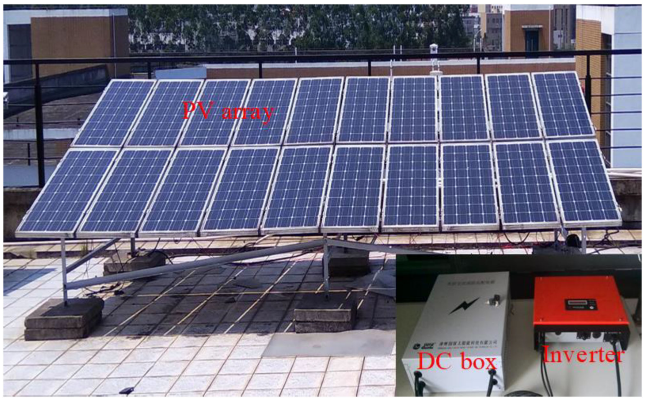


Fig. 16. Photograph of a 3 × 6 scale PV system.

TABLE X
PARAMETERS OF A 3 × 6 SCALE PV SYSTEM UNDER STANDARD TEST CONDITION

Equipment	Parameter				
PV module	P_{mpp}	V_{mpp}	I_{mpp}	V_{OC}	I_{sc}
	99.75W	17.5V	5.7A	21.5V	6.03 A
PV array	3×6 modules (Parallel and Series)				
Inverter	Model	Goodwe GW3000-NS			
	Start voltage	80V			
	MPP Voltage range	80-450V			
	Input	80-500V			
	Output	220V			
	Switching frequency	20kHz			

V. COMPARISON AND DISCUSSION

A. Performance Analysis of Each Module

For the purpose of analyzing and comparing the performance of the IEWT, the CMPE, and the TWSVM, three models are constructed, namely the EWT-CMPE-TWSVM model, the IEWT-MPE-TWSVM model, and the IEWT-CMPE-SVM model. According to the data set in Table IV, the above-mentioned models are compared with the proposed method, and the ratio of sample classification remains unchanged. The test results including detection accuracies, execution times, and relative percentages of execution times are summarized in Table XI. Note that, the

TABLE XI
TEST RESULTS FOR VARIOUS MODELS

Model	Accuracy	Execution time	Relative percent of execution time
EWT-CMPE-TWSVM	88.89%	243ms	98.78%
IEWT-MPE-TWSVM	92.38%	142ms	57.72%
IEWT-CMPE-SVM	98.10%	246ms	100%
Proposed method	98.10%	246ms	100%

relative percentage of the execution time in Table XI is calculated by the basis of the execution time required in the proposed method (246 ms).

In fact, the EWT algorithm improved by the mathematical morphology can effectively avoid too dense spectrum segmentation, prevent effective information fragmentation, and improve the accuracy of recognition. As can be seen from Table XI, the IEWT algorithm improves the accuracy by 9.21%, whereas the corresponding execution time is only increased by 1.22%. Note that, the effectiveness of the mathematical morphology in the feature extraction can also be referred to [33]. When the MPE is used in the fault feature extraction, the detection speed is increased by 42.28%, but its accuracy is 5.62% lower. This indicates that the CMPE can improve the entropy quality by solving the problem that the entropy of the MPE changes greatly due to coarse granularity. Although the SVM is very effective in dealing with the binary classification problem, it needs to solve the quadratic programming problem in the process of calculation, which requires a large amount of computation. As Jayadeva [34] stated, the TWSVM can determine two nonparallel planes by solving two related SVM-type problems, and the computation time can be greatly improved in comparison with the SVM. This time difference is mainly reflected in the training stage. The experiment shows that the running times of the TWSVM and the SVM are, respectively, 100 ms and 380 ms after one training process, which is four times faster. During the testing phase, the SVM and the TWSVM run roughly with the same amount of time. The same parameter tuning and optimization techniques are used to implement the SVM and the TWSVM, and the diagnostic accuracy of both methods is 98.10%. That is, the TWSVM do not change the classification accuracy of the SVM.

