ISSN NO: 2770-4491
May 2024

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

Silas Abraham Friday[1]
Department of Electrical/ Electronic Engineering
Adekeke University, Ede Osun State, Nigeria
Umoren Mfonobong A.[2]
Department of Electrical /Electronic Engineering,
University of Uyo, Akwa Ibom State Nigeria
Abba MaryRose Obiageli[3]
Department of Electrical and Electronic Engineering,
Enugu State University of Science and Technology, ESUT
Agbani, Enugu State, Nigeria

Abstract—
Early changes in transformer winding geometry often appear as weak mechanical and acoustic signatures before they develop into insulation breakdown, severe inter-turn short circuit or service failure. This paper presents a non-invasive vibro-acoustic deep learning framework for early winding-fault detection. Synchronized vibration and acoustic recordings were obtained from a controlled 15 kVA, 11/0.415 kV, 50 Hz oil-immersed laboratory transformer under healthy operation and four emulated early-fault states. Vibration windows were represented as continuous wavelet transform scalograms, while acoustic windows were represented as Mel-spectrograms. A dual-branch CNN-BiLSTM model with additive attention was used to learn each sensing channel before feature-level fusion and five-class classification. The dataset comprised 600 raw recording runs and 7,200 synchronized windows. The split was performed at raw-run level before segmentation-based learning to prevent leakage among overlapping windows. Across five independent training runs, the fusion-attention model achieved 98.92 ± 0.24% accuracy, 98.61 ± 0.27% precision, 98.47 ± 0.29% recall and 98.54 ± 0.26% F1-score on the held-out test set. The results indicate that paired vibration and acoustic evidence improves early recognition of winding looseness, axial displacement, radial deformation and incipient inter-turn weakness under the controlled test conditions. The method provides a laboratory-validated alarm layer whose substation use should be checked against established maintenance evidence from frequency-response analysis, dissolved-gas analysis, thermal inspection and expert review.
Keywords: power transformer; winding fault; vibration signal; acoustic signal; CNN; BiLSTM; attention; sensor fusion; condition monitoring; deep learning

I. Introduction

A. Background

Power transformers are expensive and strategically important assets in transmission and distribution networks. A major transformer failure can interrupt supply, damage nearby equipment, create fire risk and impose long replacement delays. Winding faults are especially important because the winding structure is exposed to electromagnetic force, thermal ageing, mechanical vibration, transport shock, insulation stress and load cycling. Frequency response analysis is widely used to assess winding deformation in oil-immersed transformers, but its interpretation depends on winding structure, test connection and measurement practice [1].

A winding fault rarely begins as a catastrophic event. It may start as reduced clamping pressure, a small axial shift, radial deformation or weakened inter-turn insulation. Each of these changes modifies the mechanical transfer path between winding, core and tank. Vibration monitoring is therefore useful because winding and core movement are coupled to electromagnetic force, magnetic flux, load current, tank structure and clamping condition [2]. Acoustic sensing is also useful because abnormal mechanical impacts, loosened parts and discharge-related events can produce airborne acoustic activity that is detectable without opening the transformer tank [3].

A single sensing channel can be unreliable in service. Vibration measurements are sensitive to mounting pressure, tank location and auxiliary mechanical noise, while acoustic measurements are affected by background sound, wind, fans and nearby equipment. A paired vibro-acoustic approach can reduce this weakness by combining evidence from two physical paths. Recent CNN-based transformer-diagnosis studies have shown that vibration images and time-frequency representations can support automated fault recognition [2], [4]. The remaining challenge is to combine synchronized vibration and acoustic evidence in a leakage-aware training protocol that reports data construction, model training and performance limits clearly.

B. Problem Statement

Established transformer maintenance tools include frequency response analysis, dissolved gas analysis, partial-discharge testing, insulation-resistance measurement and thermal inspection [1], [5], [6]. These methods remain valuable, but some are offline, some require specialist interpretation and some respond more clearly after the fault has progressed. Recent vibration-based and acoustic-emission methods have produced useful results, yet many reported systems still depend on one sensing channel, hand-selected features or laboratory assumptions that are difficult to reproduce in service. The problem addressed in this study is the need for a non-invasive early-warning method that links winding-related vibration changes with acoustic signatures, while using a split strategy that prevents leakage from overlapping windows.

C. Aim and Contributions

The aim of this study is to develop and evaluate a deep learning framework for early transformer winding-fault detection using synchronized vibration and acoustic measurements. The main contributions are:

i. a non-invasive vibro-acoustic monitoring workflow for early winding-fault detection;
ii. a dual-branch CNN-BiLSTM attention architecture that learns vibration and acoustic features before feature-level fusion;
iii. the use of continuous wavelet transform scalograms and Mel-spectrograms to represent fault-sensitive time-frequency patterns;
iv. run-level splitting before window-based learning so that overlapping windows from the same recording run do not appear in more than one subset;
v. an evaluation that includes single-sensor baselines, fusion baselines, repeated training runs, ablation analysis and noise-robustness testing.

II. Related Work

A. Transformer Winding Fault Diagnosis

Transformer winding condition assessment has been studied for several decades because winding deformation, looseness and displacement can develop into insulation damage and short-circuit failure. Frequency response analysis remains one of the most established tools for detecting winding deformation, although the interpretation of the response can change with transformer structure, connection arrangement and measurement practice [1], [6].

Hong et al. converted vibration monitoring data and load information into vibration images and used a CNN classifier for transformer winding condition assessment. Their study showed that image-based vibration features can support non-intrusive winding-fault diagnosis under a controlled experimental setting [2]. Li et al. later developed a CNN-based transformer fault-diagnosis method using vibration signals from dry-type transformers and reported competitive diagnostic results, confirming the value of time-frequency learning for transformer vibration diagnosis [4].

