frontiers | Frontiers in Energy Research

TYPE Original Research
PUBLISHED 10 January 2023
DOI 10.3389/fenrg.2022.1036984

OPEN ACCESS

EDITED BY
Da Xie,
Shanghai Jiao Tong University, China

REVIEWED BY
Arturo Garcia Perez,
University of Guanajuato, Mexico
Tao Rui,
Anhui University, China

*CORRESPONDENCE
Qiong Li,
powerdsp339@163.com

SPECIALTY SECTION
This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research

RECEIVED 05 September 2022
ACCEPTED 22 November 2022
PUBLISHED 10 January 2023

CITATION
Wu Y, Li Q, Long G, Chen L, Cai M and Wu W (2023), Research on arc grounding identification method of distribution network based on waveform subsequence segmentation-clustering.
Front. Energy Res. 10:1036984.
doi: 10.3389/fenrg.2022.1036984

Research on arc grounding identification method of distribution network based on waveform subsequence segmentation-clustering

Yihui Wu[1], Qiong Li[1]*, Guohua Long[2], Liangliang Chen[1], Muliang Cai[2] and Wenbao Wu[3]

[1]Nanchang Hangkong University, Nanchang, China, [2]National Network Jiangxi Electric Power Co. Ltd., Nanchang, China, [3]China Power Construction Group Jiangxi Electric Power Construction Co. Ltd., Nanchang, China

The traditional method of detecting fault current based on threshold judgment method is limited by the current size and is easily disturbed by noise, and it is difficult to adapt to the arc ground fault detection of the distribution network. Aiming at this problem, this paper proposes a single-phase arc-optic ground fault identification method based on waveform subsequence splitting fault segmentation, combined with three-phase voltage-zero sequence voltage waveform feature extraction clustering. First of all, the waveform fault segment is segmented and located, secondly, the characteristic indexes of the time domain and frequency domain of the combined three-phase voltage-zero sequence voltage waveform are established, and the multidimensional feature distribution is reduced by the principal component analysis method, and finally, the characteristic distribution after the dimensionality reduction is identified by the K-means clustering algorithm based on the waveform subsequence. Experimental results show that the arc light grounding fault identification method proposed in this paper achieves 97.12% accurate identification of the test sample.

KEYWORDS

arc grounding, waveform subsequences, K-means clustering, fault identification, sequence segmentation

1 Introduction

According to the survey, more than 80% of the power outage losses are caused by distribution network failures, so the fault diagnosis of distribution networks has always been the research object of power supply units. Among them, arc light grounding fault is not easy to find and the harm is huge, and the mechanism is complex, which is a category of grounding faults that are difficult to detect. Therefore, it is of great significance to propose a reliable and efficient arc-ray grounding identification algorithm for the operation of the distribution network.

The traditional arc-optic grounding fault identification method of the distribution network is based on the steady-state or transient electrical parameters and the set threshold (Chen et al., 2021), and in the fault identification method based on the transient electrical parameters, the characteristic parameters of the typical fault type are first extracted, including wavelet transform (WT) (Qin et al., 2018; Lin et al., 2019; Wei et al., 2020a), empirical mode decomposition (EMD) (Guo et al., 2019; Cai and Wai, 2022), and S transformation (ST) (Peng et al., 2019); Then, the arc ground fault is classified and identified by the pattern recognition method, mainly including the neural network method (Siegel et al., 2018; Du et al., 2019a), the Support Vector Machine (SVM) (Xia et al., 2019; Dang et al., 2022), the fuzzy control method (Zeng et al., 2016), the clustering (Wang et al., 2015), etc., in addition, the high-precision current transformer can be used to improve the fault identification ability (Paul, 2015), but the detection cost is also significantly increased.

Reference (Mishra et al., 2016) classifies arc ground faults through fault data, extracts five fault features, and inputs them into a fuzzy inference system for identification. Although the identification effect is remarkable, the establishment of fuzzy control rules relies on historical experience, and the ability to learn independently is poor. Reference (Gadanayak and Mallick, 2019; Wang et al., 2021) combined Variational Mode Decomposition (VMD) and support vector machine to identify arc ground faults. Different eigenmode functions were obtained by decomposing the collected ground fault signals by VMD, and the faults were extracted. The typical characteristics of the signal are found, the displacement entropy with the greatest contribution is found, and the arc ground fault is identified by the support vector machine. However, support vector machines need a large number of samples for training, and mode aliasing effects are prone to occur during empirical mode decomposition, and the parameters of variational mode decomposition need to be selected manually. Reference (Guo M. F. et al., 2018) proposed a ground fault detection method based on wavelet transform and Convolutional Neural Network (CNN). The time-frequency components were obtained through wavelet transform, and then each component was normalized. Identify fault features. However, the selection of wavelet transform basis functions has limitations, and the neural network needs to be trained on a large number of samples. Reference (Wei et al., 2020b) proposed a generalized S-transform with variable factors to detect ground faults. This method has stronger adaptability and higher detection accuracy, but local over-fitting is prone to occur in the S-transform calculation process. Reference (Zhang et al., 2019) proposed a fault identification method based on waveform feature extraction and matrix analysis and clustering. This method can identify different grounding resistances, but the efficiency and accuracy of the algorithm need to be improved.

For this reason, this paper proposes a time series characteristic analysis method combining three-phase voltage and zero-sequence voltage waveforms to solve the problems that the threshold value setting cannot be automated and requires a large number of samples training in the traditional detection of electrical parameters and threshold values. At the same time, in view of the problems of traditional waveform analysis feature dimension redundancy and large amount of calculation, a method of arc grounding fault identification for distribution network based on segmentation-clustering is proposed.

The main contributions of this paper are:

(1) Considering that there are developing faults in the field data, the direct use of the wave recorder data will cause the eigenvalues of the arc ground fault to be confused with other types of fault data. Therefore, a waveform subsequence segmentation method based on the sliding *t*-test is proposed to achieve the same Segmentation of different types of fault data in recorded wave data.
(2) A fault identification model combining the time series feature extraction of three-phase voltage and zero-sequence voltage waveforms is established. Through the analysis of experimental data, the boundary conditions of arc ground fault and other faults are obtained, which effectively improves the traditional feature extraction based on current analysis. The problem.

The first part of this paper analyzes the waveform characteristics based on waveform subsequence segmentation, the second part proposes an arc-flash grounding fault identification algorithm based on segmentation-clustering, the third part carries out numerical example simulation and analysis, and finally gives the conclusion.

2 Waveform feature analysis based on fault waveform subsequence segmentation

2.1 Waveform subsequence segmentation

The data source of the on-site arc ground fault is mainly the recorded data of the fault recorder. When a fault occurs, the fault recorder can automatically and accurately record the changes of various electrical quantities in the process before and after the fault occurs. Due to the development of ground faults, it is possible to evolve from one type of fault to another. As shown in Figure 1, no fault occurred in the time period of 1.0 s–1.08 s; arc ignition occurred many times in the time period of 1.08 s–2.7 s; resonance fault occurred in the time period of 2.7 s–3.1 s; 3.1 s–3.5 return to normal within s time period.

If the data analysis of the fault segment of the wave recorder is used directly, the analysis of fault

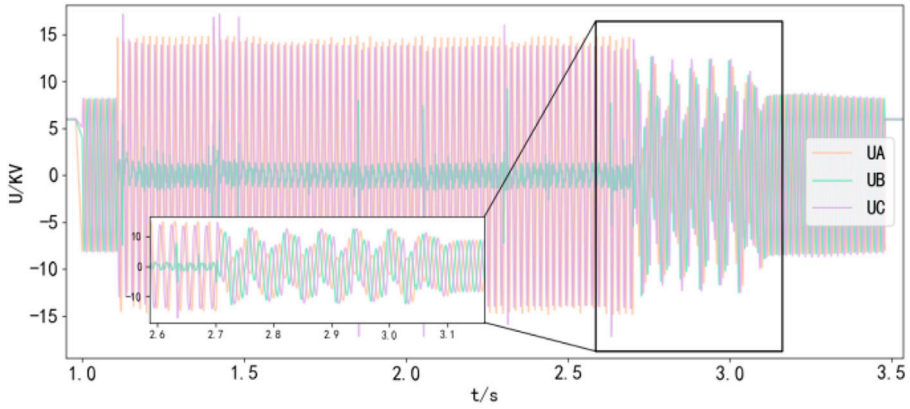


FIGURE 1
Three-phase voltage waveform diagram of developing fault.

