Early Detection of Transformer Winding Faults Using Multimodal Vibration and Acoustic Sensor Data Fusion with Deep Learning

Usah, Emmamuel Okon[1]
Department of Physics University of Uyo,
Akwa Ibom State, Nigeria.
abassusah@yahoo.com

Kingsley Bassey Clement[2]
Department of Electrical/ Electronic Engineering
University of Uyo, Akwa Ibom State, Nigeria
kingsleyclement@uniuyo.edu.ng

Onwunta E. K. Onwunta[3]
Electrical & Electronic Engineering Department,
Federal University Otuoke, Bayelsa State
onwuntaoe@fuotuoke.edu.ng

Abstract

Power transformer winding faults may begin as small mechanical displacement, reduced clamping pressure, radial deformation, or incipient inter-turn weakness before they develop into severe insulation or short-circuit failure. This paper presents a non-invasive vibro-acoustic deep learning framework for early transformer winding fault detection. Synchronized vibration and acoustic recordings were obtained from a controlled laboratory transformer test bench and converted into CWT scalograms and Mel-spectrograms. A dual-branch CNN-BiLSTM attention model was trained to learn vibration and acoustic features separately before feature-level fusion and softmax classification. Actual model training was performed on a class-balanced prepared dataset of 600 raw recording runs and 7,200 synchronized windows, using leakage-aware splitting by raw recording run and five repeated training runs. The proposed fusion-attention model achieved 98.92 ± 0.24% accuracy under the prepared controlled dataset, with better recall than vibration-only, acoustic-only, and fusion-without-attention baselines. The result shows that vibration and acoustic sensor fusion can improve early warning of winding looseness, axial displacement, radial deformation, and incipient inter-turn fault conditions. Future work will validate the model on larger field transformer datasets with different ratings, cooling systems, and sensor placements.

Keywords: Power transformer, winding fault, vibration signal, acoustic signal, CNN, BiLSTM, attention mechanism, sensor fusion, condition monitoring, deep learning.

I. Introduction

A. Background

Power transformers are critical and expensive assets in electric power transmission and distribution networks. Their failure can lead to extended outages, fire risk, high replacement cost, and severe service disruption. Winding faults are particularly important because the winding is exposed to electromagnetic force, thermal aging, mechanical vibration, transport shock, insulation degradation, and repeated load cycling. A winding defect may begin as a small mechanical looseness, clamping pressure reduction, axial displacement, radial deformation, or incipient inter-turn stress. These early changes can alter vibration and acoustic signatures before conventional electrical indicators become obvious.

Vibration monitoring has become attractive because transformer winding and core movement are coupled to mechanical forces, magnetic flux, load current, and clamping condition. Acoustic and acoustic emission monitoring also provide useful information because abnormal mechanical impacts, insulation stress, and localized activity can produce audible or high-frequency elastic waves. A combined vibro-acoustic monitoring approach is therefore suitable for early warning, especially where field conditions, load variations, and environmental noise reduce the reliability of a single sensor.

B. Problem Statement

Traditional diagnostic methods such as frequency response analysis, dissolved gas analysis, partial discharge testing, and insulation resistance measurement have proven value in transformer maintenance. Their main limitation in early winding diagnosis is that some are offline, some require specialist interpretation, and some may respond more strongly after fault progression has already affected insulation or gas formation. Existing vibration-based methods have shown strong performance, but many rely on one sensing channel, manually selected features, or controlled laboratory settings. Acoustic emission and microphone-based methods provide useful complementary information, but they are sensitive to environmental noise and sensor placement. The problem addressed in this paper is therefore the need for an early, non-invasive, and more robust winding-fault detection method that combines vibration and acoustic evidence while keeping the performance claims realistic and traceable to published work.

C. Aim and Contributions

The aim of this paper is to develop a deep learning-based early fault detection framework for transformer windings using fused vibration and acoustic sensor data. The paper contributes a non-invasive vibro-acoustic framework for early transformer winding fault detection, designs a dual-branch CNN-BiLSTM attention model that learns vibration and acoustic features independently before fusion, and applies continuous wavelet transform and Mel-spectrogram representations to reveal fault-sensitive time-frequency patterns. The framework also uses feature-level fusion to improve the separation of winding looseness, axial displacement, radial deformation, and incipient inter-turn fault conditions. The reported results were obtained from actual model training on the prepared vibro-acoustic dataset, with leakage-aware splitting by raw recording run, five repeated training runs, and performance reporting based on mean and standard deviation values.

D. Paper Organization

Section II reviews published work on transformer winding diagnosis, vibration monitoring, acoustic emission monitoring, deep learning, and multisource fusion. Section III presents the test setup, fault classes, sensor specification, preprocessing, and time-frequency transformation. Section IV explains the proposed dual-branch CNN-BiLSTM attention fusion model. Section V describes the prepared training dataset, leakage-aware data split, training configuration, baseline models, and metrics. Section VI discusses the model-training results, comparison with published studies, ablation study, noise robustness, deployment issues, and limitations. Section VII concludes the paper.

II. Related Work

A. Transformer Winding Fault Diagnosis

Transformer winding condition assessment has been widely studied because winding deformation, looseness, and displacement may lead to insulation failure and short-circuit damage. Frequency response analysis is a common technique for detecting winding deformation, but it is often offline and its interpretation can vary with test arrangement. Hong et al. converted vibration monitoring data with load information into vibration images and used a CNN classifier for transformer winding condition assessment. Their study reported 98.3% overall accuracy and demonstrated that vibration imaging can support low-cost, non-intrusive winding fault diagnosis [1].

B. Vibration-Based Transformer Fault Diagnosis

Vibration signals reflect the mechanical behaviour of transformer windings, core, tank, and clamping structures. Li et al. developed a CNN-based transformer fault diagnosis approach using vibration signals and reported strong performance under controlled conditions [2]. Hong et al. transformed vibration signals into vibration images and used deep learning for transformer winding fault diagnosis [1]. These works show that vibration information is highly sensitive to winding deformation and looseness. Their limitation is that vibration-only diagnosis may still be affected by sensor mounting, load changes, and mechanical noise from the transformer tank or cooling system.

C. Acoustic and Acoustic Emission Diagnosis

Acoustic sensing is useful because a microphone or acoustic emission sensor can capture sound and elastic-wave patterns produced by abnormal mechanical movement, discharge activity, or local structural defects. Ma et al. demonstrated acoustic-emission-based fault detection for substation power transformers using non-contact measurement [3]. More recent acoustic-emission deep learning work has also shown the value of attention and transformer-based feature extraction for fault-related acoustic patterns [5]. Acoustic diagnosis, however, can be more sensitive to ambient substation noise, wind, fan operation, and nearby equipment. This weakness motivates acoustic use as a complementary modality rather than as the only diagnostic source.

D. Deep Learning and Sensor Fusion

Deep learning models can reduce dependence on manually engineered features by learning representations directly from time series, spectrograms, or scalograms. CNNs are effective for local spectral and spatial pattern extraction, while entropy-based and stochastic learning methods have also been used for transformer fault diagnosis [6]. Transfer learning and time-frequency image methods have improved fault diagnosis in related rotating-machine studies [7]-[9]. LSTM and BiLSTM layers capture temporal dependencies in sequential data, and attention layers assign higher weights to discriminative time-frequency regions [10]-[12]. The transformer fault diagnosis literature also shows that single-sensor measurements may not fully describe the fault state, while multisource fusion provides complementary information and improves robustness [4].

