

Article

Research on a Method for Identifying Key Fault Information in Substations

Pan Zhang ¹, Lei Guo ¹, Zhicheng Huang ¹, Zhoupeng Rao ¹, Ying Zhang ¹, Zhi Sun ¹, Rui Xu ^{2,*} and Deng Li ³ 

¹ State Grid Hubei Electric Power Co., Ltd., Xiaogan Power Supply Company, Xiaogan 432000, China; zhangp8@hb.sgcc.com.cn (P.Z.); guol4@hb.sgcc.com.cn (L.G.); huzc13@hb.sgcc.com.cn (Z.H.); raozp@hb.sgcc.com.cn (Z.R.); zhangy46@hb.sgcc.com.cn (Y.Z.); sunz7@hb.sgcc.com.cn (Z.S.)

² School of Computer Science and Engineering, Central South University, Changsha 410083, China

³ School of Electronics Information, Central South University, Changsha 410083, China; d.li@csu.edu.cn

* Correspondence: 234712184@csu.edu.cn

Abstract: The identification of critical fault information plays a crucial role in ensuring the reliability and stability of power systems. However, existing fault-identification technologies heavily rely on high-dimensional sensor data, which often contain redundant and noisy information. Moreover, conventional data preprocessing approaches typically employ fixed time windows, neglecting variations in fault characteristics under different system states. This limitation may lead to incomplete feature selection and ineffective dimensionality reduction, ultimately affecting the accuracy of fault classification. To address these challenges, this study proposes a method of critical fault information identification that integrates a scalable time window with Principal Component Analysis (PCA). The proposed method dynamically adjusts the time window size based on real-time system conditions, ensuring more flexible data capture under diverse fault scenarios. Simultaneously, PCA is employed to reduce dimensionality, extract representative features, and remove redundant noise, thereby enhancing the quality of the extracted fault information. Furthermore, this approach lays a solid foundation for the subsequent application of deep learning-based fault-diagnosis techniques. By improving feature extraction and reducing computational complexity, the proposed method effectively alleviates the workload of operation and maintenance personnel while enhancing fault classification accuracy. Our experimental results demonstrate that the proposed method significantly improves the precision and robustness of fault identification in power systems.

Keywords: critical fault information identification; principal component analysis; scalable time window; feature extraction; transformer substation; power system reliability



Academic Editor: Andry Sedelnikov

Received: 25 March 2025

Revised: 22 April 2025

Accepted: 28 April 2025

Published: 6 May 2025

Citation: Zhang, P.; Guo, L.; Huang, Z.; Rao, Z.; Zhang, Y.; Xu, R.; Rui, X.; Li, D. Research on a Method for Identifying Key Fault Information in Substations. *Computation* **2025**, *13*, 109. <https://doi.org/10.3390/computation13050109>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

High-Voltage Direct Current (HVDC) technology has driven major advances in long-distance, high-capacity power transmission in China, offering lower line and converter station costs, higher throughput, reduced losses, asynchronous interconnection and precise short-circuit current control compared with traditional High-Voltage Alternating Current (HVAC) systems. HVDC substations therefore encompass extensive primary- and secondary-side equipment, generating dense streams of time-series alarms and status signals (currents, voltages, breaker trip indications, etc.) whose volume and complexity present significant fault-diagnosis challenges.

These challenges motivate the development of diagnostic methods that operate solely on such alarm and status data—without invoking any DC- or AC-specific power flow

models—and which are, thus, equally applicable to HVAC substations. To assist operation and maintenance personnel in accurately determining and locating faults, researchers have proposed various methods for identifying key fault information in intelligent substations by extracting representative features [1–4]. However, these existing approaches face notable challenges. Ke et al. [1] developed a fault diagnosis model based on cooperative game theory and cloud models, establishing an index system for key fault factors. However, the entropy-weighting approach used in this method suffers from limitations in weight assignment, reducing its accuracy. Ding et al. [2] introduced an expert system-based fault-diagnosis approach that stores expert knowledge in a structured knowledge base. While effective in small-scale applications, its efficiency diminishes when applied to large-scale power systems. Zhan et al. [3] integrated Dempster–Shafer evidence theory with fuzzy logic, utilizing historical fault data for decision making. However, the method encounters difficulties in constructing fault data range tables, defining fuzzy membership functions, and optimizing multi-source information fusion. Dai et al. [4] proposed a big data-driven approach to identifying multi-source transmission line tripping faults. While promising, this method requires extensive fault current data, which is often difficult to obtain in real-world power systems.

Furthermore, existing fault-diagnosis methods often generate an overwhelming amount of complex and redundant information, complicating the identification of critical fault-related data. The inability to effectively filter and prioritize key information hinders targeted fault resolution [5]. To address these challenges, this study proposes an enhanced method of critical fault information identification for substations. The method incorporates a scalable time window adjustment for dynamic data preprocessing, ensuring that fault information is captured based on real-time system conditions. Additionally, PCA is utilized to extract key features, reduce data dimensionality, and eliminate redundant information. By integrating these approaches, the proposed method aims to enhance fault classification accuracy and efficiency, thereby reducing the operational burden on maintenance personnel and improving the overall reliability of power systems.

The main contributions are summarized as follows:

1. A hybrid method for identifying key fault information in substations is proposed, combining a scalable time window adjustment mechanism with Principal Component Analysis (PCA). The time window adapts to the alarm density observed in real time, enabling data capture under varying fault conditions.
2. The method includes a structured alarm event grouping and timestamp alignment process, which allows temporally overlapping events to be clustered and synchronized based on correlation, improving the temporal consistency of the input data.
3. Validation on an actual substation alarm dataset showed that retaining 10 principal components can retain 90% of the total variance, which demonstrated its effective dimensionality reduction and noise filtering in a data redundancy environment while maintaining interpretability and computational efficiency.

The remainder of this paper is organized as follows: Section 2 provides a brief overview of related research on fault-diagnosis methods. Section 3 outlines the proposed method for key fault information identification. Section 4 presents experiments conducted on the collected dataset to validate the effectiveness of the approach. Finally, Section 5 provides the conclusion and suggestions for future work

2. Related Works

Technology for identification of key fault information plays a crucial role in predictive maintenance and fault diagnosis by monitoring and analyzing system data, to detect and isolate potential failures. The primary objective of this technology is to extract relevant

features from fault phenomena, enabling both quantitative and qualitative analysis, to establish a structured fault feature library. This interdisciplinary field integrates concepts from electronic engineering, statistical mathematics, modern control theory, signal processing, pattern recognition, computer science, and artificial intelligence, to develop efficient fault identification methods.

Existing approaches to identification of key fault information can be broadly categorized into three types: mathematical model-based methods, empirical knowledge-based methods, and data-driven methods [6], as illustrated in Figure 1. Each of these methodologies has distinct advantages and limitations, in terms of accuracy, computational complexity, and data dependency.