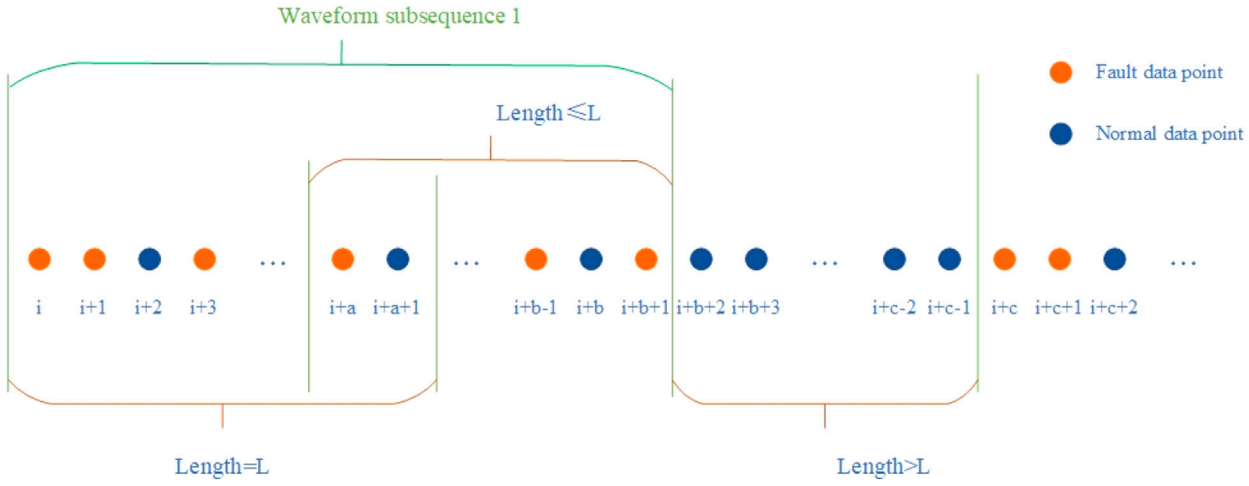


FIGURE 2
Schematic diagram of waveform subsequence segmentation.

characteristics may be confused, so each fault needs to be segmented. To this end, this paper proposes a fault subsequence segmentation method based on sliding *t*-test. Compared with other sequence segmentation algorithms, this method is simple in algorithm and less computationally expensive. Due to the strong periodicity of the fault voltage waveform, the parameter selection problem of the sliding *t*-test becomes simple, and the selected parameters are applicable to all the recorder data.

Sliding *t*-test tests for mutation by calculating whether the difference between the mean of two groups of samples is significant, t follows a distribution with $df = n_1 + n_2 - 2$ degrees of freedom, Given a significant level of $\alpha = 0.05$, the critical value $t_{0.05}$ is obtained by looking up the t distribution table, if $|t_i| > t_\alpha$,a fault is considered to have occurred at that point. For time series $X(t)$ of length n, set time $i(n_1 \le i \le n - n_2)$ as the reference point, The two subsequences x_1 and x_2 before and after the reference point define the statistic t_i at time i:

$$t_i = \frac{1}{\sqrt{\frac{(n_1 s_1^2 + n_2 s_2^2)}{(n_1 - n_2 - 2)^*}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} (\overline{x_1} - \overline{x_2}) \quad (1)$$

In the formula (Du et al., 2019b): n_1 and n_2 are the sample sizes of the two subsequences, $\overline{x_1}$ and $\overline{x_2}$ are the average of the two subsequences, s_1^2 and s_2^2 are the variances of the two subsequences.

TABLE 1 Time-domain waveform characteristics table.

Type	Mean				Variance			
	U_A/kV	U_B/kV	U_C/kV	U_0/kV	U_A	U_B	U_C	U_0
Arc grounding	−0.449	0.039	−0.409	−0.819	113.749	6.261	82.247	238.297
Ferromagnetic resonance	−0.223	0.212	0.268	0.256	39.791	40.559	46.308	62.438
Normal	0.287	−0.061	−0.234	−0.008	36.239	35.824	36.580	7.351
Type	**Kurtosis**				**Peak to peak**			
	U_A	U_B	U_C	U_0	U_A/kV	U_B/kV	U_C/kV	U_0/kV
Arc grounding	−1.647	0.089	−1.169	−1.851	29.359	12.632	29.067	44.765
Ferromagnetic resonance	−1.026	−1.023	−1.153	−0.472	23.541	24.665	25.018	35.469
Normal	−1.385	−1.450	−1.453	1.855	16.096	16.057	16.267	2.795

TABLE 2 Frequency domain waveform characteristics table.

Type	Odd harmonic content of zero sequence voltage			Center frequency		Frequency standard deviation		Root mean square frequency	
	3 times (%)	5 times (%)	7 times (%)	Fault phase voltage	Zero sequence voltage	Fault phase voltage	Zero sequence voltage	Fault phase voltage	Zero sequence voltage
Arc grounding	6.937	2.503	2.266	213.947	143.786	234.466	204.445	317.408	476.532
Ferromagnetic resonance	2.547	0.765	0.194	183.792	93.655	357.686	260.193	402.143	208.420
Normal	24.514	1.244	1.217	—	312.341	—	384.744	—	495.566

As shown in Figure 2, when the sliding *t*-test is performed on the fault recorder data in this paper, there is at least a segment of non-fault sample points before the fault sample point i and the length is at least L, the length of the sequence (i, i+a+1) is L, the length of the sequence (i+a, i+b+1) is less than or equal to L, the length of the sequence (i+b+1, i+c-1) is greater than L and all are normal sample points, the sequence (i, i+b+1) is a waveform subsequence. In this paper, the following considerations are made for subsequence segmentation:

(1) Sequence (i, i+a+1), that is, a sequence of length L at the beginning of the fault is temporarily stored as the initial waveform subsequence.
(2) If there are fault points in the sequence (i+a, i+b+1), that is, taking the last fault point of the waveform subsequence as the starting point and there are fault points in the supplementary sequence of length L, all the fault points before the last fault point of the supplementary sequence are merged with the initial waveform subsequence, and the initial waveform subsequence is replaced. If there is no fault point within the supplementary sequence, the initial waveform subsequence is split into the final waveform subsequence and labeled.
(3) Repeat step (2) until all waveform subsequences are segmented.

2.2 Waveform feature analysis

2.2.1 Time domain waveform characteristics

Time Domain Analysis enables intuitive and accurate analysis of systems in the time domain. For the arc grounding system, time domain indicators such as mean value, variance, peak-to-peak value and kurtosis coefficient are selected for analysis. The mean shows the average level of the data, the variance measures the degree of dispersion of the data, the peak-to-peak value shows the difference between the highest value and the lowest value of the signal in a period, and the kurtosis coefficient reflects the distribution characteristics of the vibration signal. The calculation of the kurtosis coefficient when the fourth power is used, the influence of noise can be reduced and the signal-to-noise ratio can be improved.

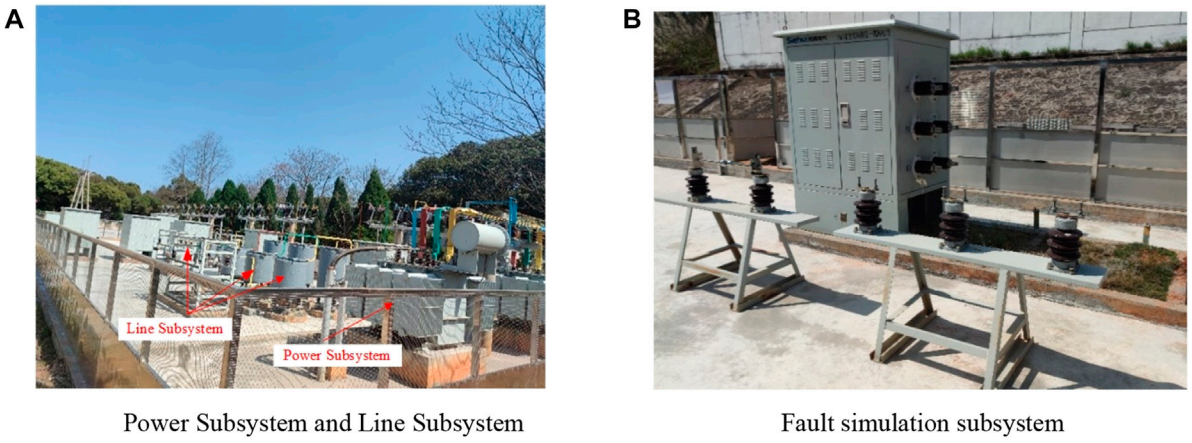


FIGURE 3
Physical experiment platform. **(A)** Power Subsystem and Line Subsystem. **(B)** Fault simulation subsystem.