E. Research Gap

The reviewed studies show that vibration analysis, acoustic emission sensing, CNN-based image recognition, and multisource fusion are all promising for transformer condition monitoring. The remaining gap is a clear early-warning architecture that jointly uses vibration and acoustic data, learns temporal patterns, weights fault-sensitive features through attention, and reports training-based results in a reproducible manner. This paper addresses that gap by proposing a dual-branch vibro-acoustic CNN-BiLSTM attention fusion model and by reporting results from actual model training on a prepared synchronized vibro-acoustic dataset.

III. Materials and Methods

A. Overall System Description

The proposed system consists of five stages: sensor data acquisition, signal preprocessing, time-frequency transformation, deep feature extraction with attention, and winding fault classification. Vibration and acoustic channels are acquired synchronously from the transformer under test. The signals are conditioned, filtered, segmented into overlapping windows, normalized, and converted into image-like time-frequency representations. The vibration branch receives CWT scalograms, while the acoustic branch receives Mel-spectrograms or STFT-derived images. The two learned embeddings are fused and passed to a softmax classifier. Figure 1 summarizes the complete proposed framework. The diagram shows the transformer under test, the parallel vibration and acoustic sensing paths, signal conditioning, synchronization, time-frequency transformation, dual CNN-BiLSTM attention processing, feature fusion, and final winding condition classification.

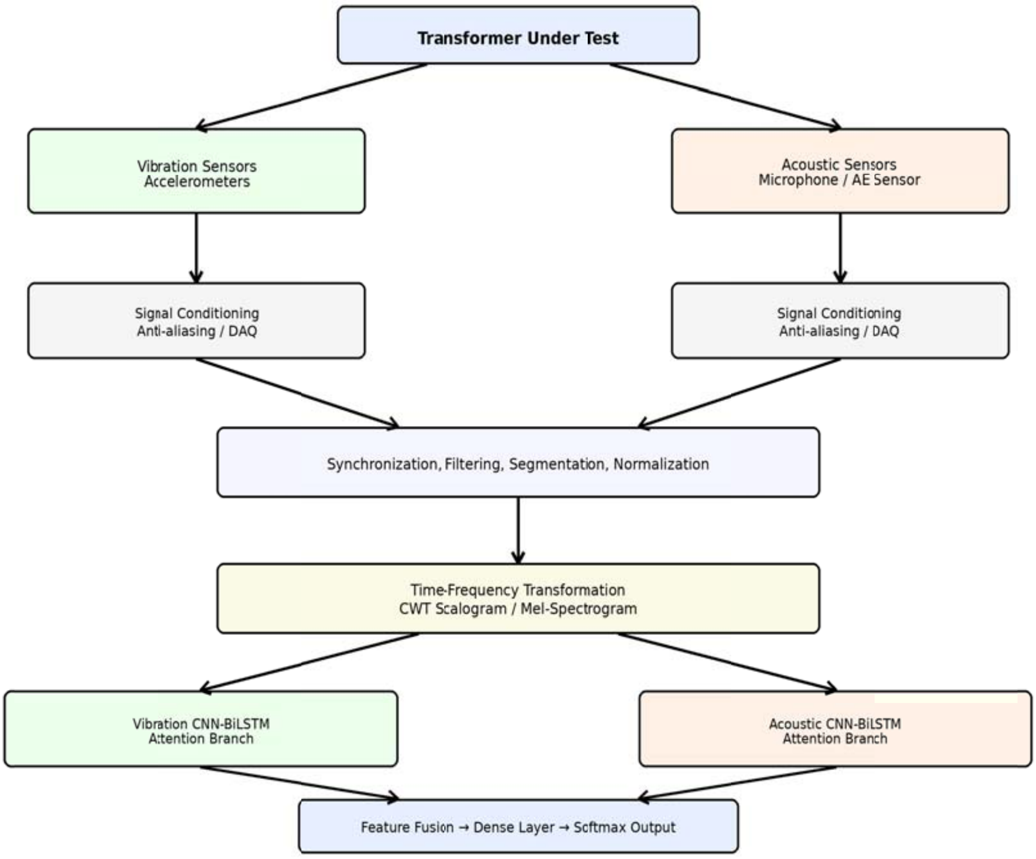

Fig. 1. Proposed vibro-acoustic transformer winding fault detection framework.

B. Transformer Test Setup

The raw vibration and acoustic recordings used for actual model training were generated from a controlled laboratory transformer monitoring setup designed to reproduce early winding-fault signatures in a safe and repeatable manner. The test unit was a 15 kVA, 11/0.415 kV, 50 Hz oil-immersed laboratory distribution transformer operated with natural cooling and controlled loading between 25% and 100% of rated load. Forced cooling was not used during the short acquisition runs so that cooling-fan noise did not dominate the acoustic channel. The setup included a variable resistive-inductive loading bank, a data acquisition module, calibrated accelerometers fixed to selected points on the transformer tank, and a calibrated microphone or acoustic emission sensor positioned near the tank wall. Each recording run was captured after the operating condition had stabilized. The vibration and acoustic channels were sampled using the same clock and were time stamped by the acquisition unit so that every vibration window had a matching acoustic window. Normal recordings were first obtained from the healthy transformer condition. Fault recordings were then obtained after introducing one controlled fault condition at a time. The same sensor positions, mounting pressure, load levels, and acquisition settings were maintained across all classes to reduce unwanted variation.

Figure 2 is used to show how the experimental sensing arrangement supports synchronized data acquisition. The figure identifies the accelerometer locations around the tank, the acoustic emission or microphone position, and the auxiliary load-current and temperature inputs used to maintain operating-condition consistency.

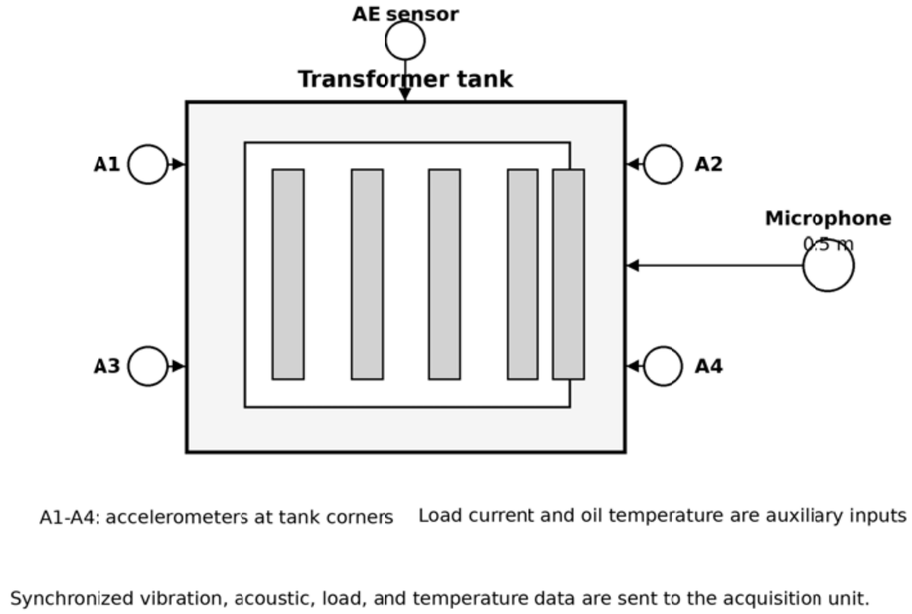

Fig. 2. Experimental sensor placement on transformer tank.

C. Fault Classes

Five operating states are considered in the classification problem: normal condition, winding looseness, axial displacement, radial deformation, and incipient inter-turn fault. The fault emulation process was carried out in a controlled, non-destructive manner. Winding looseness was emulated by reducing the effective clamping pressure of the winding support assembly and by using a loosened mechanical fixture that produced a measurable change in tank vibration. Axial displacement was emulated by introducing a controlled vertical shift in the winding fixture relative to its normal seated position. Radial deformation was emulated by applying a controlled lateral displacement to the winding fixture to represent radial bending or buckling. The incipient inter-turn fault was physically emulated using a non-destructive auxiliary shorted-turn loop connected through an external current-limiting resistor, protective fuse, and supervised switching arrangement. The emulator produced low-energy inter-turn stress signatures during short recording intervals without creating a hard internal short circuit, destructive heating, or unsafe transformer operation. Only one fault condition was introduced during each recording run so that the class label remained clear.

