Topological Data Analysis for fault classification on transmission lines

Eloi Gravot, Sergio Torregrosa, Nicolas Hascoët, Xavier Kestelyn, Francisco Chinesta

Electric Power Systems Research 248 (2025) 111915

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Topological Data Analysis for fault classification on transmission lines

Eloi Gravot [a,b,*], Sergio Torregrosa [c], Nicolas Hascoët [a], Xavier Kestelyn [a,b], Francisco Chinesta [a,d]

[a] Chimera RTE Chair @ PIMM Lab, Arts et Métiers Institute of Technology, 151 Bd de l'Hôpital, Paris, 75013, France
[b] L2EP Lab, Centrale Lille, Junia ISEN Lille, Arts et Metiers, University of Lille, Lille, 59000, France
[c] PIMM Lab, Arts et Métiers Institute of Technology, 151 Bd de l'Hôpital, Paris, 75013, France
[d] CNRS @ CREATE, 1 Create Way, 08-01 CREATE Tower, 138602, Singapore

ARTICLE INFO

Keywords:
Fault classification
Signal processing
Topological Data Analysis
Transmission grid
Machine Learning

ABSTRACT

This paper proposes a novel method for fault classification on transmission lines through a hybrid model combining Topological Data Analysis and unsupervised Machine Learning. Through persistent homology, signal topological signatures are extracted from each current's phase and residual current. The spatial properties of the signatures are then fed to a K-means clustering algorithm for fault classification. The method produces accurate and consistent results across a variety of fault records, even when tested under diverse parameterized faults and noise intensities. To investigate further, the model is applied to field records of the French transmission operator RTE (Réseau de Transport d'Electricité) without any parametrization or prior training. The accuracy reflects the generalization abilities of the approach.

1. Introduction

With societies having a growing reliance on electricity, transmission grid operators have to adapt their assets to receive more power while maintaining constant availability to their direct clients and distribution operators. This challenge is paired with the integration of renewable energy on the grid and their intermittent and decentralized electricity production, increasing operation's complexity. Optimizing the energy flux is highly linked to fault events on the grid and how they are handled. Shunt faults on transmission lines can happen due to various external factors (lighting, vegetation, construction work...) and lead to the isolation of the line through circuit breakers. Line isolation, due to Kirchhoff's law, changes the power flux on the network, especially on neighboring lines which can induce a cascade effect of faults on the system. It is therefore necessary to reduce the delay between the fault detection and its clearance. In fault diagnosis, once the fault is detected and the line isolated, a classification procedure is executed. Its goal is to identify which phase is affected by the shunt fault (and whether the ground is implied) based on the fault record (comtrade file). This step is necessary to apply one-ended impedance fault location methods that take the fault class as input. Classifying the fault can also be the first step into identifying the source of the short circuit. There are four different types of shunt faults in a three-phase alternating current system: Phase-to-Ground fault (PG), Phase-to-Phase fault (PP), Phase-to-Phase-to-Ground fault (PPG), and Three-Phase fault (3P) (including Three-Phase-to-Ground fault). For simplicity, the term "phase" is used instead of "line" throughout this paper to denote conductors in a three-phase system.

Fault classification on three-phase transmission lines is a widely studied topic. The main research axes that stand out are frequency based algorithms, and feature extraction combined with Machine Learning (ML) approaches. Frequency based algorithms combine classic frequency analysis methods with IF-THEN decision algorithms. Wavelet Transform (WT) is commonly studied for its ability to make frequency feature extraction on non-stationary time series through convolution of a mother wavelet [1] where more classical methods such as Fourier Transform (FT) fail to capture time varying characteristics without segmenting the signal [2]. That characteristic makes it therefore suitable for fault analysis and classification [3,4]. Various signal transformations can be added to study ground fault such as Clarke Transform or Karrenbaur [5]. The main drawback from such methods is that these frequency-based algorithms entirely depend on the thresholds values, which may lack robustness and generalization without thorough investigation and calibration when applied to different setups and scenarios. It can become problematic for transmission operators that face a wide variety of system topology. Furthermore, WT can be sensitive to noise and harmonics. Those issues may require parametrization of the mother wavelet according to the system topology and frequency. This

* Corresponding author.
E-mail addresses: eloi.gravot@ensam.eu (E. Gravot), sergio.torregrosa_jordan@ensam.eu (S. Torregrosa), nicolas.hascoet@ensam.eu (N. Hascoët), xavier.kestelyn@ensam.eu (X. Kestelyn), francisco.chinesta@ensam.eu (F. Chinesta).

https://doi.org/10.1016/j.epsr.2025.111915
Received 17 March 2025; Received in revised form 2 June 2025; Accepted 3 June 2025
Available online 23 June 2025

Abbreviations	Definition
PG	Phase-to-Ground fault
PP	Phase-to-Phase fault
PPG	Phase-to-Phase-to-Ground fault
3P	Three-Phase fault
3PG	Three-Phase-to-Ground fault
AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DT	Decision Trees
EMT	Electromagnetic Transient
EWT	Empirical Wavelet Transform
FT	Fourier Transform
GAF	Gramian Angular Field
GCNN	Graph Convolutional Neural Network
GS	Gradient Similarity
KNN	K-Nearest Neighbors
ML	Machine Learning
NB	Naive Bayes
PCA	Principal Components Analysis
PD	Persistence Diagram
PSO	Particle Swarm Optimization
RTE	Réseau de Transport d'Électricité
SNR	Signal to Noise Ratio
SVM	Support Vector Machine
TDA	Topological Data Analysis
WT	Wavelet Transform

leads to exhaustive studies, lowers generalization and brings higher computational time with high sampling rates [6]. However, threshold-based methods have the benefits of having very low computational cost (compared to data driven ones), high interpretability and are straightforward to implement.

To enhance generalization, Artificial Intelligence (AI) methods have greatly expanded in the field of fault classification in the last years with many different approaches, often combining frequency based feature extraction with supervised ML. Such hybrid approach allows models to capture critical frequency-domain characteristics of the fault signals, enhancing model performance and interpretability, which raw data alone may not provide effectively. Artificial Neural Networks (ANN) and their variations are heavily used for their capacities to learn complex, non-linear relationships in data through multiple layers of interconnected nodes (neurons), adjusting weights via backpropagation. A common architecture in fault classification in power system is the combination of Fourier or Wavelet Transform with an ANN [7–10]. There are many variations of ANN that have been developed, each with their own particularities to tackle specific problems. Recent studies have widely focused on Convolutional Neural Networks (CNN). Through convolution and pooling layers, CNN have proven to be efficient in extracting spatial features of images, or in the discussed topic, on frequency analysis outputs. Their uses have been studied for fault classification, with various particularities and feature extraction process [11–17]. Graph Convolutional Neural Network (GCNN) is another example designed to operate on graph-structured data, enabling the extraction of features and relationships from nodes and edges within the graph which can be used for fault diagnosis [18,19]. Other publications use Support Vector Machine (SVM) combined with WT [20,21]. Those models work by finding the hyperplane that best separates data points of different classes in a high-dimensional space, maximizing the margin between them. Some publications address the issue of parameterizing wavelet to make them more generalizable and adaptable to various configurations and topologies. In [22], Particle Swarm Optimization (PSO) is used to identify the WT mother wavelet and therefore counters the WT parametrization. The approach is combined with different supervised ML models : Decision Tree (DT), SVM, Naive Bayes (NB) and one semi-supervised K-Nearest Neighbors (KNN). The paper [23] also avoids WT parametrization through Empirical Wavelet Transform (EWT). To lower computational cost, Principal Components Analysis (PCA) is performed, followed by different supervised classification ML models : KNN, SVM, DT and NB.

