Zaria Journal of Electrical Engineering Technology
Department of Electrical Engineering, Ahmadu Bello University Zaria
ISSN: 3156-4895 | Vol. 15, No. 1, 2026

Underground Fault Detection and Classification: An Explainable, SCADA-Ready Pipeline with Adaptive DWT and Compact CNN-Transformer

Kabiru Ibrahim Dahiru[1,*], Ahmad Abubakar[2], Sulaiman Haruna Sulaiman[3], Yusuf Idris[4], Aliyu Lawan Musa[4], Ibrahim Mahdi[4]
[1]Sichuan Industrial and Intelligent Application Lab, Chengdu University of Technology, China
[2]Khalifa University, Abu Dhabi, United Arab Emirates
[3]Department of Electrical Engineering, Ahmadu Bello University, Zaria-Nigeria
[4]Department of Computer Engineering, Kano State Polytechnic, School of Technology, Kano-Nigeria
***Corresponding author**: kabirudahiruibrahim@kanopoly.edu.ng; +2348036559060; ORCID: 0009-0008-2881-4352

Keywords:
CNN-Transformer;
Discrete wavelet transform;
Explainable AI;
SCADA systems integration;
SHAP value attribution;
Wavelet-based feature extraction.

Article History:
Received: 20 August 2026
Revised: 26 August 2026
Accepted: 29 August 2026
Published: 11 September 2026

Article Info:
DOI: https://doi.org/10.xxxx/xxxx
Homepage: https://zjeet.com.ng
Volume: 15, No. 1, 2026
Pages: 35-52

ABSTRACT

Underground faults are difficult to detect and classify reliably because transient signatures are short-lived, dispersive, and often masked by operating noise. Deployment in supervisory control and data acquisition (SCADA) environments demands deterministic, sub-millisecond latency and transparent decision logic. This research presents an explainable, SCADA-ready pipeline that couples adaptive discrete wavelet transform (ADWT) feature extraction with a compact convolutional neural network-Transformer (CNN-Transformer) classifier. The ADWT adaptively selects the mother wavelet and decomposition depth using an energy-entropy criterion to obtain a concentrated representation of fault-transient energy, with db4 identified as the optimal wavelet for the evaluated dataset. Post-hoc explanations are produced using SHapley Additive exPlanations (SHAP), yielding phase- and sub-band-level attributions that align with expected fault physics (phase-to-ground impulses and inter-phase couplings). On a twelve-class test set generated from a controlled MATLAB/Simulink model of the 200-km underground-cable system, the hybrid CNN-Transformer attains the highest deep-model accuracy with a strong macro-precision and recall. The CNN-Transformer achieves a model-level CPU inference latency of 0.036 ms per sample, indicating computational feasibility for protection-adjacent and edge-oriented monitoring. The combined classification, latency, and SHAP results demonstrate an accuracy-efficiency-interpretability trade-off for underground-cable fault diagnosis.

1. INTRODUCTION

Continuous electrification and increasing urban load density drive extensive deployment of underground (UG) cables. Compared with overhead lines, UG assets reduce visual impact and exposure to weather, but locating and classifying faults remains challenging due to attenuation and dispersion in soil–cable structures. Wavelets has been effective for extracting fault transients, enabling sensitive detection and classification in noisy environments. Concurrently, deep learning has improved fault recognition across power systems, yet deployment in SCADA and protection-adjacent contexts demands compact models, deterministic latency, and transparent decisions. This work proposes a practical, SCADA-ready pipeline that leverages adaptive DWT features and a compact CNN-Transformer, augmented with SHAP explanations to meet operational requirements.

The core technical innovations and contributions of this work are as follows;

- SCADA-ready edge architecture through a deterministic, sub-millisecond CPU inference with a footprint for real-time deployment.
- Energy/entropy-guided adaptive DWT with polarity-invariant pre-processing feeding a lightweight CNN-Transformer that captures local impulses and long-range dependencies under tight latency/parameter budgets.
- Physics-aligned explainability with SHAP attributions at phase and sub-band levels that mirror expected UG-cable fault mechanisms, enabling operator-trustable decisions.
- Deployment-oriented validation by 12-class taxonomy, baseline comparisons, ROC and latency benchmarking, stress tests across resistance/inception angle/SNR, and a modular design ready for integration and hardware-in-the-loop trials.

1.1 Wavelet-based fault analysis in power cables and transmission systems

Wavelet-based time-frequency analysis remains widely used for power-system fault diagnosis because transient disturbances are non-stationary and contain localized spectral information that can be obscured in purely time or frequency-domain representations. Mallat's multiresolution framework established the theoretical basis for efficient wavelet decomposition through hierarchical filter banks [1]. In power-system applications, He et al. [2] combined wavelet analysis with singular entropy for transmission-line fault detection and classification, demonstrating the utility of multiscale transient-energy characterization. Wavelet-based approaches have also been applied to travelling-wave fault analysis, including real-world transmission-line measurements [3, 4]. These studies establish the suitability of multiresolution representations for extracting short-duration fault signatures.

Recent reviews identify time–frequency representation, incipient-fault detection, localization, intelligent classification, generalization, and deployment under realistic operating conditions as continuing research priorities in cable diagnosis [5, 6, 9, 10]. Shalby et al. [7] further demonstrated that wavelet-family and decomposition-level selection materially influence fault representation, emphasizing that these parameters should be chosen according to signal characteristics rather than assumed universally optimal. Entropy-based wavelet representations provide an additional mechanism for quantifying transient-energy concentration [5, 6, 8].

For underground-cable applications, wavelet and other transient-sensitive representations have increasingly been coupled with intelligent classifiers. Recent underground-cable studies have further investigated deep-learning-based detection and localization using resonance-frequency information [11], while multiresolution representations continue to be integrated with machine-learning and deep-learning classifiers in broader power-system fault studies [12, 13]. This supports systematic, signal-dependent wavelet configuration rather than assuming a universally optimal basis.

Many wavelet-based fault-diagnosis pipelines rely on predetermined wavelet configurations or feed handcrafted features directly to conventional classifiers. This limits their ability to jointly exploit local cross-scale relationships and broader dependencies among phase-dependent transient descriptors. The present work therefore evaluates the mother wavelet and decomposition depth using an energy-entropy criterion before organizing the resulting denoised multiscale descriptors into a structured representation for subsequent CNN-Transformer classification.

1.2 Deep-learning and Transformer-based fault diagnosis

Deep learning has extended fault diagnosis beyond conventional feature–classifier pipelines through hierarchical nonlinear representation learning. CNN-based architectures are effective at extracting localized fault signatures from electrical measurements, as demonstrated by recent convolutional fault-location frameworks [16]. Cable-specific research has further investigated incremental deep convolutional learning to improve generalization under changing operating conditions [14], while hybrid wavelet–deep-learning methods continue to exploit multiresolution fault representations [12]. Attention-based models provide a complementary mechanism for learning broader feature dependencies. Transformer models have been applied to power system fault diagnosis in transmission line fault detection, classification and localization [15,19]. Wu et al. [17] employed a Transformer for fault detection and classification in rotary systems, while Wang et al. [18] combined convolutional preprocessing with a Transformer for fault diagnosis in nuclear-power-plant rectifier-filter circuits. These studies support hybrid convolution-attention architectures in which convolution extracts localized patterns and self-attention captures broader dependencies.

However, directly applying Transformer architectures to high-rate raw waveforms can increase computational requirements and may be unnecessary when physically meaningful transient structure has already been isolated through multiresolution processing. The present study instead supplies the network with a compact sequence of phase-scale ADWT descriptors. A depthwise-separable 1-D convolutional tokenizer captures local relationships among neighboring feature tokens, while a shallow multi-head self-attention encoder models cross-token dependencies. This design targets a balance between representation capability and the computational constraints associated with protection-adjacent and SCADA-oriented deployment [19][20].

1.3 Explainability and deployment-oriented fault diagnosis

Despite improvements in predictive performance, deployment of deep fault-diagnosis models in operational power systems requires more than classification accuracy. Protection-adjacent analytics should also provide computationally tractable inference and interpretable evidence supporting model decisions. Explainable artificial intelligence has become increasingly relevant to condition monitoring and fault diagnosis. Yan et al. [21] investigated relationships between fault-characteristic frequencies and SHAP-based local explanations, illustrating how feature attributions can be related to physically meaningful fault signatures. Brito et al. [22] similarly applied explainable AI to fault diagnosis using transfer learning and augmented synthetic data, highlighting the value of post-hoc interpretation for understanding model predictions.