Through the time domain feature extraction of arc ground fault, ferromagnetic resonance fault and normal conditions, the time domain eigenvalues of the three types of data shown in Table 1 are obtained. In the table, the zero-sequence voltage variance of arc ground fault is obviously larger than that of ferromagnetic resonance fault and normal condition. And the mean value of the normal situation is closer to zero, the zero-sequence voltage kurtosis value of the normal situation is the largest, and it is distinguished from the other two faults. The peak-to-peak value of the normal case is close to zero, and the peak-to-peak value of the three-phase voltage of the arc ground fault is larger than the other two cases except for the faulty phase.

The variances and peak-to-peak values of zero sequence voltages and three-phase voltages of the four characteristics can distinguish arc grounding faults, ferromagnetic resonance faults and normal conditions.

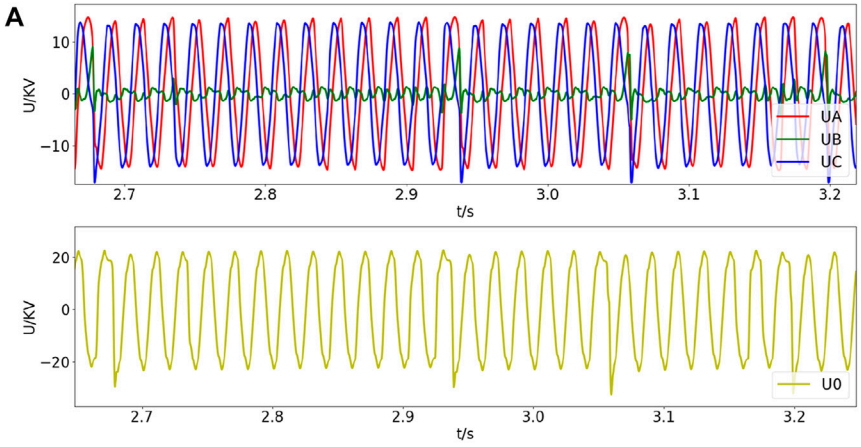
2.2.2 Frequency domain waveform characteristics

Frequency domain analysis is a method of evaluating system performance using graphical analysis in the frequency domain. It can not only reflect the steady-state performance of the system, but also can be used to study the stability and transient performance of the system. For the arc grounding system, frequency domain time scales such as odd harmonic content, spectral gravity center frequency, spectral frequency standard deviation, and spectral root mean square frequency are selected for analysis. The harmonic content is the amount obtained by subtracting the fundamental wave component from the alternating current. The voltage in the power grid is mainly 50 Hz. In some cases, a higher frequency signal will appear. When the frequency of the harmonic signal is the fundamental wave signal when the frequency is an odd multiple, the harmonic is called an odd harmonic. The center of gravity frequency can describe the frequency of the signal component with larger components in the frequency spectrum of the signal, and reflects the situation of the signal power spectrum. The frequency standard deviation describes the spread of the power spectrum energy distribution. The root mean square frequency is the arithmetic square root of the mean square frequency, which can be regarded as the radius of inertia.

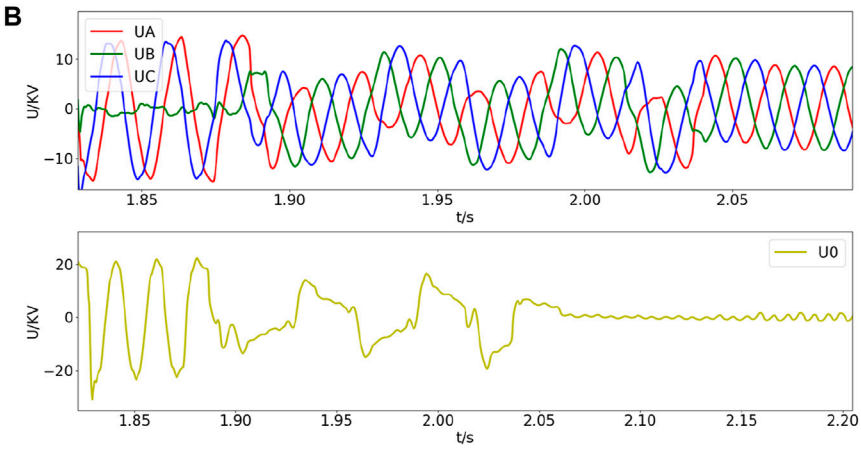
The frequency domain eigenvalues of the three types of data shown in Table 2 are obtained by extracting the frequency domain features for arc ground fault, ferromagnetic resonance fault and normal conditions. The third harmonic content of the normal condition in the table is significantly higher than the other two fault conditions. Under normal conditions, the centroid frequency of zero-sequence voltage, the standard deviation of spectral frequency and the frequency of spectral root mean square are the largest. The frequency of the spectral center of gravity of the arc ground fault is slightly larger than that of the ferromagnetic resonance fault.

The spectral centroid frequency, spectral frequency standard deviation, and spectral root mean square frequency of the frequency domain waveform feature can clearly distinguish the three cases. The arc ground fault, ferromagnetic resonance fault and the normal zero-sequence voltage odd harmonic content can also distinguish the three cases.

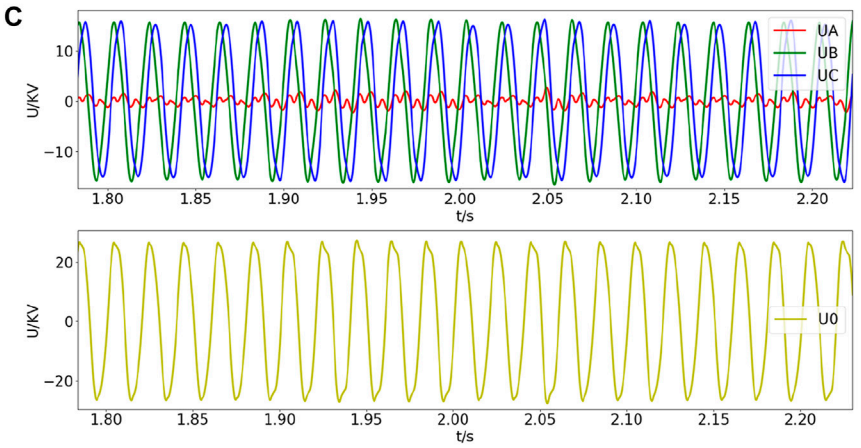
In Algorithm 1, input the recorder data CFG file, and the sliding step. After initialization, the t index is calculated. When the t index exceeds the significant interval, it is saved in A, and then the fault points in A are merged and subsequences are divided. Finally, the time domain and frequency domain eigenvalues of each subsequence are extracted, and the subsequence and the eigenvalue matrix are output.



Three-phase voltage and zero-sequence voltage waveform of arc ground fault



Three-phase voltage and zero-sequence voltage waveform of ferromagnetic resonance fault



Three-phase voltage and zero-sequence voltage waveform of general ground fault

FIGURE 4
Typical failure diagram of the test sample. **(A)** Three-phase voltage and zero-sequence voltage waveform of arc ground fault. **(B)** Three-phase voltage and zero-sequence voltage waveform of ferromagnetic resonance fault. **(C)** Three-phase voltage and zero-sequence voltage waveform of general ground fault.

TABLE 3 Variance explanation rate table.

Element	Variance explained rate		
	Characteristic root	Percent variance	Accumulation %
1	13446.99	92.34	92.34
2	1051.66	7.22	99.56
3	40.07	0.28	99.84
4	14.30	0.10	99.93
5	8.22	0.06	99.99
6	1.23	0.01	100.00
7	0.07	0.00	100.00
8	0.00	0.00	100.00
9	0.00	0.00	100.00
10	0.00	0.00	100.00

Input: CFG file D, L .
Output: subsequence z , feature matrix x.

```
input D;
initialize L, N=length(D), A;
for i=1,2,…,N-L do
    Calculate t_i by formula (1);
    if |t_i| > t_α then
        add i to A;
    end if
end for
combine z in A;
split subsequence;
extract x;
return z, x;
```

Algorithm 1. Subsequence segmentation and feature extraction.