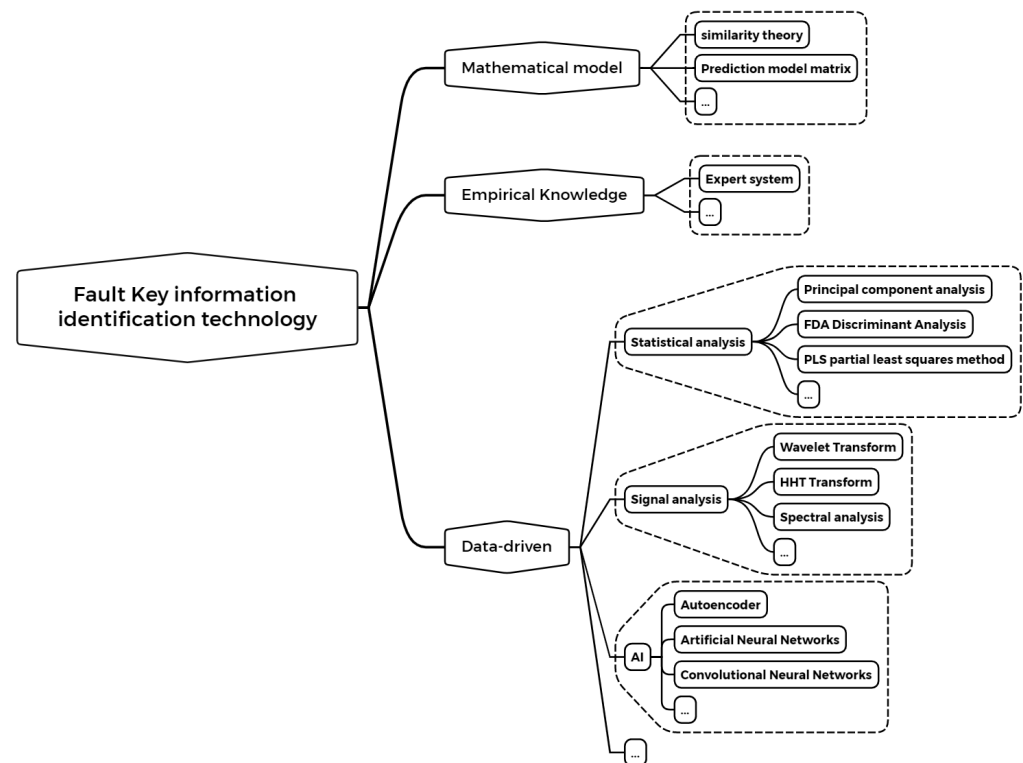


Figure 1. Technology for identification of key fault information.

2.1. Mathematical Model-Based Methods

Mathematical model-based methods rely on precisely defined mathematical representations of power system components, enabling the characterization of parameter variations and theoretical threshold ranges under different operating conditions. These models provide high recognition accuracy and strong real-time performance, as they establish explicit relationships between system states and abnormal conditions. By leveraging these models, operation and maintenance personnel can diagnose real-time system conditions with greater confidence. However, the major limitation of this approach is the substantial effort required to construct accurate models, which demands extensive domain expertise, prior knowledge, and computational resources. Additionally, real-world power systems are highly dynamic and complex, making it difficult to maintain up-to-date models that reflect evolving operational characteristics. Early studies primarily focused on this method due to its interpretability and reliability [7].

Li et al. [8] proposed a fault-identification method based on a Multi-resolution Morphological Gradient (MMG) for distribution networks with Doubly Fed Induction Generators (DFIGs). By constructing a discriminant criterion using gradient peak ratios, the method achieves precise identification of fault types and phases. In scenarios with nonlinear cur-

rent distortion caused by DFIGs, it demonstrates a 100% fault-detection rate and 99.75% recognition accuracy.

Zhang et al. [9] developed an online diagnosis method for synchronous compensators, using Variable Predictive Model-based Class Discrimination (VPMCD). This approach uses the minimum sum of squared prediction errors to distinguish between fault currents and inrush currents, addressing the limitations of traditional methods in handling small samples and high-dimensional data. The overall recognition accuracy exceeds 82%, outperforming Support Vector Machines (SVMs).

2.2. Empirical Knowledge-Based Methods

Empirical knowledge-based methods primarily rely on expert systems, where fault diagnosis rules and decision making models are established, based on the experience, knowledge, and heuristics of domain experts. These systems infer equipment operating conditions by formalizing human expertise into structured rules, allowing for fault detection without requiring complex mathematical models or large-scale historical data. The key advantages of this approach include ease of implementation and strong interpretability, making it particularly suitable for environments where expert domain knowledge is readily available. However, the method suffers from maintenance challenges, as updating and refining rule-based systems is labor-intensive. Furthermore, its performance is heavily dependent on the expertise and availability of domain specialists, limiting scalability, and accuracy compared to automated methods. This approach has been widely used in early fault diagnosis studies [10,11].

2.3. Data-Driven Methods

With advancements in big data analytics and machine learning, data-driven methods have emerged as powerful alternatives for fault diagnosis. These approaches leverage historical fault data to extract hidden failure patterns and relationships, enabling automated fault monitoring and prediction. Unlike traditional model-based or expert-driven methods, data-driven techniques do not require explicit knowledge of the physical system structure. Instead, they rely on statistical analysis, signal processing, and artificial intelligence to learn fault characteristics from available data. The primary advantages of this approach include high adaptability, automation capability, and scalability, making it well-suited for large-scale, complex power systems. However, the effectiveness of data-driven methods is heavily dependent on the availability and quality of historical data. Inaccurate, incomplete, or biased datasets can significantly impact model performance. Many recent studies have explored machine learning and deep learning techniques for fault identification, highlighting the growing importance of this approach in modern power systems [12–17].

For example, Wang et al. [17] proposed a fault-diagnosis method based on an artificial neural network, which was verified by a 110kv substation example with an accuracy of 97.68%, which is better than many traditional and intelligent methods.

In summary, the aforementioned methods still face challenges in practical substation scenarios. Model-based approaches often require complex modeling and are difficult to generalize to modern digital substations. Expert systems, while interpretable, lack scalability and adaptability. Although data-driven methods have shown promise, they typically rely on static data preprocessing steps such as fixed time windows, which may fail to capture dynamic fault behaviors. Moreover, most existing methods overlook the need to filter redundant alarm signals before applying dimensionality reduction. To address these gaps, our method integrates a scalable time window adjustment mechanism with PCA-based feature extraction, aiming to capture fault-specific data sequences dynamically and

reduce dimensionality with interpretability. This hybrid approach enables more accurate identification of key fault features in various fault scenarios in substations.

3. Scalable Time Window–PCA Key Fault Information Identification

This section presents the proposed method for identification of key fault information about substations; it is designed to enhance the accuracy of fault identification by integrating scalable time window correction and PCA-based feature extraction. This approach provides a structured preprocessing framework that supports subsequent deep learning-based fault discrimination analysis. The overall technical architecture is illustrated in Figure 2.