All the mentioned methods achieve great fault classification accuracy (above 99% in the vast majority) on their respective test datasets. However supervised ML methods rely on large datasets to train and learn nonlinear bonds between signals and fault classes, which also implies large computational costs [24]. It can become problematic for transmission line fault classification where field records are rare. As a result, models are often trained and tested on a single network setup on Electromagnetic Transient (EMT) software. Of all the mentioned methods, only two studies included some field data in their results [9,19]. In practice, power line faults are uncommon events, particularly when it comes to multiple faults occurring on the same line. This rarity creates a gap between the controlled conditions in which some algorithms are developed/trained and the challenges of using them in real-world situations. This issue highlights the potential lack of generalization and industrial application. In addition, only few studies displayed results when facing noised data [11,15–18,23] which raises concerns : real-world transmission line data is inherently noisy, and models trained on clean, noise-free data may fail to perform reliably in practical scenarios. Deep ML models also lack explainability by working as black boxes which is problematic for end users who end-up working with algorithms that are opaque in terms of decisions making. However, it is important to note that some of these methods do not solely address classification, but also either detection and/or location, which adds complexity.

This paper aims to display a novel approach for fault classification on transmission lines by combining Topological Data Analysis for feature extraction and K-means clustering algorithm for classification. With only the three current phase signals as inputs, neither a vast amount of data, signal pre-processing, nor parameter tuning is needed to achieve accurate, consistent, and real-time results across a variety of fault records. The method is tested on several EMT fault simulations with various noise levels and is compared against methodologies presented in other publications. To further demonstrate the generalization performance of the method, results on raw field data from RTE (Réseau de Transport d'Électricité) will be displayed without any parameter adjustment nor training process.

The main contributions of this article are :

- Development of a new method using Topological Data Analysis and persistent homology for feature extraction of signal which is, to authors' knowledge, a first in the domain of fault diagnosis in power system.
- Use of unsupervised Machine Learning models insuring simplicity, ease of implementation, low computational cost and better interpretability compared to deep supervised ML.
- A robust, signal processing and tuning free model that can be directly applied to field data with different sampling frequencies, system frequencies, without pre-processing or training, while maintaining highly accurate real time results and using current records only.

2. Topological data analysis

Topological Data Analysis (TDA) is the study of geometrical properties and spatial relations unaffected by continuous changes and linear transformations of shape or size of data [25]. In order to extract topological features of a dataset, persistent homology is used [26], through the so-called lower star filtration when it comes to analyzing time series. One way to graphically represent that filtration is through

a water rising on a smooth function on the $y-axis$ from the global minimum to the global maximum. Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth function with multiple non-degenerate critical points (local minimums and maximums). When the water level reaches a local minimum, it is the birth of a topological feature, and when a local maximum is reached, two features merge, and the youngest one dies. The process goes on until the water level reaches the global maximum. Topological features are then put in a Persistence Diagram (PD) where the $x-axis$ is birth ($f(t)$ values of minimums), and y-axis death ($f(t)$ values of maximums). Each point verifies $y > x$ (birth occurs before death), and the closer the point to the curve $x = y$, the smaller its lifetime ($y - x$). This entire process is done with the Python Ripser library [27].

The following example illustrates the persistent homology process for topological feature extraction on a time series Fig. 1. By following the lower star filtration, three topological features can be seen $S = \{(-2, 4), (0, 1), (1.5, 3)\}$. The features are then placed into the persistence diagram $PD(S)$ as in Fig. 2. H_0 points refer to 0-dimensional homology classes (only class represented in 1-dimensional data).

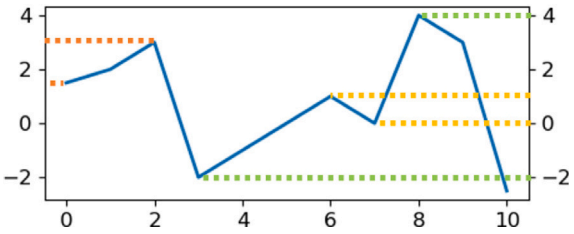


Fig. 1. Time series and its critical points.

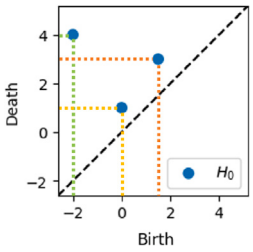


Fig. 2. Persistence diagram.

There is a growing interest towards TDA, especially for time series classification in various fields. In [28], TDA is used to classify atrial fibrillation using a piezoelectric sensor. In [29] uses TDA to classify time series that translate tape surfaces and [30] combines it with ML for structure health monitoring. The electrical engineering field is starting to apply such methods like in [31] where TDA is applied to detect and classify motor eccentricity faults by taking as inputs the three-phase currents.

It becomes clear that a time series with a variety of topological features with different births, deaths and lifespans will lead to a more spread-out cloud of points in its PD whereas a sinusoidal function will have topological features with same birth and death values and therefore a compact cloud of overlapping points in the diagram.

Through its process, it is possible to remove any topological features with a lifetime below a chosen threshold, ensuring that only persistent and meaningful structures are retained. Fig. 3 displays an example of a noised sinusoidal function. To clear the PD from noise, a threshold of 2 is applied : each point that has $death-birth < 2$ is erased. When it comes to alternative signal, positive minimums and negative maximums can also be considered as noise, therefore, topologies with a positive birth or negative death are erased.

Furthermore, TDA is unaffected by any temporal variation such as frequency change or phase shift but also sampling frequency. An example is displayed in Fig. 4, where frequency and sampling rate are changed on a sinusoidal function. As can be seen, the PD maintain the same signature's shape. The only difference is the number of points overlapping.