3 Arc ground fault identification algorithm based on segmentation-clustering

3.1 Feature dimensionality reduction based on principal component analysis

The extraction of key feature indicators is an effective method for dimensionality reduction of high-dimensional feature vectors, that is, through data correlation analysis, the original data is converted into effective parameters that are independent of each other and contain the main information. The principal component analysis method uses the knowledge of linear algebra to reduce the dimensionality of the data, and converts multiple variables into a few irrelevant comprehensive variables to more comprehensively reflect the entire data set. The comprehensive variables are called principal components, and the principal components are not correlated with each other, that is, the information they represent does not overlap. This method can effectively reduce the parameter redundancy and improve the efficiency of fault diagnosis (Wang et al., 2015). The steps of principal component analysis are as follows:

1) Input m pieces of n-dimensional data, and form the original data into a matrix $X = \{x_1, x_2, x_3, \ldots, x_n\}$ of n rows and m columns, where x_i is an m-dimensional vector.
2) Zero-means each row of matrix X, that is, subtracts the mean of that row.
3) Calculate the covariance matrix C.

$$C = \begin{pmatrix} Cov(x_1, x_1) & \cdots & Cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ Cov(x_n, x_1) & \cdots & Cov(x_n, x_n) \end{pmatrix} \tag{2}$$

In the formula: $Cov(x_i, y_j)$ represents the covariance of x_i and y_j.

4) Calculate the eigenvalues and eigenvectors of the covariance matrix, sort the eigenvalues from large to small, select the largest N, and then use the corresponding N eigenvectors as row vectors to form the eigenvector matrix P.
5) Transform the data into a new space constructed by N feature vectors, that is, $Y = PX$.

3.2 Cluster analysis model based on K-means

The K-means algorithm is a typical distance-based clustering algorithm, and the distance is used as an evaluation index for similarity, that is, the closer the distance between two samples, the greater the similarity.

First determine the value of k, which means the number of aggregated classes.

Second, randomly select k initial cluster centers. Randomly select k centroid vectors $\{\mu_1, \mu_2, \ldots, \mu_k\}$ from the data set $D = \{x_1, x_2, \ldots x_m\}$, and the coordinates of the centroid vectors are selected by the formula:

$$\begin{cases} C_{[x]} = \left(min_x + range_x * rand(\,)\right) \\ C_{[y]} = \left(min_y + range_y * rand(\,)\right) \\ C_{[z]} = \left(min_z + range_y * rand(\,)\right) \end{cases} \tag{3}$$

In the formula: min_x represents the smallest value in the X coordinate, $range_x$ represents the difference between the maximum value and the minimum value of the X coordinate, and $rand(\,)$ represents a random number between (0,1). Y and Z are the same.

Then assign sample points. Calculate the distance between the sample $x_i\,(i = 1, 2, \ldots m)$ and each centroid vector d_{ij}:

$$d_{ij} = \left\| x_i - \mu_j \right\|_2^2 \tag{4}$$

In the formula: μ_j is the mean vector of the cluster. Assuming that the cluster is divided into $(C_1, C_2, \ldots C_k)$, the x_i is credited to the class C_j with the smallest distance. At this point, the centroid of μ_j is recalculated and updated, and the expression is:

$$\mu_j = \frac{1}{|C_j|} \sum_{x \in C_j} x \tag{5}$$

Repeat the steps of allocating sample points and updating the cluster center until all the sample points are allocated, the category of all the sample points does not change or the number of iterations reaches the specified maximum value, the clustering is stopped. Output the clusters where clustering is done.

In Algorithm 2, the eigenmatrix is input first, then the mean value is removed, and the covariance matrix, eigenvalues, and eigenvectors are calculated. Then reduce the dimension of the original feature matrix. Then initialize the centroid, and when there are still cluster assignment results that change, calculate the distance between the centroid and the sample point, assign the sample point, and update the centroid. Finally, output the dimension reduction matrix, each cluster data and the cluster center.

Input: feature matrix x, k.
Output: dimensionality reduction matrix Y, category y, cluster center μ.

```
input x;
remove the average;
covariance matrix by formula (2);
calculate eigenvalue, eigenvectors;
eigenvector matrix P;
Y=P*x;
initialize k centroids, y;
repeat
for i=1 to n then
    calculate distance by formula (4);
    assign data points;
    for j=1 to k do
        calculate μ by formula (5);
    end for
end for
until y no longer changes
return Y, y, μ;
```

Algorithm 2. Feature dimensionality reduction and clustering.

3.3 A fault identification algorithm based on segmentation and clustering of waveform subsequences

Combined with the above analysis, this paper proposes an arc-ground fault identification algorithm for distribution network based on waveform subsequence segmentation-clustering. The fault identification process is as follows:

1) Step 1: Valid segment data extraction.

After the data is input, the data needs to be preprocessed to extract the data at the moment of failure, that is, the valid segment data.

2) Step 2: Segment and extract the waveform subsequence sequence.

Using the method proposed in Section 1.1, the waveform subsequences are segmented.

3) Step 3: Feature value and feature vector extraction.

The sub-waveform sequence is analyzed in time domain and frequency domain, and the characteristic index proposed in Section 1.2 is calculated and combined into a characteristic vector.

4) Step 4: Feature dimensionality reduction.

For the eigenvalues and eigenvectors extracted in the third step, the dimension is high, and most of the information is redundant, so the principal component analysis method proposed in Section 2.1 is used to reduce the dimension of the data, and the original ten-dimensional data is reduced to three-dimensional.

5) Step 5: K-means clustering.

The clustering algorithm proposed in Section 2.2 is used to perform cluster analysis on the dimensionally reduced 3D data. In this paper, the value of k is selected as 3, that is, the data is clustered into three categories.

6) Step 6: Cluster data output.
7) Step 7: Identify the fault category.

3.4 Arc ground fault safe boundary model

The arc ground fault safety boundary refers to the arc ground fault data: If the fault data falls within the safety boundary, the

TABLE 4 Feature vector table.

Index	Factor load factor		
	Principal component 1 (92.34%)	Principal component 2 (7.22%)	Principal component 3 (0.28%)
Variance	0.93	0.35	−0.06
Kurtosis	−0.01	0.01	0.09
Peak-to-Peak	0.11	0.09	0.20
Mean	−4.55E-4	0.01	0.04
Center frequency	−0.15	0.45	−0.76
Frequency standard deviation	−0.21	0.49	0.60
Rms frequency	−0.26	−0.65	−0.07
5th harmonic	−2.64E-5	−6.71E-5	1.22E-4
5th harmonic	−1.94E-5	−4.72E-5	6.83E-5
7th harmonic	−1.67E-5	−3.88E-5	5.43E-5

fault data can be accurately identified as the fault category; if the fault data falls on the safety boundary or outside the safety boundary, there are the fault data may be identified as other types of faults. And the arc ground fault safety boundary can be effectively distinguished from other types of faults.

The center of the arc ground fault safety boundary is the cluster center of the arc ground fault type data after clustering, and the equatorial radius and polar radius of the safety boundary are shown in Eq. 6.

$$\begin{cases} a = \frac{\sqrt{2}}{2} * (max(x) - min(x)) \\ b = \frac{\sqrt{2}}{2} * (max(y) - min(y)) \\ c = \frac{\sqrt{2}}{2} * (max(z) - min(z)) \end{cases} \tag{6}$$

In the formula, a, b, and c are the radius of one equator and the radius of two poles, respectively, $max(x/y/z)$ is the maximum value of the clustered arc ground fault type data along the x/y/z direction, $min(x/y/z)$ is the minimum value along the x/y/z direction.

In Algorithm 3, the safety boundary model is established by formula Eq. 6 and the arc ground fault clustering center μ through the arc ground fault data.

Input: Arc Ground Fault Data Y, μ.
Output: Security Boundary

1: **input** Y, μ;
2: calculate the polar radius and equatorial radius by formula (6);
3: the center of the circle is μ;
4: **return** Security Boundary;

Algorithm 3. Security Boundary Model.

4 Case study

4.1 Experimental conditions

The experimental data comes from the arc grounding physics experiment platform, which includes power supply subsystem, circuit subsystem, fault simulation subsystem and measurement subsystem. Figure 3A shows the power supply subsystem and the circuit subsystem.

In the training samples selected in the experiment: 540 data of arc ground fault (including high resistance ground fault) and 18 data of ferromagnetic resonance fault; in the test sample: 75 data of arc ground fault (including high resistance ground fault), ferromagnetic resonance two fault data. As shown in Figures 4A–C are several typical arc ground faults, ferromagnetic resonance faults and general ground faults of the test samples, respectively.

4.2 Analysis of waveform feature parameter distribution results

Combined with the waveform feature analysis proposed in Section 1.2, the time domain features and frequency domain features are accumulated to obtain ten eigenvalues, and a large number of features have redundancy, so feature dimension reduction is performed. The variance explanation rate of each feature index obtained by the principal component analysis method is shown in Table 3.