Deployment considerations introduce an additional constraint. Edge-oriented inference can reduce dependence on cloud communication and associated network latency, while supporting local operational-technology analytics for power systems fault analysis [23], transformers and PMU-based frameworks demonstrate increasing computational responsiveness alongside predictive performance [19, 20]. However, low inference time alone does not establish complete SCADA compliance; practical deployment additionally depends on communication acquisition, preprocessing, hardware and end-to-end timing requirements.

Recent research increasingly treats deployability and decision transparency as complementary requirements to classification performance. Reviews of machine learning for power-system protection emphasize robustness, computational requirements, and integration with operational protection functions as important barriers to practical adoption [24]. Attention-based architectures have also been applied to supervisory measurements, including Transformer-based fault detection directly from SCADA data [25]. In parallel, explainability is becoming an important component of intelligent grid diagnosis. Recent power-grid research has investigated explainable graph neural networks for fault detection [26]. SHAP-based diagnostic studies relate learned attributions to physically meaningful fault signatures. These developments support evaluation frameworks that jointly consider predictive performance, computational feasibility, and interpretable decision evidence.

The present work evaluates classification performance together with CPU inference latency and SHAP-based attribution. SHAP is used to examine phase- and sub-band-level contributions to the CNN-Transformer predictions, while the compact network architecture is designed to constrain inference cost. The resulting framework is positioned as SCADA-oriented and protection-adjacent, rather than as a fully validated protection relay or end-to-end SCADA implementation.

1.4 Research gap and positioning of the present work

Existing studies have independently demonstrated the effectiveness of wavelet representations for transient characterization, CNNs for local fault-pattern extraction, Transformers for modelling broader dependencies, and SHAP for post-hoc interpretation. However, these capabilities are commonly investigated as separate components. Wavelet-based cable-diagnosis methods often rely on fixed decomposition settings or conventional downstream classifiers, whereas deep-learning approaches frequently operate on raw or conventionally preprocessed measurements without exploiting adaptively selected multiscale transient descriptors. Furthermore, computational performance and explainability are not consistently considered together when assessing suitability for protection-adjacent deployment. The resulting gap is therefore not the absence of wavelet, CNN, Transformer, or SHAP-based fault diagnosis individually, but the lack of a compact framework that integrates data-dependent multiscale representation, local and cross-token feature modelling, multiclass fault discrimination, attribution-based interpretation, and deployment-oriented latency assessment within a common underground-cable evaluation pipeline.

Recent reviews of AI-based power-system fault detection similarly characterize data processing, representation design, learning architecture, and optimization as interdependent components of modern intelligent protection pipelines [27]. Nevertheless, underground-cable studies that jointly integrate adaptive multiscale representation, compact attention-based classification, latency assessment, and physically interpretable attribution remain comparatively limited.

To address this gap, the present study integrates an energy-entropy-guided adaptive DWT with a compact CNN-Transformer for twelve-class underground-cable fault classification. Candidate wavelets and decomposition depths are evaluated before extracting denoised phase-scale descriptors; a lightweight convolutional tokenizer captures local descriptor relationships, while shallow self-attention models broader cross-token dependencies. The resulting predictions are interpreted using SHAP and evaluated jointly in terms of classification performance and CPU inference latency. This formulation positions the contribution as an integrated accuracy-efficiency-interpretability framework rather than claiming novelty for any individual constituent technique.

2. METHOD

2.1 Signal model, dataset, and preprocessing

The dataset is generated using a MATLAB/Simulink model of a three-phase underground transmission system represented by a lumped π-section cable model as described in Figure 1. The test system comprises a 250-kV RMS, 50-Hz grounded three-phase source, a 200-km underground cable, and a balanced three-phase RLC load. The source resistance and inductance are 0.8929 Ω and 16.58 mH, respectively. Current measurements are acquired from the Phase-A, Phase-B, Phase-C, and ground-current channels under normal and faulted operating conditions. The simulation architecture is adapted from the underground-cable test configuration previously reported in [28], with the cable length extended to 200 km for the present study. The principal electrical, acquisition, fault-scenario, and signal-processing settings are summarized in Table 1. The study extends the signal-processing and classification stages through adaptive DWT feature selection and a compact CNN–Transformer.

The lumped π-section representation provides a computationally efficient approximation of the principal cable parameters required for the present controlled fault-classification study. It does not explicitly reproduce all frequency-dependent attenuation and wave-dispersion phenomena obtainable from distributed-parameter cable models; consequently, the present analysis focuses on the discriminative multiscale structure of the simulated fault-current transients rather than exact travelling-wave reconstruction [28].

To capture non-stationary, transient-rich current signatures, as described in Figure 2, an Adaptive Discrete Wavelet Transform (ADWT) framework was adopted (Section 2.2). The mother wavelet is determined through the adaptive selection procedure described in Section 2.2.3 rather than prescribed a priori. Among the evaluated candidate wavelets, db4 yielded the highest selection merit and was therefore retained for subsequent decomposition and feature extraction. Its compact support and favorable time–frequency localization are also well suited to abrupt fault-current transients [1, 2]. Wavelet features remain widely used for fault analysis and travelling-wave studies on practical power-system measurements [3, 4]. The DWT detail coefficients are first denoised and processed to derive compact level-wise descriptors, including log-energy, Teager–Kaiser energy, excess kurtosis, sub-band entropy, and adjacent-scale variations. These descriptors are subsequently normalized using training-set statistics and organized into the feature sequence supplied to the compact CNN-Transformer.

2.1.1 Simulation space and scenario grid

The simulation considers the twelve mutually exclusive operating classes listed in Table 1. Fault resistance is varied from 0.1 to 200 Ω, fault-inception angle from 0^0 to 360^0, and additive measurement noise is introduced to obtain SNR levels between 10 and 40 dB. Each simulation realization produces the four-channel current response $\{I_A, I_B, I_C, I_G\}$ at the measurement node. Such grids align with recent PMU and waveform-based detection pipelines in transmission and distribution lines studies [4, 16, 20]. The simulation campaign yielded 1,990 labeled observation windows distributed across the 12 mutually exclusive operating classes. As illustrated in Figure 3(b), the resulting class distribution is approximately balanced, thereby limiting substantial class-frequency bias in the subsequent classifier evaluation.

2.1.2 Signal acquisition and windowing

The acquired four-channel current signals are segmented into fixed 20-ms observation windows, corresponding to T=400 samples per channel at the 20-kHz sampling rate. Each observation is represented by the multichannel current matrix $W_i \in \mathrm{R}^{T\times C}$, with C=4 corresponding to the Phase-A, Phase-B, Phase-C, and ground-current channels, as defined in Equation (1). Windowing preserves the fault-onset and early post-fault transient information required for classification while maintaining a compact input representation [2, 3]. Additive white Gaussian noise (AWGN) is superimposed on the time-domain current measurements before ADWT processing to obtain prescribed per-channel SNR values within 10-40 dB. For each channel, the SNR over the corresponding observation window is defined by Equation (2).

$$W_i = \begin{bmatrix} x_A[n_i : n_i + T - 1] \\ x_B[n_i : n_i + T - 1] \\ x_C[n_i : n_i + T - 1] \\ x_G[n_i : n_i + T - 1] \end{bmatrix} \in \mathbb{R}^{T\times C} \qquad (1)$$

Where W_i denotes the *i-th* current window, n_i its starting sample index, T=400 is the number of samples per channel, and C=4 denotes the Phase-A, Phase-B, Phase-C, and ground-current channels.

$$\mathrm{SNR}_\phi = 10\log_{10}\left(\frac{\sum_n s_k[n]^2}{\sum_n \eta_k[k]^2}\right) \qquad (2)$$

Where SNRϕ denotes the signal-to-noise ratio of channel ϕ in decibels, $s_k[n]$ is the corresponding noise-free current vector, and $\eta_k[k]$ is the additive-noise vector over the observation window

2.1.3 Normalization and polarity invariance

After denoising and feature extraction, each derived ADWT descriptor is min–max normalized using statistics estimated exclusively from the training partition. Absolute-value processing of the detail coefficients is applied where required to obtain polarity-invariant magnitude-based descriptors before feature computation as described in Equation (3).

$$\hat{z} = \frac{z - z_{min}}{z_{max} - z_{min} + \varepsilon} \qquad (3)$$

Where $\hat{z}$ is the normalized feature value (bounded in [0, 1]), z is any scalar feature, z_{min} and z_{max} are the per-feature minima and maxima computed on the training set, and ε > 0is a small positive constant that prevents division by zero. This yields polarity-invariant, scale-stable inputs that improve training stability in wavelet-driven CNN-Transformer models [1, 3, 19].

Table 1: Electrical-system, simulation, acquisition, wavelet-processing, and classification settings used in the present study.