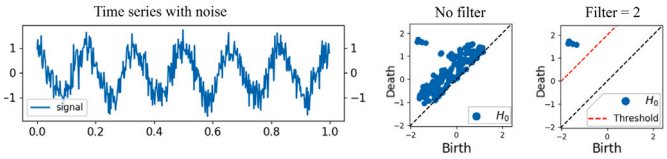


Fig. 3. Application of a TDA filter on a sinusoidal function. The threshold applied on the last PD removes all points with a lifetime below it. Only the main features are kept with birth and death around $-1,8$ and 1.8 respectively (lifetime $\approx$3.6).

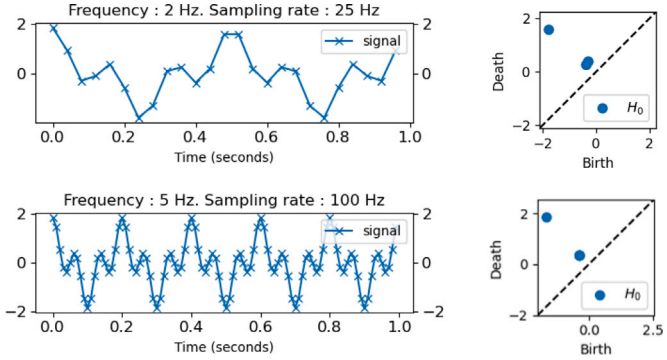


Fig. 4. Time series and persistent diagrams.

3. System modeling

The system built on EMTP Works is a 3-phase 50 Hz, 220 kV transmission line with a length of 110 km. The line model is Frequency Dependent (FD) based on [32] with the following Impedance : $Z_0 = 0.198 + i1.111\ \Omega/\text{km}$ and $Z_1 = 0.067 + i0.396\ \Omega/\text{km}$. The line current signals are recorded at one end of the line at a 100 kHz sampling frequency. 11 types of fault are simulated : 3 PG, 3 PP, 3 PPG, and one 3P. They are located from 10 to 100 km in steps of 10. Different fault resistance (R_f) values are applied : 0.01, 1, 10, 50 and 100 ohm. Finally, 6 fault inception angles are applied (from 0° to 300° in step of 60°). Fault duration is 200 ms. Table 1 summarizes the simulation scenario parameters.

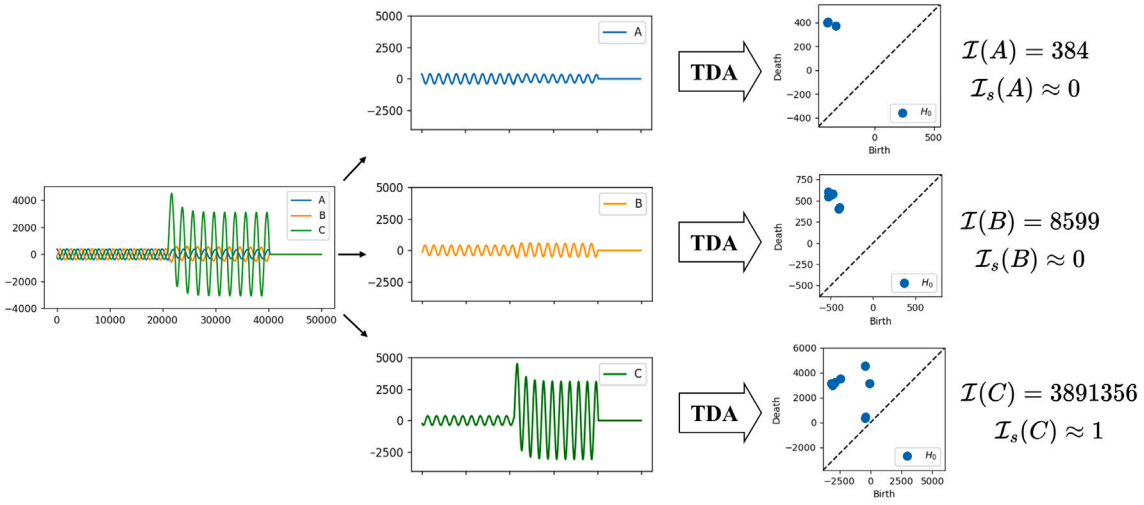
Table 1
Simulation scenarios.

Parameters	Values	Count
Fault type	A-G, B-G, C-G, AB, AC, BC, AB-G, AC-G, BC-G, ABC, ABC-G	11
Fault location	10, 20, 30, 40, 50, 60, 70, 80, 90, 100 (km)	10
Fault resistance	0.01, 1, 10, 50, 100 (Ω)	5
Fault inception angle	0, 60, 120, 180, 240, 300 (degrees)	6
Total		3300

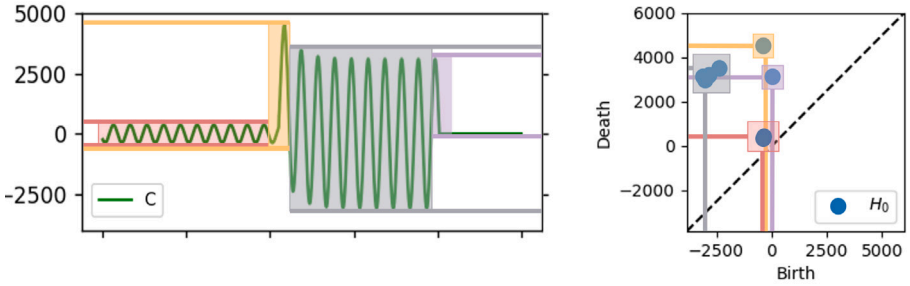
4. Method

4.1. Feature extraction with TDA

Records of faults on transmission lines are designed to start recording before the fault occurs to capture the pre-fault conditions and provide a complete analysis of the system's behavior leading up to and during the fault. Pre-fault current magnitude can therefore be extracted. Different methods are available to extract this magnitude, the following is chosen ; with known frequency and timestamps, the maximum and minimum value can be extracted from the first wavelength. TDA is applied to each phase of the fault current separately. To minimize the impact of noise, topological features with a lifespan below the pre-fault magnitude are discarded by setting the TDA threshold to this



a) Feature extraction : from time series to Inertia



b) Detailed TDA on phase C

Fig. 5. Feature extraction process with TDA on a C-G fault.

magnitude. A margin of 10% is applied by multiplying the threshold by 0.90, allowing for a slight inclusion of topological features just below the pre-fault magnitude without being overly strict. Various margin percentages were tested under different Signal to Noise Ratio (SNR) to assess robustness. This analysis will be displayed in the Result section.