As shown in the figure, the proposed fault identification framework consists of two primary components: data preprocessing and feature extraction. The data sources include switching states, protection actions, and time-sequenced information collected from substations, as well as status data from primary and secondary equipment:

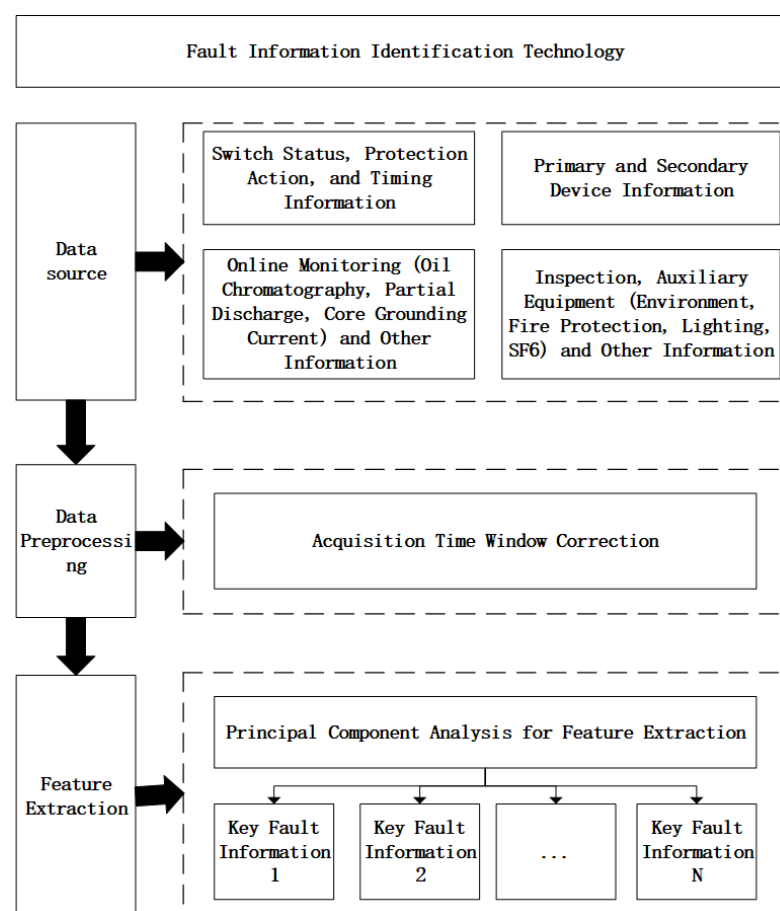


Figure 2. Technical architecture diagram for the automatic identification of key fault information.

3.1. Data Preprocessing—Scalable Time Window Correction

In the context of substation key fault information identification, the collected fault data primarily consist of one-dimensional time series signals. Consequently, effective feature extraction and noise reduction are crucial for optimizing the accuracy and reliability of fault identification. To achieve this, the proposed method adopts a scalable time window correction approach. The data preprocessing phase consists of three major steps.

3.1.1. Scalable Time Window Adjustment

Many fault diagnosis studies adopt a fixed-length acquisition window, T_W , as proposed in [14], which may introduce identification errors when the time window is too

large or too small. To mitigate this issue, this study introduces a scalable time window, T_F , dynamically adjusted based on the fault alarm density. The window size is defined as

$$T_F = T_W + T_f \quad (1)$$

where T_W represents the minimum fixed time window (default: 10 s [18]), and where T_f is the floating acquisition time. The adjustment mechanism follows this principle:

If the rate of alarm occurrences $\Delta alarm / \Delta t$ within a specific time interval Δt exceeds a predefined threshold (φ) then the floating time window Δt expands dynamically until the alarm rate falls below (φ). Specifically, $\Delta alarm / \Delta t$ is calculated by counting the number of distinct alarm events that occur within a sliding window of 1 s. If multiple alarms have identical timestamps (i.e., overlap) then each is counted separately unless they are from the same equipment and of the same type, in which case they are treated as a single logical event, to prevent redundancy. This count is updated continuously as the time window moves forward, providing a real-time estimate of the current alarm density.

$\Delta alarm / \Delta t$ will not increase indefinitely. If it reaches T_{max} , which is preset by the operator, it will stop expanding.

The scalable time window correction mechanism is illustrated in Figure 3:

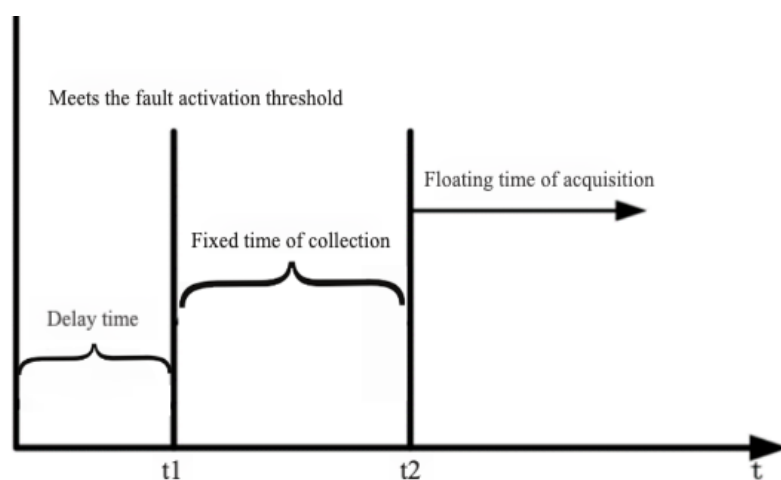


Figure 3. Scalable time window. Here t_1 represents the moment the fault-initiation threshold is met. T_W is the fixed acquisition time between t_1 and t_2 . T_f denotes the floating time extension beyond t_2 .

We performed a comprehensive sensitivity analysis by sweeping φ from 0.000 to 1.000 and measuring (a) the number of STW windows, (b) event-grouping integrity, and (c) overall STW + PCA feature fidelity. The key results are summarized in Figure 4 below:

For $\varphi \leq 0.556$, we retained 100% event integrity (all alarm clusters fell entirely within single windows). Feature fidelity (total variance retained by PCA) peaked at 99.762% when $\varphi = 0.556$ and declined only marginally beyond that; $\varphi > 0.556$ caused over-splitting (integrity dropped to 80% or less).

Based on this analysis, we set $\varphi = 0.556$ alarms/s in all subsequent experiments, to achieve the optimal balance between complete alarm coverage and maximal feature fidelity.

This dynamic adjustment ensures optimal window selection, enhancing the accuracy of event grouping and fault sequence reconstruction.