Category	**Parameter**	**Symbol**	**Value**	**Unit / Description**
Source	Nominal line voltage	V_s	250	kV RMS, three-phase
	System frequency	f_0	50	Hz
	Source configuration	—	(Y_n), grounded	Grounded three-phase source
	Source resistance	R_s	0.8929	Ω
	Source inductance	L_s	16.58	mH
Underground cable	Cable model	—	Lumped (\pi)-section	Three-phase underground transmission model
	Cable length	L_{cable}	200	km
	Cable resistance	R_{cable}	12	Ω
	Cable capacitance	C_{cable}	40	µF
Load	Load type	—	Balanced three-phase RLC	—
	Active power	P_L	334.8	kW
	Reactive power	Q_L	18	MVAR
Measurement	Current channels	—	(I_A,I_B,I_C,I_G)	Phase-A, Phase-B, Phase-C, and ground currents
	Number of channels	C	4	Current measurements
Sampling	Sampling frequency	f_s	20	kHz
	Sampling interval	T_s	50	µs
	Observation-window duration	T_w	20	ms
	Samples per window	T	400	samples/channel
	Fundamental cycles/window	—	1	At 50 Hz
Fault scenarios	Number of operating classes	K	12	11 fault classes + No-fault
	Fault classes	—	AG, BG, CG, AB, BC, CA, ABG, ACG, BCG, ABC, ABCG, No-fault	Single-label multiclass taxonomy
	Fault resistance	R_f	0.1–200	Ω
	Fault inception angle	θ_f	0–330	degrees
Noise conditions	Signal-to-noise ratio	SNR	10–40	dB
Wavelet processing	Candidate mother wavelets	ψ	db2, db4, db6, sym4, coif1	Adaptive search set
	Selected mother wavelet	ψ*	db4	Highest energy–entropy selection merit
	Denoising	—	Level-wise soft thresholding	Applied to DWT detail coefficients
	Extracted descriptors	—	Log-energy, mean TKEO, excess kurtosis, sub-band entropy, adjacent-scale variation	Per phase and decomposition level
Classification	Problem formulation	—	Single-label multiclass	12 mutually exclusive classes
	Network output	—	12 logits + softmax	CNN–Transformer classifier
	Training loss	—	Categorical cross-entropy	Consistent with 12-class formulation

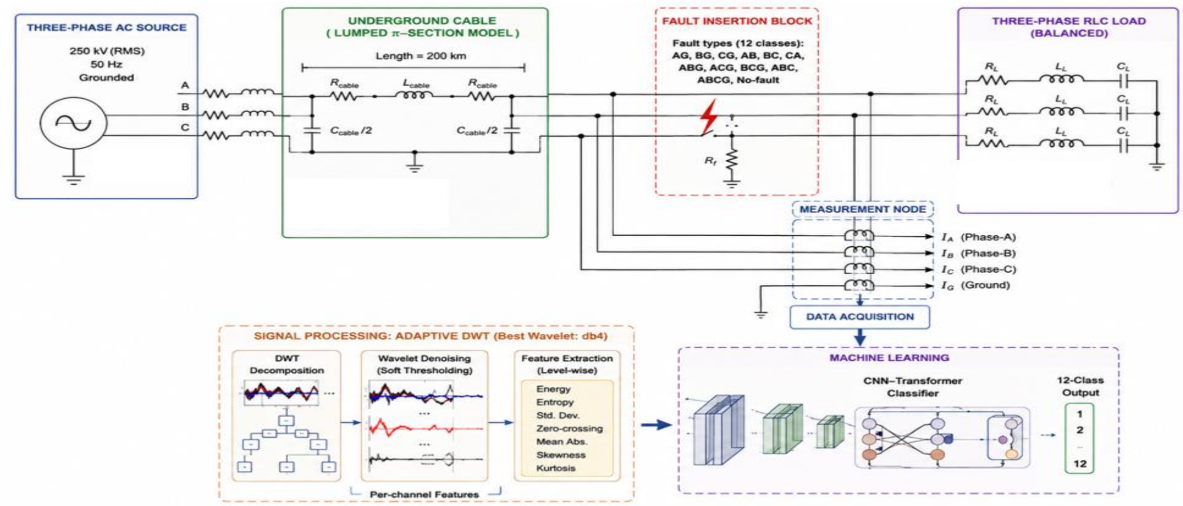


Fig. 1: MATLAB/Simulink architecture of the 200-km underground-cable test system showing the grounded three-phase source, lumped π-section cable, configurable fault-injection block, balanced three-phase load, and Phase-A, Phase-B, Phase-C, and ground-current measurement channel.

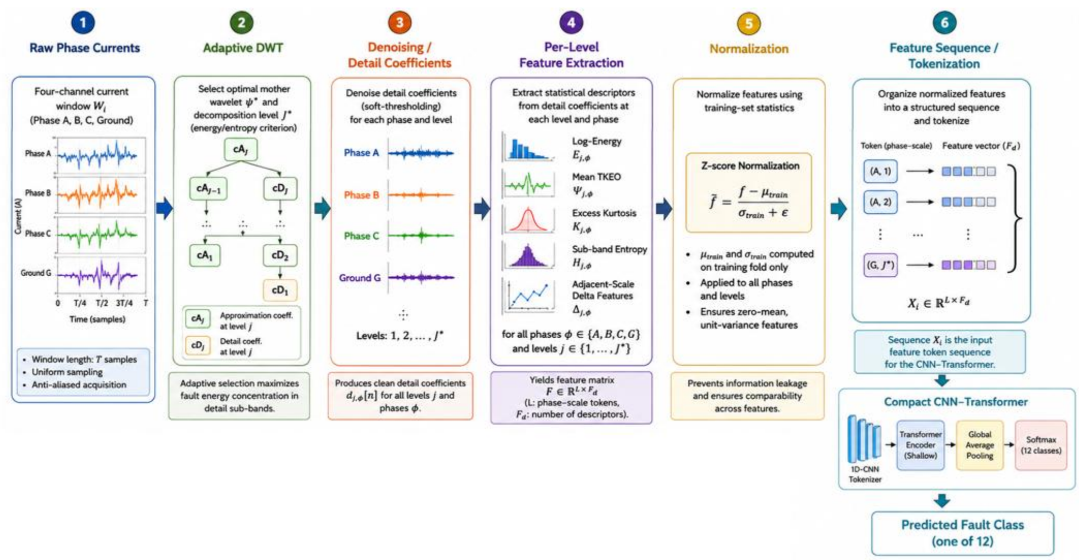


Fig. 2: Adaptive wavelet-based feature extraction pipeline showing transformation of the four-channel current windows into denoised DWT detail coefficients, extraction of level-wise statistical descriptors, training-set-based normalization, and organization of the resulting features for CNN–Transformer tokenization.

2.1.4 Labels and splits

The fault-diagnosis task is formulated as a single-label multi-class classification problem comprising K=12 mutually exclusive operating states 11 faults, and No-fault conditions. For each input window W_i, the corresponding target is represented by a one-hot vector $y_i \in \{\mathbf{0},\mathbf{1}\}^K$, where $y_{i,k} = 1$ if window i belongs to class k, and $y_{i,k} = 0$ otherwise, as defined in Equation (4).

$$y_{i,k} = \sum_{k=1}^{K} y_{i,k} = 1 \tag{4}$$

Where y_i is the label vector for sample i (one-hot in the multi-class setting), $y_{i,k}$ is the indicator for class k in sample i (1 if class k is true, else 0), K is the number of classes. The dataset is stratified into an 80% training partition and a 20% held-out test partition while preserving the relative frequency of all twelve classes. Five-fold stratified cross-validation is applied within the training partition for model selection and hyperparameter assessment, whereas the held-out test partition is reserved exclusively for final evaluation. All normalization statistics and other data-dependent preprocessing parameters are estimated from the training data only.

The ADWT-derived descriptors are subsequently arranged into the normalized feature sequence defined in Sections 2.2 and 2.3 for CNN–Transformer classification. This preserves transient timing for real-time, SCADA-ready inference while enabling explainable attributions. The use of wavelet features alongside lightweight attention is consistent with recent, high-accuracy detection/localization results on field and synthetic datasets [3, 4, 19, 20].

2.2 Adaptive DWT Feature Extraction

Adaptive DWT is employed because underground-cable fault currents are non-stationary and their transient characteristics vary with resistance, inception angle, bonding conditions, and other operating factors [29-31]. A fixed mother wavelet or decomposition depth can under or over-resolve the transient content [5, 32]. To obtain a compact and signal-dependent representation, candidate mother wavelets and decomposition depths are evaluated using the energy-entropy [5, 32] criterion described in Section 2.2.3, while level-wise denoising thresholds are estimated from the corresponding detail coefficients [5, 27, 32].