The outputs of this first process are 3 PDs, one for each phase of the simulation. As mentioned in the TDA section, a sinusoidal signal will have as many topological features as it has oscillations, with the same birth and death value for all of them. Therefore, the resulting PD will be a dense overlapping point cloud. On the contrary, a signal with variations in magnitude will result in a spread out point cloud. The idea to detect if a phase is involved in the fault is to quantify how much the PD is spread out. There are different ways of computing such. The method used here can be referred as the inertia $\mathcal{I}$ and is the average of the squared Euclidean distances between the n points $X_i = (x_i, y_j)$ and the cloud's mean/centroid $\bar{X} = (\bar{x}, \bar{y})$ (1).

$$\mathcal{I} = \frac{1}{n}\sum_{i=1}^{n} d(X_i, \bar{X})^2 \tag{1}$$

The inertia of the each phases' diagram is computed and standardized to ensure that the sum of the 3 values is equal to 1 as Eq. (2) displays for the standardized inertia of phase A. The complete feature extraction process is displayed in Fig. 5.

$$\mathcal{I}_s(A) = \frac{\mathcal{I}(A)}{\mathcal{I}(A) + \mathcal{I}(B) + \mathcal{I}(C)} \tag{2}$$

4.2. Classification with K-means

The three numerical values $\{\mathcal{I}_s(A), \mathcal{I}_s(B), \mathcal{I}_s(C)\}$ of each simulation are fed to the K-means clustering algorithm with k set to 7 ; implication of ground is not considered in a first time, two-phase and two-phase to ground are gathered together (same for three-phase and three-phase to ground). Therefore, the 7 classes are : A-G, B-G, C-G, AB, AC, BC, ABC. For a given healthy phase X_h, $\mathcal{I}_s(X_h)$ is expected to be close or equal to 0 as it can be seen on the C-G fault of Fig. 5 where $\mathcal{I}_s(A) \approx \mathcal{I}_s(B) \approx 0$. Conversely, the sum of $\mathcal{I}_s$ of faulted phases X_f is expected to be close or equal to one. For single-phase fault $\mathcal{I}_s(X_{f,1}) \approx 1$ (Phase C on Fig. 5), for two-phase faults $\mathcal{I}_s(X_{f,1}) \approx \mathcal{I}_s(X_{f,2}) \approx 0.5$ and for three-phase faults $\mathcal{I}_s(X_{f,1}) \approx \mathcal{I}_s(X_{f,2}) \approx \mathcal{I}_s(X_{f,3}) \approx 0.33$. Therefore, the cluster with the centroid closest to (1,0,0) is labeled as A-G, (0.5, 0.5, 0) as AB, (0.33, 0.33, 0.33) as ABC, etc... This post-clustering assignment allows the unsupervised K-Means to align with the predefined fault classes, enabling a meaningful interpretation of the clustering results. K-Means algorithm is applied without any specific parameter tuning or initialization bias. The initialization method is random. Finally, no random seed is specified, ensuring truly stochastic behavior for centroid initialization.

Despite being both unbalanced faults, two-phase faults (LL) display much lower residual current I_{res} (3) values than two-phase to ground ones (LL-G) [33]. Three-phase to ground are an exception, by being symmetrical faults ; their ground implication is not studied. Because load is never perfectly balanced, setting thresholds on I_{res} to indicate if the ground is implied or not is problematic and lacks robustness.

$$I_{res} = I_A + I_B + I_C \tag{3}$$

Therefore, if a record is classified as two-phase by the previous K-means, TDA is applied to I_{res}. Then, inertia $\mathcal{I}(res)$ is computed and standardized (with $\mathcal{I}(A)$, $\mathcal{I}(B)$ and $\mathcal{I}(C)$). The resulting $\mathcal{I}_s(res)$ for two phase faults is expected to be close to 0, and positive value for two-phase to ground faults. To address the clustering problem where one cluster is near zero and the other contains sparse positive values, a

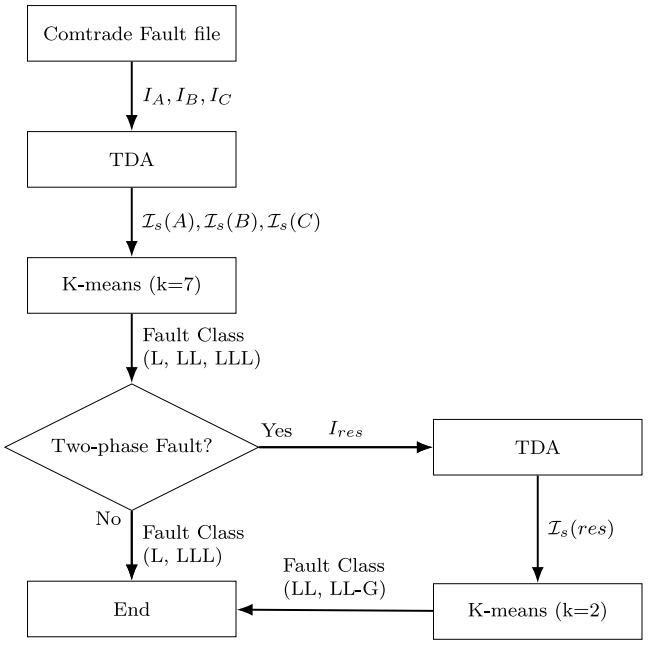


Fig. 6. Fault classification process based on TDA.

centered sigmoid transformation is applied before running K-Means (k = 2) on $\mathcal{I}_s(res)$ values (4).

$$f_\lambda(x) = \frac{1}{1 + e^{-\lambda x}} \tag{4}$$

This non-linear transformation compresses large positive values and enhances the distinction near zero, reducing the dominance of Euclidean distances in the original space. As a result, the transformed distances better reflect the cluster structure, allowing K-means to separate the two groups more effectively. The chosen value for λ is 50, but like for the TDA threshold margins, the result section will display analysis with various values of λ to assess robustness. The output of this second K-means on $\mathcal{I}_s(res)$ finishes the classification process between two-phase and two-phase to ground fault. The complete method is displayed in Fig. 6.

5. Results and discussion

In this section, this performances of the model for fault classification on transmission line are investigated under different parameter values (TDA threshold margin and sigmoid λ) and different noise levels as well as on field data. Furthermore, a comparative study is performed with various existing fault classification methods.

Since the model uses unsupervised K-means clustering, there is no need to separate test data for performance evaluation. The clustering outcome is assessed based on the structure and separability of the data itself, and performance can be evaluated directly on the entire dataset. However, it is still possible to assess its performance on new data by assigning a new point to the closest precomputed centroids. In this case, the centroids obtained during the initial clustering process remain fixed, and new data points are classified based on their distance to these centroids. This process will be applied on the field data later.