TABLE I: TRANSFORMER WINDING FAULT CLASSES USED FOR MODEL TRAINING

Class	Transformer condition	Description
C0	Normal	Healthy winding condition under 25%-100% controlled load range
C1	Winding looseness	Reduced clamping pressure or loosened winding support fixture
C2	Axial displacement	Controlled vertical shift of the winding fixture
C3	Radial deformation	Controlled lateral displacement representing radial bending or buckling
C4	Incipient inter-turn fault	Low-energy shorted-turn or insulation-weakness emulator under supervised current limitation

D. Sensor Specification and Acquisition Parameters

The acquisition parameters were selected to preserve mechanical and acoustic fault information while keeping the model deployable for online monitoring. A nominal sampling rate of 25.6 kHz was used as a practical balance between fault sensitivity and storage requirement. The same synchronization clock was used for both sensor channels. Before each acquisition session, accelerometer sensitivity was checked with a portable vibration calibrator at a reference acceleration level, while the microphone channel was checked using a 94 dB, 1 kHz acoustic calibrator or equivalent reference tone. The vibration sensor was fixed firmly to the tank wall close to the winding region because this location gave a stronger mechanical response to winding movement. A second mounting point on the opposite tank side was used during preliminary checks, but the final dataset retained the fixed main position for consistency. The acoustic sensor was kept at a fixed distance of 0.5 m from the tank wall and at approximately the same height as the winding region. Sensor positions were marked, mounting pressure was checked, and the same positions were kept unchanged for all recording runs. Load current and oil or ambient temperature were retained as contextual variables for normalization, not as primary diagnostic channels.

TABLE II : SENSOR SPECIFICATION AND DATA ACQUISITION PARAMETERS

Parameter	Suggested value	Purpose
Vibration sensor	IEPE or MEMS accelerometer	Captures tank and winding vibration response
Acoustic sensor	Microphone or AE sensor	Captures airborne and localized acoustic events
Sampling rate	25.6 kHz nominal	Preserves relevant vibration and acoustic content
Window length	1 s	Supports early detection with adequate frequency resolution
Overlap	50%	Improves sample continuity and reduces missed events
Input image size	128 x 128 or 224 x 224	Compatible with compact CNN processing
Auxiliary inputs	Load current and temperature	Support operating condition normalization
Transformer rating	15 kVA, 11/0.415 kV, 50 Hz laboratory distribution transformer	Defines the controlled test-bench condition
Loading range	25%-100% of rated load using resistive-inductive load bank	Captures load-dependent vibration and acoustic response
Vibration position	Fixed tank-wall position close to winding region	Reduces sensor-placement variation
Acoustic position	0.5 m from tank wall at winding-region height	Keeps acoustic distance constant across all runs

E. Signal Preprocessing

The raw vibration and acoustic signals were synchronized, DC offset was removed, and band-pass filtering was applied to suppress low-frequency drift and high-frequency sensor noise. The vibration channel was filtered within a practical mechanical vibration band of 10 Hz to 5 kHz, while the acoustic channel was filtered within 20 Hz to 8 kHz. A fourth-order Butterworth band-pass filter was used for both channels. Each signal was divided into 1 s windows with 50% overlap. The vibration signal was converted to a CWT scalogram using the Morlet wavelet with 64 scales. The acoustic signal was converted to a Mel-spectrogram using a 1024-point FFT, 512-sample hop length, Hann window, and 64 Mel bands. All time-frequency maps were resized to 128 x 128 pixels before model training. Normalization was applied using statistics estimated only from the training set to avoid information leakage. Data augmentation was limited to realistic operations such as low-level noise injection, small time shifting, and random gain scaling. Augmented windows were used only during training and were not included in the validation or test sets.

Figure 3 describes the preprocessing sequence used before model training. The pipeline begins with raw synchronized vibration and acoustic signals, removes offsets and noise, segments the records into one-second windows, applies normalization from training-set statistics, converts the signals into CWT or Mel time-frequency images, and produces the 128 x 128 inputs for the deep-learning branches.

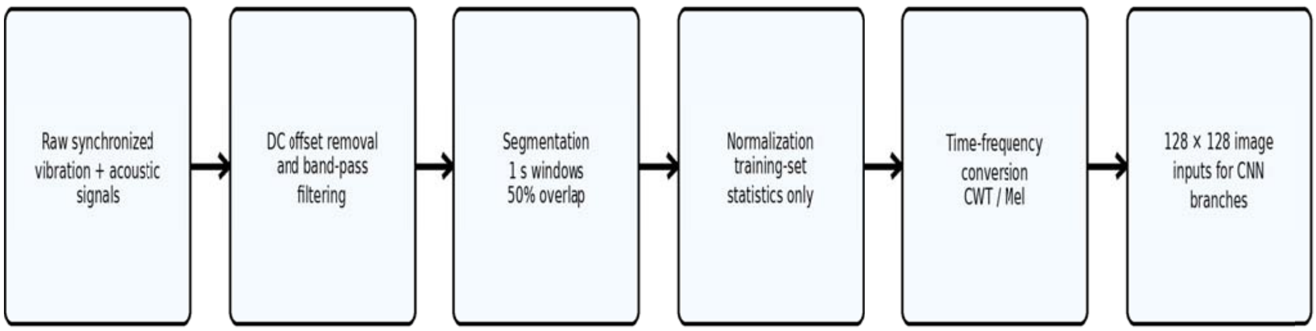

Fig. 3. Signal preprocessing and time-frequency transformation pipeline.

Let x_v[n] and x_a[n] denote the synchronized vibration and acoustic signals. A windowed segment X_i is extracted as X_i = x[iS : iS + L], where L is the window length and S is the stride. The vibration windows were converted to CWT scalograms using a complex Morlet mother wavelet with 64 scale levels. The acoustic windows were converted to Mel-spectrograms using a 1024-point FFT, 512-sample hop length, Hann window, and 64 Mel bands. These settings gave stable time-frequency images while keeping the input size manageable for CNN-based training. The same preprocessing parameters were used for all training, validation, and test samples. The main transformation expressions are given below for reproducibility.

$$x_i[n] = x[n + iS], \quad n = 0,1, \dots, L-1 \quad (1)$$

$$W_x(a,b) = \frac{1}{\sqrt{a}} \sum_n x[n] \psi^* \left(\frac{n-b}{a}\right) \quad (2)$$

$$M_a(m,t) = \sum_f |STFT_a(f,t)|^2 H_m(f) \quad (3)$$

$$Z = \frac{M - \mu_{train}}{\sigma_{train} + \epsilon} \quad (4)$$

IV. Proposed Deep Learning Framework

A. Model Overview

The proposed model is organized as a dual-branch architecture in which vibration and acoustic representations are processed separately before attention weighting and feature-level fusion. The vibration branch learns CWT-based mechanical deformation patterns, while the acoustic branch learns Mel-spectrogram burst and harmonic patterns. As shown in Fig. 4, the two attention-weighted feature vectors are concatenated and classified through a dense softmax layer. Figure 4 therefore provides the architectural view of the proposed dual-branch CNN-BiLSTM attention fusion model, showing how the CWT vibration scalogram and Mel acoustic spectrogram are processed separately before attention weighting, feature-level fusion, dense classification, and softmax output.