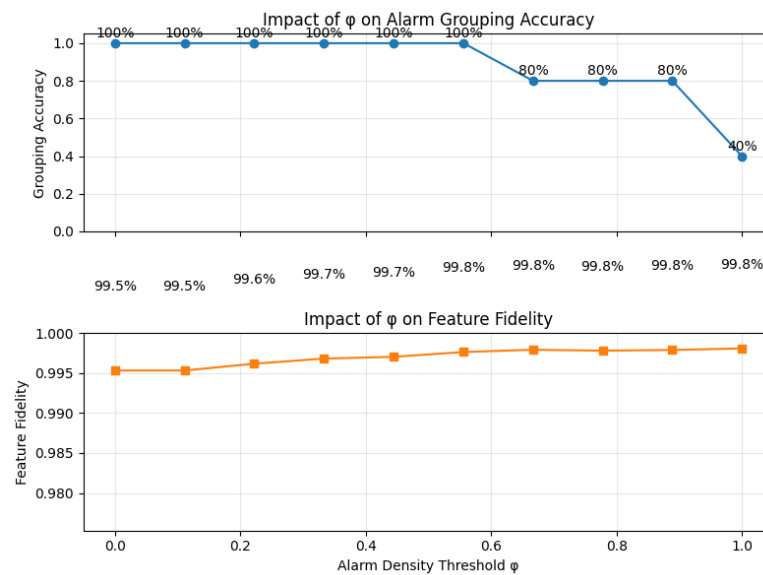


Figure 4. Determination of the ϕ value.

3.1.2. Event Grouping via Time Window Clustering

In complex substation environments, multiple faults can occur simultaneously or within short intervals, making it challenging to differentiate between related and unrelated fault events. To address this, a time window clustering strategy is introduced, to group correlated alarm events systematically. The alarm event grouping process is based on the following:

- (1) Alarm severity: Events with high severity levels are prioritized for grouping.
- (2) Temporal proximity: Events occurring within the same time window are evaluated for correlation.
- (3) Fault type similarity: Similar fault types within a specific period are clustered.

The correlation coefficient and similarity index between alarm events are computed, to determine whether they belong to the same fault cluster:

- (1) If events are correlated, they are grouped together.
- (2) If events are uncorrelated, they are assigned to separate groups.
- (3) If multiple alarm events originate from the same root cause, they are combined into a single fault group.
- (4) If multiple alarm events stem from different sources, they are processed separately for independent fault analysis.

This structured clustering process ensures that alarm event groups are well defined, facilitating accurate fault classification and diagnosis.

3.1.3. Unified Time Correction for Fault Groups

Substations often encounter multiple fault events within short timeframes, making it challenging to establish a consistent time reference across all events. To overcome this, a unified time correction process is developed, to synchronize event timestamps effectively.

The flowchart in Figure 5 illustrates the correction mechanism.

From the process in the diagram, it can be seen that if the start-up criterion is satisfied then the data of the current period will be collected, and then the collected data will be divided into clusters and groups. The combination will be named from 1, and then the correction value of the time of group i will be obtained iteratively. The main and auxiliary protection devices of the group i also perform time correction, and then i points to the group number of the next group, repeating the above steps until the correction of all the groups ends.

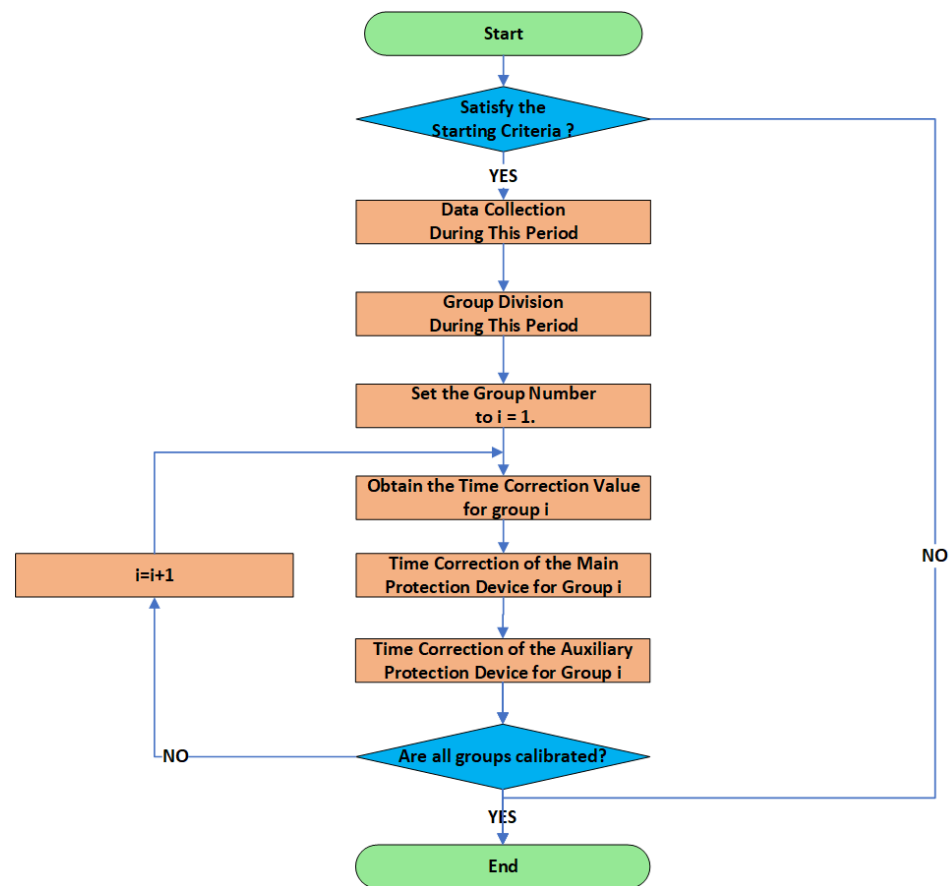


Figure 5. Flow chart of obtaining calibration reference values for each group.

The time-correction process considers three primary timestamps for comparison and adjustment: the SOE (Sequence of Events) timestamp of the group's start-up event; the first-action event timestamp for each protection device; and the event reception timestamp recorded by the master station.

3.2. Feature Extraction—Principal Component Analysis (PCA)

In this study, PCA is employed as an unsupervised dimensionality-reduction method to extract key fault features from high-dimensional substation data. PCA aims to simplify datasets, reduce feature redundancy, and enhance computational efficiency while preserving essential information for fault diagnosis. By retaining the principal components that maximize data variance, PCA effectively identifies critical fault parameters while minimizing irrelevant noise.

There are also some nonlinear dimensionality-reduction methods. Unlike nonlinear methods such as t-SNE or UMAP, PCA offers deterministic and interpretable transformation of high-dimensional fault features into a compact set of principal components. In substation fault diagnosis, sensor data like current, voltage, and gas pressure are structured and physically meaningful—and PCA allows us to trace each component back to interpretable engineering variables. This transparency is critical for maintenance engineers. In addition, PCA is computationally inexpensive and produces stable results, making it ideal for integration into real-time monitoring systems. In contrast, methods such as t-SNE and UMAP are stochastic and primarily used for visualization, while autoencoders require extensive training and hyperparameter tuning. Given that our goal is to extract the most representative fault indicators for scalable online applications, PCA is the most practical and technically appropriate choice.

Substation fault analysis involves multiple interrelated fault parameters, leading to correlations and overlaps among features. Extracting key fault information from these complex datasets is crucial for accurate fault identification and classification. To integrate PCA into substation fault diagnosis, the feature-extraction process follows these structured steps.

3.2.1. Data Standardization

Before performing PCA, it is essential to standardize the data, to ensure consistency across different feature scales. In this study, Z-score normalization is applied, which centers the mean at zero while preserving the original scale of data without altering variance. This allows the relative proportion and unit of each fault feature to remain unchanged.

