Artificial Intelligence for Engineering

WILEY

ORIGINAL RESEARCH OPEN ACCESS

From Local Patterns to Global Context: A Multimodal Deep Learning Approach for Complex Power Quality Disturbance Recognition

Jiajian Lin[1] | Hadi Nabipour Afrouzi[2] | Mehran Motamed Ektesabi[3] | Jalal Tavalaei[1]

[1]Faculty of Engineering, Computing and Science, Swinburne University of Technology, Sarawak Campus, Kuching, Malaysia | [2]Faculty of Computing, Engineering and the Built Environment, Birmingham City University, Birmingham, UK | [3]School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Hawthorn, Victoria, Australia

Correspondence: Jalal Tavalaei (jtavalaei@swinburne.edu.my)

Received: 20 November 2025 | **Revised:** 1 January 2026 | **Accepted:** 8 February 2026

Keywords: convolutional neural network | cross attention | deep learning | power quality disturbance | transformer

ABSTRACT

The increasing penetration of renewable energy sources introduces complex and mixed power quality disturbances (PQDs) that challenge traditional diagnostic approaches. To address the limitations of convolutional neural networks (CNN) in capturing long-range temporal dependencies, this study proposes an intelligent classification framework integrating fast Fourier transform, one-dimensional CNN, lightweight Bidirectional Encoder Representations from Transformers (LBERT1d) and a signal-based cross-attention (SCA) mechanism. The proposed MCNN1d-LBERT1d-SCA framework leverages multimodal time–frequency feature fusion, combining local and global representations to enhance the recognition of concurrent and nonstationary PQDs. A synthetic dataset following IEEE Std. 1159 was constructed encompassing 25 disturbance types with multiple signal-to-noise ratios, to ensure robustness and generalisation. Experimental results demonstrate acceptable performance, achieving an average accuracy of 99.30% on the synthetic dataset and maintaining better reliability under noise conditions down to 20 dB. Validation using real-world IEEE PES data and MATLAB/Simulink simulations yielded accuracies of 95.88% and 97.47%, respectively, confirming the model's strong adaptability and real-time capability. These results indicate that the proposed hybrid deep learning framework offers a practical and scalable solution for intelligent PQD monitoring contributing to the reliability and stability of modern power systems.

1 | Introduction

The environmental pollution caused by traditional fossil fuels forces humans to develop cleaner energy sources [1]. In order to promote sustainable development strategies, modern power systems extensively use renewable energy represented by wind and solar energy. The integration of various power sources may cause increased fluctuations in the power grid due to their differences. With the addition of a large amount of renewable energy, detecting a single disturbance no longer meets the current requirements for grid stability [2]. In power systems, concurrent mixed disturbances often challenge traditional PQD classification methods. To improve power quality and promote the development of smart grids, there is an urgent need for a more effective detection method for complex PQDs.

Abbreviations: β, shift parameter in batch normalisation; γ, scale parameter in batch normalisation; ε, small constant for numerical stability in batch normalisation; BERT, bidirectional encoder representations from transformers; dB, decibel; FFT, fast Fourier transform; K, key in the SCA module (frequency-domain feature); LBERT1d, lightweight one-dimensional BERT; MCNN1d, multimodal one-dimensional convolutional neural network; PQD, power quality disturbance; Q, query in the SCA module (time-domain feature); SCA, signal-based cross-attention; SNR, signal-to-noise ratio; V, value in the SCA module (frequency-domain feature).

With the advancement of CNN, automatic feature extraction has reduced errors from manual design and been widely applied in PQD classification [3]. Ensemble CNNs have improved short-term feature extraction, reduced complexity and enhanced noise resistance [4], whereas hybrid models combining CNN with other architectures improve performance. Examples include CNN-GRU with SE blocks for adaptive channel weighting [5], CNN-LSTM for joint spatial–temporal feature learning with high accuracy on synthetic and simulated datasets [6], and 1DCNN-BiGRU with time–frequency fusion and attention for improved feature diversity [7]. However, most CNN-based models primarily focus on local or short-term temporal patterns within a limited receptive field making them less effective in capturing long-sequence dependencies that reflect the global temporal evolution of PQD signals. In practical power systems, many disturbances span multiple cycles and exhibit nonstationary characteristics over time. Accurately modelling these long-range dependencies is essential for distinguishing between transient and steady-state events, improving feature continuity and enhancing the classifier's generalisation under complex disturbance conditions. Therefore, there is a pressing need for architectures capable of learning and integrating long-sequence contextual information to advance complex PQD recognition.

Transformer, introduced by Vaswani et al., has greatly advanced deep learning by enabling the efficient modelling of long-range dependencies in sequential data through self-attention mechanisms [8]. Unlike recurrent neural networks (RNNs) that process signals sequentially, the Transformer captures global temporal correlations in parallel making it highly suitable for time-series signal analysis. Recent studies have demonstrated its superiority in handling one-dimensional nonstationary signals like PQDs. For example, a lightweight temporal-channel attention convolutional transformer (TCAC-Transformer) was proposed for rotating machinery fault signal diagnosis, which effectively captured both local and global dependencies in time-series data while significantly reducing model parameters and computational cost, demonstrating the potential of Transformer architectures for efficient one-dimensional industrial signal analysis [9]. Similarly, A multimodal Transformer-based framework, TFCA-TransNet, was proposed to integrate fast Fourier transform with spatial convolution and channel attention for joint time–frequency–spatial feature learning of EEG motor imagery signals, achieving superior decoding accuracy and demonstrating the Transformer's effectiveness in modelling complex one-dimensional time-series signals [10]. In the power domain, Transformers have also been utilised for transient stability prediction [11] and short-term load forecasting [12], showing strong potential in modelling the dynamic behaviour of electrical signals. These studies demonstrate that the Transformer's self-attention mechanism can effectively learn global and contextual dependencies in one-dimensional signals essential for accurately distinguishing complex PQDs that exhibit overlapping, transient and steady-state characteristics. Therefore, applying a Transformer-based framework to PQD recognition offers significant advantages in capturing multiscale temporal dependencies that conventional CNN or RNN architectures often overlook.

Building upon the insights above, enhancing the ability to capture global features within CNN-based deep learning networks has become a pivotal research direction. Transformer and attention mechanisms have been proven in other fields to excel at capturing global features effectively. Inspired by these advancements, we proposed a multiscale complex PQD automatic recognition and classification model based on MCNN1d-LBERT1d-SCA. By combining these architectures, our approach aims to improve recognition accuracy and address the research gap in long-term feature extraction within PQD recognition tasks. In Table 1, we contrast our work with the literature.

Our main contributions are listed as follows.

1. To overcome CNN's limitation in capturing global features, a new neural network that integrates CNN1d and lightweight one-dimensional BERT (LBERT1d) is proposed. With its superior multihead attention mechanism, the Transformer's encoder is crucial in understanding global information.
2. To enhance the fusion of time-domain and frequency-domain features, this study introduces a signal-based cross-attention (SCA) mechanism in the backend of the classification model, which allows the model to establish stronger correlations between multiscale features, significantly reducing information loss and improving its ability to capture complex disturbance characteristics.
3. A complex PQD automatic recognition framework has been proposed based on the multimodal one-dimensional CNN, lightweight BERT1d model and signal-based cross-attention mechanism (MCNN1d-LBERT1d-SCA). This framework eliminates the tedious procedures of manual extraction and setting required features based on experience.

2 | Model Explanation

The successful identification and classification of single PQDs can no longer meet the current requirements for power quality. To achieve effective and reliable detection of complex PQDs, this study used MCNN1d-LBERT1d-SCA framework for automatic feature extraction and classification. The architecture is shown on Figure 1.

2.1 | Fast Fourier Transform

The core idea of FFT was to break down the overall problem into several small problems to solve separately and then merge them to obtain the result. The most used algorithm was Cooley–Tukey [13]. Decompose the discrete Fourier transform with length N into two subtransformations with length $\frac{N}{2}$. Firstly, assuming that N is an integer power of 2, $N = 2^m$, divided the signal $x[n]$ into odd and even parts:

$$x[n] = x_{\text{even}}\left[\frac{n}{2}\right] + x_{\text{odd}}\left[\frac{n}{2}\right]. \tag{1}$$

Performed DFT transformation on these two signals with a length of $\frac{N}{2}$ to obtain the following:

TABLE 1 | Contrasting our work with the literature.