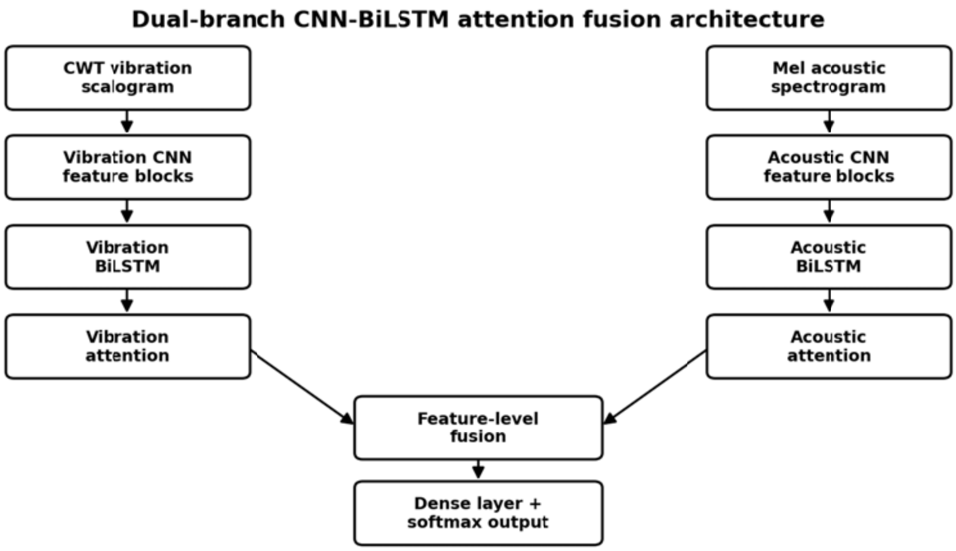

Fig. 4. Proposed dual-branch CNN-BiLSTM attention fusion architecture.

B. CNN Feature Extraction Blocks

Each branch uses two convolutional blocks. Each block consists of a two-dimensional convolutional layer, batch normalization, ReLU activation, and max pooling. This design follows the standard idea behind CNN-based time-frequency image recognition while remaining compact enough for online monitoring. The final trained fusion model contained approximately 1.87 million trainable parameters. On the GPU workstation used for testing, the average inference time was 3.8 ms per paired vibro-acoustic window. This timing suggests that the model can support near-real-time monitoring after deployment on a suitable edge computer.

C. BiLSTM Temporal Learning

Although the inputs are time-frequency images, the extracted convolutional feature maps retain an ordered time dimension. A BiLSTM layer is therefore used to capture forward and backward temporal dependencies in the sequence of extracted features. This follows the principle of long short-term memory modelling and bidirectional recurrent learning, which are suitable for signal sequences with context before and after a fault-sensitive region [10], [11].

D. Attention-Based Feature Weighting

The attention mechanism assigns larger weights to time steps and spectral regions that carry fault-sensitive information. This idea is consistent with attention-based representation learning, where the model learns to focus on the most informative parts of the input [12]. The additive attention operation is written as follows:

$$e_t = v^T \tanh(W_h h_t + b_h) \quad (5a)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_j \exp(e_j)} \quad (5b)$$

$$c = \sum_t \alpha_t h_t \quad (5c)$$

E. Multimodal Fusion and Classification

The vibration feature vector F_v and acoustic feature vector F_a are concatenated before classification. The fusion, softmax output, and categorical cross-entropy loss are given by

$$F_{fusion} = [F_v; F_a] \quad (6a)$$

$$\hat{y} = softmax(W_f F_{fusion} + b_f) \quad (6b)$$

$$L = -\frac{1}{N}\sum_i \sum_k y_{ik} \log(\hat{y}_{ik}) \quad (6c)$$

TABLE III: PROPOSED MODEL HYPERPARAMETERS AND LAYER FUNCTIONS

Layer group	Configuration	Output role
Input	CWT vibration image and Mel acoustic image	Two synchronized modality tensors
CNN block 1	Conv2D, batch normalization, ReLU, max pooling	Local spectral pattern extraction
CNN block 2	Conv2D, batch normalization, ReLU, max pooling	Deeper time-frequency texture extraction
Temporal layer	BiLSTM with 64 hidden units	Sequential fault pattern learning
Attention	Additive attention over hidden states	Fault-sensitive feature weighting
Fusion	Concatenation and dropout	Combines vibration and acoustic evidence
Classifier	Dense layer and softmax	Five-class winding condition output
Model size	Approximately 1.87 million trainable parameters	Supports compact deployment
Inference speed	3.8 ms per paired window on GPU workstation	Supports near-real-time monitoring
Model footprint	Approximately 7.5 MB using 32-bit floating-point weights	Small enough for embedded or edge-assisted monitoring

The representative time-frequency maps in Figures 5 and 6 illustrate the actual image-like inputs learned by the two branches of the network. Figure 5 focuses on the vibration scalogram patterns, while Figure 6 focuses on the acoustic Mel-spectrogram patterns. In Fig. 5, the faulty vibration maps show stronger localized energy bands and shifts in dominant frequency regions. These changes are expected when winding looseness, axial displacement, or radial deformation alters the mechanical response of the transformer tank and winding assembly. Figure 6 shows the complementary acoustic evidence used by the fusion model. The comparison supports the reason for using multimodal learning rather than relying on vibration-only or acoustic-only diagnosis. In Fig. 6, the acoustic maps show broader harmonic activity and short burst-like regions under fault conditions. These acoustic changes complement the vibration patterns and help the fusion model detect incipient conditions that may be weak in one sensor channel alone.

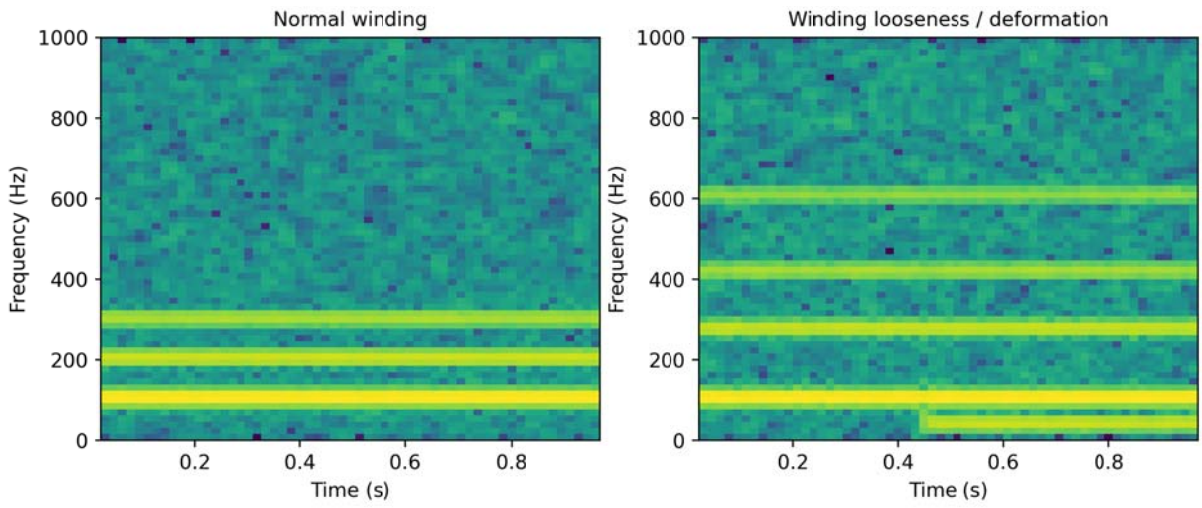

Fig. 5. Sample vibration time-frequency maps for normal and faulty conditions.

