

Unsupervised Clustering for Fault Analysis in High-Voltage Power Systems Using Voltage and Current Signals

Julian Oelhaf*, Andreas Maier, Siming Bayer

Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg

Georg Kordowich, Johann Jäger

Institute of Electrical Energy Systems, Friedrich-Alexander-University Erlangen-Nuremberg

*corresponding author: julian.oelhaf@fau.de

Presented at the
27th Annual Georgia Tech Fault and Disturbance Analysis Conference
Atlanta, Georgia
May 5–6, 2025

Unsupervised Clustering for Fault Analysis in High-Voltage Power Systems Using Voltage and Current Signals

Julian Oelhaf^{1*}, Georg Kordowich², Andreas Maier¹, Johann Jäger², Siming Bayer¹

¹*Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg*

²*Institute of Electrical Energy Systems, Friedrich-Alexander-University Erlangen-Nuremberg
Erlangen, Germany*

*corresponding author: julian.oelhaf@fau.de

Abstract—The widespread use of sensors in modern power grids has led to the accumulation of large amounts of voltage and current waveform data, especially during fault events. However, the lack of labeled datasets poses a significant challenge for fault classification and analysis. This paper explores the application of unsupervised clustering techniques for fault diagnosis in high-voltage power systems. A dataset provided by the Réseau de Transport d'Électricité (RTE) is analyzed, with frequency domain features extracted using the Fast Fourier Transform (FFT). The K-Means algorithm is then applied to identify underlying patterns in the data, enabling automated fault categorization without the need for labeled training samples. The resulting clusters are evaluated in collaboration with power system experts to assess their alignment with real-world fault characteristics. The results demonstrate the potential of unsupervised learning for scalable and data-driven fault analysis, providing a robust approach to detecting and classifying power system faults with minimal prior assumptions.

Index Terms—unsupervised learning, fault analysis, power system protection, clustering, voltage and current analysis

I. INTRODUCTION

The increasing deployment of sensors in modern power grids has led to the generation of vast amounts of data related to voltage and current waveforms, particularly during fault events. This growing availability of high-resolution measurements presents a unique opportunity for enhancing fault detection and diagnosis. However, a significant challenge lies in the fact that much of this data remains unlabeled, limiting its direct applicability in supervised learning-based fault classification. Extracting meaningful insights from these large-scale, unlabeled datasets is critical for ensuring grid stability and preventing cascading failures.

Accurate and timely fault diagnosis plays a pivotal role in power system operation, as undetected or misclassified faults can lead to severe system disturbances, increased downtime, and economic losses. Traditional fault classification methods typically rely on supervised machine learning models, which require extensive labeled datasets for training. However, generating such labeled datasets is both time-consuming and resource-intensive. This challenge is exacerbated in industrialized nations such as Germany, where a shortage of skilled professionals further hinders the manual labeling of power

system faults. Consequently, there is a growing need for data-driven approaches that can extract insights from raw measurements without relying on expert-annotated datasets.

Unsupervised machine learning provides a promising alternative by enabling the automatic discovery of hidden patterns and structures within data. Clustering techniques, in particular, allow for the identification and categorization of fault events without predefined labels, making them well-suited for real-world power system applications. This paper presents a novel approach that applies unsupervised clustering to analyze voltage and current signals, facilitating the identification and classification of fault types in high-voltage power systems.

To this end, a dataset provided by Réseau de Transport d'Électricité (RTE) is analyzed, with frequency-domain features extracted using the Fast Fourier Transform (FFT). The K-Means clustering algorithm is then employed to group fault events based on their underlying similarities. The resulting clusters are evaluated in collaboration with power system experts to assess their alignment with real-world fault characteristics. The findings highlight the potential of unsupervised learning in power system fault analysis, demonstrating its capability to reduce reliance on labeled data while providing scalable and adaptive fault detection.

The remainder of this paper is organized as follows: Section 2 describes the dataset and preprocessing steps. Section 3 provides an overview of fault types in transmission lines. Section 4 details the methodology, including feature extraction and clustering. Section 5 presents the clustering analysis, while Section 6 reports the results. The discussion interprets the findings and their implications, followed by the conclusion summarizing key insights and future directions.

II. DATASET

Understanding the characteristics of fault events in high-voltage power systems is crucial for developing reliable protection mechanisms. This study leverages a real-world dataset of voltage and current waveform recordings from fault events in the French transmission network. The dataset provides valuable insights into transient behavior and serves as the basis for applying unsupervised clustering techniques to fault

analysis. In this section, we describe the data acquisition process, key signal characteristics, and preprocessing steps.

A. Data Source and Acquisition

The dataset used in this study is provided by RTE¹, the French transmission system operator. It is available through the RTE Fault Recording Database (DFRDB)² and comprises 12 053 voltage and current waveform recordings captured during fault events in France’s high-voltage transmission network [1]. As the operator of more than 105 000 kilometers of transmission lines and 50 interconnections with neighboring countries, RTE plays a key role in ensuring grid stability, making this dataset invaluable for improving fault analysis and protection strategies.

Each fault event in the dataset is represented by 3-phase-voltage and current signals. The signals are recorded at a high sampling rate of 6400 Hz, enabling precise capture of transient voltage and current variations, which are essential for accurate fault analysis, classification, and diagnosis. Analyzing these signals is crucial for understanding fault dynamics and enhancing the reliability of protection mechanisms.

B. Signal and Sample Characteristics

The dataset consists of 12 053 fault event recordings, each containing six time-synchronized signals: three phase voltages (V_1, V_2, V_3) and three phase currents (I_1, I_2, I_3). Each signal is recorded for 21 000 timesteps, corresponding to 3.281 25 s of data. The data is structured as a three-dimensional array with dimensions (12 053, 6, 21 000), where:

- 12 053 represents the number of recorded fault events,
- 6 denotes the number of signals per fault event (three voltage and three current signals),
- 21 000 corresponds to the number of timesteps per signal.

The nominal system frequency is 50 Hz, and the nominal voltage level is 90 V, reflecting the operating conditions of the French high-voltage transmission network. The dataset was preselected based on waveform characteristics to include only significant fault events. This selection was performed by the dataset providers, and we have no insight into their specific criteria or process.

The signals are recorded at a sampling frequency of 6400 Hz, providing high temporal resolution for capturing rapid voltage and current transients during faults. They are quantized using 16-bit encoding, with amplitude values ranging from $-32\,767$ to $32\,767$, ensuring precise representation of both minor and severe disturbances. The quantization step sizes are 18.310 V for voltage signals and 4.314 A for current signals, ensuring accurate representation of signal variations critical for fault analysis.