B. Compared With Other Methods

The comparison with other similar methods in [11], [12], [17], and [35] via current data of PV arrays to realize the arc fault detection is summarized in Table XII. In [11], the current time-domain sequence was collected at a sampling frequency of 250 kHz; the current data were decomposed by the method of the wavelet packet transform, and the wavelet coefficients and energy were used as the input data of the SVM classifier for training. However, not all interferences phenomena, especially the most common shading and inverter startup, are considered. In [12], a complex VMD algorithm was adopted to calculate the change of entropy in a time window of 50 ms, so that whether an arc fault occurred can be determined. In other words, the signal used in [12] must contain transient processes, and the effective entropy can be obtained based on the amount of mutation. On the contrast, the proposed method in this study focuses on the steady-state signal of the arc fault; additionally, the early stages of the development of the fault are also taken into account. In [17], the current sequence waveform was folded into a 2-D matrix as input data, and an end-to-end domain adaptation combined with DC-GAN (DA-DCGAN) was constructed to directly detect arc faults. As for [17], it is inevitable that a lot of network layer parameters must be set manually, and the quantity of computation of training the network will be quite large. Besides, considering that there is difficulty in obtaining the measured signal, a sampling frequency of 20 kHz, and a time window of 20 ms is set in [17] as the acquisition standard. On this basis, 25 000 sets of normal data and 5000 sets of arc fault data collected from the PV simulation system are used as source-domain data, and the obtained data is enhanced to arc data of the real environment by transfer learning. In other words, to achieve higher diagnostic accuracy, a large amount of data is required to train the identification model in [17]. In [35], the WPD was performed as Rbio3.1 was applied as the mother wavelet. Moreover, the wavelet coefficient energy value of a specific frequency band was taken as the characteristics, and the random forest (RF) was used for classification and identification to detect the arc faults of different PV systems. However, the change in switching frequency during the inverter startup is not taken into account in [35], which is easily to cause misjudgment. Besides, only when the mutation of the wavelet coefficients (transient process) is detected, the occurrence of the fault can be found and the identification cannot be realized for arc fault signals at the steady state.

To compare the advantage of the proposed method, this study will make a quantitative comparison with [11], [12], [17], [20], and [35]. In [20], the mean, the median, the variance, the root-mean-square value, and the difference between the maximum and minimum values of the bus current are used as feature quantities, and an integrated machine-learning algorithm is applied as a classifier for the dc arc recognition. Although the method in [20] is not applied for a PV system, it is a typical time-domain-based analytic method for arc faults, which meets the requirements of the quantitative comparison with the proposed method in this study.

The measured data collected in Tables IV, VI, and VIII are used as input data, the ratio of the training samples to the testing ones is 4:1, and the classification accuracy and the execution time are adopted as evaluation indicators. The comparisons of the detection accuracy and the execution time of the proposed method with the ones in [11], [12], [17], [20], and [35] are summarized in Table XIII. As can be seen from Table XIII, the defects of [11], [12], and [35] are the lack of the anti-interference ability, and various normal interference samples are recognized as fault samples, which leads to an accuracy of only 61.4%, 60.7%, and 65.61%, respectively. In [12], due to the use of the VMD algorithm, the calculation time is up to 1.7 s, which is much larger than other methods. In [17], the deep learning

TABLE XII
COMPARATIVE RESULTS BETWEEN THE PROPOSED METHOD AND METHODS IN [11], [12], [17], AND [35]