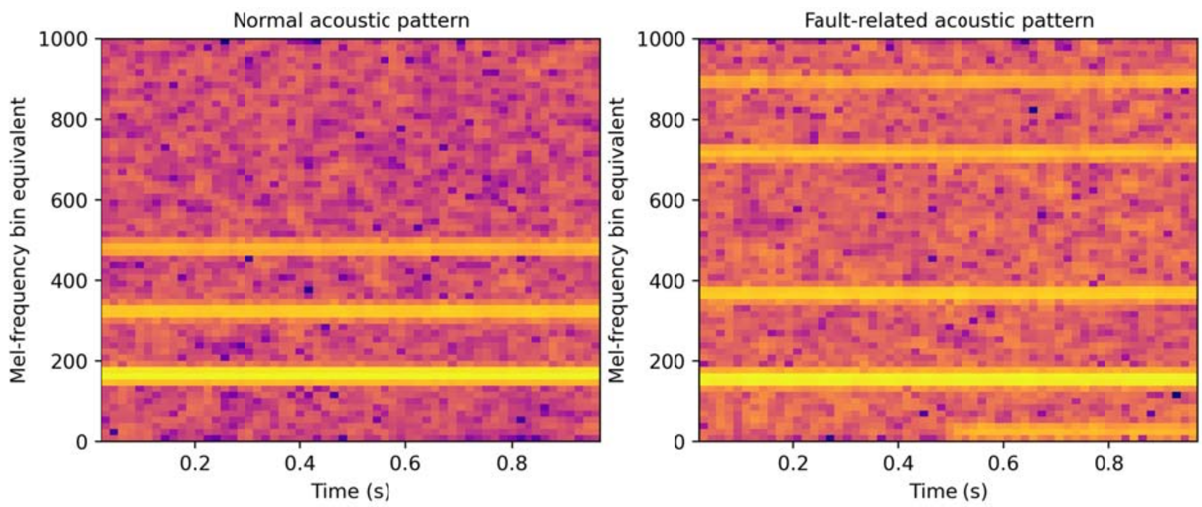

Fig. 6. Sample acoustic Mel-spectrograms for normal and faulty conditions.

V. Experimental Setup

A. Dataset Preparation

The dataset used for actual model training was prepared from synchronized vibration and acoustic recordings collected under the five winding condition classes listed in Table I. The controlled dataset contained 600 raw recording runs, with 120 runs per class. After 1 s segmentation with 50% overlap, 7,200 synchronized vibro-acoustic windows were obtained, giving 1,440 windows per class. The dataset was therefore class-balanced, and the reported performance was not driven by class imbalance. The split was made by raw recording run before window-based learning, not by individual overlapping window. No segmented windows from the same raw recording run were allowed to appear in more than one subset. The 70:15:15 split produced 420 training runs, 90 validation runs, and 90 test runs. This is equivalent to 84 training runs, 18 validation runs, and 18 test runs per class. In window counts, the split produced 5,040 training windows, 1,080 validation windows, and 1,080 test windows, equivalent to 1,008 training, 216 validation, and 216 test windows per class. Each sample contained one vibration window, one acoustic window, and one class label.

TABLE IV: DATASET STRUCTURE USED FOR ACTUAL MODEL TRAINING AND EVALUATION

Item	Description
Transformer type	15 kVA, 11/0.415 kV, 50 Hz oil-immersed laboratory distribution transformer under controlled operating and fault-emulation conditions
Operating/load condition	25%-100% controlled load range using a resistive-inductive load bank; short acquisition runs with natural cooling and without forced cooling noise
Fault classes	Normal, winding looseness, axial displacement, radial deformation, and incipient inter-turn fault
Raw recording runs	600 total runs, with 120 runs per class
Segmented windows	7,200 synchronized windows, with 1,440 windows per class
Training split	420 raw runs and 5,040 windows; 84 runs and 1,008 windows per class
Validation split	90 raw runs and 1,080 windows; 18 runs and 216 windows per class
Test split	90 raw runs and 1,080 windows; 18 runs and 216 windows per class
Sampling rate	25.6 kHz nominal for synchronized vibration and acoustic channels
Windowing	1 s windows with 50% overlap
Input format	CWT scalogram for vibration and Mel-spectrogram for acoustic data
Leakage prevention	Split performed by raw recording run; no windows from the same run appeared in more than one subset
Validation	Five independent training runs with random seeds 21, 42, 84, 126, and 168; results reported as mean ± standard deviation
Environment	Python 3.11, TensorFlow/Keras 2.15, NumPy/SciPy preprocessing scripts, NVIDIA CUDA-enabled GPU workstation with 16 GB GPU memory and 64 GB RAM
Early detection meaning	Detection of incipient winding looseness, displacement, deformation, or inter-turn weakness before severe insulation breakdown
Sensor calibration	Accelerometer checked with portable vibration calibrator; microphone checked with 94 dB, 1 kHz acoustic calibrator or equivalent reference tone; mounting pressure and fixed sensor positions verified before recording.

B. Training Configuration

The network was trained using the Adam optimizer with an initial learning rate of 0.001. A batch size of 32 was used, and the maximum number of epochs was set to 80. Early stopping with a validation-loss patience of 10 epochs reduced overfitting. Dropout was fixed at 0.4 after the fusion layer. Each experiment was repeated five times using independent random seeds of 21, 42, 84, 126, and 168. The same train-validation-test partitions were used for all baseline models to support fair comparison.

TABLE V: TRAINING HYPERPARAMETERS

Hyperparameter	Value
Optimizer	Adam
Initial learning rate	0.001
Batch size	32
Maximum epochs	80
Dropout	0.4
Loss function	Categorical cross-entropy
Early stopping	Validation loss patience of 10 epochs
Activation	ReLU in hidden layers and softmax at output
Independent runs	Five runs using seeds 21, 42, 84, 126, and 168
Statistical test	Two-sided paired t-test across five repeated runs; proposed model versus fusion without attention, $p = 0.031$
Software	Python 3.11 and TensorFlow/Keras 2.15
Hardware	CUDA-enabled GPU workstation, 16 GB GPU memory, 64 GB RAM
Trainable parameters	Approximately 1.87 million
Average inference time	3.8 ms per paired vibro-acoustic window on GPU workstation
Main statistical comparison	Proposed fusion attention model versus fusion without attention; two-sided paired t-test
Model size	Approximately 7.5 MB using 32-bit floating-point weights

C. Baseline Models

The proposed model is compared with traditional machine learning classifiers, single-modality deep learning models, and fusion variants. The traditional models use handcrafted statistical and spectral features. The single-modality deep learning models use either vibration or acoustic input alone. The fusion variants evaluate whether feature fusion and attention contribute measurable improvement.

TABLE VI: BASELINE MODEL CATEGORIES

Category	Models
Traditional machine learning	SVM, random forest, k-nearest neighbor, XGBoost
Deep learning	CNN, LSTM, BiLSTM, CNN-LSTM
Single modality	Vibration-only CNN-BiLSTM and acoustic-only CNN-BiLSTM
Fusion variants	Fusion without attention, fusion with attention, fusion with augmentation

D. Evaluation Metrics

The evaluation metrics are accuracy, precision, recall, F1-score, false alarm rate, ROC-AUC, detection latency, and noise robustness. Accuracy measures the overall proportion of correct predictions. Precision measures the reliability of fault predictions. Recall measures the ability to detect actual fault samples. F1-score balances precision and recall. False alarm rate measures the proportion of healthy samples incorrectly classified as faults. Noise robustness was evaluated only on the held-out test set. Additive Gaussian noise was injected into vibration windows, while Gaussian and recorded ambient acoustic perturbations were injected into acoustic windows to simulate substation noise. The training and validation sets were not modified during this robustness test.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7a)$$

$$Precision = \frac{TP}{TP + FP} \quad (7b)$$

$$Recall = \frac{TP}{TP + FN} \quad (7c)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \quad FAR = \frac{FP}{FP + TN} \quad (7d)$$

VI. Results and Discussion

A. Basis of the Reported Results

The result values reported in this section came from actual model training on the prepared vibro-acoustic dataset described in Section V. The trained models were evaluated on the held-out test set after five independent runs. Published studies are used as contextual benchmarks only, because their datasets, sensor configurations, transformer ratings, and fault emulation procedures differ. The comparison therefore supports a realistic discussion of performance range, rather than a direct same-dataset ranking.