According to the cumulative variance contribution rate shown in Table 1, the cumulative explanation rate of the first three principal components is 99.84%, so the first three principal components can be considered to represent the original variables.

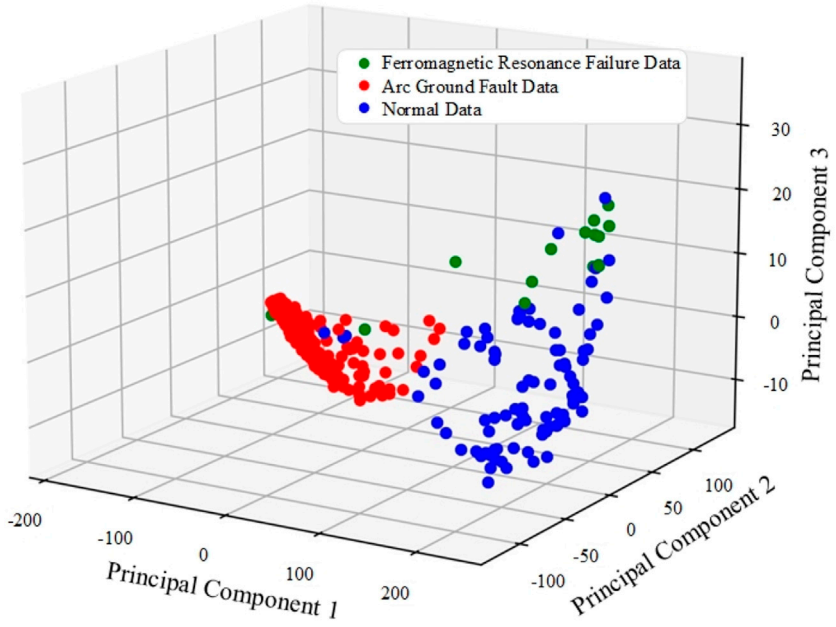


FIGURE 5
Distribution map of arc grounding, ferromagnetic resonance, and normal conditions.

TABLE 5 Distance table between training samples and cluster centers.

Training data	Distance from class 1	Distance from class 2	Distance from class 3	Judgment
1	10.22	112.46	243.55	Class 1
2	9.83	113.51	242.40	Class 1
3	13.02	114.86	242.28	Class 1
4	14.85	116.37	241.49	Class 1
5	130.36	59.65	349.47	Class 2
6	212.92	321.10	35.78	Class 3
7	276.04	386.51	35.85	Class 3
8	202.56	312.47	38.31	Class 3
9	219.14	330.72	29.11	Class 3
10	195.82	308.16	59.08	Class 3

Table 4 is the eigenvector table. The indicators in the table are the standardized variance, kurtosis, peak-to-peak value, mean value, barycentric frequency, frequency standard deviation, root mean square frequency, third harmonic, fifth harmonic, and seventh harmonic. It can be seen from the table that the principal component 1 has a large positive correlation with the variance; the principal component 2 has a large negative correlation with the root mean square frequency, and has a large positive correlation with the frequency standard deviation and the center of gravity frequency; There is a negative correlation with the frequency standard deviation and a large positive correlation with the frequency standard deviation.

The waveform characteristic parameter distribution of the training samples is shown in Figure 5. In the figure, the arc ground fault data is concentrated in the vicinity of (−50, 15, 0), while the ferromagnetic resonance fault data and normal data are scattered in (125, 50, 0), respectively. 15) and (200,-75,-10). Therefore, the data after dimensionality reduction can better describe the arc ground fault, and can effectively distinguish the arc ground fault from the other two faults.

TABLE 6 Distance table between test sample and cluster center

Test data	Distance from class 1	Distance from class 2	Distance from class 3	Judgment
1	42.81	226.33	238.69	Class 1
2	254.80	14.58	79.18	Class 2
3	282.92	26.26	62.33	Class 2
4	275.28	18.04	62.24	Class 2
5	264.57	14.12	78.47	Class 2
6	268.78	12.14	72.73	Class 2
7	283.49	27.20	61.63	Class 2
8	180.39	79.03	95.43	Class 2
9	201.62	67.54	64.52	Class 3
10	243.36	88.53	31.89	Class 3

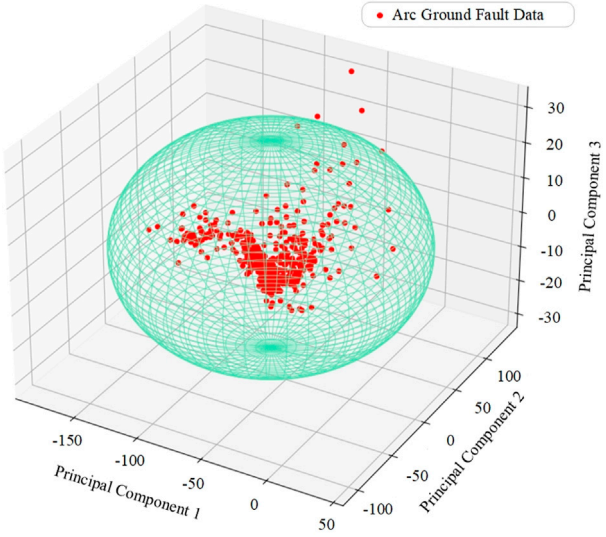


FIGURE 6
Arc grounding safety boundary.

4.3 Model identification verification

1) After the model is trained with training samples, the distance from each sample to each cluster center is shown in Table 5, and the last column is the fault type discrimination. Type 1 is arc ground fault, type 2 is ferromagnetic resonance fault, and type 3 is normal condition. Calculate the distance between the training data and each cluster center, and divide the data into the closest classes.
2) After the test sample data is identified by the model, the distance and attribution type of each test sample from each cluster center are shown in Table 6. It can be concluded from the table that the sample points are always closer to one of the cluster centers, and farther away from the other two types of sample centers. Type 1 is arc ground fault, type 2 is normal, and type 3 is ferromagnetic resonance fault.
3) Combined with the safety boundary proposed by Eq. 6, the equatorial radius and polar radius of the safety boundary are calculated. After calculation, the equatorial radius of the safety boundary is 111.0 and 119.5, and the polar radius is 29.7. The cyan spherical area in Figure 6 is the safety boundary of arc ground fault, and the red sample points are the data classified as arc ground fault after clustering. It can be seen from the figure that most of the arc ground fault data falls within the safety boundary, that is, the safety boundary can more accurately distinguish arc ground faults from other faults.
4) For the arc grounding system of the distribution network, the identification of high resistance grounding faults is a difficult point (Kavaskar and Mohanty, 2019). The arc high-resistance fault has obvious intermittent, the phase voltage is basically unchanged, the zero-off time is long, and it lasts for several cycles intermittently. At the same time, the zero-sequence voltage has nonlinear distortion (Zhang et al., 2021). The identification of single-phase ground fault is mainly realized by detecting the zero-sequence voltage. When the zero-sequence voltage suddenly increases, it is judged that a ground fault occurs.

For arc high-resistance grounding faults, the method proposed in this paper is used to segment the fault waveform sub-sequence, extract features, and reduce the dimension to obtain the principal component components of four types of samples after dimension reduction as shown in Table 7. Among them, the high-resistance grounding samples And the low resistance ground samples are within the safety boundary, while the normal case and ferromagnetic resonance fault samples are outside the safety boundary. Combining with Figure 7, it is obvious that the method proposed in this paper can also accurately identify the arc high-resistance ground fault.

TABLE 7 Principal component components after dimension reduction for four types of samples.

Type	Principal component 1	Principal component 2	Principal component 3	Relative position
Low resistance ground fault	−45.08	101.58	−17.82	Inside
High resistance ground fault	20.62	−9.30	−1.81	Inside
Ferromagnetic resonance	−60.90	59.21	73.66	Outside
Normal	−249.37	9.80	−40.94	Outside

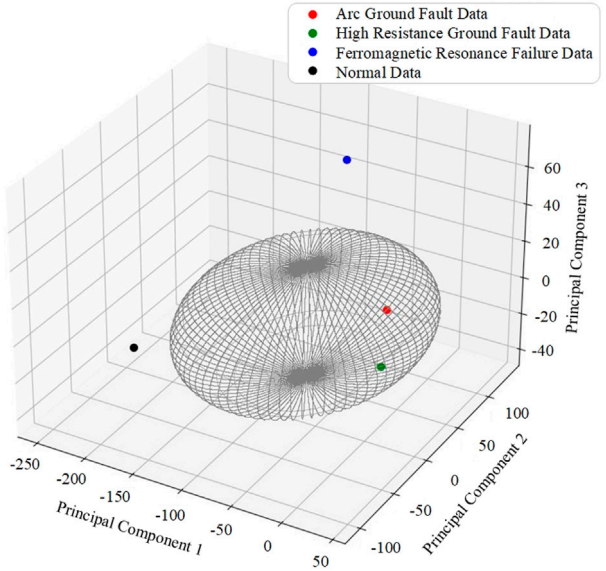


FIGURE 7
The relative positions of the four types of sample data and the security boundary.