Case Study		Proposed method	[11]	[12]	[17]	[35]
Year		2021	2019	2019	2019	2019
Capacity		6.4kW	5.0kW	0.96kW	1.4kW	10kW
Diagnosis technology	Signal source	DC-bus current	DC-bus current	DC-bus current	DC-bus current	DC-bus current
	Signal characteristics	Transient or steady state	Transient or steady state	Transient	Transient or steady state	Transient
	Time window	5ms	4ms	50ms	20ms	8ms
	Sample rate	200kHz	250 kHz	200kHz	20kHz	187.6 kHz
	Feature extract	IEWT-CMPE	WPD	VMD	2D-transform	WPD
	Classifier	SSA-TWSVM	SVM	Threshold value method	DA-DCGAN	RF
Are different fault positions considered?		Yes	Yes	Yes	Yes	Yes
Is shadow occlusion considered?		Yes	No	Yes	Yes	Yes
Is long distance considered?		Yes	No	No	No	No
Is the process of inverter startup considered?		Yes	No	Yes	Yes	No
Is the impact of strong wind considered?		Yes	No	No	No	No
Is MPPT adjustment considered?		Yes	No	Yes	Yes	Yes
Is action of power switches considered?		Yes	No	No	Yes	No

TABLE XIII
COMPARISON OF TEST RESULTS OF DIFFERENT METHODS

Method	Proposed method	[11]	[12]	[17]	[20]	[35]
Accuracy	98.94%	61.4%	60.7%	84.56%	68.07%	65.61%
Execution time	246ms	14ms	1.7s	174ms	2ms	15ms

method is adopted, and good detection results can be obtained. However, it required many data as support, and the parameter adjustment process takes a lot of time. In [20], time-domain data are taken as features such that the corresponding execution time is shortest. Moreover, there are more features captured in [20], which makes it slightly more accurate than that in [11]. Because the fault feature in [20] cannot be effectively expressed by the features in the time domain, the accuracy is only 68.07%. In addition, the method in [20] is also prone to misjudgment when it is faced with interference samples. Although the EWT is adopted by the proposed method to result in a slower speed, a high quality of feature expression can be obtained via the time–frequency feature extraction, and the holistic accuracy is superior to the methods in [11], [12], [17], [20] and [35].

VI. CONCLUSION

A new type of PV arc fault diagnostic method is proposed in this study by investigating the time–frequency characteristics of PV arrays under normal and arc fault conditions. The time–frequency domain information matrix of the signals can be obtained by using the EWT algorithm of mathematical morphology modification. Moreover, the fault features can be characterized by the CMPE, and the binary classification can be realized by the TWSVM. The major contributions of this study are summarized as follows.

1) Because the Hankel-SVD can eliminate the specified switching frequency and the background noise, it can avoid the influence of the inverter noise and other electrical noise on the extraction of arc fault characteristics.
2) The frequency spectrum decomposed by the EWT is smoothed by the closed operation in the mathematical morphology, which solves the problem of too dense frequency spectrum division of dc arc signals by the EWT.
3) Compared with the SVM, the utilization of the TWSVM greatly improves the training time of the classification algorithm. By using SSA to optimize the hyperparameters in the TWSVM, the corresponding optimal values can be found quickly, and the recognition accuracy can be improved.
4) Arc faults in the transient state or the steady state can be effectively detected by the proposed method.

The proposed method can effectively resist the interference of dynamic fast shading, the inverter startup, and the strong wind, which has been proved by experimental verification. Additionally, the detection accuracy of early faults is relatively high. In the case of long-line faults, single string array, and multiply strings array of different sizes, the adaptability of the proposed method is also relatively strong. In this study, the proposed algorithm applies the EWT and the CMPE, respectively, for the time–frequency decomposition and the eigenvalue calculation. It takes a longer time and the single execution time reaches 246 ms. In the future research, the diagnostic speed can be improved by optimizing the time–frequency decomposition algorithm, screening the effective fault features and adopting a fast classification algorithm.

REFERENCES

[1] REN21, "Renewables 2019 global status report," REN21, Paris, France, May 2019. [Online]. Available: https://www.ren21.net/wp-content/uploads/2019/05/gsr_2019_full_report_en.pdf

[2] Q. Xiong *et al.*, "Arc fault detection and localization in photovoltaic systems using feature distribution maps of parallel capacitor currents," *IEEE J. Photovolt.*, vol. 8, no. 4, pp. 1090–1097, Jun. 2018.