Given a fault dataset containing m instances and n fault features, the fault matrix is denoted as

$$X_{mn} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \quad (2)$$

The mean fault matrix is computed as

$$\bar{X}_{mn} = \frac{1}{m} \sum_{i=1}^m x_{in} \quad (3)$$

Finally, data standardization is performed by subtracting the mean:

$$X = X_{mn} - \bar{X}_{mn} \quad (4)$$

This step ensures that all fault parameters are centered at zero, allowing PCA to identify the most influential fault features effectively.

3.2.2. Covariance Matrix Computation

The covariance matrix captures the correlation structure between fault features, which is fundamental for identifying principal components. The covariance matrix C is computed as

$$C = \frac{1}{n-1} XX^T \quad (5)$$

where X is the standardized fault matrix. The covariance matrix provides a quantitative measure of how different fault parameters co-vary, ensuring that redundant and highly correlated features can be effectively reduced.

3.2.3. Eigenvalue and Eigenvector Computation

To identify principal components, eigenvalue decomposition is performed on the covariance matrix C :

$$C = V\Lambda V^T \quad (6)$$

where V is the eigenvector matrix, representing the principal component directions; V^T is the transpose of the eigenvector matrix; and Λ is a diagonal matrix containing the eigenvalues, which indicate the variance explained by each principal component.

Solving the equation

$$(C - \lambda I)V = 0 \quad (7)$$

yields eigenvalues λ_i and their corresponding eigenvectors ω_i . The eigenvalues are sorted in descending order, where

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0 \quad (8)$$

This ranking ensures that the first principal components capture the most significant fault variations.

3.2.4. Principal Component Selection

The top k eigenvectors corresponding to the largest k eigenvalues are selected to form the transformation matrix

$$W_{kn} = \begin{bmatrix} V_1^T \\ \vdots \\ V_k^T \end{bmatrix} \quad (9)$$

where W_{kn} organizes the principal components row by row based on their contribution. The selection criterion follows:

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^n \lambda_i} \geq t \quad (10)$$

where t is a preset threshold (typically 90%) that determines the number of retained components. A higher threshold retains more features but may introduce redundant information, while a lower threshold reduces dimensionality but may lose critical fault information. The threshold is set based on the complexity of the substation fault data.

3.2.5. Data Transformation and Dimensionality Reduction

To achieve dimensionality reduction, the original fault data matrix X_{mn} is projected onto the selected principal components:

$$Y_{km} = W_{kn} X_{mn}^T \quad (11)$$

where Y_{km} is the low-dimensional representation of the original fault data, and where W_{kn} ensures that only the most informative features are retained.

This transformation reduces computational complexity while preserving the most critical fault information.

3.2.6. Selection of Key Fault Features

The contribution rate of each principal component vector is calculated sequentially, to determine its relative importance in fault identification. The contribution rate y_i of the q^{th} principal component is given by

$$y_i = \frac{\lambda_q}{\sum_{i=1}^k \lambda_i} \quad (12)$$

The principal components are ranked in descending order, and only those reaching a cumulative contribution rate of 90% are retained as key fault features (the reason is given in the parameter discussion of part (3) in Section 4.2 below). This ensures that the most significant fault indicators are preserved while eliminating redundant information.

The feature extraction process is summarized in Figure 6, which visually represents the PCA-based fault feature selection mechanism.

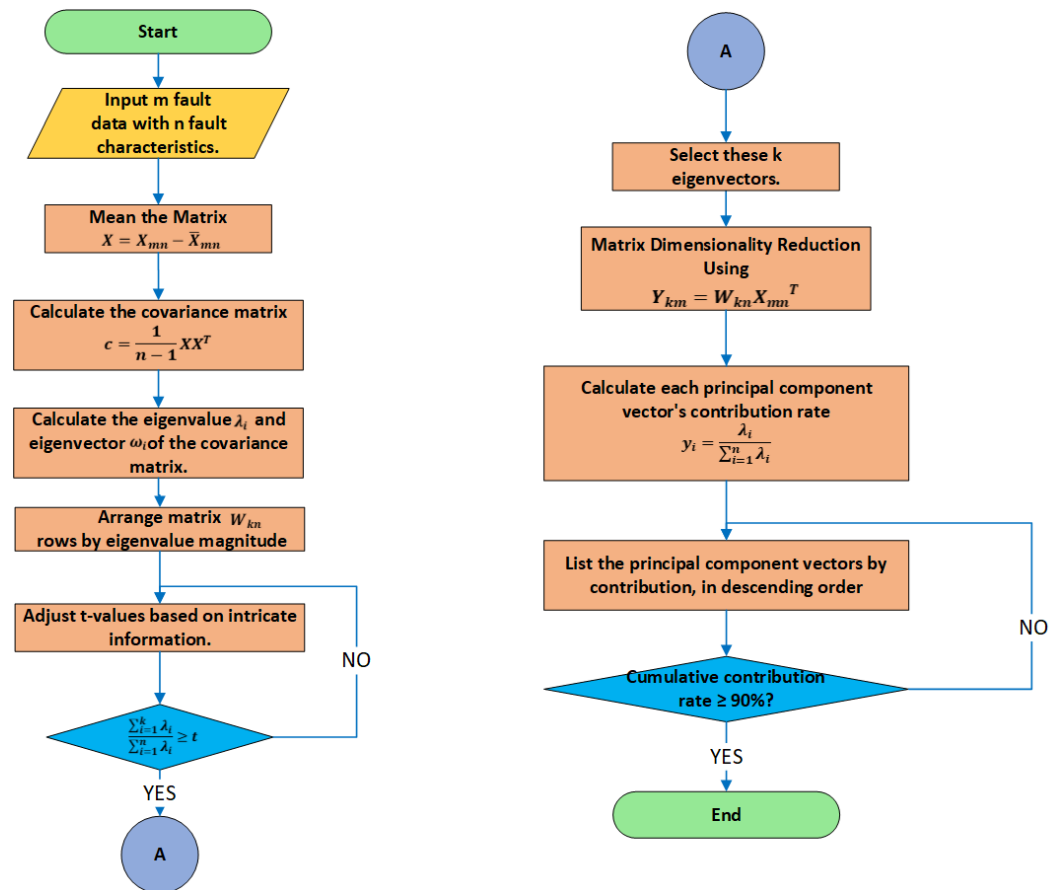


Figure 6. PCA feature extraction steps diagram.

4. Case Analysis and Experiments

4.1. Case Analysis

To demonstrate the effectiveness of the proposed method of key fault information identification method, we conducted a case analysis using sample data collected from third-generation intelligent substation equipment. The case study focused on a 500 kV substation, where we analyzed alarm events triggered during a fault occurrence.

The dataset was generated through simulation, based on the operational logic and failure characteristics of a typical 500 kV third-generation intelligent substation. The simulation framework incorporated the behavior of protection relays, SCADA systems, and sensor monitoring units, and it was designed to mimic real-world fault and alarm patterns under diverse conditions. The fault categories (D1 to D8) were defined by domain experts, referencing actual substation event types and failure modes. All data were labeled, anonymized, and processed following general standards for data security and privacy.