5.1. Preliminary results

This first subsection displays results on noise-free data with a 10% TDA threshold margin and $\lambda = 50$ on the sigmoid. Fig. 7 shows the classification results of the first K-means (k = 7). Each point represents

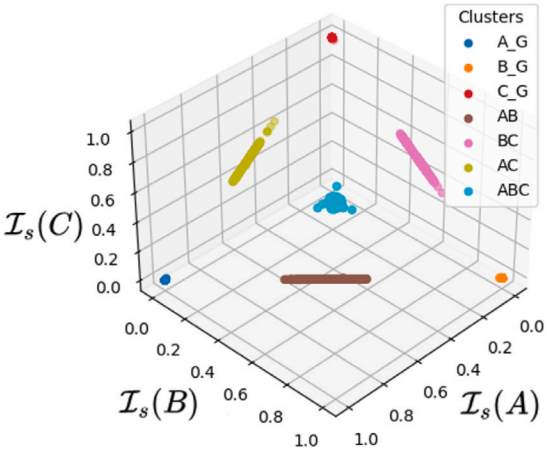


Fig. 7. Feature extraction and classification results.

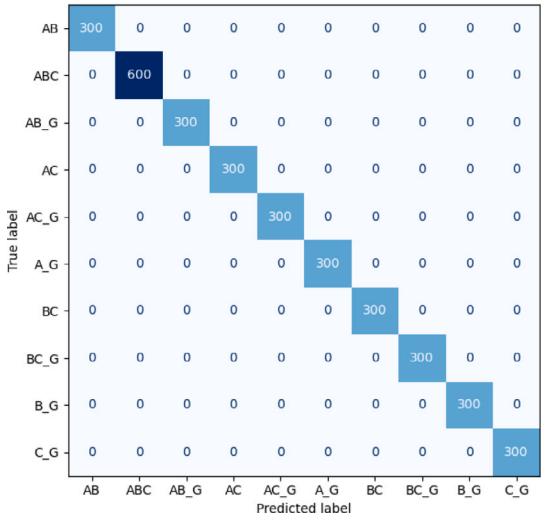


Fig. 8. Confusion matrix displaying 100% accuracy.

one simulation (there are therefore 3300 points here). The distinctiveness of the clusters observed in the visualization is a direct result of the efficient feature extraction process. The extracted features exhibit high separability, meaning that data points corresponding to different fault classes are well-separated in the feature space. This clear separation allows the K-Means algorithm to converge rapidly, typically in less than 5 iterations, as the centroids quickly stabilize near their optimal positions. The efficiency of this convergence further highlights the quality of the features in representing the underlying fault patterns, making the clustering process both effective and computationally efficient. The confusion matrix (Fig. 8) shows a 100% accuracy for fault classification.

5.2. Effect of λ on two-phase and two-phase to ground classification

Different values of λ are tested to assess the sensitivity of the parameter on the method. Results in Fig. 9 display the clustering and sigmoid function application process between AB and AB-G faults based on $\mathcal{I}_s(res)$, on noise-free records. Different λ values are tested ranging from 10 to 200. All of them allow a perfect classification of the faults even though clusters are more distinct starting at $\lambda = 50$.

5.3. Effect of TDA threshold margin

In the proposed method, getting rid of all variation with amplitude below pre-fault's one seems ideal, by setting the TDA threshold to

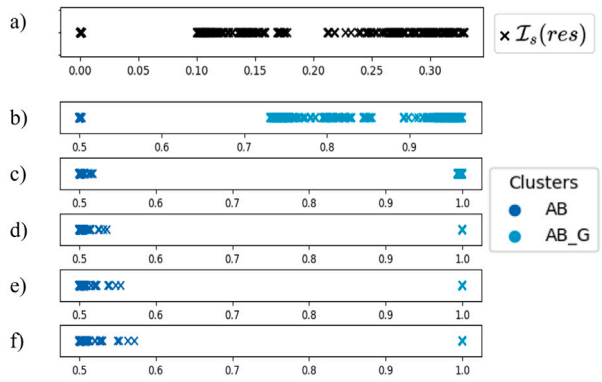


Fig. 9. Classification between AB and AB-G process with various λ values for the sigmoid. (a) Original $\mathcal{I}_s(res)$ results, (b) $\lambda = 10$, (c) $\lambda = 50$, (d) $\lambda = 100$, (e) $\lambda = 150$, (f) $\lambda = 200$. In each case, accuracy reaches 100%.

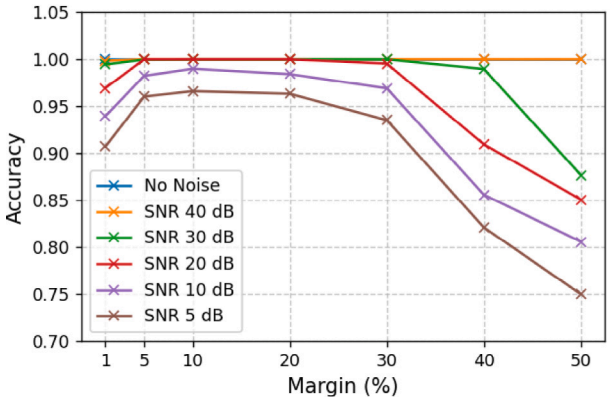


Fig. 10. Accuracy of the method under different SNR noise levels and TDA threshold margins.

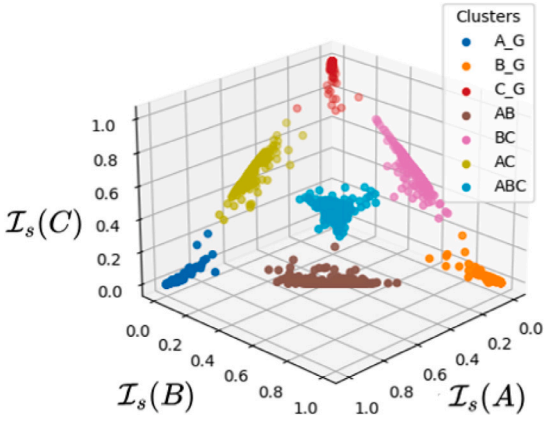


Fig. 11. Feature extraction and clustering on phases (k = 7) under 5 dB SNR.

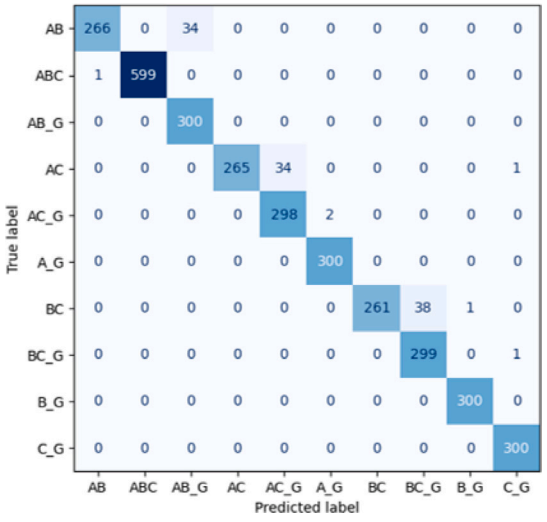


Fig. 12. Confusion Matrix under 5 dB SNR.