4.4 Algorithm comparison

4.4.1 Segmentation accuracy comparison

Through the above analysis, the algorithm proposed in this paper is compared with the wavelet transform algorithm and the variational mode decomposition algorithm. The waveform subsequence segmentation algorithm proposed in this paper only needs to determine the sliding step size L and the 95% significant interval range when selecting parameters. Since the fault recorder data has strong periodicity, and the period is 80, L = 80 is selected. For the selection of L, the detection effects corresponding to different L values are shown in Figure 8. The figure shows the detection effect of the sliding *t*-test when L = 40, 80, and 120, respectively. When L = 40 and L = 120, the entire recorded wave data is judged as fault, when L = 80, the fault data can be detected accurately.

Wavelet transform can decompose the signal into a series of signal sub-sequences, which has the characteristics of multi-resolution analysis. In practical applications, discrete wavelet transform with less computational complexity and higher accuracy is often used. The Symlet wavelet function is an approximately symmetrical wavelet function after the improvement of the db function. The support range of the symN wavelet is 2N-1, the vanishing moment is N, and it also has good regularity. Compared with the dbN wavelet, the wavelet is consistent with the dbN wavelet in terms of continuity, support length, filter length, etc., but the symN wavelet has better symmetry, that is, it can reduce the time to analyze and reconstruct the signal to a certain extent. Phase distortion. For the fault recorder data, this paper selects the wavelet function sym8 as the fundamental wave function, and the number of decomposition layers is 9. The decomposition results are shown in Figure 9. The figure is the approximate coefficient CA9 and the detail coefficient CD9-CD1 after wavelet transformation. The threshold detection is performed on the five detail coefficients CD1 and CD2 with obvious fault characteristics. As the threshold increases, the arc ground fault detection is more obvious, but only the start time of the fault can be found. When the waveform subsequence is divided, the end time of the fault needs to be set as the start time of the next fault. In CD7, the fault detection is more accurate, but the arc ground fault and ferromagnetic resonance fault between the sampling points (60, 80) cannot be accurately divided, and it needs to be handled manually. If the threshold is increased, some arc ground faults cannot be accurate detection.

The variational modal decomposition method is an adaptive, completely non-recursive modal variation and signal processing method, which is suitable for non-stationary sequences, and decomposes to obtain relatively stable subsequences containing multiple different frequency scales. The VMD algorithm decomposes the original non-stationary signal f into k relatively stationary sub-signals with different center frequencies w_k and priority bandwidths. Each sub-signal, as a modal component of the original signal, can reflect the original signal at different time scales Structure. As shown in Figure 10, IMF2 and IMF3 are more accurate in detecting faults. Although IMF3 can detect the fault segment, it is greatly affected by the threshold value. If the threshold value increases, although the arc ground fault can be distinguished from the ferromagnetic resonance fault, But the fault end-point detection of Ferromagnetic resonance faults becomes inaccurate. No matter whether the threshold of IMF4 is increased or decreased, the fault point cannot be accurately detected. For

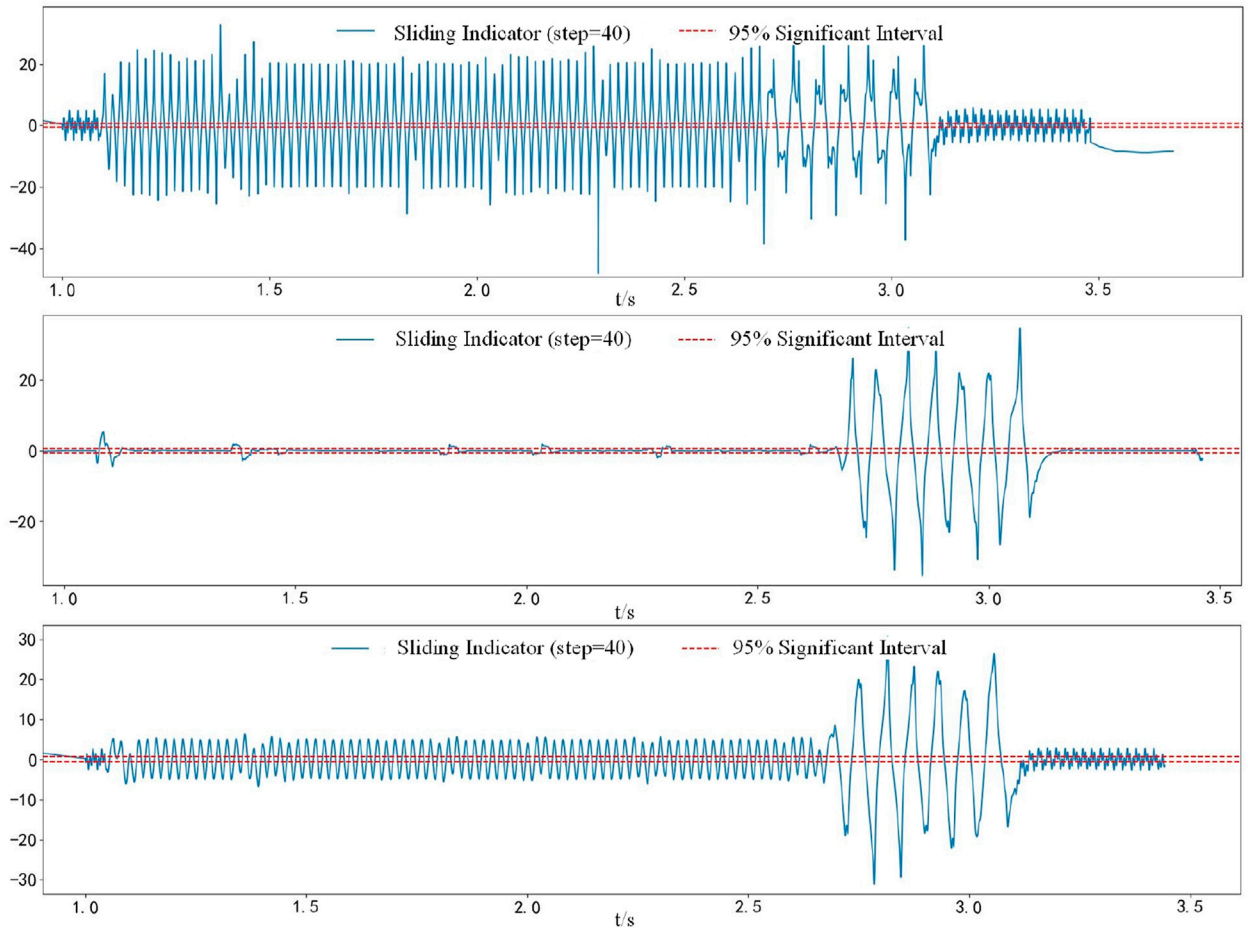


FIGURE 8
Detection results of different sliding steps.

the fault detection of IMF5-IMF7, although the fault location can be detected, the fault detection cannot be completed, and the detection of arc ground fault is invalid.

Table 8 compares the detection and segmentation of faults by the method proposed in this paper, wavelet transform and variational modal decomposition for different experimental samples. Variational modal decomposition can only detect ferromagnetic resonance faults, and wavelet transform cannot accurately detect general ground faults. The main reason is that the end point of arc ground fault cannot be accurately detected during wavelet transform detection. Arc ground faults start as an end point, resulting in general ground faults being divided into arc ground faults.

4.4.2 Comparison of recognition accuracy

After the fault data is segmented and eigenvalue extracted, the fault type needs to be identified. Comparing the algorithm proposed in this paper with WT-CNN and VMD-SVM, the CNN model structure consists of input layer, convolution layer, pooling layer, activation function layer, fully connected layer and output layer group layer. For the processed data, the output layer is the final result, and the convolution layer, pooling layer and activation function layer together form a hidden layer of CNN. The structure and parameters of the CNN models introduced in this paper for comparison are shown in Table 9.