References	Automatic feature extraction	Time–frequency representation	Long-range temporal dependency modelling	Attention/ transformer mechanism	Capability for complex PQDs
[5]	√	×	×	×	√
[6]	√	√	×	×	√
[7]	√	√	×	√	√
[9]	√	×	√	√	×
[10]	√	√	√	√	×
[11]	√	×	√	√	×
[12]	√	×	√	√	×
Our work	√	√	√	√	√

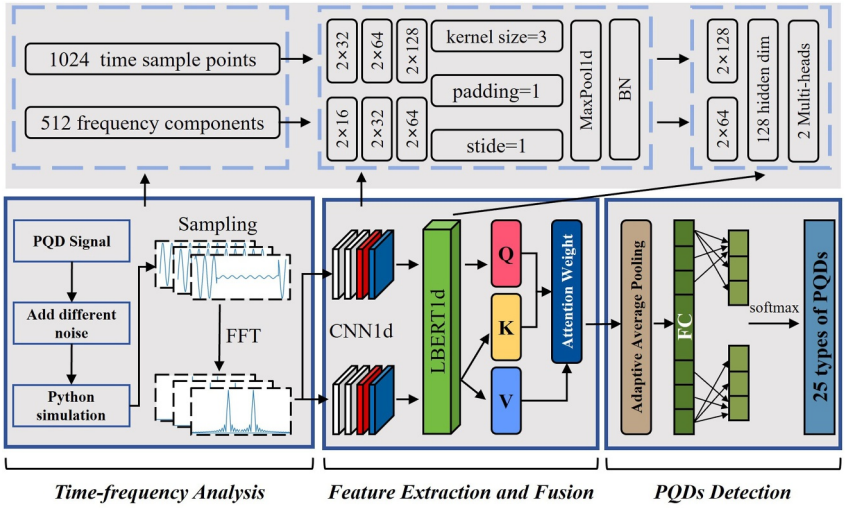

FIGURE 1 | Architecture of the MCNN1d-LBERT1d-SCA for complex PQDs classification.

$$X_{\text{even}}[k] = \sum_{n=0}^{\frac{N}{2}-1} x_{\text{even}}[n] \cdot e^{-j\frac{2\pi}{\frac{N}{2}}kn}, \tag{2}$$

$$X_{\text{odd}}[k] = \sum_{n=0}^{\frac{N}{2}-1} x_{\text{odd}}[n] \cdot e^{-j\frac{2\pi}{\frac{N}{2}}kn}. \tag{3}$$

Merged the results of the above two DFT subtransformations to obtain the following:

$$X[k] = X_{\text{even}}[k] + e^{-j\frac{2\pi}{N}k} \cdot X_{\text{odd}}[k], \tag{4}$$

$$X\left[k + \frac{N}{2}\right] = X_{\text{even}}[k] - e^{-j\frac{2\pi}{N}k} \cdot X_{\text{odd}}[k]. \tag{5}$$

Through the above splitting and recursive methods, FFT significantly improved the computational efficiency of converting disturbance signals from the time domain to the frequency domain, made it easier to handle large-scale disturbance signal data.

2.2 | Multimodal Convolutional Neural Network

CNN, widely applied in computer vision, has also proven effective for one-dimensional classification and regression tasks [14]. VGGNet, a representative CNN, stacks convolutional and pooling layers into blocks typically using 3 × 3 convolutions with padding 1 and stride 1 and 2 × 2 pooling with stride 2, achieving strong classification performance in various applications [15]. To adapt VGG for PQD recognition, a one-dimensional CNN with max pooling and batch normalisation was employed to simplify the network and reduce parameters. Each disturbance signal contained 1024 time-domain samples and 512 frequency components with the time-domain network structured as (2,32) (2,64) (2,128) and the frequency-domain network as (2,16) (2,32) (2,64). Two parallel channels, each comprising three convolutional blocks with two convolutional layers, one pooling and one normalisation layer, formed the improved VGG structure illustrated in Figure 2.

The batch normalisation layer of each convolutional layer block can reduce the internal covariate shift of the neural network while also improving the training speed of the model, meeting the need to quickly identify power quality disturbances. Batch

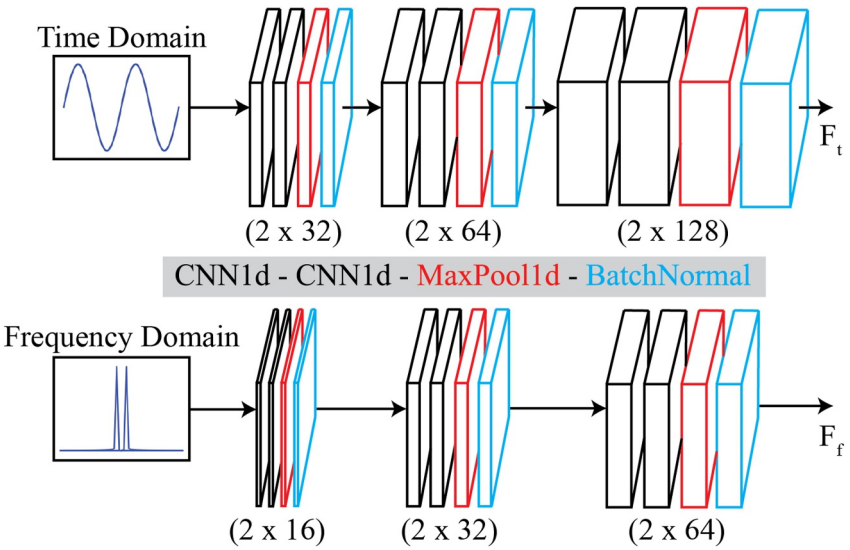

FIGURE 2 | Structure of modified VGG.

normalisation effectively solves the problems of slow training and overfitting in VGG networks when facing complex data features [16]. For a t-dimensional input $x = (x^{(1)}, x^{(2)}, ..., x^{(t)})$, batch normalisation first normalised each dimension as follows:

$$\hat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}] + \epsilon}}, 1 \le k \le t, \tag{6}$$

where ε is a constant added to the batch data variance to maintain numerical stability with both the expectation and variance coming from the training set. After normalisation, the values were scaled and shifted by introducing parameters $\gamma^{(k)}$ and $\beta^{(k)}$:

$$y^{(k)} = \gamma^{(k)}\hat{x}^{(k)} + \beta^{(k)}, 1 \le k \le t. \tag{7}$$

The parameters $\gamma^{(k)}$ and $\beta^{(k)}$ were set based on the parameters of the original model. In order to stabilise the distribution of inputs, normalisation was performed on each batch of data to make its mean 0 and variance 1, thereby improving the convergence speed of the neural network and shortened training time.

2.3 | Lightweight-BERT1d

The Transformer model overcomes RNN limitations such as sequential computation and information loss by providing effective context encoding [8]. BERT, built on bidirectional Transformer encoders, achieves strong performance in diverse NLP tasks through pretraining on unlabelled text with its bidirectional structure enhancing contextual understanding [17]. Inspired by BERT's success and the characteristics of PQDs, this study proposes a lightweight one-dimensional BERT (LBERT1d) with two encoder layers, 128 hidden units and two attention heads, balancing training efficiency and recognition accuracy while meeting power system security requirements as illustrated in Figure 3.

Each encoder layer comprised positional encoding, multihead self-attention, residual connections, layer normalisation and a feedforward network (FFN). Multihead attention generated multiple attention weights through linear transformations, concatenated them and applied a final linear projection. Residual connections enabled direct signal transmission across layers, mitigating gradient vanishing and accelerating training, whereas layer normalisation stabilised training by standardising inputs for serialised and variable-length data. The FFN contained two linear layers with a ReLU activation, mapping the 128-dimensional input to a higher dimension and back, thereby enhancing nonlinear representation and model expressiveness.