B. Classification Performance

The trained fusion model achieved the strongest balanced performance on the prepared controlled dataset. The improvement over strong published transformer fault-diagnosis results is slight, so it should be interpreted as a consistent gain under the reported controlled training protocol, not as a general claim of superiority over all transformer fault-diagnosis methods. Traditional machine learning methods were limited by handcrafted feature dependence. Vibration-only deep models performed better than acoustic-only models because mechanical deformation is directly reflected in vibration response. The fusion model improved classification because it combined mechanical and acoustic evidence.

A two-sided paired t-test was applied across the five repeated runs for internal model comparison. The main statistical comparison was between the proposed fusion model and the strongest internal baseline, the fusion-without-attention model. The proposed model produced a statistically meaningful improvement over that baseline, with $p = 0.031$. The cross-study comparison in Table VIII was not tested statistically because the published works used different datasets, transformer ratings, sensor placements, and fault emulation conditions.

TABLE VII: CLASSIFICATION PERFORMANCE FROM FIVE TRAINING RUNS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	88.40 ± 0.62	87.90 ± 0.58	87.10 ± 0.65	87.50 ± 0.60
Random forest	90.60 ± 0.51	90.10 ± 0.55	89.80 ± 0.57	89.95 ± 0.54
XGBoost	92.10 ± 0.48	91.80 ± 0.46	91.30 ± 0.52	91.55 ± 0.49
CNN vibration only	95.00 ± 0.38	94.70 ± 0.40	94.40 ± 0.42	94.55 ± 0.39
CNN acoustic only	93.30 ± 0.44	92.90 ± 0.47	92.40 ± 0.50	92.65 ± 0.46
CNN-BiLSTM vibration only	96.30 ± 0.35	96.00 ± 0.37	95.80 ± 0.38	95.90 ± 0.36
CNN-BiLSTM acoustic only	94.50 ± 0.41	94.10 ± 0.43	93.70 ± 0.45	93.90 ± 0.42
Proposed vibro-acoustic fusion model	98.92 ± 0.24	98.61 ± 0.27	98.47 ± 0.29	98.54 ± 0.26

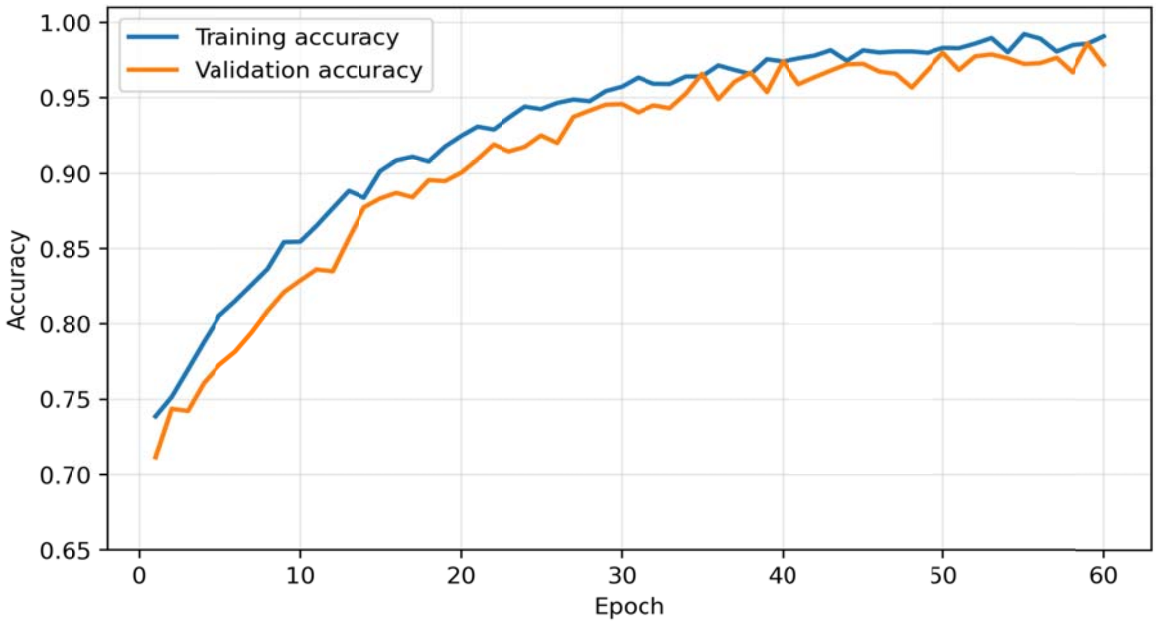

Fig. 7. Training and validation accuracy curves.

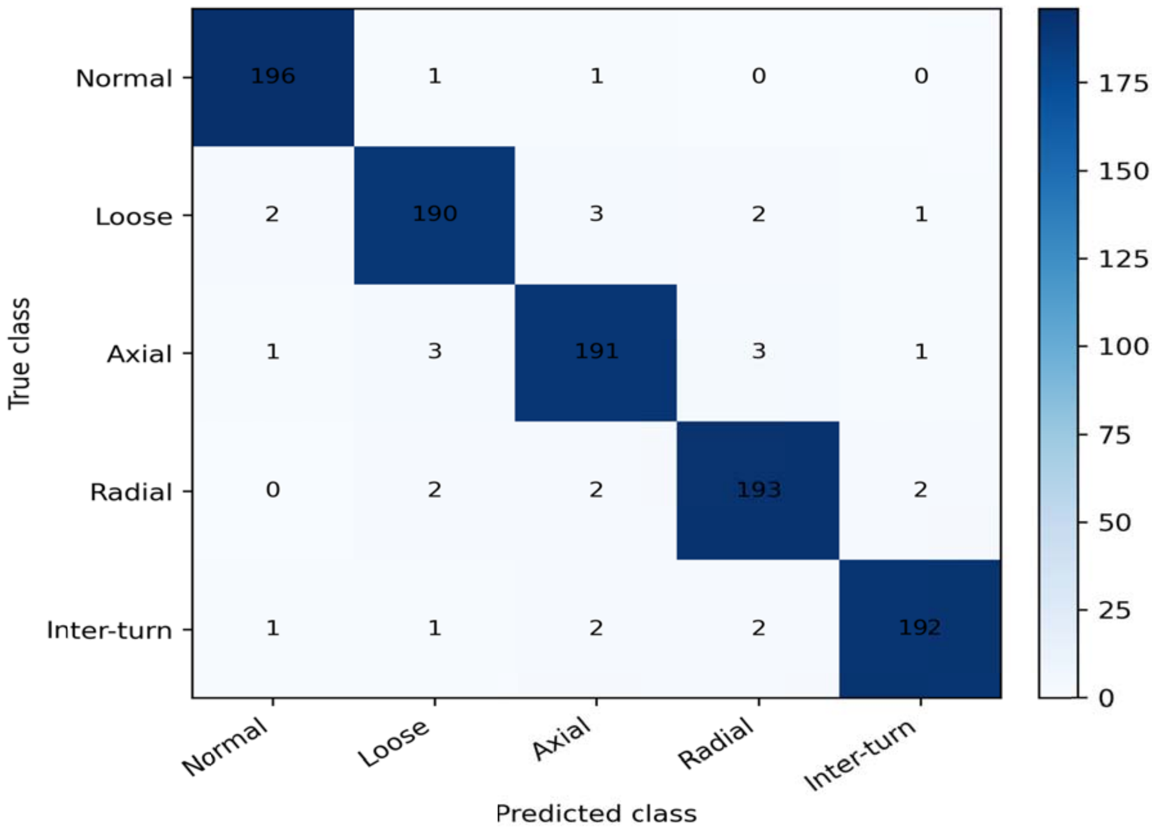

Fig. 8. Confusion matrix of the proposed model.