[3] National Electrical Code(R) (NEC) Edition, NFPA70, Nat. Fire Protection Assoc., Quincy, MA, USA, 2014.

[4] S. Dhar, R. K. Patnaik, and P. K. Dash, "Fault detection and location of photovoltaic based DC microgrid using differential protection strategy," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4303–4312, Sep. 2018.

[5] S. B. Lu, B. T. Phung, and D. Zhang, "A comprehensive review on DC arc faults and their diagnosis methods in photovoltaic systems," *Renewable Sustain. Energy Rev.*, vol. 89, pp. 88–98, Jun. 2018.

[6] Q. Lu *et al.*, "A DC series arc fault detection method using line current and supply voltage," *IEEE Access*, vol. 8, pp. 10134–10146, Jan. 2020.

[7] S. Chen, Q. Lv, Y. Meng, X. Li, and N. Xu, "Hardware implementation of series arc fault detection algorithm for different DC resistive systems," in *Proc. IEEE Holm Conf. Elect. Contacts*, 2019, pp. 245–249.

[8] M. Ahmadi, H. Samet, and T. Ghanbari, "Series arc fault detection in photovoltaic systems based on signal-to-noise ratio characteristics using cross-correlation function," *IEEE Trans. Ind. Inform.*, vol. 16, no. 5, pp. 3198–3209, May 2020.

[9] S. Chae, J. Park, and S. Oh, "Series DC arc fault detection algorithm for DC microgrids using relative magnitude comparison," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 4, no. 4, pp. 1270–1278, Dec. 2016.

[10] H. Zhu, Z. Wang, and R. S. Balog, "Real time arc fault detection in PV systems using wavelet decomposition," in *Proc. IEEE 43rd Photovolt. Spec. Conf.*, 2016, pp. 1761–1766.

[11] K. Xia *et al.*, "Wavelet packet and support vector machine analysis of series DC arc fault detection in photovoltaic system," *IEEJ Trans. Electr. Electron. Eng.*, vol. 14, no. 2, pp. 192–200, Feb. 2019.

[12] S. Liu *et al.*, "Application of the variational mode decomposition-based time and time-frequency domain analysis on series DC arc fault detection of photovoltaic arrays," *IEEE Access*, vol. 7, pp. 126177–126190, Sep. 2019.

[13] W. Miao, Q. Xu, K. H. Lam, P. W. T. Pong, and H. V. Poor, "DC arc-fault detection based on empirical mode decomposition of arc signatures and support vector machine," *IEEE Sens. J.*, vol. 21, no. 5, pp. 7024–7033, Mar. 2021.

[14] C. H. Wu, W. X. Xu, Z. H. Li, L. J. Xu, and T. Y. Bai, "Study on detection method and its anti-interference of dc arc fault for photovoltaic system," *Proc. CSEE*, vol. 38, no. 12, pp. 3546–3555, Jan. 2018.

[15] R. D. Telford, S. Galloway, B. Stephen, and I. Elders, "Diagnosis of series DC arc faults—A machine learning approach," *IEEE Trans. Ind. Inform.*, vol. 13, no. 4, pp. 1598–1609, Aug. 2017.

[16] A. Khamkar and D. D. Patil, "Arc fault and flash signal analysis of DC distribution system sing artificial intelligence," in *Proc. Int. Conf. Renewable Energy Integr. Smart Grids, Multidisciplinary Approach Technol. Model. Simul.*, 2020, pp. 10–15.

[17] S. Lu, T. Sirojan, B. T. Phung, D. Zhang, and E. Ambikairajah, "DA-DCGAN: An effective methodology for DC series arc fault diagnosis in photovoltaic systems," *IEEE Access*, vol. 7, pp. 45831–45840, Apr. 2019.