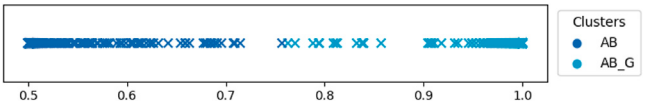


Fig. 13. Feature extraction and clustering on residual current of AB and AB-G faults (k = 2) under 5 dB SNR.

that value. However, it could be overly strict, especially when facing noise. To investigate it, the model is re-run with various SNR levels, from 40 dB to 5 dB (note: lower dB indicates higher noise, worse signal quality), and different margin values from 1% to 50%. As can be seen on Fig. 10, the method reaches at accuracy of 100% for all margin values on Noise-free and 40 dB SNR level. With stronger noise, strict margins such as the 1% or more lenient ones as 40% and 50% under-perform. Nevertheless, values ranging from 5% to 20% perform particularly on all noise levels with accuracy above 0.95. It is important to note that, as mentioned in the state of the art, the impact of noise on methods is rarely studied, and when it is, it is rarely considered under 20 dB SNR except for [16] that displays results with noise up to 10 dB SNR.

To gain a deeper insight and understanding on the impact of noise on the method, Fig. 11 displays the results of the feature extraction and confusion matrix (Fig. 12) with 5 dB SNR. The clusters become more spread out on phase inertias and residual ones compared to Fig. 7. Therefore, clusters become less distinct, particularly on $\mathcal{I}_s(res)$ (Fig. 13). As a result, the main error that can be seen in the confusion matrix are some miss-classifications between LL and LL-G faults. The model does remain accurate and robust.

5.4. Field data test

To present the generalization and potential industrial application, examples of the model on raw field records from RTE are shown in this section. No parameters are tuned, and no pre-processing is done. By using the results of previous K-means on simulated data, the records are instantly classified without any convergence process. The field dataset consists of 15 records (Fig. 14) captured on various lines with different parameters : voltage level, line length, record length and sampling frequencies (Table 2). There are 8 single-phase faults, 4 two-phase ones, 2 two-phase to ground and just one three-phase (rare event). Lastly, current field records are always measured on the secondary circuit of current transformers, to be then multiplied by the transformer's ratio. Our dataset does however contain raw records of the secondary circuit that have not been multiplied by the ratio (ID 11) but it does not impact the methodology nor requires parametrization.

Table 3 exposes the feature extraction results along the model prediction for each record. The accuracy is 100%. Even though the dataset only contains 15 records, it does display a high variety in terms of line (voltage, length), fault type and record characteristics (sampling frequency, duration). The accuracy of the proposed method supports its robustness, generalization and potential industrial application. These

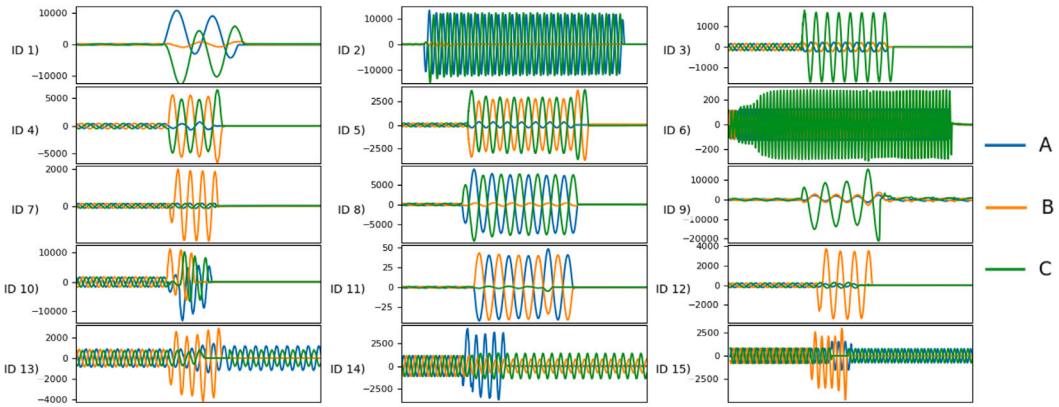


Fig. 14. Current records of faults on RTE network.

Table 2
Field data parameters.

ID	Type	Voltage (kV)	Line length (km)	Duration (s)	Sampling (Hz)
1	AC-G	400	39	0.88	1280
2	AC-G	225	14	2.12	6400
3	C-G	63	15	1.94	6400
4	BC	225	64	1.87	2000
5	BC	63	15	1.94	6400
6	C-G	63	27	1.90	1250
7	B-G	63	16	1.94	1250
8	AC	63	9	2.13	1500
9	C-G	225	14	2.12	6400
10	ABC	400	45	3.47	2000
11	AB	63	22	0.93	1600
12	B-G	63	28	3.38	6400
13	B-G	400	173	7.50	825
14	A-G	400	173	7.50	2000
15	B-G	400	173	7.50	825

Table 3
Fault classification results on field records.

ID	Type	$I_s(A)$	$I_s(B)$	$I_s(C)$	$I_s(res)$	Prediction
1	AC-G	0.61	0.00	0.39	0.39	AC-G
2	AC-G	0.47	0.00	0.53	0.26	AC-G
3	C-G	0.00	0.00	1.00	NaN	C-G
4	BC	0.00	0.54	0.45	0.00	BC
5	BC	0.00	0.47	0.53	0.00	BC
6	C-G	0.00	0.00	1.00	NaN	C-G
7	B-G	0.00	1.00	0.00	NaN	B-G
8	AC	0.50	0.00	0.50	0.00	AC
9	C-G	0.01	0.03	0.96	NaN	C-G
10	ABC	0.29	0.40	0.31	NaN	ABC
11	AB	0.53	0.47	0.00	0.00	AB
12	B-G	0.00	1.00	0.00	NaN	B-G
13	B-G	0.04	0.94	0.02	NaN	B-G
14	A-G	0.84	0.00	0.12	NaN	A-G
15	B-G	0.04	0.95	0.00	NaN	B-G

results also demonstrate the hybrid nature of the method: centroid convergence was performed on simulated data, which enabled the classification task to be carried out on real data. This approach is particularly useful for fault classification, where field data are scarce.

6. Comparative study

The proposed method (PM) is compared with different approaches in literature in Table 4. It stands out from the other by using unsupervised ML and TDA. By using such architecture, the method is much less opaque than supervised ones as it relies on simpler and more interpretable classification while also requiring less computational cost than supervised models. The performance of ML models, whether supervised or unsupervised, entirely depends on the data they are trained and tested on. It ensues accuracy above 95% on all the mentioned methods. Nevertheless, by testing our method under strong noise scenarios (up to 5 dB SNR) and on different system configurations with field data, its generalization capacity is better put to the test than the literature methods. While [5] does not require any training phase by being threshold-based, making it lightweight and straightforward to implement, its robustness is difficult to evaluate as it was not tested under varying SNR levels or fault resistance conditions. Publication [17] was tested on higher fault resistance. However, their method targets distribution systems that are more likely to experience higher fault resistance than transmission grids. Since our method was developed for transmission systems, additional studies would be required to assess its performance on distribution networks.