Notably, the dataset lacks explicit labels or background information on fault types, motivating the use of an unsupervised approach for fault analysis and classification. The only available information comes from the RTE Fault Recording Database (DFRDB) [1].

¹<https://www.rte-france.com/en/home>

²<https://dfrdb.rte-france.com/>

C. Data Preprocessing

Prior to analysis, the voltage and current waveforms undergo preprocessing to enhance interpretability and improve the effectiveness of clustering techniques. This process consists of two primary steps: normalization and decomposition-based anomaly detection.

To ensure consistency across all recorded fault events, each signal is independently normalized to a fixed range. For a given raw discrete-time signal $X[n]$, its normalized counterpart $\tilde{X}[n]$ is computed as:

$$\tilde{X}[n] = 2 \frac{X[n] - X_{\min}}{X_{\max} - X_{\min}} - 1, \quad (1)$$

where X_{\min} and X_{\max} represent the minimum and maximum values of the signal, respectively. This transformation scales the signal to the range $[-1, 1]$, ensuring numerical stability and enabling comparison across different recordings.

Each signal is then decomposed into three components: the trend $T[n]$, which represents long-term variations; the seasonal component $S[n]$, capturing periodic fluctuations related to the nominal power system frequency; and the residual $e[n]$, which contains transient deviations and noise. The decomposition follows an additive model:

$$Y[n] = T[n] + S[n] + e[n], \quad (2)$$

where $Y[n]$ is the original sampled signal. The trend component $T[n]$ is estimated using a moving average filter, which smooths long-term variations. After removing the trend, the seasonal component $S[n]$ is computed as the mean of the de-trended series over one full cycle of the power system’s fundamental frequency. Given the system’s nominal frequency of 50 Hz and a sampling rate of 6400 Hz, the decomposition period is 128 time steps. Seasonal decomposition is performed using the `seasonal_decompose` class provided in the `statsmodels` library [2]. In addition to the decomposition, zero-voltage $Z_V[n]$ and zero-current $Z_I[n]$ indicators are computed for each sample.

Finally, anomaly detection is applied to the voltage signal, specifically targeting the trend $T[n]$, residual $e[n]$, and zero-voltage $Z_V[n]$ components, in order to identify time intervals with significant deviations from normal operation. Thresholds based on statistical measures are defined for each component to highlight regions where abnormal transients or zero-voltage events occur. This enables the isolation of periods where the signal deviates from expected behavior, indicating potential faults or disturbances in the system. While decomposition and anomaly detection are not used in the clustering process, the detected anomalies support the manual visual analysis by highlighting regions of interest.

This preprocessing pipeline structures the dataset to emphasize significant variations while filtering out noise and redundant information, ensuring that clustering techniques can effectively distinguish between different fault behaviors based on intrinsic signal characteristics.

III. OVERVIEW OF FAULT TYPES IN TRANSMISSION LINES

Faults in transmission lines are undesirable short-circuit conditions that can lead to significant damage in power systems. These faults can be classified based on their electrical characteristics and underlying causes. A thorough understanding of fault types is essential for the development of effective diagnostic and prognostic techniques in high-voltage and medium-voltage power networks [3].

Overvoltages: Overvoltages occur when the system voltage exceeds the nominal design value. They can be either external or internal in nature. External overvoltages are primarily caused by lightning strikes, which can lead to insulation failure and catastrophic short-circuits. Internal overvoltages are caused by switching operations, load variations, or resonance phenomena, and they can lead to insulation breakdown within the system.

Overcurrents: Overcurrents occur when the electrical current surpasses the system's designed limits. These can be transient, permanent, or fault-induced in nature. Transient overcurrents, such as those arising from motor startups or transformer energization, are typically short-lived and do not require circuit breaker intervention. Permanent overcurrents, on the other hand, occur when sustained overloads lead to conductor heating and potential long-term damage. Fault-induced overcurrents result from faults like insulation failure or accidental contact, generating high short-circuit currents that can cause extensive damage to system components.

Short-Circuit Faults: Short-circuit faults are among the most critical types of faults in transmission systems. They involve unintended low-impedance connections between conductors or between a conductor and the ground. Short-circuit faults can be categorized into four main types. The most common is the single-line-to-ground (SLG) fault, which occurs when one phase conductor comes into contact with the ground. Line-to-line (LL) faults involve a short-circuit between two phase conductors, typically due to insulation failure or mechanical damage. Double-line-to-ground (DLG) faults, where two phase conductors simultaneously make contact with the ground, are more severe and result in high fault currents. The most catastrophic fault is the three-phase (LLL) fault, which involves all three phases and typically results from equipment failure or severe external disturbances.

Open Circuit Faults: Open circuit faults occur when the conducting path is broken, often due to conductor failure or cable joint issues. These faults can lead to unbalanced currents within the system, which can cause overheating of rotating machinery and other operational disturbances.

Power Quality Disturbances: In addition to faults, power transmission systems are subject to various power quality disturbances that can affect system reliability. These disturbances include voltage fluctuations, which arise from changing load conditions or renewable energy integration, and flicker, which is caused by short-term voltage variations from industrial loads or renewable generation. Harmonic distortion, resulting from nonlinear loads and power electronic converters, can also

cause significant waveform distortions, interfering with grid operation and equipment performance [3].

Effective detection, diagnosis, and protection against faults and power quality disturbances are essential for maintaining the stability and resilience of power transmission systems. The integration of advanced fault diagnosis techniques, including artificial intelligence-based approaches, plays a vital role in ensuring operational efficiency and grid stability.

IV. METHODOLOGY

The proposed approach consists of three main stages: (i) feature extraction in the frequency domain using the Fourier Transform, (ii) dimensionality reduction through two distinct techniques, and (iii) clustering of the reduced feature representations using the K-Means algorithm.

A. Feature Extraction

To transform raw time-series data for machine learning, we apply the *Fast Fourier Transform (FFT)* to extract frequency-domain features. The FFT converts a signal $x(n)$ from the time domain into its frequency components, capturing periodic patterns that may indicate faults. The FFT is computed as:

$$X(f) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi fn/N} \quad (3)$$

We retain only the positive frequencies (first $N/2$ coefficients), as the FFT output is symmetric for real-valued signals. FFT is chosen for its efficiency in revealing spectral patterns associated with power system disturbances, such as impedance changes, arcing, and harmonics from faults.