2.2.1 Discrete Wavelet Transform (DWT)

Let $x_\varphi[n]$ denote the discrete-time current of phase $\varphi \in \{A, B, C, G\}$. Given an orthonormal wavelet system with analysis low/high-pass filters h[n], g[n] associated with a mother wavelet ψ(t) and scaling function ϕ(t). The level-(j+1) approximation and detail coefficients are obtained from the level-j approximation through the low- and high-pass analysis filters, respectively, as defined in Equation (5) [7].

$$A_{j+k}[k] = \sum_n h\,[n-2k]cA_j[n],$$

$$cD_{j+k}[k] = \sum_n g\,[n-2k]cA_j[n] \qquad (5)$$

where $cA_j[n]$ are approximation coefficients at level j; $cD_{j+k}[k]$ are detail coefficients at level j+1; h[n] and g[n] are the analysis low-/high-pass quadrature mirror filters (QMFs) for ψ; k is the sub-sampled time index; n is the convolution index; and the factor 2 reflects dyadic down sampling [1]. The wavelet representation corresponds to a multiresolution analysis (MRA), where the time support is halved and the frequency support is doubled at each level. It is computed efficiently via Mallat's pyramidal filter-bank algorithm [1].

2.2.2 Pre-denoising by level-wise wavelet shrinkage

To improve the signal-to-noise ratio (SNR) before feature computation, we apply level-wise soft-thresholding of detail coefficients using the universal threshold expressed in Equation (6) [33]:

$$\lambda_{j,\phi} = \hat{\sigma}_{j.\phi}\sqrt{2ln}N_j, \quad \hat{\sigma}_{j,\phi} = \frac{median(\left|cD_{j,\phi})\right|}{0.6745} \qquad (6)$$

where $\lambda_{j,\phi}$ is the soft threshold for phase φ and level *j*; $\hat{\sigma}_{j,\phi}$ is the robust noise estimate from the Median Absolute Deviation (MAD); N_j is the number of coefficients at level j; and $cD_{j,\phi}$ are detail coefficients at level j for phase φ.

2.2.3 Adaptive selection of mother wavelet and decomposition depth

Let W = {db2, db4, db6, sym4, coif1} denote the candidate mother-wavelet set and J=4 is the admissible decomposition-depth set. For each candidate pair (ψ, J) ∈ W × J, the DWT is applied independently to the four measurement channels ϕ∈{A,B,C,G} [5, 6, 8]. The configuration maximizing the aggregated merit is retained for subsequent feature extraction. The Shannon entropy of the subband energy and energy-to-entropy merit is shown in Equations (7)-(9).

$$E_{j,\phi} = \sum_k \left|cD_{j,\phi}[k]\right|^2, \quad p_{j,\phi}[k] = \frac{\left|cD_{j,\phi}[k]\right|^2}{\sum_k \left|cD_{j,\phi}[k]\right|^2} \qquad (7)$$

where $E_{j,\phi}$ is the subband energy for phase φ at level j, $p_{j,\phi}[k]$ is the normalized energy distribution over coefficients at that subband, and k is the index of the detail coefficients at level j.

$$H_{j,\phi} = -\sum_{k} p_{\phi,j}[k]\ln(p_{\phi,j}[k] + \varepsilon) \tag{8}$$

Where $H_{j,\phi}$ is the Shannon entropy of the subband energy distribution; ε>0 avoids log singularity.

$$\psi *, \mathrm{J} *) = arg_{\psi\in\mathcal{W}}\max_{\mathrm{j}\in \mathrm{J}}\mathcal{M}(\psi, \mathrm{J}) \tag{9}$$

Where M(ψ, J) denotes the aggregated energy–entropy merit associated with candidate mother wavelet ψ and decomposition depth J. The optimal configuration (ψ∗, J∗) is selected by maximizing M over the prescribed search space. For the present dataset, the optimization selected db4 as ψ∗.

2.2.4 *Per-level feature set*

The detail coefficients after the level-wise shrinkage is described in Section 2.2.2. From these coefficients, a compact per-level feature vector is extracted to characterize transient magnitude, impulsiveness, energy dispersion, and cross-scale evolution.

i. Log-energy (captures transient strength): For each phase ϕ and retained decomposition level j, the detail-coefficient energy is computed using Equation (10), and its logarithmic representation is obtained using Equation (11).

$$E_{j,\phi} = \sum_{k=1}^{N_j} u_{\emptyset,j}[k]^2 \tag{10}$$

$$f_{\phi,j}^{(1)} = \ln(E_{\phi,j} + \varepsilon \tag{11}$$

Where $f_{\phi,j}^{(1)}$ is the log-scaled energy feature, $E_{\phi,j}$ emphasizes transient magnitude at scale j, and ε > 0 avoids log singularity.

ii. Mean Teager–Kaiser Energy (tracks instantaneous power/impulsiveness): The Teager–Kaiser Energy Operator (TKEO) emphasizes short-duration, high-energy variations characteristic of fault onset. The mean TKEO of the denoised detail coefficients is computed using Equation (12), with boundary samples handled by symmetric padding.

$$f_{\phi,j}^{(TKEO)} = \frac{1}{N_j - 2}\sum_{k=2}^{N_j-1}\left(u_{\emptyset,j}[k]^2 - u_{\phi,j}[k-1]u_{\phi,j}[k+1]\right) \tag{12}$$

where $f_{\phi,j}^{(TKEO)}$ is the mean Teager-Kaiser Energy over the subband; Ψ[·] is the discrete TKEO; and $u_{\varphi,j}$[k] are detail coefficients.

iii. Excess kurtosis (measures impulsiveness/heavy tails): Excess kurtosis measures the impulsiveness and heavy-tailed character of the wavelet coefficients and increases in the presence of sparse, high-amplitude transient events. For each phase and decomposition level, the excess kurtosis is defined in Equation (13).

$$f_{\phi,j}^{(k)} = \frac{\frac{1}{N_j}\sum_{k=1}^{N_j}(u_{\phi,j}[k] - \mu_{\phi,j})^4}{(\sigma_{\phi,j}^2)^2} - 3 \tag{13}$$

Where $f_{\phi,j}^{(k)}$ is the excess-peakedness of subband coefficients, μ is the sample mean.

iv. Sub-band entropy (coefficient-energy dispersion): The normalized coefficient-energy distribution is used to compute the Shannon entropy defined in Equation (14). Lower entropy indicates greater energy concentration within a small number of coefficients, which is characteristic of localized transient events.

$$p_{\emptyset,j}[k] = \frac{u_{\phi,j}[k]^2}{\sum_{\ell=1}^{N_j} u_{\emptyset,j}\ell^2}, \quad \sum_{k} p_{\emptyset,j}[k] = 1 \tag{14}$$

v. Adjacent-scale features: Differences between neighboring decomposition levels characterize the cross-scale evolution of each base descriptor while reducing sensitivity to global amplitude variation and background drift. For a base descriptor f, the adjacent-scale difference is defined in Equation (15).

$$f_{\phi,j}^{(m,\Delta)} = f_{\phi,j}^{(m)} - f_{\phi,j+1}^{(m)}, \quad j \geq 2, m\epsilon\{1,2,3,4\} \tag{15}$$

where $f_{\phi,j}^{(m,\Delta)}$ captures scale-to-scale changes to improve robustness to amplitude scaling and background drift. The final feature vector for each record is formed by concatenating the extracted descriptors across all phases and decomposition levels, as defined in Equation (16).

$$F = concat_{\phi\epsilon(A,B,C,G)} [\left\{f_{\phi,j}^{(\log E)}, f_{\phi,j}^{(TKEO)}, f_{\phi,j}^{(k)}, f_{\phi,j}^{(H)}\right\}_{j=1}^{J}, \left\{f_{\phi,j}^{(m,\Delta)}\right\}_{m,j=1}^{J-1}] \quad (16)$$

2.3 Compact CNN-Transformer

The normalized ADWT feature sequence derived in Section 2.2 constitutes the input to the compact CNN–Transformer. The architecture combines a lightweight convolutional tokenizer with a shallow self-attention encoder to model local cross-scale patterns and broader dependencies among the phase–scale feature tokens. The model is designed to: (1) tokenize the structured ADWT feature sequence using compact 1-D convolutions; (2) capture cross-token dependencies through a shallow multi-head self-attention encoder; and (3) maintain a constrained parameter and latency budget for edge-oriented implementation. Transformer-based architectures have demonstrated competitive performance in industrial fault diagnosis and related safety-critical monitoring applications [17, 18, 21].