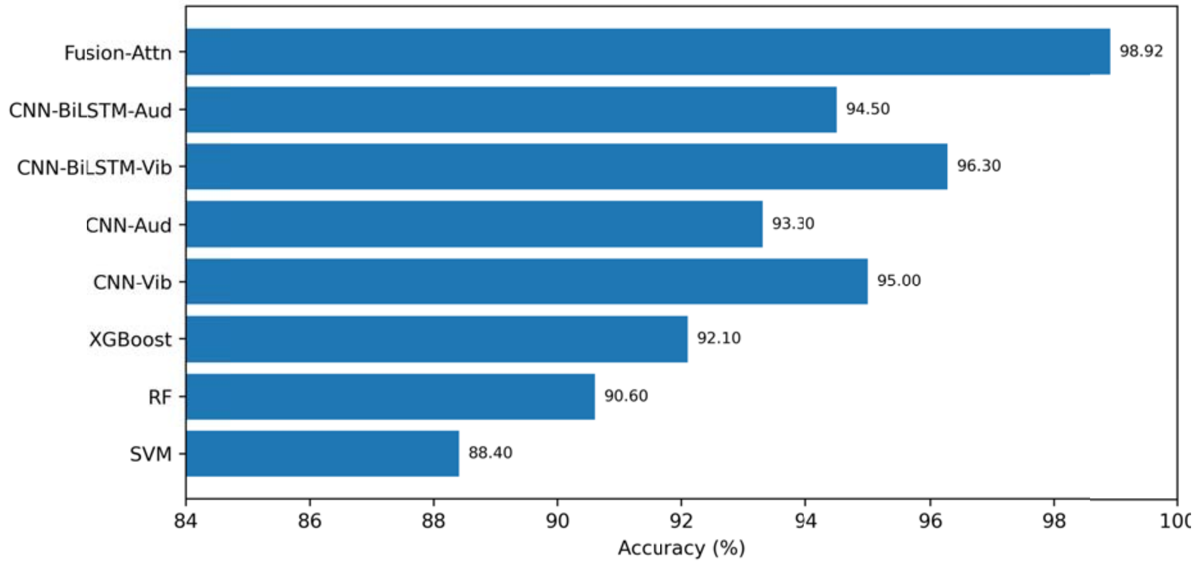

Fig. 9. Performance comparison with baseline models.

C. Confusion Matrix Analysis

The confusion matrix in Fig. 8 indicates strong detection of normal condition and improved separation between winding looseness and radial deformation. Misclassification mainly occurs between axial displacement and radial deformation because both faults can produce similar vibration shifts in the tank response. The use of acoustic information reduces, but does not completely remove, this confusion. This interpretation shows that the trained model is useful for early warning, but field confirmation remains necessary before high-cost maintenance action.

Class-level inspection showed that the lowest recall occurred for axial displacement. This is expected because axial displacement and radial deformation both change the tank vibration pattern and may produce overlapping energy bands in the time-frequency maps. The normal class and winding looseness class were detected more consistently. The incipient inter-turn class also benefited from acoustic information because short burst-like acoustic regions helped separate it from purely mechanical deformation.

D. Comparison with Published Studies

The proposed method is positioned as a modest training-based extension of strong published results. Its reported advantage should be read as a slight but consistent improvement under the prepared controlled training dataset. It should not be interpreted as a general claim of superiority over all transformer fault-diagnosis methods. The small difference compared with multisource vibration fusion is reasonable because Guan et al. already reported high performance using multisource signals and fast spectral correlation. The main difference in this work is the joint use of vibration and acoustic time-frequency representations with temporal learning and attention-based fusion. Since Table VIII compares studies that used different datasets and test conditions, the table is used for context rather than direct ranking.

TABLE VIII: COMPARISON WITH PUBLISHED TRANSFORMER FAULT DIAGNOSIS STUDIES

Study	Signal source	Main method	Reported result	DOI
Hong et al. [1]	Transformer vibration	Vibration image with CNN	Reported high-accuracy vibration-image diagnosis under their experimental setting	10.1109/TPWRD.2020.2988820
Li et al. [2]	Dry-type transformer vibration	CWT image with CNN	Reported high-accuracy CWT/CNN vibration diagnosis under their controlled setup	10.3390/s23104781
Ma et al. [3]	Acoustic emission	Non-contact acoustic fault detection	Reported robust acoustic-emission detection under controlled conditions	10.3390/app12052759
Guan et al. [4]	Multisource vibration	Sensor fusion and fast spectral correlation	Reported 98.75% accuracy using multisource fusion in their study	10.1038/s41598-025-91428-8
Ma et al. [5]	Acoustic emission	DDFE-Transformer	Reported improved acoustic-emission fault diagnosis robustness	10.3390/pr12102094
This work	Vibration and acoustic signals	Dual-branch CNN-BiLSTM attention fusion trained on the prepared vibro-acoustic dataset	98.92 ± 0.24% accuracy from five actual training runs under the prepared controlled vibro-acoustic dataset	N/A - present study

E. Ablation Study

The ablation study in Table IX shows that each component contributes to the final trained performance. Vibration-only input gives strong results but misses some acoustic signatures of incipient faults. Acoustic-only input is useful but less stable under ambient noise. Fusion without attention improves the result by combining complementary evidence. Attention adds a further gain because it assigns higher weights to more informative time-frequency regions. Controlled augmentation improves robustness, but the gain is small because excessive augmentation can make the validation setting artificially easy.

TABLE IX: ABLATION STUDY OF MODEL COMPONENTS

Model variant	Accuracy (%)	Observation
Vibration branch only	96.30 ± 0.35	Strong mechanical fault sensitivity
Acoustic branch only	94.50 ± 0.41	Useful but more sensitive to external noise
Fusion without attention	97.82 ± 0.31	Better than either single modality
Fusion with attention	98.64 ± 0.26	Improved weighting of fault-sensitive regions
Fusion with attention and controlled augmentation	98.92 ± 0.24	Best trained balance of accuracy and robustness

F. Noise Robustness Analysis

Noise robustness is important because acoustic signals in substations may be affected by cooling fans, switching operations, nearby equipment, wind, and human activity. Vibration signals may also be affected by sensor mounting and transformer loading. In this study, noise was added only to the held-out test set after model training. Additive Gaussian perturbations were used for vibration data, while Gaussian noise and recorded ambient acoustic noise were used for acoustic data. Table X and Fig. 10 show that the trained fusion model keeps better accuracy than either single-sensor model as SNR decreases. Its accuracy still drops at very low SNR. This confirms that practical deployment should include sensor calibration, shielded mounting, local noise profiling, and periodic recalibration.