Feature Selection and Normalization: To standardize the features, FFT magnitudes are normalized by dividing each component by the maximum value:

$$X_{\text{norm}}(f) = \frac{X(f)}{\max(X(f))} \quad (4)$$

Normalization minimizes the impact of amplitude variations and improves model generalization, as fault classification relies on *relative spectral characteristics*. However, it may suppress information on absolute magnitude changes.

Domain-Specific Considerations: FFT aligns with established power system research, where faults exhibit distinct frequency-domain patterns:

- **Short circuits** show high-frequency transients due to sudden current flow changes.
- **Open-circuit faults** produce dominant low-frequency components from power loss.
- **Unbalanced faults** cause asymmetries captured as *negative sequence components*.

These insights make the extracted features meaningful representations of fault behaviors in electrical networks.

B. Feature Reduction

Principal Component Analysis (PCA): PCA is a linear technique that transforms the original feature space into a new coordinate system defined by the principal components, which are directions of maximum variance in the data [4]. It reduces the dimensionality by projecting the original feature matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ onto a matrix of principal components $\mathbf{P} \in \mathbb{R}^{D \times K}$, where $K \ll D$:

$$\mathbf{X}_{\text{reduced}} = \mathbf{X}\mathbf{P} \quad (5)$$

PCA retains the most significant components, typically preserving 95%–99% of the variance, helping to reduce noise and irrelevant features. The implementation uses the `PCA` class from `scikit-learn` [5].

t-Distributed Stochastic Neighbor Embedding (t-SNE):

While PCA handles linear reductions, t-SNE addresses non-linear relationships in high-dimensional data. t-SNE visualizes data by converting pairwise distances into conditional probabilities and minimizing the Kullback-Leibler (KL) divergence between high- and low-dimensional probability distributions [6]. t-SNE is implemented in the `TSNE` class of `scikit-learn` [5]. The algorithm uses a Gaussian distribution to model similarities:

$$p_{ij} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma^2)} \quad (6)$$

It then maps points to a lower-dimensional space $\mathbf{Y} \in \mathbb{R}^{N \times 2}$ or $\mathbb{R}^{N \times 3}$, aiming to preserve the data's local structure. The similarity in the low-dimensional space is given by:

$$q_{ij} = \frac{1}{1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2} \quad (7)$$

The KL divergence quantifies how well the local structure is preserved:

$$D_{\text{KL}} = \sum_{i=1}^N \sum_{j=1}^N p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (8)$$

Application in this Study: PCA is initially applied to reduce the dimensionality of the FFT features while retaining most of the variance. Following PCA, t-SNE is used to further reduce the data to two or three dimensions, enabling effective visualization and revealing underlying patterns and clusters. This combination aids in the identification and interpretation of fault-related structures within the data.

C. Clustering Algorithm

Clustering is essential for identifying patterns in high-dimensional data. K-Means Clustering, known for its simplicity, efficiency, and interpretability, is widely used for partitioning fault patterns in high-dimensional feature spaces. First proposed by Stuart Lloyd in 1957 for pulse-code modulation, the term "K-Means" was introduced by James MacQueen in 1967. The algorithm is also called the Lloyd–Forgy method, as Edward W. Forgy independently derived it in 1965 [7].

The K-Means algorithm aims to partition a dataset $\mathbf{X} \in \mathbb{R}^{N \times D}$ into K distinct clusters. The objective is to minimize the within-cluster sum of squared distances between data points and the centroids of their respective clusters. Given a set of data points, K-Means performs the following steps iteratively:

- **Initialization**: Randomly select K initial cluster centroids.
- **Assignment Step**: Assign each data point to the closest centroid based on a distance metric (typically Euclidean distance).
- **Update Step**: Recompute the centroids by calculating the mean of all points assigned to each centroid.
- **Convergence**: Repeat the assignment and update steps until convergence, typically when the centroids no longer change or the maximum number of iterations is reached.

Mathematically, the objective function that K-Means seeks to minimize is given by:

$$J = \sum_{k=1}^K \sum_{i=1}^{N_k} \|\mathbf{x}_i - \mu_k\|^2 \quad (9)$$

where \mathbf{x}_i represents the i -th data point, μ_k is the centroid of the k -th cluster, and N_k is the number of data points assigned to cluster k . The term $\|\mathbf{x}_i - \mu_k\|^2$ represents the squared Euclidean distance between the data point and the centroid.

Implementation and Parameters: The K-Means algorithm is implemented using the `KMeans` class in the `scikit-learn` library [5]. The algorithm requires the specification of the number of clusters K , which is often determined using methods such as the elbow method or silhouette analysis. The elbow method involves plotting the sum of squared distances (within-cluster inertia) for different values of K and selecting the value at which the rate of decrease slows down. The silhouette analysis measures how similar a point is to its own cluster compared to other clusters, providing a quantitative way to assess the quality of clustering.

Applications and Limitations: K-Means is widely used due to its computational efficiency, especially for large datasets. It is particularly well-suited for scenarios where the clusters are approximately spherical and have roughly equal sizes. This makes K-Means effective for partitioning fault patterns that exhibit well-separated, homogeneous characteristics in feature space.

However, K-Means has several limitations. One major limitation is its sensitivity to the initial placement of centroids. Poor initialization can lead to suboptimal clustering results. Additionally, K-Means assumes that clusters are spherical and equally sized, which may not hold true for all datasets. To mitigate these limitations, variations of K-Means, such as K-Means++ for centroid initialization, have been proposed.

Usage in this Study: In this study, K-Means is applied to the extracted features to identify fault patterns in the high-dimensional feature space. After dimensionality reduction using PCA and t-SNE, K-Means is employed to partition the data into distinct clusters that correspond to different fault types.

The resulting clusters are analyzed to identify key patterns and relationships between the fault occurrences, which can then inform further diagnostic and predictive modeling.

V. CLUSTERING ANALYSIS

In this study, we adopt a two-stage approach to examine the clustering of fault signals. First, we quantitatively assess the clustering output using metrics such as cluster separability and sample distribution. This analysis helps verify that the resulting groups are statistically meaningful. Second, we perform an expert-driven validation where a power system specialist reviews representative samples from each cluster. This step enables us to interpret cluster compositions, confirm correspondence with actual fault types, and identify any overlaps among fault categories.

A. Cluster Visualization

Figure 1 shows a two-dimensional t-SNE projection of the feature space, where the 15 clusters are color-coded to highlight distinct fault patterns. This visualization provides an intuitive assessment of cluster separability, helping to identify both well-separated clusters that suggest unique fault characteristics and overlapping regions that may indicate transitional fault states or feature ambiguities. For further evaluation, we consider two key aspects:

Cluster Separability: This assesses the spatial distinctiveness of the clusters. Well-separated clusters suggest distinct fault types, while overlap may indicate similarities or transitional conditions between fault states.