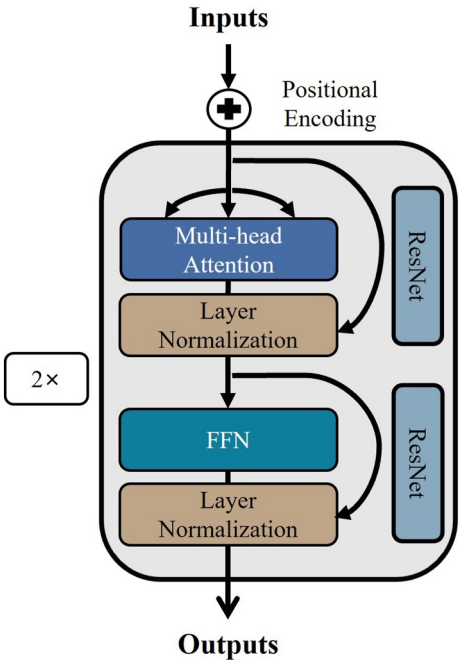

FIGURE 3 | Structure of lightweight-BERT1d.

2.4 | Signal-Based Cross-Attention

This study proposed a signal-based cross-attention (SCA) module. The time-domain features of the disturbance signal were

used as Query, and the frequency-domain features were used as Key and Value. After feature fusion, the spatial dimension was fixed through an adaptive average pooling layer. The output of the adaptive average pooling layer was flattened into a two-dimensional tensor, and a fully connected layer generated the classification results. The SCA module consisted of linear transformation layers, multihead attention, addition and normalisation and a feedforward network as shown in Figure 4.

After multiple feature extraction using CNN1d and LBERT1d, time-domain and frequency-domain features F_Q^{time}, F_K^{fre} and F_V^{fre} were obtained. The time-domain features were used as Query, and the frequency-domain features were used as Key and Value. Further performed a linear transformation on the features as shown below:

$$Q_{\text{input}}^{\text{time}} = M_Q \otimes F_Q^{\text{time}}, \tag{8}$$

$$K_{\text{input}}^{\text{fre}} = M_K \otimes F_K^{\text{fre}}, \tag{9}$$

$$V_{\text{input}}^{\text{fre}} = M_V \otimes F_V^{\text{fre}}, \tag{10}$$

where M_Q, M_K and M_V are linear transformation matrices, $Q_{\text{input}}^{\text{time}}$, $K_{\text{input}}^{\text{fre}}$ and $V_{\text{input}}^{\text{fre}}$ are transformed attention input features and '⊗' represents matrix multiplication. Then, $Q_{\text{input}}^{\text{time}}$, $K_{\text{input}}^{\text{fre}}$ and $V_{\text{input}}^{\text{fre}}$ were input into the multihead attention mechanism for operation:

$$\mathcal{T}\left(Q_{\text{input}}^{\text{time}}, K_{\text{input}}^{\text{fre}}, V_{\text{input}}^{\text{fre}}\right) = \sigma\left(\frac{Q_{\text{input}}^{\text{time}}\left(K_{\text{input}}^{\text{fre}}\right)^{\text{T}}}{\sqrt{C}}\right) \otimes V_{\text{input}}^{\text{fre}}, \tag{11}$$

where $\sigma(\cdot)$ is the Softmax activation function, and C is the feature dimension that evenly distributes input features to each attention head. This study used two attention heads. $\mathcal{T}(\cdot)$ represents the attention-weighted feature function. Then, perform feature fusion on the multihead attention output:

$$F_{\text{fusion}} = \sigma\left(\mathcal{A}\left(\mathcal{F}\left(\mathcal{A}\left(\mathcal{T}\left(Q_{\text{input}}^{\text{time}}, K_{\text{input}}^{\text{fre}}, V_{\text{input}}^{\text{fre}}\right)\right)\right)\right)\right), \tag{12}$$

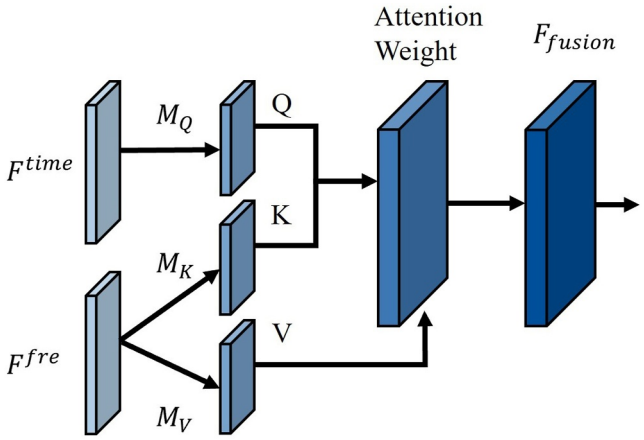

FIGURE 4 | Structure of signal-based cross-attention.

where F_{fusion} is the final fused attention feature, $\mathcal{A}(\cdot)$ is the sum and normalisation of ResNet and $\mathcal{F}(\cdot)$ is an FFN consisting of 2 linear layers and one activation function.

3 | Methodology

Figure 5 illustrates the overall workflow of the proposed MCNN1d–LBERT1d–SCA framework for recognising power quality disturbances. The flowchart outlines the main processing stages, including data generation and preprocessing, time–frequency feature extraction, global feature modelling, cross-attention-based feature fusion and final classification and validation, providing a clear overview of the methodological pipeline adopted in this study.

3.1 | PQDs Dataset

To ensure credibility, a synthetic PQD dataset was generated in Python based on IEEE Std. 1159, comprising 25 disturbance types with 1000 waveforms each, totalling 25,000 samples [18]. Each waveform contained 10 cycles at 50 Hz sampled at 5120 Hz for 0.2 s with 1024 points. To simulate real grid conditions, Gaussian white noise with different signal-to-noise ratios of 20, 30 and 40 dB was added to all disturbance signals.

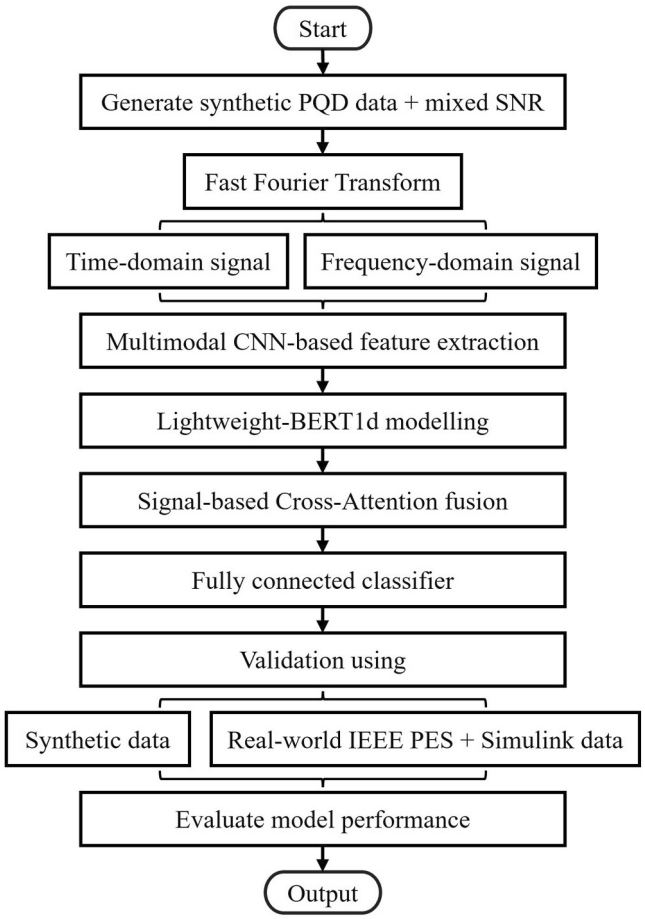

FIGURE 5 | Flowchart of the proposed MCNN1d–LBERT1d–SCA framework.