2.3.1 Input representation and tokenization

For each current window Wi, the normalized ADWT descriptors defined in Section 2.2.4 are organized across the four measurement channels and retained decomposition levels to form the structured phase–scale feature sequence Xi defined in Equation (17).

$$X_i = [x_{i,1}, x_{i,2}, \dots, x_{i,L}]^T \quad \in \quad R^{L\times F} \quad (17)$$

Where L denotes the number of phase–scale feature tokens and F denotes the number of descriptors associated with each token. The feature sequence $\boldsymbol{X_i}$ is subsequently processed by a depthwise-separable 1-D convolutional tokenizer that captures local relationships among neighboring phase–scale tokens and projects them into a d-dimensional embedding space, as defined in Equation (18).

$$Z_i = GELU(DSConv1D(X_i))W_p + b_p \quad (18)$$

Where Z_i is the embedded token sequence, DSConv1D(·) denotes the depthwise-separable convolutional operator, W_p is the learnable projection matrix, and b_p is the corresponding bias vector.

2.3.2 Lightweight transformer encoder

The final encoder adopts a pre-normalization Transformer configuration with M=2 encoder blocks, h=4 attention heads, and an embedding dimension of d=128. For each encoder block, Layer Normalization precedes the self-attention operation, as defined in Equation (19). Scaled dot-product attention for each head r ∈ {1, . . . , h} is computed using Equation (20), and the resulting head representations are combined through the multi-head attention operation in Equation (21).

$$\widetilde{H}^{(\ell)} = LN\left(\widetilde{H}^{(\ell-1)}\right), Q = \widetilde{H}^{(\ell)}W_Q, \quad K = \widetilde{H}^{(\ell)}W_K, \ V = \widetilde{H}^{(\ell)}W_V \quad (19)$$

$$Attn(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (20)$$

$$MNSA(Q,K,V) = Concat(Attn_1, \dots Attn_h), \quad \widetilde{H}^{(\ell)} = \widetilde{H}^{(\ell-1)} + Droupout(MHSA) \quad (21)$$

Where $\widetilde{H}^{(\ell)}$ is the normalized hidden state at block, Q, K, and V are learned projections of the normalized hidden states, and d_k denotes the dimensionality of each attention head and determines the scaling factor used in the dot-product attention operation. A position-wise two-layer feed-forward network (FFN) with GELU activation expands the representation and introduces non-linearity, as described in Equation (22), where $H^{I^{(\ell)}}$ is the first residual output, b_1, b_2 are biases vectors, and W_1, W_2 are learnable feed-forward weights.

$$FNN(x) = W_2.GELU(W_1x + b_1) + b_2, \ H^{I^{(\ell)}} = \left(H^{(\ell)} + Droupout(FNN(LN(H^{(\ell)})))\right) \quad (22)$$

2.3.3 Classification head

After the final Transformer encoder block, global-average pooling aggregates the token representations into a latent feature vector $h_i \epsilon R^d$, as presented in Equation (23) where T′ denotes the number of encoded tokens and $H_{i,t}$ is the representation of token t for sample i. Because the fault-diagnosis task comprises K=12 mutually exclusive operating states, a fully connected classification layer maps the pooled representation to a K-dimensional logit vector described in Equation (24). The corresponding class probabilities are obtained using the softmax function as shown in Equation (25).

$$h_i = \frac{1}{T'}\sum_{t=1}^{T'} H_{i,t} \tag{23}$$

$$O_i = W_c h_i + b_c\,, \quad O_i \epsilon R^K \tag{24}$$

$$P_{i,k} = \frac{\exp(O_{i,k})}{\sum_{j=1}^{K}\exp(O_{i,k})} \quad , k = 1, \dots, K \tag{25}$$

where O_i denotes the logit vector for sample i, $O_i \epsilon R^K$ and $b_c \epsilon R^K$ are the learnable classifier parameters, and K=12 is the number of mutually exclusive operating classes. $p_{i,k}$ is the predicted probability that sample *i* belongs to class k, $O_{i,k}$ is the corresponding class logit. The softmax operation normalizes the class probabilities such that $\sum_{k=1}^{K} P_{i,k}$=1. The predicted operating state is determined as $\hat{y}_i = max_k P_{i,k}$.

2.3.4 Training protocol

Training is performed using the AdamW optimizer with an initial learning rate of $\boldsymbol{\eta_0}$=1×10^{-3}, a weight-decay coefficient of λ=1×10^{-4}, and a mini-batch size of B=32. The network is trained for a maximum of 100 epochs, with the learning rate adjusted using a cosine-annealing schedule. Categorical cross-entropy is used as the objective function for the K=12 mutually exclusive operating classes. For a mini-batch of B samples, the classification loss is defined in Equation (26).

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{B}\sum_{k=1}^{K}\left(y_{i,k}\log p_{i,k} + \in\right) \tag{26}$$

Where $y_{i,k}$ is the one-hot target for sample i, $p_{i,k}$ is the corresponding softmax probability, K=12, and ϵ>0 is a small numerical constant introduced to avoid evaluating log(0).

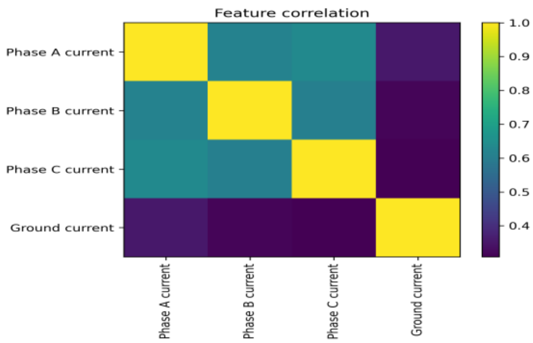
2.3.5 Model size, SCADA deployability and Explainability

The architecture is constrained to fewer than 0.5 million parameters, ensuring compatibility with resource-limited industrial edge hardware. Model-level CPU inference latency is evaluated experimentally and reported in Section 3.3; end-to-end acquisition, preprocessing, communication, and jitter are outside the present timing assessment. This edge-centric deployment significantly reduces network round-trip latency compared with cloud-only configurations, in accordance with contemporary best practices for low-latency and resilient operational-technology (OT) analytics [23]. Model interpretability is achieved through Deep SHAP applied to the penultimate layer, quantifying feature contributions across wavelet levels and phases. Prior investigations have correlated SHAP-derived attributions with physically interpretable fault signatures in the time–frequency domain, thereby strengthening engineering trust and enabling systematic alarm triage [21, 22].

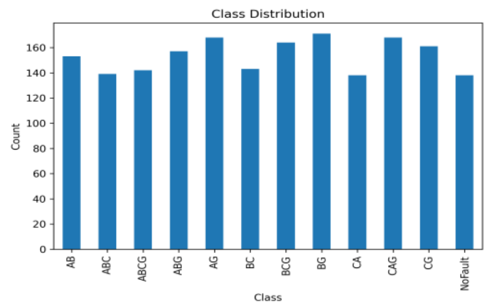
3. RESULTS and DISCUSSION

3.1 Dataset characterization and feature relationships

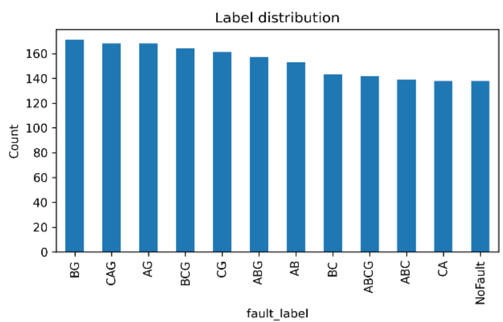
The dataset exhibits moderate inter-phase coupling (Phase pairs r≈0.6–0.8) and weaker phase–ground correlation (r<0.5), consistent with three-phase UG behaviour as shown in Figure 3(a). Label histograms presented in Figure 3(b-c) confirm near-balance across the twelve classes, limiting class-imbalance bias and ensuring rare composite faults (e.g., ABCG, ACG) are represented.