Cluster Size Distribution: This examines the number of samples within each cluster. Larger clusters typically correspond to more common fault patterns, while smaller clusters may represent rare or anomalous events.

The 2D scatter plot in Fig.1 offers valuable insights into the clustering results. Combined with quantitative metrics, these visual tools allow for a comprehensive evaluation of clustering quality.

B. Manual Expert Labeling

To further validate the clustering results, a power system expert manually examines a subset of signals from each cluster. During this process, the expert assigns fault labels based on established fault characteristics, comparing the unsupervised clusters with real-world fault types such as three-phase short circuits (3PH-SC), line-to-ground (LG) faults, and line-to-line (LL) faults.

For this evaluation, randomly selected samples from each cluster are reviewed. The expert assesses cluster coherence by visualizing the voltage and current signals over multiple time windows, enabling a detailed analysis of transient phenomena, waveform distortions, and phase interactions. Each labeled sample is annotated with the following details:

- **Fault Class:** A broad category (e.g., short circuit, open circuit, transient disturbance).
- **Fault Type:** A specific classification (e.g., single-phase-to-ground short circuit, three-phase fault).

TABLE I
EXAMPLE OF LABELS ASSIGNED BY THE EXPERT.

Sample ID	Fault Class	Fault Type	Phase
27	Normal	Normal	N/A
2654	Other	Off - No Switch	N/A
3335	Short-circuit	1-P-SC	C
5125	Transients	Transients	N/A
8578	Switching	Switch On	N/A
11491	Short-circuit	1-P-SC	B

- **Phase:** The impacted phase(s).
- **Comment:** Additional observations or clarifications regarding the fault characteristics.

For example, a sample might be labeled as: *Fault Class: Short Circuit, Fault Type: 1-P-SC, Phase: A*, accompanied by comments describing waveform irregularities or recovery patterns. An excerpt of the labels assigned by the expert is shown in Table I.

Additionally, the expert is provided with an interactive feature that allows them to *scroll* through the signal data. This enables seamless switching between detailed, zoomed-in views and comprehensive full-sequence plots, ensuring that localized anomalies are evaluated within the broader context of the fault event. This multi-scale assessment enhances labeling accuracy and provides a valuable reference for improving fault classification models in high-voltage power systems. In the Appendix Fig. 9 illustrates an example highlighting transient effects and the subsequent post-fault stabilization.

VI. CLUSTERING RESULTS

In this work, clustering is employed to identify distinct fault and event patterns within high-voltage power systems. The number of clusters is set to 15, reflecting the diversity of known fault types and system disturbances. This choice balances interpretability and granularity, enabling meaningful distinctions between different event categories.

A. Overview of Clustering Performance

Tables II and III present the distribution of samples across clusters, obtained using K-Means with PCA and t-SNE for dimensionality reduction, respectively. The PCA-based approach results in highly imbalanced cluster sizes, with the largest cluster containing 7364 samples (61.1%) and the smallest only 29 (0.2%). The standard deviation in cluster sizes is 1794 (14.89% of the total dataset), indicating that a few large clusters dominate the partitioning.

Conversely, the t-SNE-based clustering produces a more uniform distribution, with cluster sizes ranging from 471 (3.9%) to 1169 (9.7%) samples, and a significantly lower standard deviation of 184 (1.53%). These differences underscore the impact of dimensionality reduction: while PCA preserves global variance leading to imbalanced clusters, t-SNE emphasizes local structure, resulting in a more balanced clustering outcome.

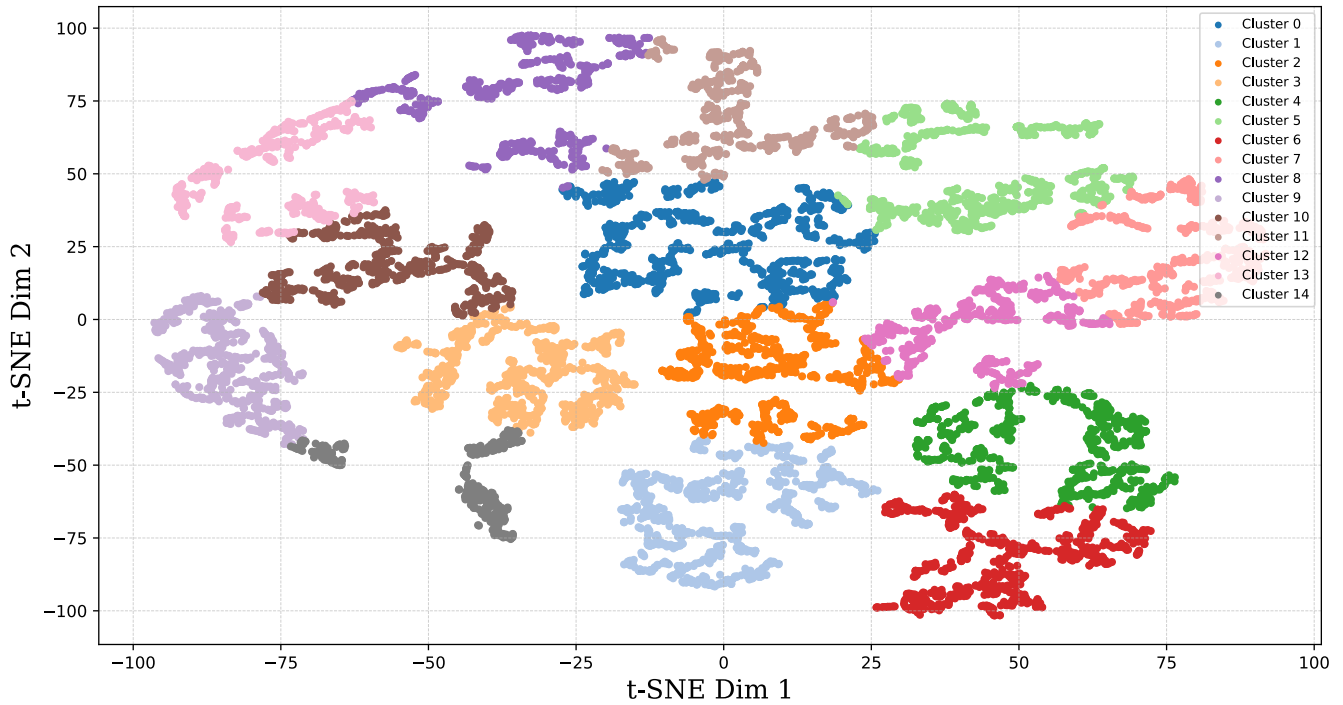


Fig. 1. t-SNE visualization of clustering results in two-dimensional space. Each point represents a sample, with colors indicating clusters identified by K-Means. This visualization provides insight into the cluster distribution and separability of fault events.