The dataset included alarm records from line protection devices, circuit breaker protection devices, and other related equipment, such as switch signals, WAMS waveform data, and recorded data. These alarm events were collected within a time window T_F following the initial fault detection threshold t_1 . Table 1 presents a sample of the recorded alarm events within this timeframe.

To efficiently process the collected alarm data, we utilized the time window grouping technique, which organizes alarm events into clusters based on temporal and topological correlations. The event grouping followed three key steps:

Table 1. Part of 500 kV substation alarm dataset.

Alarm Time	Event Type	Related Equipment	Alarm Description	Data Source
t1	Line protection action	500 kV Line A	Line protection device tripped	Line protection device
t2	Switch protection action	Station 1 Switch B	Station 1 switch protection device tripped	Switch protection device
t3	Switch protection action	Station 2 Switch C	Station 2 switch protection device tripped	Switch protection device
t4	Abnormal current	500 kV Line A	The line current is abnormal and exceeds the rated value	WAMS system
t5	Voltage sag	500 kV Line A	The voltage drops rapidly and is momentarily lost	Wave recording device
t6	Abnormal vibration signal	Station 1 Transformer	Transformer vibration abnormal, possible mechanical failure	Monitoring Sensors
t7	Abnormal waveform	500 kV Line A	Recorded waveform has abnormal frequency	Wave recording device
t8	Abnormal switch signal	Station 2 Switch C	Switch signal abnormal, possible malfunction	Switch signal system
t9	Switch signal recovery	Station 1 Switch B	Switch signal returns to normal	Switch signal system
t10	Current recovery	500 kV Line A	Line current returns to normal	WAMS system
t11	Voltage recovery	500 kV Line A	Voltage returns to normal level	Wave recording device

(1) Alarm Event Filtering:

Filter out irrelevant alarms and retain only those directly related to the 500 kV line fault. Discard events associated with unrelated lines or equipment.

(2) Formation of Event Clusters and Groups:

Construct event clusters by utilizing all relevant alarms within the designated time window. Organize events into groups based on equipment relevance and network topology:

Group G1: Line protection events (e.g., main protection action signal A6, circuit breaker position signal A1).

Group G2: Circuit breaker operation abnormalities (e.g., zero circuit breaker current A2, zero circuit breaker voltage A3).

Group G3: Pressure-related alarms (e.g., low oil/gas pressure A10, pressure lockout signal A12).

Group G4: SF6 gas anomalies (e.g., SF6 gas pressure alarm A14, SF6 density relay abnormality A15).

(3) Correlation Analysis:

Intra-group correlation: Identify multiple alarms occurring in the same equipment.

Inter-equipment correlation: Establish relationships between circuit breaker actions and line protection responses.

Topological correlation: Analyze connectivity between switch protection devices and line protection systems.

The identified event groups served as the foundation for feature extraction and fault classification in the subsequent experimental phase.

To ensure the reliability of the proposed method, we selected 500 sets of representative fault data, including 350 real training samples and 150 test samples, as summarized in Table 2. Since real substation fault data are relatively scarce, we generated 60 synthetic samples from the original 150 test samples based on real datasets, using the `make_classification` function in the Scikit-learn library to enhance the diversity and coverage of the test set. All these samples could fully characterize different fault scenarios and help to more robustly verify the proposed method.

Following the event-grouping process, we conducted preliminary feature selection to extract the most relevant indicators contributing to fault diagnosis. Table 3 presents a comprehensive summary of the general fault information features identified across all the data samples. Additionally, it highlights the key features specifically related to the D6

fault, ensuring that the redundant and less relevant features were filtered out to improve diagnostic accuracy.

Table 2. Fault data sample.

Fault Number	Substation Status	Training Set	Test Set
D1	Normal	55	25
D2	Circuit breaker tripping	48	22
D3	Control circuit disconnection fault	34	15
D4	Low oil (gas) pressure → lockout and closing fault	45	15
D5	Hydraulic mechanism oil pump pressure timeout	42	22
D6	Circuit breaker arc chamber explosion	70	30
D7	Protection action → circuit breaker refuses to open	32	12
D8	SF ₆ gas low pressure fault	24	9

Table 3. General fault information features and D6 fault selection.

Feature Number	Fault Characteristics	D6 Relevance
A1	Circuit breaker position	YES
A2	Circuit breaker current is zero	YES
A3	Circuit breaker voltage is zero	YES
A4	Circuit breaker position closed	—
A5	Fault interval protection action signal	—
A6	Main protection action	YES
A7	Control circuit disconnection	—
A8	Monitoring circuit breaker control circuit signal light is off	—
A9	No protection action signal received	—
A10	Low oil (gas) pressure alarm of operating mechanism	YES
A11	Locking reclosing	—
A12	Low pressure lockout signal	YES
A13	Oil pump pressure timeout alarm	—
A14	SF ₆ gas low pressure alarm signal	YES
A15	SF ₆ density relay indicates abnormality	YES
A16	Partial discharge level exceeds threshold	YES

Specifically, we now explain that the features marked “YES” were selected based on their high relevance to the D6 fault, as determined by the following:

- ① Their contribution rates in PCA analysis (e.g., A3 and A2 having high eigenvalue loadings).
- ② And expert knowledge from substation engineers confirming their diagnostic significance.

Conversely, features marked with “—” were excluded due to low variance, weak correlation with D6, or redundancy with other features.

To verify the efficiency of the feature-extraction process, we applied the complete method to all the training set samples.

4.2. Experiment

The D6 circuit breaker arc chamber explosion fault is used as an example to illustrate the experimental steps.

(1) PCA-Based Feature Extraction and Analysis

After event grouping, we applied PCA to extract the key fault features. The following steps were performed:

- ① Data standardization: The selected fault feature dataset was normalized to ensure consistency.

- ② Mean calculation: The dataset was averaged.
 - ③ Covariance matrix computation: The covariance matrix was derived from the standardized dataset, as described by the formula $C = \frac{1}{n-1}XX^T$.
 - ④ Eigenvalue and eigenvector calculation: The eigenvalues λ_i and eigenvectors ω_i of the covariance matrix were computed.
 - ⑤ Dimensionality reduction: The original data matrix X_{mn} was projected onto a lower-dimensional space, using the transformation $Y_{km} = W_{kn}X_{mn}^T$.
 - ⑥ Principal component contribution analysis: The contribution rates of all the principal components were calculated and arranged in descending order. The PCA outcomes are illustrated in Figure 7.
- (2) Principal Component Contribution Analysis

Figure 7 presents a bar chart alongside a cumulative contribution curve. The blue bars indicate the contribution rate of each individual principal component, while the red line illustrates the cumulative contribution rate.