SVM is a shared supervised learning method suitable for small sample, nonlinear and high-dimensional data. For a given collinear classifiable training dataset, a kernel function is used to map the data from the original feature space to a high-dimensional feature space, so that the linear inner product is nonlinear, and then the classification interval is maximized in the high-dimensional feature space. Optimal hyperplane. Penalty factor C and RBF kernel function parameters are two important parameters in SVM. Penalty factor $C > 0$, the larger C is, the greater the penalty for misclassification, but overfitting is easy; the smaller C is, the less the penalty for misclassification is, the complexity of

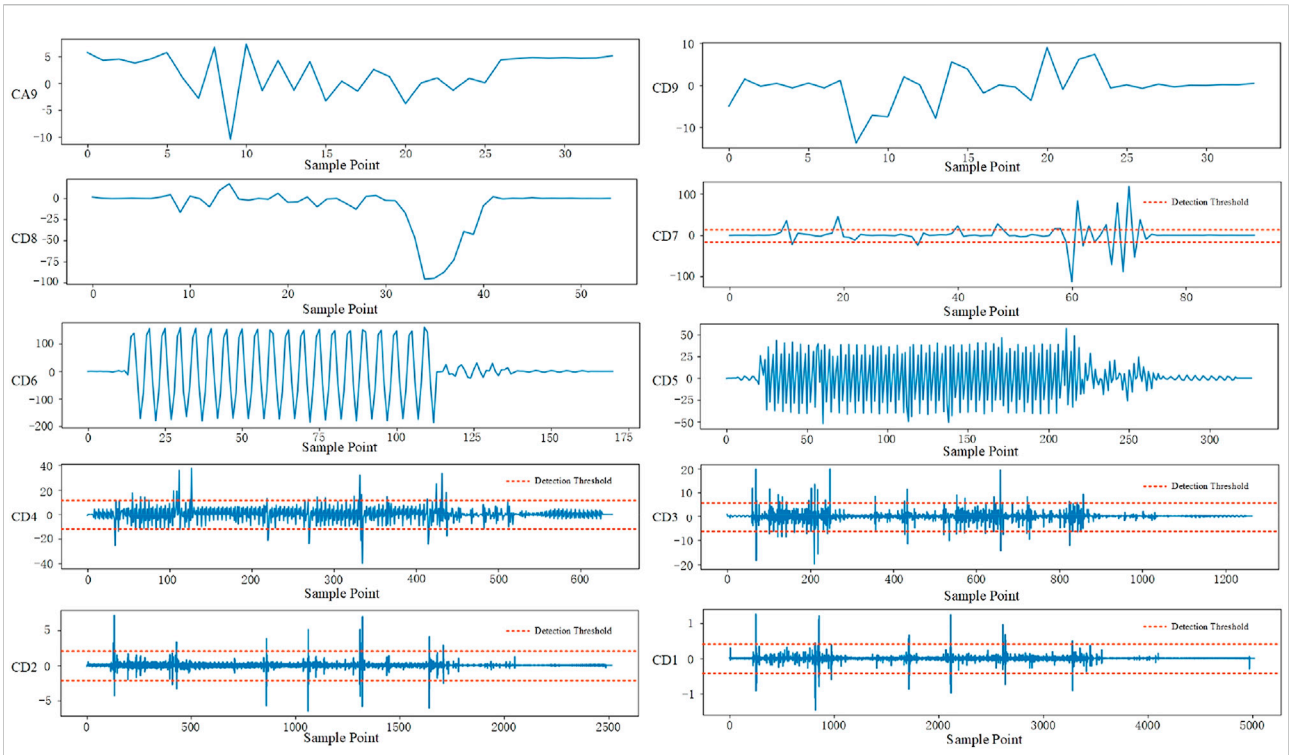


FIGURE 9
Wavelet analysis components of each layer.

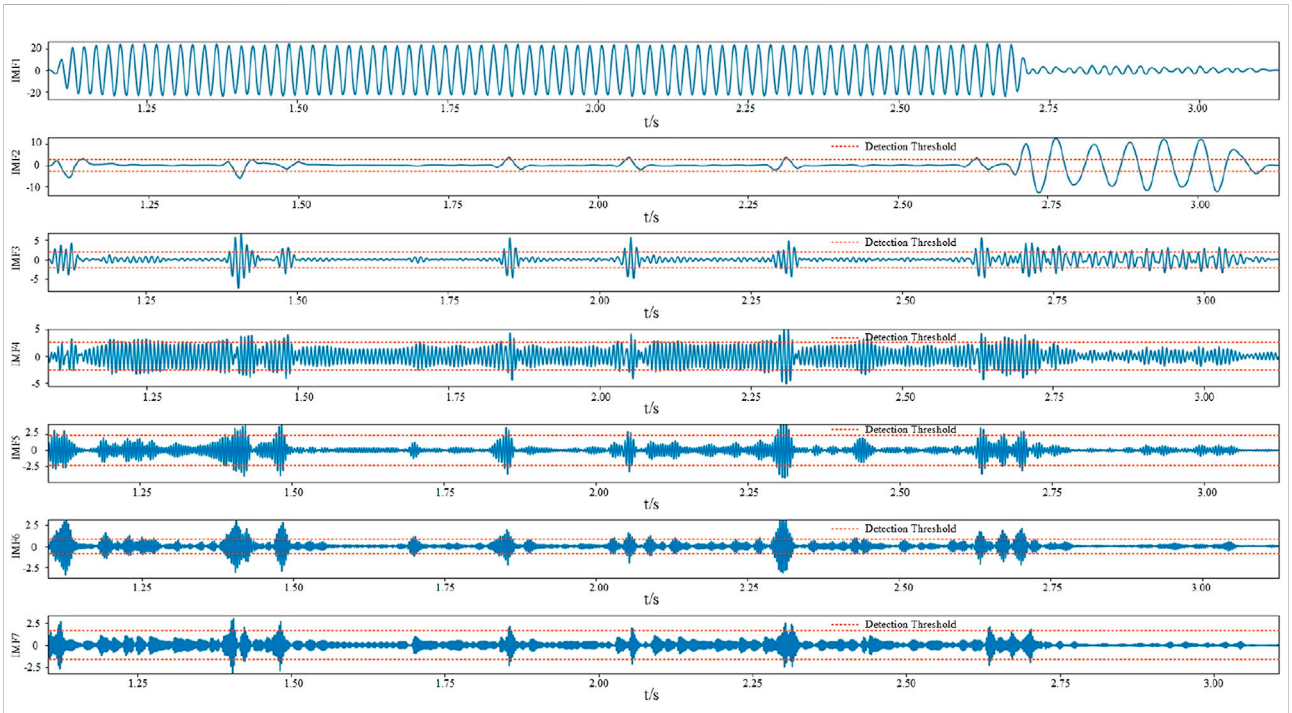


FIGURE 10
Variational modal decomposition of components at each layer.

TABLE 8 Comparison of fault segmentation accuracy of different methods.

Sample number	Actual number of failures (ferromagnetic resonance/arc light/general ground)	Ours	WT	VMD
Sample 1	1/6/5	1/6/5	0/6/1	0/6/5
Sample 2	1/4/3	1/4/3	1/3/2	0/4/3
Sample 3	0/2/1	0/2/1	0/2/1	0/2/1
Sample 4	0/4/3	0/4/3	0/4/1	0/4/3
Sample 5	0/1/1	0/1/1	0/1/1	0/1/1
Segmentation accuracy	—	100%	71.875%	93.75%

TABLE 9 CNN model structure and parameters.

Layers	Structural layer	Parameter
1	Input layer	—
2	Convolutional layer 1	3 × 3,8
3	Pooling layer 1	2 × 2
4	Convolutional layer 2	3 × 3,16
5	Pooling layer 2	2 × 2
6	Fully connected layer	3 nodes
7	Output layer	3 classes

TABLE 10 Identification performance of different algorithm models.

Algorithm model	Recognition accuracy (%)
WT-CNN	93.81
VMD-SVM	94.74
Ours	97.12

the model is reduced, and underfitting is prone to occur. γ determines the distribution of the data mapped to the new feature space. The smaller γ, the more support vectors, the greater the smoothing effect of the model, and the easier it is to underfit; easy to overfit.

For the problem that the arc ground fault sub-sequence cannot be accurately segmented in the methods WT-CNN and VMD-SVM, the fault merging method proposed in this paper is used to determine the fault end point. For ferromagnetic resonance faults, the waveform sub-sequence is manually segmented segmentation. It can be seen from Table 10 that the identification accuracy rates of several algorithms listed in the table are all greater than 90%. The recognition accuracy of the algorithm model proposed in this paper is 97.12%, and the recognition accuracy ratio is 3.31% and 2.38% higher than that of WT-CNN and VMD-SVM, respectively.