TABLE II
NUMBER OF SAMPLES PER CLUSTER USING K-MEANS PCA LABELS ($N = 12,053$)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Count	7364	1626	704	439	411	300	222	207	164	144	143	109	98	93	29
%	61.1%	13.5%	5.8%	3.6%	3.4%	2.5%	1.8%	1.7%	1.4%	1.2%	1.2%	0.9%	0.8%	0.8%	0.2%

TABLE III
NUMBER OF SAMPLES PER CLUSTER USING K-MEANS T-SNE LABELS ($N = 12,053$)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Count	1169	1141	938	916	890	849	821	753	752	741	725	682	626	579	471
%	9.7%	9.5%	7.8%	7.6%	7.4%	7.0%	6.8%	6.3%	6.3%	6.2%	6.0%	5.7%	5.2%	4.8%	3.9%

TABLE IV
GLOBAL CLUSTERING EVALUATION METRICS ($N = 204$)

Clustering Method	Purity (\uparrow)	Entropy (\downarrow)	Silhouette (\uparrow)
K-Means PCA	0.608	0.767	-0.744
K-Means t-SNE	0.721	1.095	-0.399

Figure 2 shows the distribution of event types within the manually labeled subset ($N = 204$), providing context for the fault categorization.

B. Quantitative Evaluation of Clustering

We evaluate the clustering quality using several standard metrics ([8]–[12]):

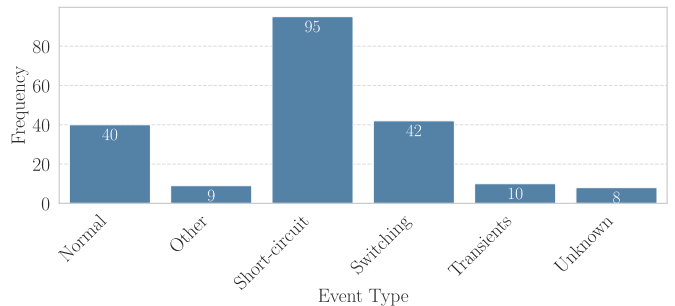


Fig. 2. Distribution of the event types in the manually labeled subset ($N = 204$).

Purity measures how well each cluster consists of a single

fault type:

$$\text{Purity} = \frac{1}{N} \sum_k \max_j |C_k \cap T_j|, \quad (10)$$

where C_k is the set of samples in cluster k , T_j is the set of samples of fault type j , and N is the total number of samples. Higher values (closer to 1) indicate better purity.

Entropy quantifies uncertainty within a cluster:

$$H(C_k) = - \sum_j p_j \log_2 p_j, \quad (11)$$

where p_j is the proportion of fault type j in cluster C_k . Lower values are better, with 0 indicating a homogeneous cluster.

Silhouette Score measures the compactness and separation of clusters, ranging from -1 to 1. Higher values indicate well-defined, distinct clusters, while values closer to 0 suggest overlapping clusters, and negative values indicate poor clustering.

These metrics were computed for both PCA and t-SNE clustering methods, with results summarized in Tab. IV. Detailed results for each cluster are available in Appendix Tab. VI.

C. Analysis of Manually Labeled Clustering Results

Before clustering, 100 randomly selected samples are manually labeled by an expert to gain an initial understanding of the signals.. Additionally, after clustering, 7 samples from each of the 15 clusters are randomly selected (one sample was excluded as it was unusable). This results in $N = 204$ manually labeled samples, enabling further analysis of the clusters.

Fig. 3 and Fig. 4 present heatmaps showing the percentage distribution of event types within each cluster for the PCA-based and t-SNE-based approaches, respectively. In each case, the percentages per cluster sum to 100%, providing a visual representation of how faults are distributed across clusters.

The results highlight the impact of the chosen dimensionality reduction technique on clustering performance and provide insights into the internal consistency and separation of the identified clusters.

VII. DISCUSSION

In this study, we applied unsupervised clustering to a large, unlabeled dataset to extract meaningful insights without prior knowledge of the data. In real-world applications, experts often lack the time and resources to manually review each individual sample or signal. The goal of our approach is to provide a general understanding of the dataset and identify a relevant subset for further investigation. To the best of our knowledge, no similar methods have been applied in this context, and we propose a systematic approach to address this gap.

Our methodology involved using the FFT to extract frequency-domain features, followed by PCA for dimensionality reduction. We also applied t-SNE for visualization, and K-Means clustering was employed.

From the initial clustering results, two key observations emerged. First, clustering directly on PCA features resulted in highly imbalanced cluster sizes (see Table II), while applying

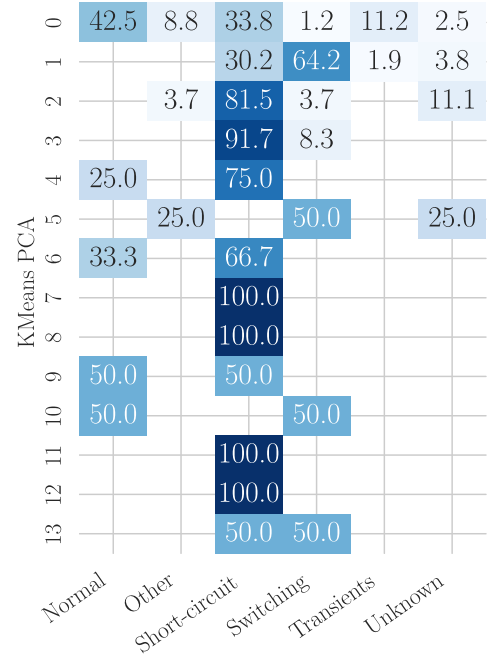


Fig. 3. Percentage distribution of event types per K-Means PCA cluster in the manually labeled subset ($N = 204$). Each cluster sums to 100%.