As shown in the figure, the cumulative contribution rate of the top 10 principal components reached approximately 93%. Specifically, PC1 contributed about 28%, with its top two loading features being A2 and A3; PC2 contributed about 14%, with A12 and A6 as the dominant load features; PC3–PC5 each contributed 7–8%; and the remaining components contributed approximately 5% or less. This analysis demonstrates that selecting the top 10 principal components allowed us to retain more than 90% of the information while significantly reducing computational complexity (see Section 4.2, part 2, paragraph 2).

This experiment verified that the proposed PCA-based feature-extraction method successfully identifies the key indicators of the D6 fault, reducing dimensionality while preserving critical information.

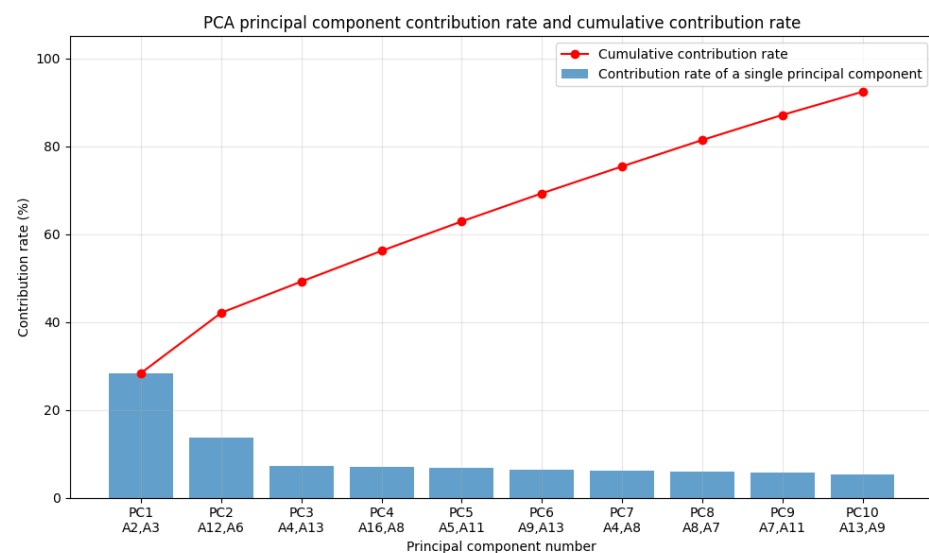


Figure 7. Results of Principal Component Analysis.

(3) Parameter Discussion

To explore how to choose the threshold t , we conducted an experimental study; the results are summarized in Table 4.

From these results, we observe the following: When $t = 85$, only 9 principal components were retained, and the fidelity was only 87.09%, which had lost some fault information; if it was increased to 90%, 10 principal components were needed to recover 92.37% of the variance, and the compression rate was still 62.50%; but if it was increased to 95, 11 principal components were needed, and the fidelity was increased to 96.39%, and the compression rate

would rise to 68.75%. After exceeding 90%, although Fidelity and MSE were still improved, the gain was relatively marginal ($\approx 4\%$ vs. $\approx 5\%$ in the previous section), but more storage and processing overhead was required. Therefore, we chose $t = 90\%$ as the practical optimum: it captured more than 92% of the variance with 10 principal components (achieving a 62.5% compression ratio), while having less storage and processing overhead than 95%.

Table 4. Comparison table with existing methods.

Threshold (%)	PCs	Fidelity (%)	MSE	Global Compression (%)	Time (ms)
80	8	81.38	0.1862	50.00	1.40
80	9	87.09	0.1291	56.25	0.96
80	10	92.37	0.0763	62.50	0.85
80	11	96.39	0.0361	68.75	0.86

(4) Comparative Study

We conducted comparative experiments on the method proposed in this paper and some other methods. The results of the comparison with the existing Fixed TW + PCA and Raw Features methods are shown in Table 5:

Table 5. Comparison table with existing methods.

Method	Event Integrity (%)	Feature Fidelity (%)	Avg. Window Compression (%)
Raw Features	100.0	100.0	100.0
Fixed TW + PCA	40.0	99.6	35.2
STW + PCA	100.0	99.6	34.4

Raw Features trivially captured all the alarm clusters (100% integrity) but offered no compression. Fixed TW + PCA covered only 40% of the clusters entirely within single windows, though it retained 99.6% of the variance at 35.2% compression. STW + PCA achieved both 100% event integrity and the same 99.6% fidelity at 34.4% compression—demonstrating that aligning windows to alarm groups preserves complete events while discarding slightly more redundancy.

In addition to comparing the same type of linear dimensionality reduction methods, we also compared the nonlinear dimensionality reduction methods. Five data-processing processes are included in Figure 8 below, and we report two key metrics:

Feature Fidelity—the percentage of total variance (or trustworthiness) retained.

Compression Ratio—the dimensionality after embedding divided by the original 16 features.

The original data retained all the variance, but no compression was achieved. Both PCA variants achieved 99.6% fidelity, but STW + PCA offered a slight improvement in compression (34.4% vs. 35.2%) and, as shown in our event completeness analysis, guaranteed 100% coverage of alarm clusters, unlike fixed time windows. Uniform Manifold Approximation and Projection (UMAP) and t-Distributed Stochastic Neighbor Embedding (t-SNE) achieved very high compression (12.5% from 16 to 2 dimensions), but their measured “fidelity” of 92.0% and 96.4%, respectively, already included a quadratic inversion based on K-Nearest Neighbors (KNNs), so the variance actually preserved by the embeddings alone was even lower. Furthermore, their stochastic nature and lack of analytical inversion or characteristic loadings make them difficult to interpret or reproduce in real-world substation processes. In contrast, Principal Component Analysis (PCA) provided a single, deterministic transformation with an exact inverse operation and a clear variance preservation measure (99.6%), ensuring high fidelity and full reconstructability.

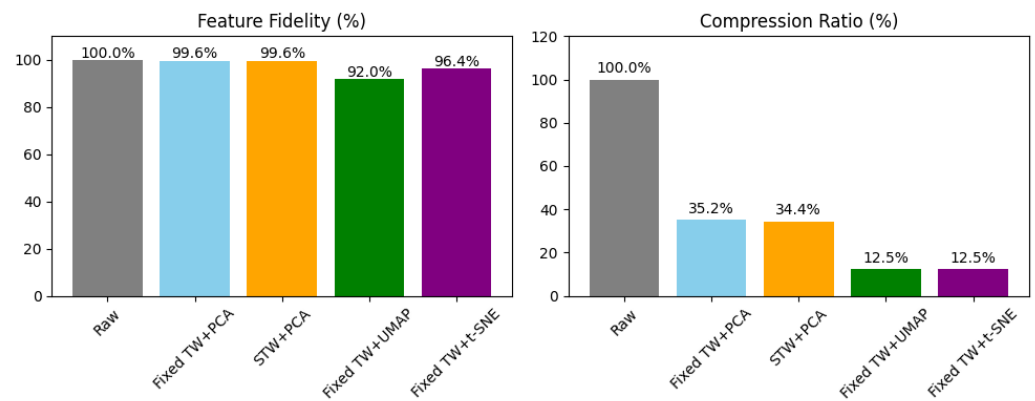


Figure 8. Comparison chart of five methods.