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right), \tag{13}$$

where P_{signal} and P_{noise} denote the average power of the signal and noise, respectively. In signal processing, the signal-to-noise ratio (SNR), expressed in decibels (dB), quantifies the relative strength of a signal with respect to background noise and is defined as the logarithmic ratio of the signal power to the noise power. A larger SNR value in dB indicates that the signal power dominates the noise power corresponding to cleaner signals and weaker noise interference. Conversely, a smaller SNR value implies that the noise power becomes comparable to or exceeds the signal power resulting in more severe signal corruption. Therefore, the dB value does not represent the absolute magnitude of the noise itself but rather reflects the balance between the signal and noise energies.

Unlike earlier studies that train separate models under each SNR, this work employed a unified mixed-SNR training strategy where samples from all noise levels were randomly combined during training and validation. This design enables the model to learn noise-invariant representations and improves its robustness against varying noise intensities. The dataset was divided into 70% for training, 20% for validation and 10% for testing. Model implementation in TensorFlow with Keras used 50 training epochs, the Adam optimiser and a 50% dropout rate in each layer to mitigate overfitting and enhance generalisation performance.

Table 2 summarises the key architectural and training parameters of the proposed model. It reports the essential configurations of each core module, including the multimodal CNN, lightweight BERT1d and signal-based cross-attention mechanism as well as the main training settings such as optimiser, learning rate, number of epochs and dropout rate. By explicitly listing these critical hyperparameters, Table 2 provides a clear and transparent reference for the model training process, thereby enhancing the reproducibility and reliability of the proposed approach.

3.2 | Training Process

Figure 6 shows the learning curves under noise levels noiseless, 20, 30 and 40 dB. With a batch size of 64 and a learning rate of 0.0003, the curves stabilised after 10 iterations. After 50 iterations, the training accuracies reached 99.70%, 98.50%, 99.80% and 99.85%, whereas validation accuracies were 99.80%, 98.05%, 99.60% and 99.70% with only a 0.95% gap between training and validation, confirming robust feature extraction. However, validation loss at 20 dB was notably higher than training loss. After the 17th iteration, it increased, indicating slight overfitting under strong noise, suggesting the need for early stopping or regularisation.

4 | Result and Analysis

4.1 | Performance Verification Under Different Noises

Table 3 reports the recognition accuracy of the proposed model under different noise levels with averages of 99.78%, 98.08%, 99.61% and 99.73% at noiseless, 20, 30 and 40 dB, demonstrating excellent classification performance.

At Noiseless, 14 PQDs were fully recognised; the rest exceeded 99.5% accuracy. At an SNR of 20 dB, the accuracy dropped by

TABLE 2 | Key architectural and training parameters of the proposed model.

Module	Parameter	Value
Input	Time-domain signal length	1024 samples
	Frequency-domain features	512-point FFT magnitude
MCNN1d (time domain)	Convolution blocks	3
	Channel configuration	(2,32) → (2,64) → (2128)
MCNN1d (frequency domain)	Convolution blocks	3
	Channel configuration	(2,16) → (2,32) → (2,64)
LBERT1d	Encoder layers	2
	Hidden dimension	128
	Attention heads	2
Signal-based cross-attention (SCA)	Query	Time-domain features
	Key/value	Frequency-domain features
	Attention heads	2
Feature aggregation	Pooling	Adaptive average pooling
Classifier	Output classes	25
Training	Optimiser	Adam
	Learning rate	0.0003
	Epochs	50
	Dropout rate	0.5

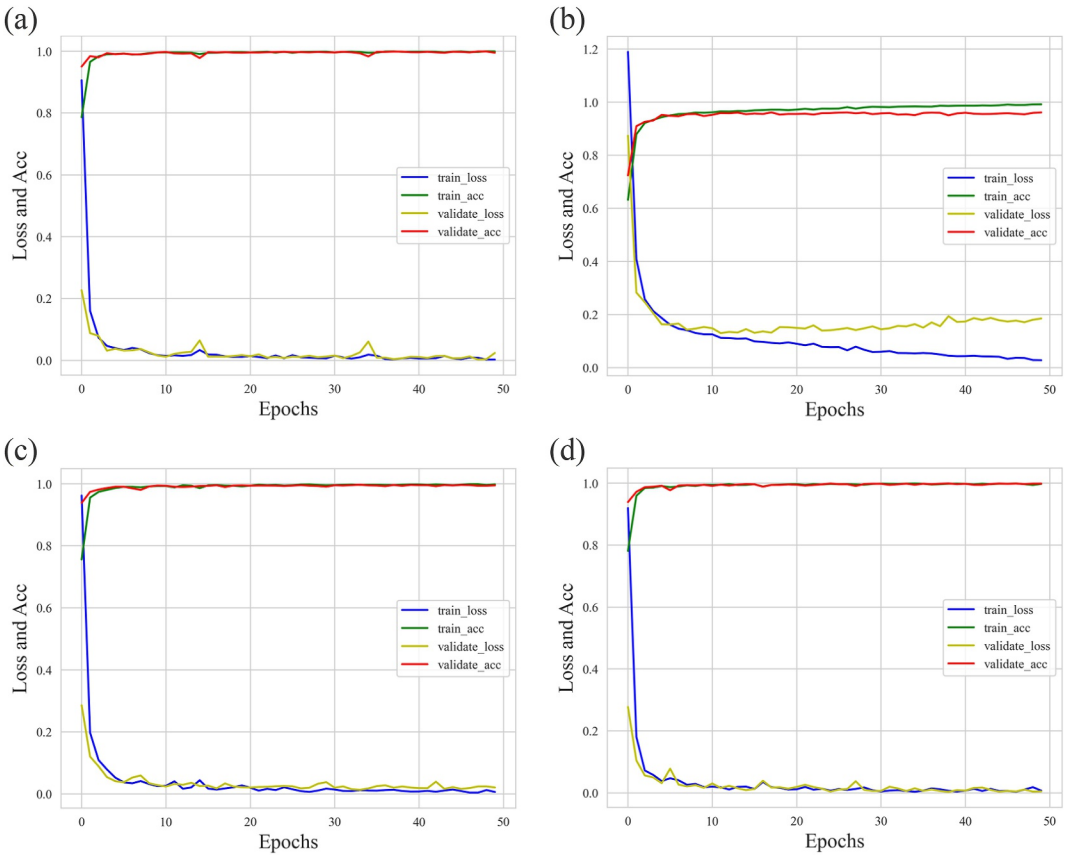

FIGURE 6 | Learning curve diagrams in different noise levels. (a) Noiseless, (b) 20 dB, (c) 30 dB, (d) 40 dB.

1.7% with two disturbance types falling below 96%. Nevertheless, the overall accuracy of 98.08% demonstrates strong robustness under relatively low SNR conditions. At 30 and 40 dB, recognition accuracy was restored for most PQDs with the lowest values of 97.19% and 97.10% for C18 showing consistent robustness.

The confusion matrix of the proposed model in 20 dB shows in Figure 7. It can be observed that the model faces difficulties in identifying C1 (Normal), C9 (Notching), C10 (Spike) and C23 (Harmonics + Transients oscillatory + Swell) with < 900 identified samples. Among them, C10 (Spike) has the lowest recognition rate. Among the 1000 samples in the test set, 113 were identified as C1 (Normal), 6 were identified as C8 (Transitional oscillatory) and 46 were identified as C9 (Notching).

As shown in Figure 8, the waveforms of C1 (Normal) and C10 (Spike) are very similar in the dataset created. The waveforms of both disturbance signals exhibit periodic changes in overall trend with amplitudes ranging from −1 to 1. The heights and positions of the peaks and valleys of the two waveforms are also very similar without any significant deviation. For C10, it can be observed that there are peaks at certain specific points, but these peaks are not obvious. Although the waveform appears similar overall, the interference signal in the C10 plot has higher frequency components in certain local areas. The waveforms of these two PQD signals exhibit high similarity in overall trend, frequency, amplitude and phase resulting in incorrect recognition of the proposed model.

4.2 | Classification Performance Comparison

To validate the classification performance of our proposed MCNN1d-LBERT1d-SCA model, it was compared with several recently proposed classification frameworks as listed in Table 4. It is worth mentioning that 'Manual' denotes feature extraction methods requiring domain expertise and predefined rules, whereas 'Automatic' refers to frameworks that autonomously perform feature extraction during the classification process.