5 Conclusion

Aiming at the problems of inaccurate fault detection and redundant feature extraction in traditional detection based on electrical parameters and thresholds, this paper proposes a segmentation-clustering-based arc-ground fault identification method for distribution networks. First, a sliding *t*-test was used to segment the waveform subsequences, considering the presence of developing faults in the recorder data. Secondly, extract eigenvalues in time domain and frequency domain for the segmented waveform subsequences, and reduce the dimension of the eigenmatrix by using principal component analysis method. Then, cluster analysis is carried out on the characteristic parameter distribution after dimension reduction, and the identification accuracy of the algorithm is verified by using the safety boundary model. Finally, compared with the traditional arc ground fault identification method, the following conclusions are drawn:

(1) Compared with the traditional arc ground fault identification method, the segmentation-clustering algorithm proposed in this paper can more accurately segment the fault. The influence of different types of fault data on waveform characteristic values is reduced.
(2) The combined three-phase voltage and zero-sequence voltage waveform eigenvalue extraction and principal component analysis dimensionality reduction established in this paper reduce the problem of insufficient feature extraction based on traditional electrical parameter analysis, and reduce the redundancy of feature data.
(3) Compared with WT-CNN and VMD-SVM, the identification accuracy of the identification method proposed in this paper is improved by 3.31% and 2.38%, respectively.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

GL, MC, and WW provide data and technical support, QL and LC provide research methods and experimental guidance, and YW conduct experimental research.

Funding

This work was supported by China Power Construction Group Jiangxi Electric Power Construction Co., Ltd., Jiangxi Provincial Department of Science and Technology Project "Development of Wind Farm Dynamic Operation and Maintenance System Based on AR Smart Glasses", and national Natural Science Foundation project "Research on Multi-classification and Efficiency Loss Prediction of Photovoltaic Module Defects for Unbalanced Distribution." The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Conflict of interest

Authors GL and MC were employed by National Network Jiangxi Electric Power Co. Ltd. Author WW was employed by 3China Power Construction Group Jiangxi Electric Power Construction Co. Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Cai, X., and Wai, R. J. (2022). Intelligent DC arc-fault detection of solar PV power generation system via optimized VMD-based signal processing and PSO–SVM classifier. *IEEE J. Photovolt.* 12 (4), 1058–1077. doi:10.1109/JPHOTOV.2022.3166919

Chen, J., Li, H., Deng, C., and Wang, G. (2021). Detection of single-phase to ground faults in low-resistance grounded MV systems. *IEEE Trans. Power Deliv.* 36 (3), 1499–1508. doi:10.1109/TPWRD.2020.3010165

Dang, H. L., Kwak, S., and Choi, S. (2022). Parallel DC arc failure detecting methods based on artificial intelligent techniques. *IEEE Access* 10, 26058–26067. doi:10.1109/ACCESS.2022.3157298

Du, R., Shang, F., and Ma, N. (2019). Automatic mutation feature identification from well logging curves based on sliding t test algorithm. *Clust. Comput.* 22, 14193–14200. doi:10.1007/s10586-018-2267-z

Du, Y., Liu, Y., Shao, Q., Luo, L., Dai, J., Sheng, G., et al. (2019). Single line-to-ground faulted line detection of distribution systems with resonant grounding based on feature fusion framework. *IEEE Trans. Power Deliv.* 34 (4), 1766–1775. doi:10.1109/tpwrd.2019.2922480

Gadanayak, D., and Mallick, R. (2019). Interharmonics based high impedance fault detection in distribution systems using maximum overlap wavelet packet transform and a modified empirical mode decomposition. *Int. J. Electr. Power & Energy Syst.* 112, 282–293. doi:10.1016/j.ijepes.2019.04.050

Guo, M. F., Yang, N. C., and Chen, W. F. (2019). Deep-Learning-based fault classification using hilbert-huang transform and convolutional neural network in power distribution systems. *IEEE Sens. J.* 19, 6905–6913. doi:10.1109/jsen.2019.2913006

Guo, M. F., Zeng, X. D., Chen, D. Y., and Yang, N. C. (2018). Deep -Learning-Based earth fault detection using continuous wavelet transform and convolutional neural network in resonant grounding distribution systems. *IEEE Sens. J.* 18, 1291–1300. doi:10.1109/jsen.2017.2776238

Kavaskar, S., and Mohanty, N. K. (2019). Detection of high impedance fault in distribution networks. *Ain Shams Eng. J.* 101, 5–13. doi:10.1016/j.asej.2018.04.006

Lin, C., Gao, W., and Guo, M. F. (2019). Discrete wavelet transform-based triggering method for single-phase earth fault in power distribution systems. *IEEE Trans. Power Deliv.* 34 (5), 2058–2068, Oct. doi:10.1109/TPWRD.2019.2913728

Mishra, M, Routray, P., and Rout, P. K. (2016). A universal high impedance fault detection technique for distribution system using S-transform and pattern recognition. *Technol. Econ. Smart Grids Sustain. Energy* 1, 9. doi:10.1007/s40866-016-0011-4

Paul, D. (2015). High -resistance grounded power system. *IEEE Trans. Ind. Appl.* 51, 5261–5269. doi:10.1109/tia.2015.2422825

Peng, N., Ye, K., Liang, R., Hou, T., Wang, G., Chen, X., et al. (2019). Single-Phase-to-Earth faulty feeder detection in power distribution network based on amplitude ratio of zero-mode transients. *IEEE Access* 7, 117678–117691. doi:10.1109/ACCESS.2019.2936420

Qin, X., Wang, P., Liu, Y., Guo, L., Sheng, G., and Jiang, X. (2018). Research on distribution network fault recognition method based on time-frequency characteristics of fault waveforms. *IEEE Access* 6, 7291–7300. doi:10.1109/ACCESS.2017.2728015

Siegel, J. E., Pratt, S., Sun, Y., and Sarma, S. E. (2018). Real-time deep neural networks for internet-enabled arc-fault detection. *Eng. Appl. Artif. Intell.* 74, 35–42. doi:10.1016/j.engappai.2018.05.009

Wang, L., Qiu, H., and Yang, P. (2021). Arc Fault detection algorithm based on variational mode decomposition and improved multi-scale fuzzy entropy. *Energies* 14, 4137. doi:10.3390/en14144137

Wang, Y., Zhou, J., Li, Z., Dong, Z., and Xu, Y. (2015). Discriminant-analysis-based single-phase earth fault protection using improved PCA in distribution systems. *IEEE Trans. Power Deliv.* 30 (4), 1974–1982. doi:10.1109/TPWRD.2015.2408814

Wei, M., Shi, F., Zhang, H., Jin, Z., Bao, H., Zhou, J., et al. (2020). High impedance arc fault detection based on the harmonic randomness and waveform distortion in

the distribution system. *IEEE Trans. Power Deliv.* 35 (2), 837–850. doi:10.1109/TPWRD.2019.2929329

Wei, Z., Mao, Y., Yin, Z., Sun, G., and Zang, H. (2020). fault detection based on the generalized S-transform with a variable factor for resonant grounding distribution networks. *IEEE Access* 8, 91351–91367. doi:10.1109/ACCESS.2020.2994139

Xia, K., He, S., Tan, Y., Jiang, Q., Xu, J., and Yu, W. (2019). Wavelet packet and support vector machine analysis of series DC ARC fault detection in photovoltaic system. *IEEJ Trans. Elec. Electron. Eng.* 14, 192–200. doi:10.1002/tee.22797

Zeng, X., Yu, K., Wang, Y., and Xu, Y. (2016). A novel single phase grounding fault protection scheme without threshold setting for neutral ineffectively earthed power systems. *CSEE J. Power Energy Syst.* 2 (3), 73–81. doi:10.17775/CSEEJPES.2016.00038

Zhang, H., Wang, J., Liu, Z., Han, J., Liu, J., and Zhu, H. "Comparative analysis of ferroresonance and arc high impedance grounding fault in high voltage transmission line," in Proceedings of the 2021 International Conference on Power System Technology (POWERCON), Haikou, China, December 2021, 2331–2335. doi:10.1109/POWERCON53785.2021.9697581

Zhang, L., Wang, Y., Yan, H., and Shi, F. "Single-phase-to-ground fault diagnosis based on waveform feature extraction and matrix analysis," in Proceedings of the 2019 9th International Conference on Power and Energy Systems (ICPES), Perth, Australia, December 2019, 1–6. doi:10.1109/ICPES47639.2019.9105483