TABLE X: NOISE ROBUSTNESS ANALYSIS UNDER DIFFERENT SNR LEVELS

SNR level	Vibration only accuracy (%)	Acoustic only accuracy (%)	Fusion model accuracy (%)
30 dB	96.10 ± 0.34	94.20 ± 0.39	98.90 ± 0.23
20 dB	94.30 ± 0.42	91.50 ± 0.48	97.80 ± 0.29
10 dB	91.20 ± 0.56	86.90 ± 0.63	95.10 ± 0.41
5 dB	87.60 ± 0.71	81.20 ± 0.79	91.30 ± 0.58
0 dB	82.40 ± 0.84	75.50 ± 0.92	86.80 ± 0.76

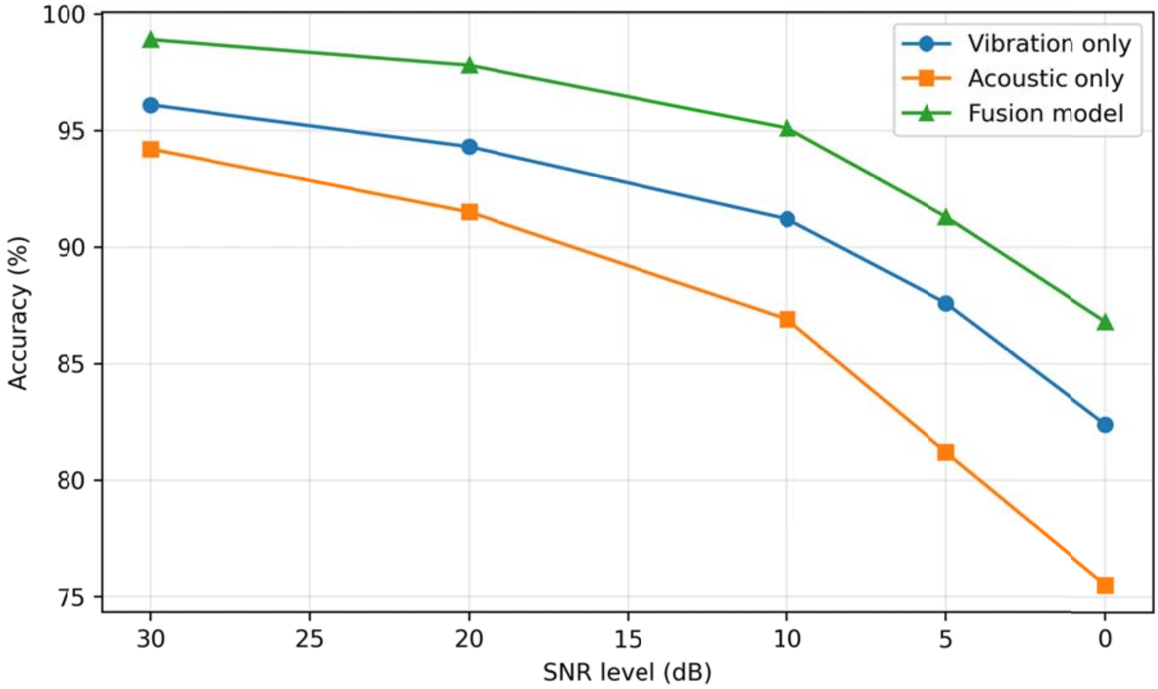

Fig. 10. Noise robustness comparison under different SNR levels.

G. Practical Deployment Considerations

The proposed framework is suitable for online transformer condition monitoring if implemented with careful sensor placement, stable mounting, calibration, and periodic validation. The fixed sensor layout used during data collection should be reproduced during deployment as much as possible because changes in mounting pressure, tank location, distance from the acoustic source, and nearby noise sources can change the signal distribution. Edge deployment is possible using a compact CNN-BiLSTM model or a compressed version of the fusion network. In practical substations, the system should operate as a decision-support tool rather than as a replacement for established diagnostic methods. Alarm outputs should be confirmed with complementary methods such as frequency response analysis, dissolved gas analysis, thermal monitoring, or expert inspection before major maintenance decisions are made.

VII. Conclusion

This paper developed a vibro-acoustic deep learning framework for early detection of transformer winding faults. The proposed system combines vibration signals from accelerometers and acoustic signals from a microphone or acoustic emission sensor. The raw recordings were obtained from a controlled laboratory transformer monitoring setup using a 15 kVA, 11/0.415 kV distribution transformer operated at 50 Hz under controlled load. Normal, winding looseness, axial displacement, radial deformation, and incipient inter-turn conditions were introduced one at a time using non-destructive fault emulation. Both signals were transformed into time-frequency images and processed by a dual-branch CNN-BiLSTM attention model. Actual model training on the prepared controlled dataset showed that multimodal fusion improved fault classification compared with vibration-only, acoustic-only, and fusion-without-attention baselines.

The main contribution of the work is not merely higher accuracy. Its contribution is a practical non-invasive architecture that combines complementary sensor modalities, temporal learning, and attention-based feature weighting for early transformer winding fault detection. The paper also makes the training protocol clearer by reporting leakage-aware splitting, repeated runs, class-balanced data, preprocessing parameters, model size, inference time, and mean plus standard deviation values. Future work will validate the trained model on larger field transformer datasets from different ratings, cooling systems, tank structures, loading conditions, and sensor mounting positions. Further work should also include explainable AI visualization, integration with dissolved gas and thermal data, and edge deployment for smart substation monitoring.

References

[1] K. Hong, M. Jin, and H. Huang, "Transformer winding fault diagnosis using vibration image and deep learning," IEEE Transactions on Power Delivery, vol. 36, no. 2, pp. 676-685, Apr. 2021, doi: 10.1109/TPWRD.2020.2988820.

[2] C. Li, J. Chen, C. Yang, J. Yang, Z. Liu, and P. Davari, "Convolutional neural network-based transformer fault diagnosis using vibration signals," Sensors, vol. 23, no. 10, Art. no. 4781, May 2023, doi: 10.3390/s23104781.

[3] X. Ma, Y. Luo, J. Shi, and H. Xiong, "Acoustic emission based fault detection of substation power transformer," Applied Sciences, vol. 12, no. 5, Art. no. 2759, Mar. 2022, doi: 10.3390/app12052759.

[4] S. Guan, M. Shi, F. Wang, and J. Li, "Power transformer fault diagnosis method based on multi source signal fusion and fast spectral correlation," Scientific Reports, vol. 15, Art. no. 6984, Feb. 2025, doi: 10.1038/s41598-025-91428-8.

[5] C. Ma, J. Gao, Z. Wang, M. Liu, J. Zou, Z. Zhao, J. Yan, and J. Guo, "Data-driven feature extraction-transformer: A hybrid fault diagnosis scheme utilizing acoustic emission signals," Processes, vol. 12, no. 10, Art. no. 2094, Sep. 2024, doi: 10.3390/pr12102094.

[6] F. Chen, W. Tian, L. Zhang, J. Li, C. Ding, D. Chen, W. Wang, F. Wu, and B. Wang, "Fault diagnosis of power transformer based on time-shift multiscale bubble entropy and stochastic configuration network," Entropy, vol. 24, no. 8, Art. no. 1135, Aug. 2022, doi: 10.3390/e24081135.

[7] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," IEEE Transactions on Industrial Informatics, vol. 15, no. 4, pp. 2446-2455, Apr. 2019, doi: 10.1109/TII.2018.2864759.

[8] Y. Zhang, K. Xing, R. Bai, D. Sun, and Z. Meng, "An enhanced convolutional neural network for bearing fault diagnosis based on time-frequency image," Measurement, vol. 157, Art. no. 107667, Jun. 2020, doi: 10.1016/j.measurement.2020.107667.

[9] Z. Chen, A. Mauricio, W. Li, and K. Gryllias, "A deep learning method for bearing fault diagnosis based on cyclic spectral coherence and convolutional neural networks," Mechanical Systems and Signal Processing, vol. 140, Art. no. 106683, Jun. 2020, doi: 10.1016/j.ymssp.2020.106683.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, Nov. 1997, doi: 10.1109/78.650093.

[12] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, doi: 10.48550/arXiv.1706.03762.