(a)



(b)

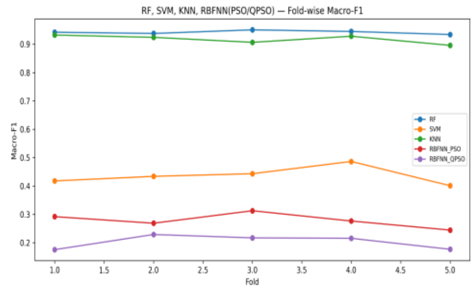


(c)

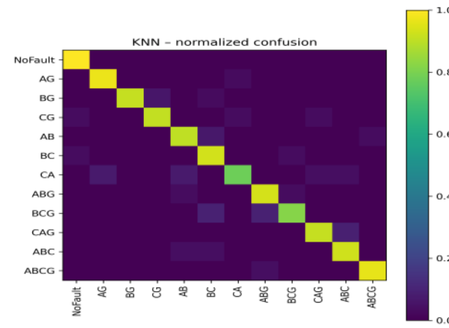
Fig. 3: Feature correlation and class distribution. (a) Correlation matrices for Phase A, B, C and Ground currents showing moderate inter-phase coupling and weaker phase–ground correlation. (b-c) Label distributions across twelve classes confirm near-balanced sampling of single, double and composite faults, including the no-fault condition.

3.2 Classical machine-learning benchmarks

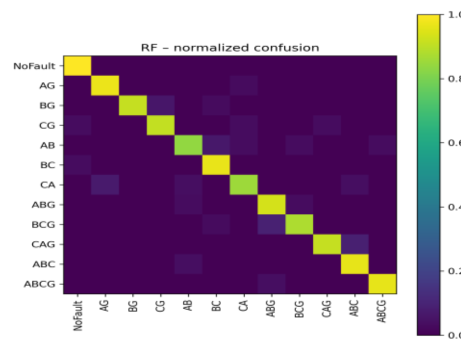
Across five-fold cross-validation, RF and KNN exhibit the strongest performance, whereas SVM and the RBFNN variants show substantially weaker class separation. The complete cross-validated results are summarized in Table 2**.**



(a)



(b) KNN



(c) RF

Fig. 4: Baseline classical performance. (a) Fold-wise Macro-F1 scores for RF, SVM, KNN, RBFNN–PSO and RBFNN–QPSO over five-fold cross-validation. (b-c) Normalized confusion matrices for KNN and RF, respectively, illustrating dominant diagonal entries and residual confusion between physically similar fault combinations.

Table 2: Cross-validated performance of baseline classical models (RF, SVM, KNN, RBFNN-PSO, and RBFNN-QPSO). Values report mean ± standard deviation of Accuracy and Macro-F1 over five folds.

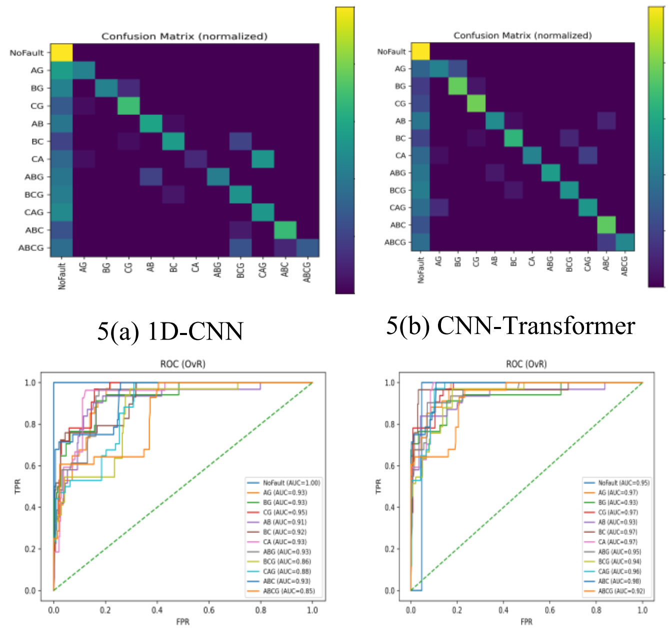
Model	**Accuracy (mean)**	**Accuracy (std)**	**Macro-F1 (mean)**	**Macro-F1 (std)**
KNN	0.918	0.015	0.917	0.015
RBFNN_PSO	0.278	0.016	0.279	0.025
RBFNN_QPSO	0.216	0.021	0.203	0.025
RF	0.942	0.006	0.942	0.006
SVM	0.412	0.028	0.437	0.032

As presented in Table 2, RF delivers the best cross-validated performance (Accuracy 0.942 ± 0.006, Macro-F1 0.942 ± 0.006), with KNN close behind (0.918 ± 0.015 / 0.917 ± 0.015). The RF-KNN gap is modest (Δ Accuracy ≈ +0.024, Δ Macro-F1 ≈ +0.025), and the near-equality of Accuracy and Macro-F1 for both models indicates balanced per-class behaviour. In contrast, SVM underperforms markedly (0.412 ± 0.028 / 0.437 ± 0.032), trailing RF by ≈ 53 percentage points (pp) in Accuracy and ≈ 50 pp in Macro-F1, with higher variance across folds. The RBFNN variants are weak baselines here, with RBFNN-PSO (0.278 ± 0.016 / 0.279 ± 0.025) outperforming RBFNN-QPSO (0.216 ± 0.021 / 0.203 ± 0.025) by 6.2 pp Accuracy and 7.6 pp Macro-F1, suggesting that, on these WT features, their training heuristics do not yield competitive decision boundaries. Overall, classical baselines establish a strong reference via RF/KNN, against which deep models will be compared.

3.3 Deep-learning models and Runtime

Two deep architectures were investigated: a compact one-dimensional convolutional neural network (1D-CNN) and a hybrid CNN-Transformer that augments local convolution with global self-attention. Both deep architectures were evaluated using the same preprocessed feature representation derived from the four-channel current windows, enabling a consistent comparison of their classification capabilities. As summarized in Table 3, the CNN–Transformer consistently outperforms the compact 1D-CNN on the held-out test set, with the largest gains observed in recall and macro-F1. The hybrid architecture provides the strongest test-set performance and will serve as the deep baseline for SCADA-oriented comparisons. RF achieves the strongest cross-validated classical-model performance, whereas the CNN–Transformer achieves the strongest held-out performance among the evaluated deep-learning models. Because the reported classical and deep-learning metrics arise from different evaluation summaries, they are interpreted within their respective protocols rather than ranked directly.

More importantly, the deep networks exhibit strong discriminative power in terms of ROC characteristics. Figure 5(a-b) shows normalized confusion matrices for the compact 1D-CNN and the hybrid CNN-Transformer, respectively, while Figures 5(c-d) illustrate one-vs-rest ROC curves. For the CNN-Transformer, most classes achieve AUC values above 0.95, including complex composite faults such as ABCG and ACG.



5(a) 1D-CNN 5(b) CNN-Transformer

5(c) 1D-CNN 5(d) CNN-Transformer

Fig. 5: Deep-learning performance. (a) Normalized confusion matrix for the 1D-CNN model. (b) Normalized confusion matrix for the CNN-Transformer, showing clearer separation for several composite faults. (c-d) One-vs-rest ROC curves for 1D-CNN and CNN-Transformer, respectively, with most classes achieving AUC above 0.9 and several exceeding 0.95.

Table 3: Test-set accuracy and macro-averaged precision, recall and F1-score for the deep learning models (1D-CNN and CNN-Transformer). Metrics derived from per-class classification reports.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
1D-CNN	0.520	0.785	0.520	0.553
CNN-Transformer	0.612	0.810	0.615	0.653

Table 4 shows that all evaluated classifiers achieve sub-millisecond per-sample inference on the target platform. The CNN–Transformer remains substantially faster than RF while retaining the strongest held-out macro-F1 among the evaluated deep models. All evaluated models exhibit sub-millisecond model-level inference on the tested hardware. KNN and 1D-CNN are the fastest at 0.006 ms (~ 6 µs) per sample, followed by the CNN-Transformer at 0.036 ms (~ 36 µs). RF is the slowest at 0.262 ms (~ 262 µs), yet it remains well within the sub-millisecond model-level inference range observed on the tested hardware. Relative to RF, the CNN-Transformer is ~ 7.3 times faster, while KNN/1D-CNN is ~ 44 times faster; the latter reflects the tiny feature dimension and efficient BLAS operations for distance/conv. kernels. Combined Tables 2-4, these results show that the Hybrid CNN-Transformer offers a favourable speed-accuracy trade-off: microsecond-scale model-level latency with the strongest held-out macro-F1 among the evaluated deep-learning models.

Table 4: Model-level CPU inference latency per sample for the evaluated classical and deep-learning classifiers.