Table 4 indicates that classification frameworks with automatic feature extraction achieve consistently higher accuracy than manual methods with models such as 3D-VSC-ER + MVCNN and IMTF + AFF-ResNetXt50 reaching 98.23% and 98.50%. In comparison, most manual approaches fall below 98%. The proposed MCNN1d-LBERT1d-SCA attains the highest accuracy of 99.30% across 25 PQD types, highlighting its superior feature extraction, robustness and generalisation. In comparison, CNN-GRU-P achieved 99.00% but was tested on only 12 disturbance types limiting its practical applicability.

4.3 | Model Ablation Experiment Performance

In order to further compare the performance of the proposed model, this study conducted an ablation experiment to evaluate its effectiveness in classifying PQDs. By gradually removing the components of the proposed model and comparing it with the base model, the contribution of each module was evaluated.

TABLE 3 | Proposed model performance on four different noise levels.

Disturbance term		Accuracy (%)			
		Noiseless	20 dB	30 dB	40 dB
C1	Normal	100.00	96.19	99.90	100.00
C2	Swell	100.00	96.89	99.50	100.00
C3	Sag	99.40	96.90	99.60	97.80
C4	Harmonics	100.00	99.30	100.00	100.00
C5	Flicker	100.00	99.20	99.90	100.00
C6	Interruption	100.00	99.60	98.80	100.00
C7	Transient impulsive	100.00	99.28	100.00	100.00
C8	Transient oscillatory	100.00	99.28	100.00	100.00
C9	Notching	99.60	98.39	98.90	100.00
C10	Spike	99.90	95.27	99.60	99.90
C11	Harmonics + swell	100.00	98.60	99.50	100.00
C12	Harmonics + sag	99.90	99.60	100.00	100.00
C13	Harmonics + interruption	100.00	99.70	100.00	100.00
C14	Harmonics + flicker	100.00	99.30	99.80	100.00
C15	Harmonics + transient impulsive	99.40	98.60	99.80	99.90
C16	Harmonics + transient oscillatory	100.00	96.49	99.90	99.80
C17	Flicker + swell	98.00	97.80	99.10	99.40
C18	Flicker + sag	99.50	96.90	97.19	97.10
C19	Flicker + transient oscillatory	100.00	98.20	100.00	100.00
C20	Flicker + transient impulsive	99.80	99.60	99.80	99.90
C21	Transient oscillatory + swell	100.00	97.98	100.00	99.90
C22	Transient oscillatory + sag	99.70	97.30	99.80	99.90
C23	Harmonics + transients oscillatory + swell	100.00	95.75	99.40	99.90
C24	Harmonics + transients oscillatory + sag	99.60	97.19	100.00	99.80
C25	Harmonics + transients oscillatory + flicker	99.80	98.79	100.00	100.00
Average		99.78	98.08	99.61	99.73

This experiment was conducted at different SNRs to evaluate noise robustness and classification accuracy. The experimental results are shown in Table 5.

Table 5 presents the ablation results under different noise conditions, allowing a systematic evaluation of the individual and combined contributions of lightweight BERT1d and signal-based cross-attention. Among the single-module baselines, MCNN1d consistently outperforms CNN1d and CNN2d across all signal-to-noise ratios, indicating that multiscale convolution enhances robustness to noise by capturing complementary local features. Introducing lightweight BERT1d into the convolutional backbone yields a substantial performance improvement, as evidenced by the increase from 95.51% to 97.46% in average accuracy when moving from MCNN1d to MCNN1d–LBERT1d. This gain confirms that lightweight BERT1d effectively models global temporal dependencies that cannot be fully captured by convolution alone. When signal-based cross-attention is applied without LBERT1d, MCNN1d–SCA further improves the average accuracy to 98.01%, demonstrating that attention-based feature refinement contributes additional robustness under noisy conditions. Notably, the full model MCNN1d–LBERT1d–SCA achieves the best performance across all noise levels with an average accuracy of 99.30% significantly surpassing all ablated variants. This result indicates that lightweight BERT1d and signal-based cross-attention provide complementary benefits where the former establishes strong global contextual representations and the latter selectively enhances informative signal components.

4.4 | Discussion

The motivation for adopting a lightweight BERT architecture in this study is grounded in recent theoretical and empirical advances in efficient Transformer design. Existing studies on lightweight BERT variants have demonstrated that full-scale BERT models are often over-parameterised for domain-specific sequence modelling tasks and that carefully constrained architectures can preserve contextual representation capability while substantially reducing computational cost. In particular, the lightweight BERT framework demonstrates that reducing encoder depth and hidden dimensionality does not necessarily degrade sequence-level semantic modelling, provided that self-

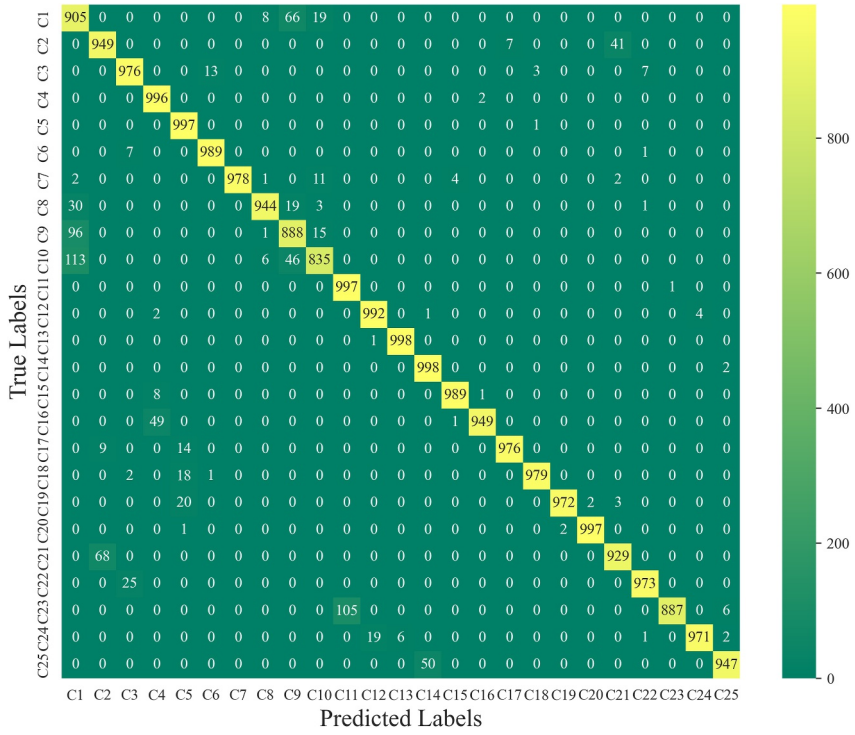

FIGURE 7 | Confusion matrix results in 20 dB.

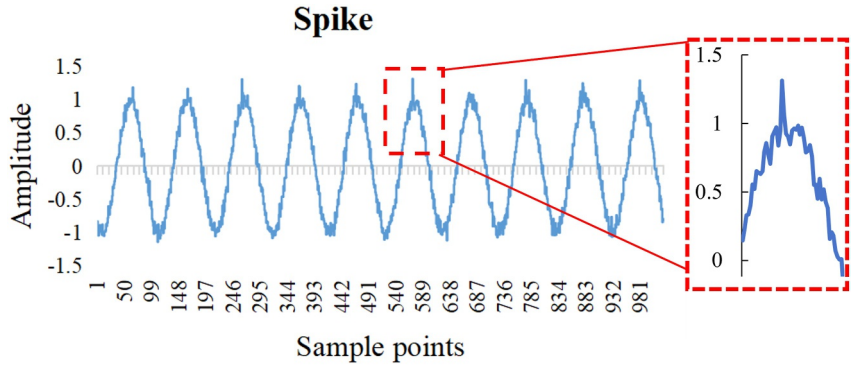

FIGURE 8 | Spike signal waveform.

TABLE 4 | Performance comparison with other techniques.