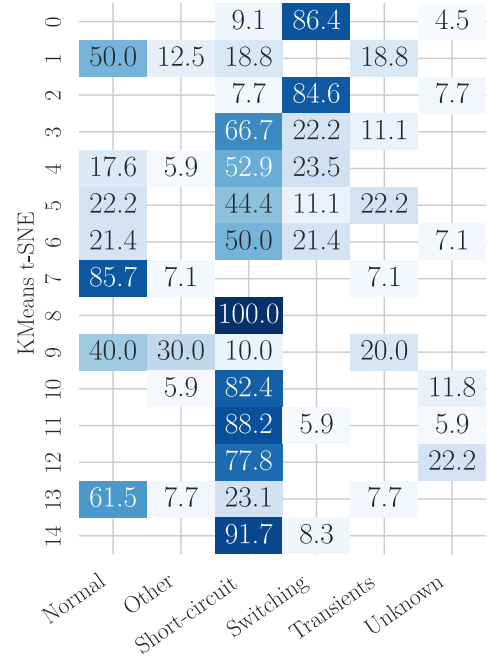


Fig. 4. Percentage distribution of event types per K-Means t-SNE cluster in the manually labeled subset ($N = 204$). Each cluster sums to 100%.

TABLE V
CLUSTERING METRICS AND WEIGHTED STATISTICS FOR K-MEANS PCA AND T-SNE.

Clustering Method	Purity	Entropy	Silhouette
K-Means PCA (Raw)	0.730 ± 0.216	0.767 ± 0.585	-0.564 ± 0.510
K-Means t-SNE (Raw)	0.708 ± 0.189	1.095 ± 0.572	-0.396 ± 0.151
K-Means PCA (Weighted)	0.608 ± 0.182	1.336 ± 0.555	-0.744 ± 0.222
K-Means t-SNE (Weighted)	0.721 ± 0.184	1.073 ± 0.557	-0.399 ± 0.153

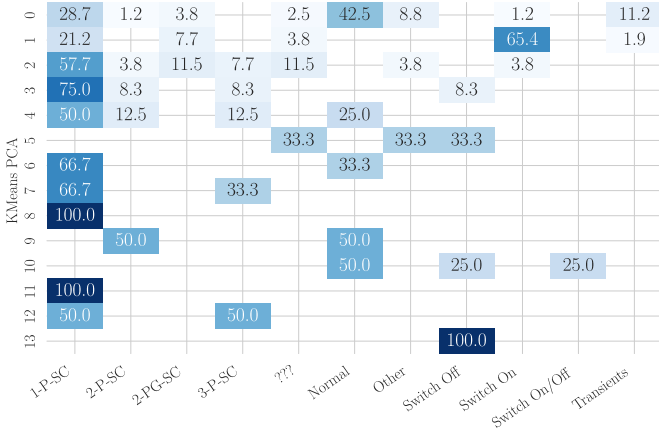


Fig. 5. Percentage distribution of fault classes per K-Means PCA cluster in the manually labeled subset ($N = 200$). Each cluster sums to 100%.

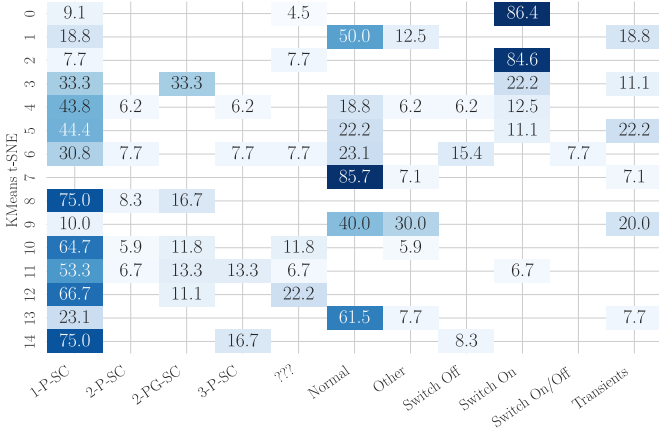


Fig. 6. Percentage distribution of fault classes per K-Means t-SNE cluster in the manually labeled subset ($N = 200$). Each cluster sums to 100%.

K-Means to the t-SNE features—after additional reduction via PCA—yielded more balanced cluster sizes (see Table III).

For a quantitative evaluation, we computed several clustering performance metrics, including purity, entropy, and silhouette score. The purity scores were relatively high, with values of 0.608 for PCA and 0.721 for t-SNE, indicating that the clusters are somewhat homogeneous with respect to the fault types. However, the silhouette scores were lower, with 0.113 for PCA and 0.115 for t-SNE, suggesting that the clusters are not well-separated and exhibit significant overlap. It is important to note that these scores were calculated on a small subset of the entire dataset ($N=204$), and may not fully

reflect the clustering performance on the entire dataset.

When analyzing the individual clusters, it is essential to consider the varying cluster sizes. For PCA-based clusters, the largest cluster, containing 80 samples, performed poorly, with a purity score of 0.425. In contrast, for t-SNE-based clusters, the largest cluster (cluster 0) achieved a much higher purity score of 0.864, despite containing only 22 samples.

The summary statistics for K-Means clustering using both PCA and t-SNE features are provided in Table V. As shown, the raw clustering metrics for K-Means with PCA features resulted in a purity score of 0.730 ± 0.216 , an entropy of 0.767 ± 0.585 , and a silhouette score of -0.564 ± 0.510 . For K-Means with t-SNE features, the raw clustering metrics yielded a purity of 0.708 ± 0.189 , an entropy of 1.095 ± 0.572 , and a silhouette score of -0.396 ± 0.151 .

When considering the weighted statistics based on cluster sizes, the purity for K-Means clustering using PCA features decreased to 0.608 ± 0.182 , while the entropy and silhouette scores increased to 1.336 ± 0.555 and -0.744 ± 0.222 , respectively. For K-Means clustering with t-SNE features, the weighted purity increased to 0.721 ± 0.184 , while the entropy and silhouette scores were 1.073 ± 0.557 and -0.399 ± 0.153 , respectively.

These statistics suggest that, although the t-SNE-based clusters tend to have smaller sizes, they generally exhibit better purity compared to PCA-based clusters, emphasizing the importance of dimensionality reduction and feature transformation in improving clustering performance.

Figures 3 and 4 show that some clusters (in both methods) consist predominantly of a single fault type, such as short-circuit or switching events. The most notable difference between the two feature reduction methods is that the largest PCA cluster, which contains 80 samples from all fault types, fails to capture the distinct fault categories. In contrast, K-Means clustering based on the t-SNE features better distinguishes between different fault types. For further reference, the absolute numbers are provided in Appendix Figures 7 and 8.

VIII. CONCLUSION

This study demonstrates that K-Means clustering can effectively identify fault patterns in voltage and current signals, providing a structured approach to analyzing large-scale, unlabeled datasets. Expert validation of a subset of 204 samples out of 12 053 confirmed that the identified clusters partially align with known fault types. These results highlight the potential of unsupervised learning for fault analysis in power systems.