These results demonstrate that STW+PCA strikes the best balance among interpretability, fidelity, compression, and alarm-event integrity, thus validating our choice of combining a scalable time window with PCA for substation fault feature extraction.

(5) Limitation

The effectiveness of the scalable time window relies on the presence of meaningful variations in alarm event density. Sparse or uniformly distributed data may reduce its effectiveness. Threshold sensitivity: Although we conducted sensitivity analysis for the PCA threshold (e.g., 90%), different substations or fault types may still require re-calibration for optimal performance. PCA may not fully capture nonlinear correlations among features, which could affect feature compression in some cases, and there may be situations where the interference of certain noise characteristics cannot be completely removed. The method was evaluated on real-world data from a single substation. Further work is needed, to validate its robustness across different substations and operational contexts.

5. Conclusions

In this paper, we proposed a scalable time window-based method of identifying key fault information, which we combined with Principal Component Analysis (PCA) to address the challenges of excessive redundant information in existing fault-diagnosis techniques. The proposed approach enables more effective extraction of critical fault-related features, reducing the interference of irrelevant data while preserving essential fault indicators. By introducing a dynamically adjustable time window, our method optimally segments alarm events according to real-time system conditions, ensuring that critical fault information is accurately captured. The integration of PCA further enhances feature selection by reducing data dimensionality, removing redundancy, and retaining the most representative fault indicators. The effectiveness and applicability of the proposed method were validated using real-world datasets from third-generation intelligent substations, demonstrating its ability to improve fault identification accuracy. The experimental results indicate that combining adaptive time window adjustment with PCA can enhance the quality of fault feature representation in alarm event datasets. This demonstrates the potential of integrating dynamic preprocessing strategies with classical feature-extraction techniques, especially in systems with heterogeneous and asynchronous alarms. The proposed method is particularly suited to substation environments where interpretability, stability, and computational efficiency are critical. The results also show that PCA effectively reduces the feature space while preserving over 90% of fault-related information, thereby minimizing computational complexity and improving diagnostic efficiency.

This study provides a practical and scalable solution for intelligent substation fault diagnosis, offering valuable insights into improving the efficiency of Operation and Maintenance (O&M) systems in modern power grids. Future research will focus on extending this methodology by incorporating deep learning-based fault diagnosis models to further improve automation and adaptability. Additionally, exploring multi-source data fusion techniques could enhance the robustness of the proposed method in complex substation environments.

Author Contributions: P.Z., L.G. and Z.H. designed the project and drafted the manuscript, as well as collecting the data. Z.R., Y.Z., R.X. and D.L. wrote the code and performed the analysis. Z.S. designed the additional experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Project of State Grid Hubei Electric Power Co., Ltd., Xiaogan Power Supply Company (research on proactive warning method of abnormal operation of substation equipment based on deep learning).

Data Availability Statement: The data presented in this study are available in this article.

Conflicts of Interest: Authors Pan Zhang, Lei Guo, Zhicheng Huang, Zhoupeng Rao, Ying Zhang and Zhi Sun were employed by the company State Grid Hubei Electric Power Co., Ltd. Xiaogan Power Supply Company. The remaining authors (Rui Xu and Deng Li) 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.

References

1. Ke, T.; Tian, S.; Zhou, H. Research on Fault Diagnosis Model of Small Substation Based on Cooperative Game and Cloud Model. *Electr. Drive* **2024**, *54*, 82–87.
2. Ding, W. Research on Fault Diagnosis of Intelligent Substation Protection System. Master's Thesis, Shandong University, Jinan, China, 2023.
3. Zhan, Z.; Gong, D.; Wang, C.; Yang, D.; Wang, J.; Hao, J. Transformer Fault Diagnosis Method Integrating D-SEvidence Theory and Fuzzy Logic. *Transformer* **2023**, *60*, 10–16.
4. Dai, Y.; Wang, J.; Tong, X. Line Trip Fault Identification Method Based on Big Data and Multi-Information Sources. *Microcomput. Appl.* **2024**, *40*, 221–224+232.
5. Zhou, W. Application of Intelligent Technology in Substation Fault Diagnosis. *Integr. Circuit Appl.* **2024**, *41*, 130–131.
6. Du, J.Q.; Zhao, M.; Yin, J.; Gu, W. Review of fault diagnosis methods for power plant. *Yunnan Electr. Power* **2018**, *46*, 88–96.
7. Chang, J.; Gao, M. Failure prognostic system of power generating equipment based on similarity modeling. *J. Mech. Electr. Eng.* **2012**, *29*, 576–579.
8. Li, A.; Zhang, L.; Li, C. Mathematical Morphology-Based Fault Identification for Distribution Network with DFIG. In Proceedings of the 2024 The 9th International Conference on Power and Renewable Energy (ICPRE), Guangzhou, China, 20–23 September 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 752–757.
9. Zhang, C.; Xiong, D.; Xia, X.; Pan, J.; Yan, Y.; Xu, A. Study on the Diagnostic Method of Fault Current and Excitation inrush Current. In Proceedings of the 2023 4th International Conference on Smart Grid and Energy Engineering (SGEE), Zhengzhou, China, 24–26 November 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 361–365.
10. Feng, Y.; Yang, W.; Jiang, D. Fault Recognition Method for Gear Case of Wind Power Generator Based on Decision Tree Classification Algorithm and Expert System. *Guangdong Electr. Power* **2013**, *26*, 17–21.
11. Ji, Y. Research and Implementation of Power Grid Fault Diagnosis Expert System Based on Fuzzy Recognition. Master's Thesis, Jilin University, Changchun, China, 2011.
12. Li, B.; You, J.; Liu, X.; Wang, R.; Leng, G.; Ma, J. Research on Substation Fault Diagnosis Method Based on Alarm Information and CNN. *Coal Mine Mach.* **2024**, *45*, 161–164.
13. Zhou, W. Research on Fault Identification Method for Secondary Circuit of Relay Protection in Intelligent Substations. *Electr. Technol. Econ.* **2024**, 324–327.
14. Huang, H.; Ji, H.; Meng, F. Fault Identification and Location of Shore Power Systems in Oilfield Clusters Based on Key Data Information. *Equip. Manag. Maint.* **2024**, 171–175.

15. Yu, Y.; Wang, T.; Wang, W.; Yang, R. Research on intelligent fault detection and identification method of relay protection device in digital substation. *Electron. Des. Eng.* **2024**, *32*, 113–117.
16. Li, Z.; Ding, X.; Du, W. Research on potential fault location and early warning of substation based on digital twin. *Electr. Eng.* **2022**, 35–38.
17. Wang, H.; He, H.; Huang, R.; Gao, P.; Zhou, N.; Ren, B. Fault diagnosis of intelligent substation protection system based on artificial neural network. *New Technol. Electr. Eng. Energy* **2023**, *42*, 97–104.
18. Lozano, J.; Koneru, K.; Castellanos, J.H.; Cardenas, A.A. Timing Analysis of GOOSE in a Real-World Substation. In Proceedings of the 2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Singapore, 25–28 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 160–165.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.