Classification method	No. of PQD	Feature extraction	Average accuracy (%)
WT-SVM [19]	6	Manual	96.37
TQWT-MSVM [20]	14	Manual	97.60
2 × 1DCNN-BiGRU [7]	20	Automatic	97.91
3D-VSC-ER + MVCNN [21]	16	Automatic	98.23
IMTF + AFF-ResNetXt50 [22]	14	Automatic	98.50
ACMP + GOA-SVM [23]	16	Manual	98.69
CNN-LSTM [6]	16	Automatic	98.95
CNN-GRU-P [5]	12	Automatic	99.00
MCNN1d-LBERT1d-SCA	**25**	**Automatic**	**99.30**

Note: Bold text is used to highlight its superior performance compared to the other models.

attention mechanisms sufficiently capture global dependencies [24]. Furthermore, MicroBERT provides a systematic theoretical basis for compressing BERT-style architectures by emphasising that shallow Transformer encoders can effectively retain sentence or sequence-level contextual information while avoiding redundancy introduced by deep multilayer stacks [25]. The study

demonstrates that lightweight Transformer configurations achieve competitive performance by prioritising contextual alignment and efficient representation learning rather than architectural scale.

When transferred to engineering signal analysis, these theoretical insights are particularly relevant. One-dimensional engineering signals are typically characterised by limited sequence length, strong local-global coupling and high sensitivity to noise. Under such conditions, excessively deep or wide Transformer architectures tend to introduce unnecessary parameters and increase the risk of overfitting without improving dependency modelling. Therefore, a lightweight BERT configuration is theoretically sufficient to capture long-range temporal dependencies while maintaining computational efficiency and robustness.

To further elucidate the sources of performance improvement, a paired *t*-test was conducted to examine the statistical significance of accuracy differences between model variants as summarised in Table 6. The paired *t*-test evaluates whether the mean difference between two paired samples is significantly different from 0 and is therefore well-suited for assessing performance variations among models trained and tested under identical experimental settings. In this analysis, the time–frequency-based MCNN architecture was adopted as the baseline, and the effects of incorporating different functional modules were systematically compared. The paired comparisons between MCNN1d and MCNN1d–LBERT1d as well as between MCNN1d–SCA and MCNN1d–LBERT1d–SCA indicate that the introduction of lightweight BERT1d yields the most significant performance gains with accuracy improvements of 1.95% and 1.29%, respectively. Notably, the corresponding 95% confidence intervals, namely [0.81%, 3.09%] and [0.38%, 2.20%], do not include 0, further confirming that the observed improvements are statistically significant. Together with *p* values below 0.01, these results provide robust statistical evidence that lightweight BERT1d contributes most substantially to overall performance enhancement by enabling the effective modelling of global temporal dependencies beyond the capabilities of time–frequency convolution alone.

To further investigate the influence of LBERT1d on the overall model performance, a more detailed comparative study was conducted by systematically varying the number of encoder layers, hidden units and attention heads as summarised in Table 7. This experiment aims to examine how different LBERT1d configurations affect model size in terms of parameters and memory footprint, computational efficiency measured by FLOPs and inference speed as well as predictive performance reflected by average accuracy.

From the efficiency perspective, Table 7 shows that increasing the encoder depth or hidden dimensionality leads to a substantial growth in parameters, computational complexity and inference latency. For instance, configurations with 256 hidden units or 3 encoder layers exhibit significantly higher FLOPs and memory consumption, whereas the corresponding accuracy gains become marginal or even decline. Conversely, configurations with a single encoder layer or small hidden dimensions suffer from insufficient representational capacity resulting in noticeably lower classification accuracy. These observations indicate that neither overly small nor excessively large LBERT1d configurations are optimal for the considered task.

Among all evaluated settings, configuration M9, which employs 2 encoder layers with 128 hidden units and 2 attention heads, achieves the highest average accuracy while maintaining a

TABLE 5 | Performance of ablation experiment.

	Accuracy (%)				
Classification method	**Noiseless**	**20 dB**	**30 dB**	**40 dB**	**Avg**
LBERT1d	86.57	65.89	75.70	82.07	77.56
CNN2d	95.24	92.43	94.04	94.28	93.99
CNN1d	96.36	92.99	95.01	95.66	95.00
MCNN1d	96.79	93.14	95.84	96.28	95.51
CNN1d-LBERT1d	98.31	93.21	97.36	97.35	96.56
MCNN1d-LBERT1d	98.86	94.34	98.20	98.46	97.46
MCNN1d-SCA	99.14	95.72	98.46	98.73	98.01
MCNN1d-LBERT1d-SCA	**99.78**	**98.08**	**99.61**	**99.73**	**99.30**

Note: Bold text is used to highlight its superior performance compared to the other models.

TABLE 6 | Paired *t*-test results of the proposed model components.

Paired *t*-test	**Accuracy difference (%)**	***p* value**	***T* value**	**Standard error**	**95% confidence interval of difference**
MCNN1d: MCNN1d-LBERT1d	1.95	0.002	3.58	0.54	[0.81%, 3.09%]
MCNN1d-LBERT1d: MCNN1d-SCA	0.55	0.031	2.20	0.25	[0.04%, 1.06%]
MCNN1d-SCA: MCNN1d-LBERT1d-SCA	1.29	0.006	3.01	0.43	[0.38%, 2.20%]

TABLE 7 | Performance and efficiency comparison of different LBERT1d configurations.

Model	Encoder layers	Hidden units	Attention heads	Parameters (M)	Computational complexity (GFLOPs)	Inference speed (ms)	Memory footprint (FP32)	Average accuracy (%)
M1	1	64	2	0.037	0.87	0.06	0.15 MB	94.91
M2	1	64	4	0.037	0.87	0.07	0.15 MB	95.05
M3	1	128	2	0.150	3.49	0.23	0.60 MB	97.80
M4	1	128	4	0.150	3.49	0.24	0.60 MB	97.95
M5	1	256	2	0.598	13.96	0.92	2.40 MB	98.40
M6	1	256	4	0.598	13.96	0.97	2.40 MB	98.45
M7	2	64	2	0.075	1.75	0.12	0.30 MB	96.60
M8	2	64	4	0.075	1.75	0.13	0.30 MB	96.75
M9	**2**	**128**	**2**	**0.299**	**6.98**	**0.46**	**1.20 MB**	**99.30**
M10	2	128	4	0.299	6.98	0.48	1.20 MB	99.05
M11	2	256	2	1.196	27.92	1.84	4.80 MB	98.85
M12	2	256	4	1.196	27.92	1.93	4.80 MB	98.90
M13	3	64	2	0.112	2.62	0.17	0.45 MB	96.30
M14	3	64	4	0.112	2.62	0.18	0.45 MB	96.45
M15	3	128	2	0.448	10.47	0.69	1.80 MB	99.00
M16	3	128	4	0.448	10.47	0.72	1.80 MB	98.95
M17	3	256	2	1.794	41.88	2.76	7.20 MB	98.70
M18	3	256	4	1.794	41.88	2.90	7.20 MB	98.65
Standard BERT	12	768	12	110	22.14	6.27	440 MB	98.90

Note: Bold text indicates that the structural configuration of M9 achieves the best results in the proposed model.

compact parameter size, a moderate computational cost and a fast inference speed. Compared with deeper or wider alternatives, M9 provides a superior balance between model complexity and predictive performance, demonstrating that effective global dependency modelling can be achieved without resorting to excessive architectural scale. This result justifies the selection of M9 as the final LBERT1d configuration used in the proposed framework.

It is also observed that varying the number of attention heads from two to four results in nearly identical parameter counts, computational complexity and memory footprints. This phenomenon arises because multihead self-attention partitions a fixed hidden representation into multiple subspaces without increasing the overall dimensionality of the projection. As a consequence, the theoretical number of parameters and FLOPs remains unchanged when the hidden dimension is fixed, and only minor differences in inference speed are observed due to implementation-level overhead rather than structural complexity.