A key advantage of this approach is its ability to segment large datasets into smaller, structured subsets, facilitating targeted analysis. Despite its limitations, clustering enables engineers to efficiently filter signals of interest. Unsupervised clustering reduces the number of samples requiring detailed inspection, thereby improving diagnostic efficiency and enabling a more systematic evaluation of fault characteristics.

However, several limitations must be considered. The absence of prior knowledge about the dataset limits the interpretability of the clustering results, as the extracted features may not fully capture the underlying fault characteristics. Additionally, the validation process is restricted by the labor-intensive nature of manual labeling, with only 7 to 10 samples per cluster—approximately 1% of the dataset—being inspected. While these results offer an initial assessment, their statistical power remains limited, highlighting the need for larger labeled datasets for comprehensive evaluation.

Methodologically, K-Means assumes spherical cluster structures and equal cluster sizes, which may not align with the true distribution of fault patterns. The low silhouette scores suggest poor cluster separation, indicating that alternative methods, such as DBSCAN or spectral clustering, may provide a more accurate representation of fault groupings.

A key challenge in this study is the assumption that fault characteristics dominate the clustering process. In reality, hidden variables such as variations in recording devices, sensor locations, or environmental conditions may introduce significant, unmeasured influences. Without explicit control over these factors, their impact on clustering performance remains unknown, potentially introducing bias and reducing interpretability.

Our contribution lies in proposing a systematic approach for unsupervised clustering of voltage and current signals to identify fault patterns without prior knowledge of the dataset. By applying K-Means clustering to a large, unlabeled dataset and validating the results with expert feedback, we demonstrate the feasibility of unsupervised learning in fault analysis. This approach provides a scalable and structured method for analyzing large datasets, which can be further enhanced by integrating simulated fault scenarios for controlled validation.

Future research should explore hybrid models that integrate unsupervised and supervised learning to improve clustering accuracy. Incorporating simulated fault scenarios could refine the clustering process by providing controlled conditions for validation.

While unsupervised clustering cannot replace expert-driven fault classification, it offers a systematic, scalable approach for analyzing large datasets, contributing to more reliable data-driven approaches for fault analysis in high-voltage power systems. Addressing the identified challenges will further enhance the robustness of clustering techniques in this context.

ACKNOWLEDGMENT

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 535389056.

REFERENCES

- [1] C. Presvôts and T. Prevost, "Database of Voltage and Current Samples Values from the French Electricity Transmission Grid, Réseau de Transport d'Electricité (RTE), France," 2024. [Online]. Available: <https://github.com/rte-france/digital-fault-recording-database/> (<https://github.com/rte-france/digital-fault-recording-database/>)
- [2] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.
- [3] M. Bindi, M. C. Piccirilli, A. Luchetta, and F. Grasso, "A Comprehensive Review of Fault Diagnosis and Prognosis Techniques in High Voltage and Medium Voltage Electrical Power Lines," *Energies*, vol. 16, no. 21, p. 7317, Oct. 2023. [Online]. Available: <https://www.mdpi.com/1996-1073/16/21/7317>
- [4] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 100, Dec. 2022. [Online]. Available: <https://www.nature.com/articles/s43586-022-00184-w>
- [5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [6] L. v. d. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: <http://jmlr.org/papers/v9/vandermaaten08a.html>
- [7] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982. [Online]. Available: <http://ieeexplore.ieee.org/document/1056489/>
- [8] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008. [Online]. Available: <https://nlp.stanford.edu/IR-book/>
- [9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.
- [10] N. X. Vinh, J. Epps, and J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," *The Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010. [Online]. Available: <http://www.jmlr.org/papers/volume11/vinh10a/vinh10a.pdf>
- [11] L. Hubert and P. Arabie, "Comparing Partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.
- [12] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

BIOGRAPHIES

Julian Oelhaf is a PhD student at the Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), under the supervision of Dr.-Ing. Siming Bayer and Prof. Dr.-Ing. habil. Andreas Maier. He began his doctoral studies in 2024 after completing a Master’s degree in Computer Science at FAU. His research focuses on machine learning applications for fault detection and protection coordination.

Georg Kordowich studied Electrical Energy Technologies at the Friedrich-Alexander-University Erlangen-Nürnberg. After completing his Master’s Degree in 2021 he started working as a PhD Student at the Institute of Electrical Energy Systems at FAU under the supervision of Prof. Dr. Johann Jäger. His main research focus is the application of machine learning for fault detection and protection coordination.

Andreas Maier (M’05–SM’20) received a M.Sc. in computer science from the University of Erlangen–Nuremberg, Erlangen, Germany, in 2005, and the Ph.D. degree in computer science from the same institution in 2009. He is currently a Professor and Head of the Pattern Recognition Lab at the University of Erlangen–Nuremberg, where he has been a faculty member since 2015. His research interests include medical imaging, image and audio processing, digital humanities, and interpretable machine learning. Prof. Maier has developed significant tools in medical signal processing, such as the PEAKS online speech intelligibility assessment tool, and has contributed to over 4,000 patient and control subject analyses. He has received numerous awards, including an ERC Synergy Grant for “4D Nanoscope”.

Johann Jäger received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from Friedrich-Alexander-University Erlangen- Nuremberg (FAU), Erlangen, Germany, in 1990 and 1996, respectively. From 1996, he was with Power System Planning Department, Siemens, Erlangen, Germany. He was working on different fields of FACTS devices, network planning and protection systems in worldwide projects. Since 2004, he is in charge of a Full Professorship for Electrical Power Systems with the FAU University Erlangen-Nürnberg. He is author of numerous publications and books as well as inventor of several patents.

Siming Bayer received her Dr.-Ing. in Computer Science from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) in 2022, where she developed machine learning algorithms for vascular structure registration in collaboration with Siemens Healthineers. She also holds an M.Sc. in Biomedical Engineering from FAU. She is currently a Research Scientist at FAU, leading the “Data Processing for Utility Infrastructure” group at the Pattern Recognition Lab, and a Strategic Research & Collaboration Manager at Siemens Healthineers, focusing on cross-business research in digital health. Her research interests include machine learning for medical imaging, industrial data processing, and innovation management in healthcare and utility infrastructures.