Model	**Latency [ms/sample]**
RF	0.262
KNN	0.006
1D-CNN	0.006
CNN-Transformer	0.036

3.4 Explainability analysis (SHAP-based interpretation)

To expose the internal decision logic of both the RF baseline and the CNN-Transformer, SHAP analysis was conducted at the class level. Figure 6 aggregates SHAP bar and beeswarm plots for RF across all twelve classes. The global importance bars reveal that ground current is the dominant descriptor for line-to-ground faults, whereas Phase A and Phase B currents alternate as leading predictors for inter-phase faults such as AB, BC and CA. However, the beeswarm plots show relatively compact SHAP distributions, indicating that RF primarily relies on a small number of high-gain splits rather than modelling richer phase interactions. The SHAP results indicate class-dependent phase and ground-current contributions that are qualitatively consistent with the simulated fault topology. These attribution patterns complement the predictive and latency results by providing an interpretable view of the CNN–Transformer decision process.

The comparative study demonstrates that classical ensemble methods such as RF remain highly competitive for steady-state fault-current patterns. RF relies on a few high-gain splits evident by its compact SHAP clouds (Figure 6), while the CNN-Transformer reveals richer, class-specific attributions and phase couplings (Figure 7), showing class-dependent attribution patterns that are qualitatively consistent with the simulated phase participation. The CNN-Transformer offers compelling advantages: it delivers strong ROC performance for all classes, maintains sub-millisecond latency, and provides interpretable class-specific attributions that complement the predictive results. From a deployment perspective, these properties make the deep architecture particularly attractive for integration in modern SCADA environments, where evolving operating regimes and the need for explainable automation favour models that can both generalize and justify their decisions.

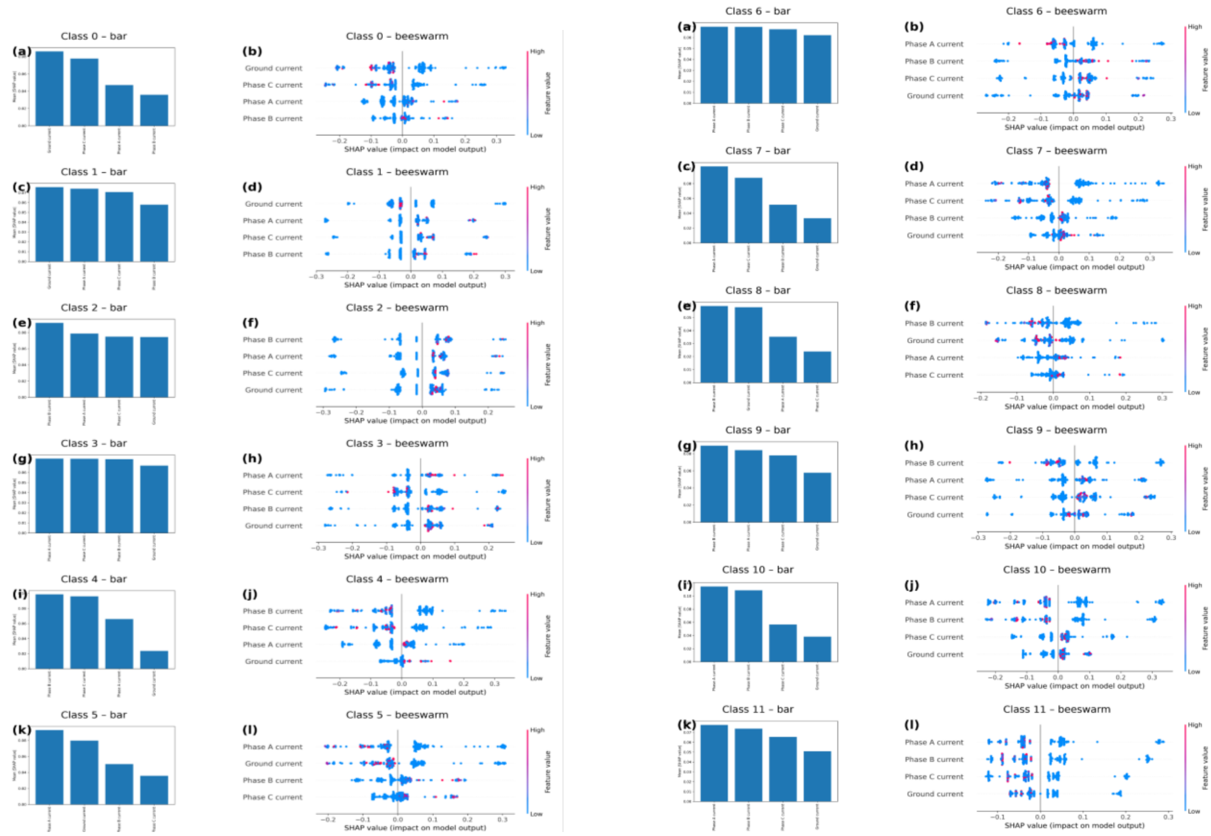


Fig. 6: Random Forest SHAP interpretation. Combined bar and beeswarm plots for classes 0–11 summarizing global feature importance and local contributions of Phase A, B, C and Ground currents to the RF decision function.

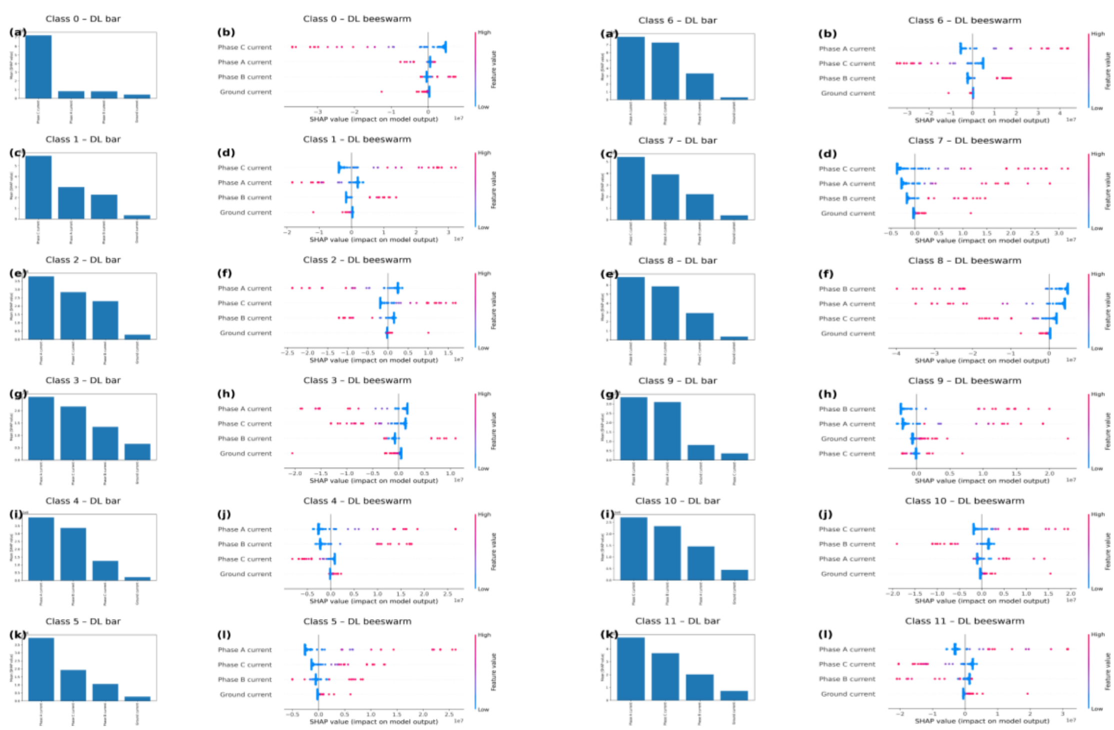


Fig. 7: CNN-Transformer SHAP interpretation. Combined bar and beeswarm plots for classes 0–11 reveal richer phase coupling and class-specific attributions than the RF baseline, consistent with the model's ability to capture richer cross-phase and cross-scale feature dependencies.

4. CONCLUSION

This study presented an explainable, SCADA-ready pipeline for underground-cable fault detection and classification that couples adaptive DWT features with a compact CNN-Transformer. On the held-out 12-class test set generated from the controlled 200-km MATLAB/Simulink underground-cable model, the CNN-Transformer achieved 0.612 accuracy, 0.810 macro-precision, 0.615 macro-recall, and 0.653 macro-F1. SHAP attributions at phase and sub-band levels were qualitatively consistent with expected phase participation in the simulated fault classes, providing post-hoc interpretability of the model predictions. The approach offers a favourable speed-accuracy-interpretability trade-off suitable for protection-adjacent analytics and edge deployment. Key limitations include reliance on synthetic data, residual confusability among topology-similar classes, the post-hoc (non-causal) nature of SHAP explanations, and incomplete end-to-end latency/jitter accounting and compliance validation.