Finally, although the configuration M18 exhibits a higher computational complexity than Standard BERT, this does not indicate a heavier architecture in general. Instead, this behaviour reflects the quadratic dependence of self-attention complexity on the effective sequence length. For LBERT1d-based models, FLOPs are computed under task-specific, longer effective sequences derived from dense one-dimensional signal representations. In contrast, the Standard BERT complexity is reported for a reference sequence length of 128 tokens. Consequently, models with fewer layers and smaller hidden dimensions may still incur higher FLOPs when operating on longer sequences. Taken together, the results in 6 demonstrate that the proposed lightweight design achieves superior performance through an appropriate configuration choice rather than increased model scale with M9 representing the optimal LBERT1d setting for this task.

5 | Experiment Validation

5.1 | Real-World Data Validation

A dataset from the IEEE PES database was used to evaluate the proposed approach on real-world data [26], comprising diverse power system disturbances with 1536 data points sampled at 256 points per cycle. Since the proposed model was trained on synthetic PQD sequences with 1024 sampling points per record, a preprocessing and resampling procedure was applied to ensure dimensional consistency before inference. Specifically, each real-world signal was first normalised and then resampled using linear interpolation to 1024 points, preserving the relative temporal and amplitude characteristics of the original waveform while matching the model input dimension. This step ensured that both datasets shared a consistent temporal resolution allowing the model to generalise effectively to real-world conditions without retraining. After preprocessing with packages of

dplyr and caret in R, the model achieved an overall recognition accuracy of 95.88%, as shown in Table 8, confirming its reliability and practical applicability.

5.2 | Simulation Data Validation

To further verify the generalisation performance of the proposed MCNN1d-LBERT1d-SCA under experimental signals, this study used MATLAB/Simulink to simulate and model power quality disturbances [27].

For clarity, the simulation parameters adopted in this study are based on the validated MATLAB/Simulink models reported in Ref. [27]. This reference presents a comprehensive framework for modelling a wide range of power quality disturbances, including distribution line faults, induction motor starting and transformer energising for voltage sag events, capacitor bank switching for oscillatory transient disturbances, lightning impulse models for impulsive transients, nonlinear load models for triplen harmonics and voltage notching as well as electric arc furnace models for flicker disturbances. Owing to its extensive coverage and established use in power quality analysis, research and educational applications, Ref. [27] provides a reliable and authoritative basis for parameter selection in the present simulation study.

The 5 simulation models designed for specific PQDs presents in Figure 9. Model 1, the Line Fault Model, simulated various line faults, including single line-to-ground, double line-to-ground, line-to-line, three-phase and multistage faults. Model 2, the capacitor bank energising model, simulated voltage oscillatory transients caused by capacitor bank energising for power factor correction. Model 3, the three-phase nonlinear load model, simulated voltage notches and harmonics caused by a 6-pulse three-phase rectifier. Model 4, the electric arc furnace model, simulated flicker disturbances caused by electric arc furnaces. Model 5, the lightning impulse model, simulated impulsive transients caused by lightning strikes near transmission lines. Table 9 shows the parameters in the simulation model and the single PQD types that can be simulated.

Simulations were conducted on 22 PQDs with a sampling frequency of 5120 Hz, 1024 points per signal and 200 ms duration, though C15, C18 and C20 could not be modelled due to structural limitations. For each disturbance, 100 samples were generated, and the results in Table 10 show that the proposed model achieved an average accuracy of 97.47%. The testing time per sample was 154 ms, below the 200 ms threshold, confirming that the MCNN1d-LBERT1d-SCA model provides accurate and efficient real-time POD detection.

6 | Conclusion

This study developed a hybrid deep learning framework, MCNN1d-LBERT1d-SCA, for the automatic recognition of complex power quality disturbances in modern power systems with high renewable energy penetration. By integrating fast Fourier transform for frequency-domain transformation, one-dimensional convolutional neural networks for local feature extraction, lightweight Bidirectional Encoder Representations from Transformers for global contextual learning, and a signal-based cross-attention mechanism for multimodal fusion, the proposed model effectively captures both short-term and long-range temporal dependencies. Experimental analyses on synthetic datasets constructed according to IEEE Std. 1159 demonstrated outstanding robustness and generalisation, achieving an average classification accuracy of 99.30% under

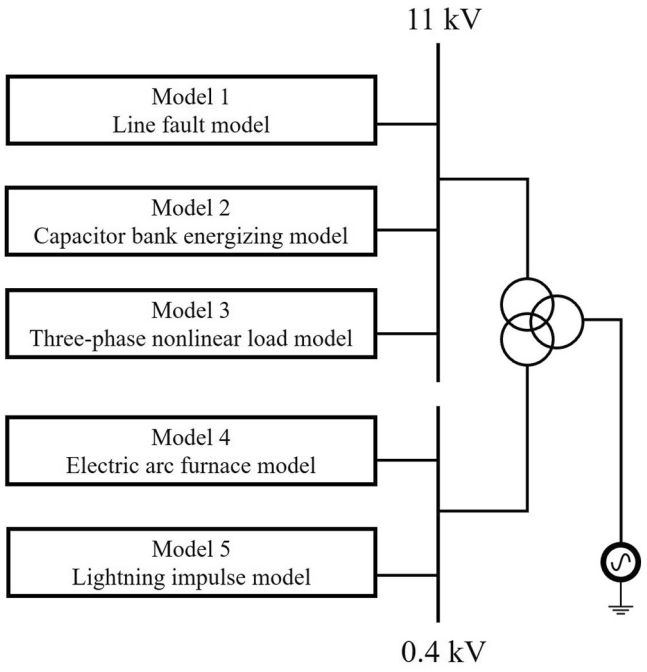

FIGURE 9 | Electrical diagram of the proposed simulation model.

TABLE 8 | Classification results of actual PQDs.

Types	Identification accuracy (%)	Average accuracy (%)
C4	95.00	95.88
C11	93.33	
C12	92.86	
C13	100.00	
C16	90.00	
C23	100.00	
C24	100.00	

TABLE 9 | Parameters of the simulation models.

Model	Three-phase source	Resistive load	Inductive load	Transformer	Types of PQDS
Line fault model	11 kV, 30 MVA, 50 Hz	10 kW	100 Var	11/0.4 kV, 1 MVA	C1 C3 C6
Capacitor bank energising model	11 kV, 30 MVA, 50 Hz	100 kW	100 kVar	11/0.4 kV, 1 MVA	C2 C8
Three-phase nonlinear load model	11 kV, 30 MVA, 50 Hz	600 V, 10 kW	600 V, 1 kVA	11/0.4 kV, 1 MVA	C2 C4 C9
Electric arc furnace model	0.4 kV, 1 MVA, 50 Hz				C5
Lightning impulse model	0.4 kV, 1 MVA, 50 Hz	10 kW	10 kVar		C7 C10

TABLE 10 | Accuracy under the experimental signals.

Disturbance term	Accuracy (%)	Average accuracy (%)	Test time per sample (ms)
C1	96.30	97.47	154
C2	98.12		
C3	90.67		
C4	99.65		
C5	99.58		
C6	97.45		
C7	97.14		
C8	98.37		
C9	95.62		
C10	96.42		
C11	99.75		
C12	95.88		
C13	99.88		
C14	99.88		
C16	98.80		
C17	93.63		
C19	99.40		
C21	97.52		
C22	97.45		
C23	96.70		
C24	97.22		
C25	98.88		

mixed signal-to-noise ratios. Comparative and ablation studies confirmed the complementary roles of CNN-based local representations and Transformer-based global attention in improving noise resilience and recognition accuracy. Validation on real-world IEEE PES and MATLAB/Simulink-generated datasets further verified the framework's adaptability, yielding accuracies of 95.88% and 97.47%, respectively, with inference times below 200 ms meeting the real-time requirements for PQD monitoring. The MCNN1d-LBERT1d-SCA framework demonstrates strong potential for deployment in smart grid environments, offering a reliable, scalable and efficient solution for intelligent power quality analysis.

Future work will aim to further extend the practical value of the proposed MCNN1d–LBERT1d–SCA framework. One important direction is to improve model interpretability by analysing attention weights and feature contributions, which can help link the learnt representations to physical characteristics of power quality disturbances. In addition, efforts will be made to optimise the model for real-time and edge-device deployment through lightweight design, model compression and efficient inference strategies. The framework can also be expanded to incorporate multi-sensor and multimodal measurements, such as combined voltage and current signals from distributed monitoring devices, to enhance robustness under more complex

grid conditions. Finally, adaptive learning strategies will be explored to enable the model to cope with changing operating environments and previously unseen disturbance patterns, further supporting its application in practical smart grid systems.