7. Fault detection and evolving faults

In transmission line fault classification, certain faults evolve progressively, starting from single-phase, advancing to two-phase, and culminating in three-phase faults for example. However, because the method displayed here is based on TDA, it lacks temporal segmentation capabilities, and may tend to "accumulate" faulted phases. A RTE field record starts as a three-phase faults for 40 wavelengths and then switch to two-phase fault for 10 wavelengths before circuit breakers are triggered. Such faults are rare but should still be taken into account when working on fault classification. Depending on end user's design brief of the fault diagnostic, the method could be modified to highlight that fault switch but also the fault start and end ; for example, by combining the pre-fault persistence diagram with the one of a non overlapping window of n wavelengths. Such a method is applied this fault record. The idea here is to display the possibilities of TDA concerning evolving fault and its detection, not to deliver a robust model proven on many noise levels or field data.

The record is 150 wavelengths long, the window size chosen is 5 wavelengths. The record is therefore divided in 30 sections. The first one being before the fault occurs, it can be referred as the "healthy" section. For each phase, TDA is applied to every section independently, with the same parameters as mentioned before. The resulting persistence diagrams are concatenated with the one of the healthy section. Inertia of the combined diagrams is computed and standardized. Exception is made for diagrams that have equal or lower inertia than the healthy section where the values are forced to 0 and put aside (it concerns all sections before and after the fault). That straightforward detection method allows to focus only on the faulted area of the record.

Table 4
Comparative studies with other methods.

Ref	Model architecture	Type of ML	Inputs	Max R_f (ohm)	Min SNR (dB)	Field data
[5]	WT + threshold	None	I	Not studied	Not studied	No
[11]	WT + CNN	Supervised	V + I	100[a]	Not studied[b]	No
[13]	GAF + CNN	Supervised	V + I	0.001	Not studied	No
[14]	CNN	Supervised	I	50	Not studied	No
[16]	WT + CNN	Supervised	V + I	20	10	No
[17]	GS + CNN	Supervised	V + I	500	20	No
[23]	EWT + KNN	Supervised	V + I	100	20	No
PM	TDA + K-means	Unsupervised	I	100	5	Yes

PM - Proposed Method

[a] Results below 90% accuracy with higher fault resistance.

[b] Gaussian noise used has a 10e–4 variance, which is negligible compared to SNR studied.

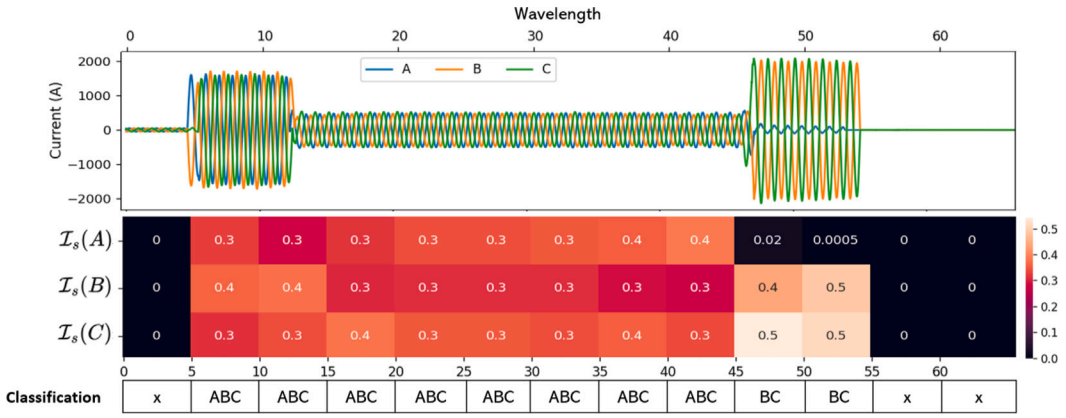


Fig. 15. Window TDA applied to a field evolutive fault. Results display the first 13 sections of the record and their standardized phases' inertia in the table below. Classification of each section is accurate as well as fault segmentation.

K-means can then be applied, the algorithm will consider each section as a record on its own and make the fault classification section by section using the centroids from the simulated data. Fig. 15 illustrates the complete process on the first 65 wavelengths of fault ID 17. The fault is three-phase starting the 5th wavelength until the 45th, and becomes BC one the following 2 sections. The algorithm computes the 3 inertia values for each sections and correctly classifies each of them on the faulted area.

8. Conclusion

This paper displayed a new approach for fault classification applied to transmission lines. The method relies of Topological Data Analysis to extract topological features of currents records in Persistence Diagrams. Through inertia computation, the feature extraction strongly highlights faulted phases and implication or not of ground in two-phase faults. The effectiveness of topological feature extraction leads to rapid convergence of the K-means algorithm, rendering the classification process highly computationally efficient. The method was tested on 3300 simulated fault records and demonstrated a perfect classification and highly accurate results under various SNR levels without any preprocessing. To further assess its generalization, the model was tested on field records of RTE, demonstrating again its performance with a 100% accuracy on raw and unprocessed Comtrade files, displaying its ease of implementation within an industrial application, with an adaptation to different system topology and record device technology. The paper also proposed a strategy that could be adopted when it comes to evolving faults and fault detection in general where, depending on end users need, outputs could be used to highlight a fault switch.

CRediT authorship contribution statement

Eloi Gravot: Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Sergio Torregrosa:** Writing – review & editing, Methodology. **Nicolas Hascoët:** Writing – review & editing, Supervision, Formal analysis. **Xavier Kestelyn:** Writing – review & editing, Supervision, Formal analysis. **Francisco Chinesta:** Writing – review & editing, Supervision, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors acknowledge the support of the Chimera RTE research chair at Arts et Métiers Institute of Technology and Porgram Descartes at CNRS @ CREATE Singapore.

Data availability

Data will be made available on request.

References

[1] P.S. Addison, The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance, Second Edition, second ed., CRC Press, Boca Raton, 2016.

[2] M. Sifuzzaman, M.R. Islam, M.Z. Ali, Application of wavelet transform and its advantages compared to Fourier transform, J. Phys. Sci. 13 (2009) 121–134.

[3] O. Youssef, Fault classification based on wavelet transforms, in: 2001 IEEE/PES Transmission and Distribution Conference and Exposition. Developing New Perspectives (Cat. No.01CH37294), Vol. 1, IEEE, Atlanta, GA, USA, 2001, pp. 531–536.