Future work will prioritize real data evaluations and robustness to domain shift; expansion to incipient and high-impedance faults and joint classification-localization with calibrated uncertainty, including hardware-in-the-loop testing and measurement of complete latency/jitter budgets.

CRediT authorship contribution statement:

Kabiru Dahiru Ibrahim: Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization. Ahmad Abubakar: Writing – review & editing, Resources, Formal analysis. Sulaiman Haruna Sulaiman: Writing – review & editing, Conceptualization, Formal Analysis. Yusuf Idris: Writing – review & editing, Visualization, Validation. Aliyu Lawal Musa: Writing – review & editing, Formal Analysis, Visualization Conceptualization. Ibrahim Mahdi: Writing – review & editing, Formal analysis, Conceptualization.

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. All authors read and approved the final manuscript.

Data availability:

Data will be made available on request.

REFERENCES

[1] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, 2002.

[2] Z. He, L. Fu, S. Lin, and Z. Bo, "Fault detection and classification in EHV transmission line based on wavelet singular entropy," *IEEE Trans. Power Deliv.*, vol. 25, no. 4, pp. 2156–2163, 2010.

[3] A. P. Oliveira, F. A. Moreira, and A. F. Picanço, "Accuracy analysis using the EMD and VMD for two-terminal transmission line fault location based on traveling wave theory," *Electr. Power Syst. Res.*, vol. 224, p. 109667, 2023.

[4] M. Parsi, P. Crossley, P. L. Dragotti, and D. Cole, "Wavelet based fault location on power transmission lines using real-world travelling wave data," *Electr. Power Syst. Res.*, vol. 186, p. 106261, 2020.

[5] Z. Liu, Y. Cui, and W. Li, "Combined power quality disturbances recognition using wavelet packet entropies and S-transform," *Entropy*, vol. 17, no. 8, pp. 5811–5828, 2015.

[6] J. Chen, Y. Dou, Y. Li, and J. Li, "Application of Shannon wavelet entropy and Shannon wavelet packet entropy in analysis of power system transient signals," *Entropy*, vol. 18, no. 12, p. 437, 2016.

[7] E. M. Shalby, A. Y. Abdelaziz, E. S. Ahmed, and B. Abd-Elhamed Rashad, "A comprehensive guide to selecting suitable wavelet decomposition level and functions in discrete wavelet transform for fault detection in distribution networks," *Sci. Rep.*, vol. 15, no. 1, p. 1160, 2025.

[8] R. Jadhav and A. Mahajan, "Smart grid data

compression and reconstruction by wavelet packet transform," *MethodsX*, vol. 13, p. 102872, 2024.

[9] M. Lopez-Bonilla, A. R. Herrera-Orozco, and A. Molina-Cabrera, "An overview of methods for detecting and locating incipient faults in underground cables," *Electr. Power Syst. Res.*, vol. 245, p. 111631, 2025.

[10] Q. Shao *et al.*, "Artificial Intelligence in Cable Fault Detection and Localization: Recent Advances and Research Challenges," *Energies*, vol. 18, no. 14, p. 3662, 2025.

[11] H. Fu, L. Qiu, Y. Ai, J. Tu, and Y. Yan, "Deep learning-based fault detection and location in underground power cables using resonance frequency analysis," *Electr. Eng.*, vol. 107, no. 4, pp. 4051–4062, 2025.

[12] Z. M. Ali and E. M. Esmail, "Deep learning and wavelet packet transform for fault diagnosis in double circuit transmission lines," *Sci. Rep.*, vol. 15, no. 1, p. 30145, 2025.

[13] A. Cano, P. Arévalo, D. Benavides, and F. Jurado, "Integrating discrete wavelet transform with neural networks and machine learning for fault detection in microgrids," *Int. J. Electr. Power Energy Syst.*, vol. 155, p. 109616, 2024.

[14] P. Chi, R. Liang, C. Hao, G. Li, and M. Xin, "Cable fault diagnosis with generalization capability using incremental learning and deep convolutional neural network," *Electr. Power Syst. Res.*, vol. 241, p. 111304, 2025.

[15] B. Baadji, S. Belagoune, and S. E. Boudjellal, "Transformer-based deep learning networks for fault detection, classification, and location prediction in transmission lines," *Netw. Comput. Neural Syst.*, pp. 1–21, 2024.

[16] H. Mirshekali, A. Keshavarz, R. Dashti, S. Hafezi, and H. R. Shaker, "Deep learning-based fault location framework in power distribution grids employing convolutional neural network based on capsule network," *Electr. Power Syst. Res.*, vol. 223, p. 109529, 2023.

[17] H. Wu, M. J. Triebe, and J. W. Sutherland, "A transformer-based approach for novel fault detection and fault classification/diagnosis in manufacturing: A rotary system application," *J. Manuf. Syst.*, vol. 67, pp. 439–452, 2023.

[18] F. Wang *et al.*, "Convolutional preprocessing Transformer-based fault diagnosis for rectifier-filter circuits in nuclear power plants," *Reliab. Eng. Syst. Saf.*, vol. 249, p. 110198, 2024.

[19] D.-H. Yoon and J. Yoon, "Development of a real-time fault detection method for electric power system via transformer-based deep learning model," *Int. J. Electr. Power Energy Syst.*, vol. 159, p. 110069, 2024.

[20] M. R. Shadi, M.-T. Ameli, and S. Azad, "A real-time hierarchical framework for fault detection, classification, and location in power systems using PMUs data and deep learning," *Int. J. Electr. Power Energy Syst.*, vol. 134, p. 107399, 2022.

[21] T. Yan, X. Xing, T. Xia, and D. Wang, "Relation between fault characteristic frequencies and local interpretability shapley additive explanations for continuous machine health monitoring," *Eng. Appl. Artif. Intell.*, vol. 136, p. 109046, 2024.

[22] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte, "Fault Diagnosis using eXplainable AI: A transfer learning-based approach for rotating machinery exploiting augmented synthetic data," *Expert Syst. Appl.*, vol. 232, p. 120860, 2023.

[23] C. Song, H. Zheng, G. Han, P. Zeng, and L. Liu, "Cloud edge collaborative service composition optimization for intelligent manufacturing," *IEEE Trans. Ind. Informatics*, vol. 19, no. 5, pp. 6849–6858, 2022.

[24] G. Porawagamage, K. Dharmapala, J. S. Chaves, D. Villegas, and A. Rajapakse, "A review of machine learning applications in power system protection and emergency control: opportunities, challenges, and future

directions," *Front. Smart Grids*, vol. 3, p. 1371153, 2024.

[25] J. Maldonado-Correa, J. Torres-Cabrera, S. Martín-Martínez, E. Artigao, and E. Gómez-Lázaro, "Wind turbine fault detection based on the transformer model using SCADA data," *Eng. Fail. Anal.*, vol. 162, p. 108354, 2024.

[26] R. Bosso, C. Chang, M. Zarif, and Y. Tang, "Explainable Graph Neural Networks for Power Grid Fault Detection," *IEEE Access*, vol. 13, pp. 129520–129533, 2025, doi: 10.1109/ACCESS.2025.3591604.

[27] Y. Liu, P. Li, Y. Si, and L. Ma, "A Comprehensive Review of AI Integration for Fault Detection in Modern Power Systems: Data Processing, Modeling, and Optimization," *Energies*, vol. 18, no. 18, p. 4983, 2025.

[28] K. D. Ibrahim, C. Dongsheng, A. Abubakar, C. C. Ukwuoma, and Q. Huang, "Multiscale wavelet and hybrid lazy rule-guided network for SCADA-compatible underground fault detection," *Electr. Power Syst. Res.*, vol. 264, p. 113850, 2027, doi: https://doi.org/10.1016/j.epsr.2026.113850.

[29] A. Saber, A. Emam, and H. Elghazaly, "A backup protection technique for three-terminal multisection compound transmission lines," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 5653–5663, 2017.

[30] C. Liu, Y. Hou, Z. Duan, R. Shang, K. Wang, and D. Chen, "Underground power cable fault detection and localization based on the spectrum of propagation functions," *IEEE Trans. Instrum. Meas.*, 2024.

[31] A. Mukherjee, P. K. Kundu, and A. Das, "A differential signal-based fault classification scheme using PCA for long transmission lines," *J. Inst. Eng. Ser. B*, vol. 102, pp. 403–414, 2021.

[32] R. J. J. Molu *et al.*, "Enhancing power quality monitoring with discrete wavelet transform and extreme learning machine: a dual-stage pattern recognition approach," *Front. Energy Res.*, vol. 12, p. 1435704, 2024.

[33] I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. R. Stat. Soc. Ser. B (statistical Methodol.*, vol. 59, no. 2, pp. 319–351, 1997.