APPENDIX

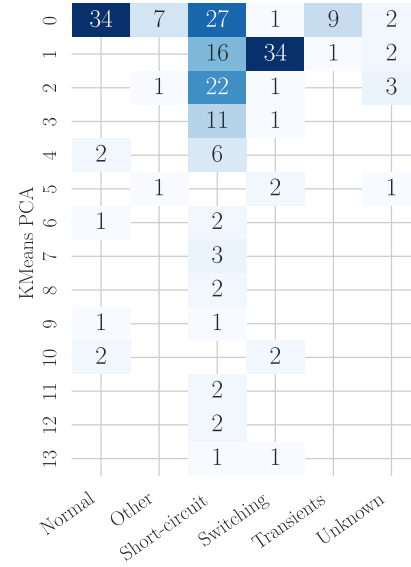


Fig. 7. Absolute number of event types per K-Means PCA cluster in the manually labeled subset ($N = 204$)

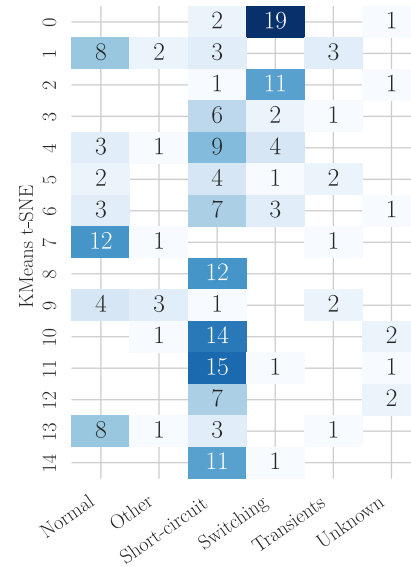


Fig. 8. Absolute number of event types per K-Means t-SNE cluster in the manually labeled subset ($N = 204$)

TABLE VI
PER-CLUSTER CLUSTERING EVALUATION METRICS ($N = 204$)

Clustering Method	Cluster	Count	Purity (\uparrow)	Entropy (\downarrow)	Silhouette (\uparrow)
K-Means PCA	0	80	0.425	1.928	-0.725
	1	53	0.642	1.219	-0.826
	2	27	0.815	0.945	-0.852
	3	12	0.917	0.414	-0.762
	4	8	0.750	0.811	-0.893
	5	4	0.500	1.500	-0.469
	6	3	0.667	0.918	-0.823
	7	3	1.000	0.000	0.207
	8	2	1.000	0.000	0.991
	9	2	0.500	1.000	-0.898
	10	4	0.500	1.000	-0.707
	11	2	1.000	0.000	-0.656
	12	2	1.000	0.000	-0.634
	13	2	0.500	1.000	-0.855
K-Means t-SNE	0	22	0.864	0.700	-0.499
	1	16	0.500	1.781	-0.439
	2	13	0.846	0.773	-0.528
	3	9	0.667	1.224	-0.519
	4	17	0.529	1.659	-0.006
	5	9	0.444	1.837	-0.115
	6	14	0.500	1.724	-0.481
	7	14	0.857	0.735	-0.398
	8	12	1.000	0.000	-0.463
	9	10	0.400	1.846	-0.496
	10	17	0.824	0.834	-0.386
	11	17	0.882	0.640	-0.515
	12	9	0.778	0.764	-0.354
	13	13	0.615	1.489	-0.480
14	12	0.917	0.414	-0.256	

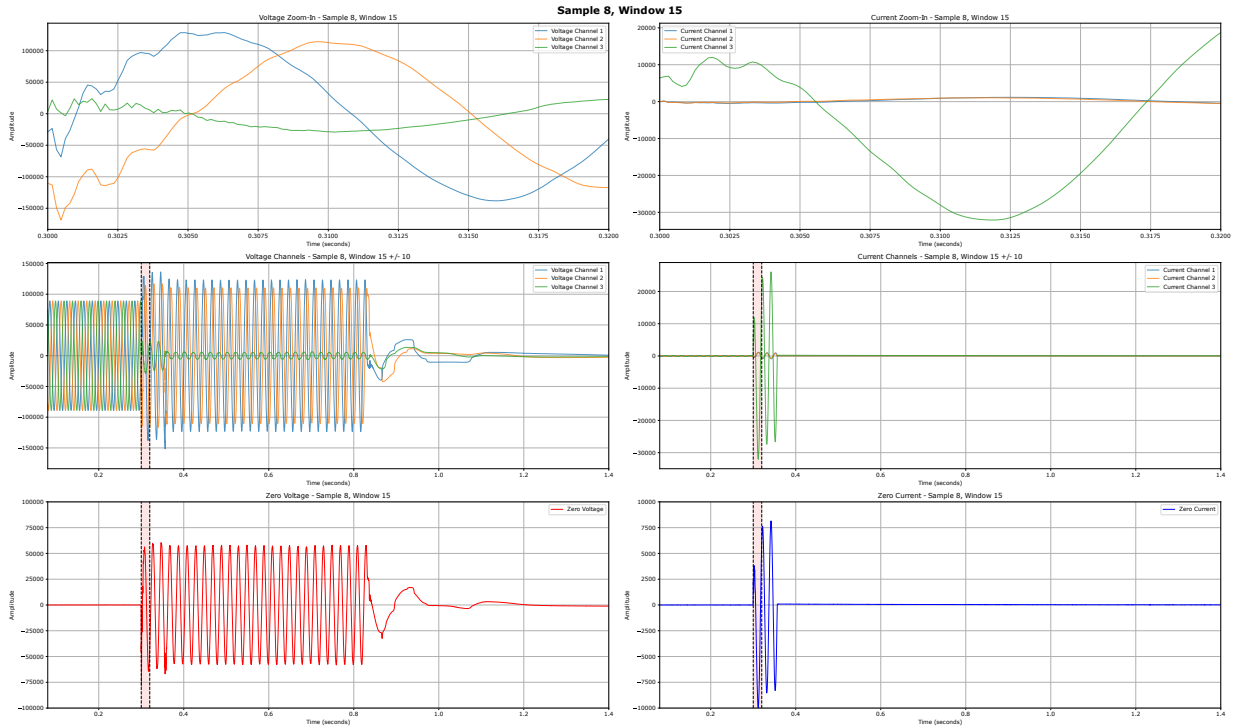


Fig. 9. Visualization of voltage and current signals for Sample 8, Window 15. The figure presents a detailed time-series analysis of three-phase voltage and current waveforms, capturing fault characteristics in a high-voltage power system. The top row provides zoomed-in views of the voltage (left) and current (right) signals within the fault event window. The middle row shows a broader perspective of the voltage (left) and current (right) waveforms surrounding the fault event, marked by vertical dashed lines. The bottom row highlights the zero-sequence voltage (left) and current (right), which help in detecting fault signatures. These signals are used to cluster and analyze fault events in an unsupervised manner.