[4] R. Salim, K. De Oliveira, A. Filomena, M. Resener, A. Bretas, Hybrid fault diagnosis scheme implementation for power distribution systems automation, IEEE Trans. Power Deliv. 23 (4) (2008) 1846–1856.
[5] Y. Usama, X. Lu, H. Imam, C. Sen, N.C. Kar, Design and implementation of a wavelet analysis-based shunt fault detection and identification module for transmission lines application, IET Gener. Transm. Distrib. 8 (3) (2014) 431–441.
[6] R. Godse, S. Bhat, Mathematical morphology-based feature-extraction technique for detection and classification of faults on power transmission line, IEEE Access 8 (2020) 38459–38471.
[7] Y. Aslan, An alternative approach to fault location on power distribution feeders with embedded remote-end power generation using artificial neural networks, Electr. Eng. 94 (3) (2012) 125–134.
[8] A. Asuhaimi Mohd Zin, M. Saini, M.W. Mustafa, A.R. Sultan, Rahimuddin, New algorithm for detection and fault classification on parallel transmission line using DWT and BPNN based on Clarke's transformation, Neurocomputing 168 (2015) 983–993.
[9] K. Silva, B. Souza, N. Brito, Fault detection and classification in transmission lines based on wavelet transform and ANN, IEEE Trans. Power Deliv. 21 (4) (2006) 2058–2063.
[10] A. Abdollahi, S. Seyedtabaii, Comparison of fourier & wavelet transform methods for transmission line fault classification, in: 2010 4th International Power Engineering and Optimization Conference (PEOCO), IEEE, 2010, pp. 579–584.
[11] V. Rizeakos, A. Bachoumis, N. Andriopoulos, M. Birbas, A. Birbas, Deep learning-based application for fault location identification and type classification in active distribution grids, Appl. Energy 338 (2023) 120932.
[12] P. Rai, N.D. Londhe, R. Raj, Fault classification in power system distribution network integrated with distributed generators using CNN, Electr. Power Syst. Res. 192 (2021) 106914.
[13] B. Khabaz, M. Saad, H. Mehrjerdi, Fault classification in distribution system utilizing imaging time-series, convolutional neural network and adaptive relay protection, Electr. Power Syst. Res. 238 (2025) 111143.
[14] M.N.I. Siddique, M. Shafiullah, S. Mekhilef, H. Pota, M. Abido, Fault classification and location of a PMU-equipped active distribution network using deep convolution neural network (CNN), Electr. Power Syst. Res. 229 (2024) 110178.
[15] S. Biswas, P.K. Nayak, B.K. Panigrahi, G. Pradhan, An intelligent fault detection and classification technique based on variational mode decomposition-CNN for transmission lines installed with UPFC and wind farm, Electr. Power Syst. Res. 223 (2023) 109526.
[16] S.R. Fahim, Y. Sarker, S.K. Sarker, M.R.I. Sheikh, S.K. Das, Self attention convolutional neural network with time series imaging based feature extraction for transmission line fault detection and classification, Electr. Power Syst. Res. 187 (2020) 106437.
[17] J. Han, S. Miao, Y. Li, W. Yang, H. Yin, Faulted-phase classification for transmission lines using gradient similarity visualization and cross-domain adaption-based convolutional neural network, Electr. Power Syst. Res. 191 (2021) 106876.
[18] H. Tong, R.C. Qiu, D. Zhang, H. Yang, Q. Ding, X. Shi, Detection and classification of transmission line transient faults based on graph convolutional neural network, CSEE J. Power Energy Syst. (2021).
[19] J. Fang, K. Chen, C. Liu, J. He, An explainable and robust method for fault classification and location on transmission lines, IEEE Trans. Ind. Inform. 19 (10) (2023) 10182–10191.
[20] S. Panigrahy, A. Chandel, A hybrid framework for fault classification and location in power distribution system using wavelet and support vector machine, in: Innovation in Electrical Power Engineering, Communication, and Computing Technology, Springer Nature Singapore Pte Ltd., 2020, pp. 339–349.
[21] J.-A. Jiang, C.-L. Chuang, Y.-C. Wang, C.-H. Hung, J.-Y. Wang, C.-H. Lee, Y.-T. Hsiao, A hybrid framework for fault detection, classification, and location—part i: concept, structure, and methodology, IEEE Trans. Power Deliv. 26 (3) (2011) 1988–1998.
[22] T.S. Abdelgayed, W.G. Morsi, T.S. Sidhu, A new approach for fault classification in microgrids using optimal wavelet functions matching pursuit, IEEE Trans. Smart Grid 9 (5) (2018) 4838–4846.
[23] A.K. Pandey, N. Kishor, S.R. Mohanty, P. Samuel, Intelligent fault detection and classification for an unbalanced network with inverter-based DG units, IEEE Trans. Ind. Inform. 20 (5) (2024) 7325–7334.
[24] P. Stefanidou-Voziki, N. Sapountzoglou, B. Raison, J.L. Dominguez-Garcia, A review of fault location and classification methods in distribution grids, Electr. Power Syst. Res. 209 (2022) 108031.
[25] A.B. El-Yaagoubi, M.K. Chung, H. Ombao, Topological data analysis for multivariate time series data, Entropy 25 (11) (2023) 1509.
[26] J.E. Goodman, Surveys on Discrete and Computational Geometry: Twenty Years Later : AMS-IMS-SIAM Joint Summer Research Conference, June 18-22, 2006, Snowbird, Utah, American Mathematical Soc., 2008.
[27] C. Tralie, N. Saul, R. Bar-On, Ripser.py: A lean persistent homology library for python, J. Open Sour. Softw. 3 (29) (2018) 925.
[28] F. Jiang, B. Xu, Z. Zhu, B. Zhang, Topological data analysis approach to extract the persistent homology features of ballistocardiogram signal in unobstructive atrial fibrillation detection, IEEE Sensors J. 22 (7) (2022) 6920–6930.
[29] T. Frahi, C. Argerich, M. Yun, A. Falco, A. Barasinski, F. Chinesta, Tape surfaces characterization with persistence images, AIMS Mater. Sci. 7 (4) (2020) 364–380.
[30] A. Lejeune, N. Hascoët, M. Rébillat, E. Monteiro, N. Mechbal, An enhanced topological analysis for lamb waves based SHM methods, Struct. Heal. Monit. (2023).
[31] B. Wang, C. Lin, H. Inoue, M. Kanemaru, Topological data analysis for electric motor eccentricity fault detection, in: IECON 2022 – 48th Annual Conference of the IEEE Industrial Electronics Society, IEEE, Brussels, Belgium, 2022, pp. 1–6.
[32] J.R. Marti, Accurate modelling of frequency-dependent transmission lines in electromagnetic transient simulations, IEEE Trans. Power Appar. Syst. PAS-101 (1) (1982) 147–157.
[33] A. Aljohani, I. Habiballah, High-impedance fault diagnosis: a review, Energies 13 (23) (2020) 6447.