[18] Q. Xiong, S. Ji, L. Zhu, L. Zhong, and Y. Liu, "A novel DC arc fault detection method based on electromagnetic radiation signal," *IEEE Trans. Plasma Sci.*, vol. 45, no. 3, pp. 472–478, Mar. 2017.

[19] M. K. Alam, F. H. Khan, J. Johnson, and J. Flicker, "PV arc-fault detection using spread spectrum time domain reflectometry (SSTDR)," in *Proc. IEEE Energy Convers. Congr. Expo.*, 2014, pp. 3294–3300.

[20] V. Le, X. Yao, C. Miller, and B. Tsao, "Series DC arc fault detection based on ensemble machine learning," *IEEE Trans. Power Electron.*, vol. 35, no. 8, pp. 7826–7839, Aug. 2020.

[21] Y. G. Yue *et al.*, "Suppression of periodic interference during tunnel seismic predictions via the Hankel-SVD-ICA method," *J. Appl. Geophys.*, vol. 168, pp. 107–117, Sep. 2019.

[22] J. Gilles, "Empirical wavelet transform," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3999–4010, Aug. 2013.

[23] A. Q. Zhang, T. Y. Ji, M. S. Li, Q. H. Wu, and L. L. Zhang, "An identification method based on mathematical morphology for sympathetic inrush," *IEEE Trans. Power Del.*, vol. 33, no. 1, pp. 12–21, Feb. 2018.

[24] M. Salehi and F. Namdari, "Fault location on branched networks using mathematical morphology," *IET Gener. Transmiss. Distrib.*, vol. 12, no. 1, pp. 207–216, Jan. 2018.

[25] S. Nalband, A. Prince, and A. Agrawal, "Entropy-based feature extraction and classification of vibroarthographic signal using complete ensemble empirical mode decomposition with adaptive noise," *IET Sci. Meas. Technol.*, vol. 12, no. 3, pp. 350–359, May 2018.

[26] X. J. Chen, Y. M. Yang, Z. X. Cui, and J. Shen, "Wavelet denoising for the vibration signals of wind turbines based on variational mode decomposition and multiscale permutation entropy," *IEEE Access*, vol. 8, pp. 40347–40356, Feb. 2020.

[27] Z. Q. Huo, Y. Zhang, L. Shu, and M. Gallimore, "A new bearing fault diagnosis method based on fine-to-coarse multiscale permutation entropy, Laplacian score and SVM," *IEEE Access*, vol. 7, pp. 17050–17066, Jan. 2019.

[28] R. K. Jayadeva and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 905–910, May 2007.

[29] S. Mirjalili *et al.*, "Salp swarm algorithm: A bio-inspired optimizer for engineering design problems," *Adv. Eng. Softw.*, vol. 114, pp. 163–191, Dec. 2017.

[30] I. Jebli, F. Belouadha, M. I. Kabbaj, and A. Tilioua, "Prediction of solar energy guided by Pearson correlation using machine learning," *Energy*, vol. 224, Jun. 2021, Art. no. 120109.

[31] N. L. Georgijevic, M. V. Jankovic, S. Srdic, and Z. Radakovic, "The detection of series arc fault in photovoltaic systems based on the arc current entropy," *IEEE Trans. Power Electron.*, vol. 31, no. 8, pp. 5917–5930, Aug. 2016.

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

[33] A. Q. Zhang, T. Y. Ji, M. S. Li, Q. H. Wu, and L. L. Zhang, "An identification method based on mathematical morphology for sympathetic inrush," *IEEE Trans. Power Del.*, vol. 33, no. 1, pp. 12–21, Feb. 2018.

[34] R. K. Jayadeva and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 905–910, May 2007.

[35] S. L. Chen, X. W. Li, Y. Meng, and Z. M. Xie, "Wavelet-based protection strategy for series arc faults interfered by multicomponent noise signals in grid-connected photovoltaic systems," *Sol. Energy*, vol. 183, pp. 327–336, May 2019.