Author Contributions

Jiajian Lin: methodology, visualization, writing – original draft, writing – review and editing. **Hadi Nabipour Afrouzi:** conceptualization, methodology, supervision, writing – review and editing. **Mehran Motamed Ektesabi:** supervision, writing – review and editing. **Jalal Tavalaei:** conceptualization, data curation, methodology, project administration, supervision, writing – original draft, writing – review and editing.

Funding

The authors have nothing to report.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. B. Kroposki, "Integrating High Levels of Variable Renewable Energy Into Electric Power Systems," *Journal of Modern Power Systems and Clean Energy* 5, no. 6 (2017): 831–837, https://doi.org/10.1007/s40565-017-0339-3.

2. S. Ćosić and I. Vokony, "Deep Learning-Based Dynamic State Estimation for Frequency Stability Monitoring in Power Systems With High Penetration of Renewable Generation," in *International Conference on Electronics, Engineering Physics and Earth Science (EEPES'23)* (MDPI, 2023), 16, https://doi.org/10.3390/engproc2023041016.

3. S. Wang and H. Chen, "A Novel Deep Learning Method for the Classification of Power Quality Disturbances Using Deep Convolutional Neural Network," *Applied Energy* 235 (2019): 1126–1140, https://doi.org/10.1016/j.apenergy.2018.09.160.

4. M. Wang, Z. Deng, Y. Zhang, and Z. Zhu, "An Automatic Identification Framework for Complex Power Quality Disturbances Based on Ensemble CNN," *IEEE Access* 11 (2023): 56550–56560, https://doi.org/10.1109/ACCESS.2023.3273294.

5. J. Cai, K. Zhang, and H. Jiang, "Power Quality Disturbance Classification Based on Parallel Fusion of CNN and GRU," *Energies (Basel)* 16, no. 10 (2023): 4029, https://doi.org/10.3390/en16104029.

6. S. Cen, D. O. Kim, and C. G. Lim, "A Fused CNN-LSTM Model Using FFT With Application to Real-Time Power Quality Disturbances Recognition," *Energy Science & Engineering* 11, no. 7 (2023): 2267–2280, https://doi.org/10.1002/ese3.1450.

7. L. Chen, S. Chen, J. Xu, and C. Zhou, "Power Quality Disturbances Identification Based on Deep Neural Network Model of Time-Frequency Feature Fusion," *Electric Power Systems Research* 231 (2024): 110283, https://doi.org/10.1016/j.epsr.2024.110283.

8. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, Vol. 30, ed. I. Guyon, U. V. Luxburg, S. Bengio, et al. (Curran Associates, Inc., 2017).

9. T. Miao, L. Sha, K. Huang, Y. Li, and B. Liu, "TFCA-TransNet: Convolutional Time–Frequency–Spatial Feature Fusion With Channel Attention Transformer Network for EEG-MI Signal Decoding," *Biomedical Signal Processing and Control* 112 (2026): 108692, https://doi.org/10.1016/j.bspc.2025.108692.

10. W. Wu, N. Zhou, X. Liang, W. Gui, C. Yang, and Y. Liu, "TCAC-Transformer: A Fast Convolutional Transformer With Temporal-Channel Attention for Efficient Industrial Fault Diagnosis," *Expert Systems with Applications* 297 (2026): 129473, https://doi.org/10.1016/j.eswa.2025.129473.

11. H. Li, D. Ma, L. Wang, W. Zhang, and W. Mo, "Ultra-Short Sequence-Augmented Power System Transient Stability Assessment Using Transformer-Based Deep Learning," *Sustainable Energy Technologies and Assessments* 81 (2025): 104432, https://doi.org/10.1016/j.seta.2025.104432.

12. A. Ahmad, X. Xiao, H. Mo, and D. Dong, "TFTformer: A Novel Transformer Based Model for Short-Term Load Forecasting," *International Journal of Electrical Power & Energy Systems* 166 (2025): 110549, https://doi.org/10.1016/j.ijepes.2025.110549.

13. J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series," *Mathematics of Computation* 19, no. 90 (1965): 297–301, https://doi.org/10.1090/S0025-5718-1965-0178586-1.

14. L. Alzubaidi, J. Zhang, A. J. Humaidi, et al., "Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions," *Journal of Big Data* 8, no. 1 (2021): 53, https://doi.org/10.1186/s40537-021-00444-8.

15. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," (2014).

16. A. A. Syahrin, D. O. Anggriawan, E. Prasetyono, et al., "Real-Time Detection of Power Quality Disturbance Using Fast Fourier Transform and Adaptive Neuro-Fuzzy Inference System," *Jurnal Rekayasa Elektrika* 20, no. 1 (2024), https://doi.org/10.17529/jre.v20i1.33695.

17. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, and Google AI Language, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," (n.d.).

18. Committee of the IEEE Power D, Society E, *IEEE Recommended Practice for Monitoring Electric Power Quality* (IEEE Power & Energy Society, 2009). 1159 TM, https://doi.org/10.1109/IEEESTD.1994.8684637.

19. B. Choudhary, "An Advanced Genetic Algorithm With Improved Support Vector Machine for Multi-Class Classification of Real Power Quality Events," *Electric Power Systems Research* 191 (2021): 106879, https://doi.org/10.1016/j.epsr.2020.106879.

20. K. Thirumala, M. S. Prasad, T. Jain, and A. C. Umarikar, "Tunable-Q Wavelet Transform and Dual Multiclass SVM for Online Automatic Detection of Power Quality Disturbances," *IEEE Transactions on Smart Grid* 9, no. 4 (2018): 3018–3028, https://doi.org/10.1109/TSG.2016.2624313.

21. J. Song, J. Wu, X. Wang, Z. Duan, X. Wang, and S. Lu, "Accurate Classification of Power Quality Disturbance Based on 3D Visualized Spiral Curve and Hybrid ER-MVCNN Model," *Measurement* 231 (2024): 114654, https://doi.org/10.1016/j.measurement.2024.114654.

22. Z. Duan, Z. Peng, J. Song, and S. Lu, "An Intelligent Complex Power Quality Disturbance Recognition Method Based on Two Dimension Encoding Conversion and Machine Vison," *Electric Power Systems Research* 232 (2024): 110413, https://doi.org/10.1016/j.epsr.2024.110413.

23. Z. T. Motlagh S and A. Akbari Foroud, "Power Quality Disturbances Recognition Using Adaptive Chirp Mode Pursuit and Grasshopper Optimized Support Vector Machines," *Measurement* 168 (2021): 108461, https://doi.org/10.1016/j.measurement.2020.108461.

24. Z. Wang, J. Li, S. Yang, X. Luo, D. Li, and S. Mahmoodi, "A Lightweight IoT Intrusion Detection Model Based on Improved BERT-of-Theseus," *Expert Systems with Applications* 238 (2024): 122045, https://doi.org/10.1016/j.eswa.2023.122045.

25. D. Zheng, J. Li, Y. Yang, Y. Wang, and P. C.-I. Pang, "MicroBERT: Distilling MoE-Based Knowledge From BERT Into a Lighter Model," *Applied Sciences* 14 (2024): 6171, https://doi.org/10.3390/app14146171.

26. T. Y. Ji, M. S. Li, and Q. H. Wu, "Disturbance Detection Using Hit-or-Miss Wavelet Singular Entropy for Power Quality Monitoring," in *2014 IEEE PES General Meeting | Conference & Exposition* (IEEE, 2014), 1–5, https://doi.org/10.1109/PESGM.2014.6939148.

27. R. H. G. Tan and V. K. Ramachandaramurthy, "A Comprehensive Modeling and Simulation of Power Quality Disturbances Using MATLAB/SIMULINK," in *Power Quality Issues in Distributed Generation* (InTech, 2015), https://doi.org/10.5772/61209.