B. Acoustic and Multimodal Diagnosis

Acoustic sensing captures sound patterns produced by abnormal mechanical movement, localized impacts and discharge-related activity. Ma et al. presented a non-contact acoustic-emission system for substation power-transformer fault detection using single-microphone measurements and automated discrimination [3]. Acoustic diagnosis is attractive because it is non-invasive, but it is sensitive to environmental noise and

ISSN NO: 2770-4491
May 2024

microphone placement. This makes acoustic information better suited as a complementary channel than as the only diagnostic source.

C. Deep Learning for Condition Monitoring

Deep learning reduces the need for hand-crafted features by learning representations directly from time series, spectrograms, scalograms or diagnostic images. CNNs are effective for local time-frequency pattern extraction, while recurrent networks can capture temporal dependencies. Long short-term memory and bidirectional recurrent networks are well suited to signals in which useful events are intermittent or distributed across time [7], [8]. Attention mechanisms further improve representation learning by assigning higher weight to the most informative hidden states [9], [10]. Similar CNN and time-frequency methods have been effective in other rotating-machine diagnosis problems, where fault energy is distributed across frequency and time [11]-[13].

D. Research Gap

The reviewed literature shows that vibration analysis, acoustic sensing, CNN-based image recognition and time-frequency learning are useful for transformer condition monitoring. A remaining need is an early-warning architecture that jointly uses synchronized vibration and acoustic data, preserves temporal information, applies attention to learned features and evaluates the model with a run-level split that prevents leakage from overlapping windows. This study addresses that need through a dual-branch vibro-acoustic CNN-BiLSTM attention fusion model trained and evaluated on a synchronized laboratory dataset.

III. Materials and Methods

A. Overall System Description

The proposed system has 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 corrected for offset, filtered, segmented, normalized with training-set statistics and converted into image-like time-frequency representations. The vibration branch receives CWT scalograms, while the acoustic branch receives Mel-spectrograms. The learned embeddings from both branches are concatenated and passed to a softmax classifier.

ISSN NO: 2770-4491
May 2024

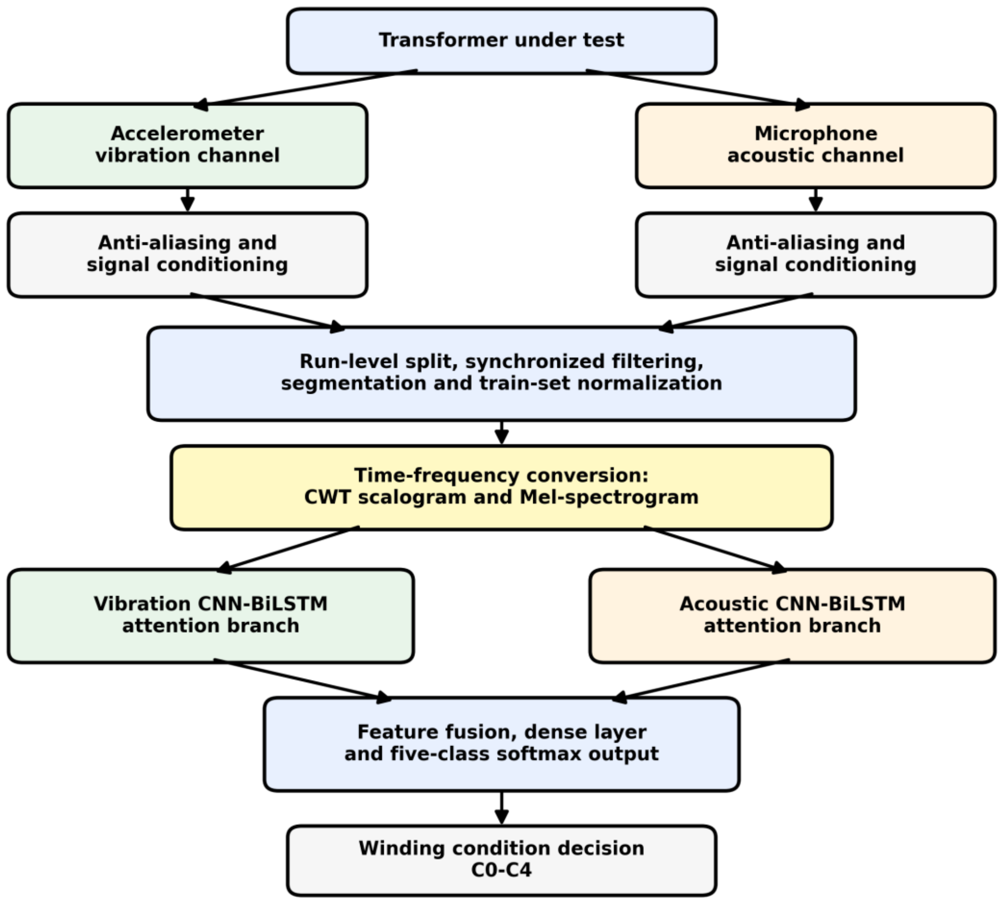

Fig. 1. Proposed vibro-acoustic transformer winding-fault detection framework.

B. Transformer Test Setup

Recordings 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. The transformer was loaded between 25% and 100% of rated load using a resistive-inductive loading bank. Forced cooling was not used during the short acquisition runs, reducing the influence of cooling-fan noise on the acoustic channel.

The setup included a synchronized data-acquisition module, calibrated accelerometers fixed to marked points on the transformer tank and a calibrated airborne microphone positioned 0.5 m from the tank wall at approximately winding-region height. Each recording was captured after the operating condition had stabilized. Vibration and acoustic channels used the same acquisition clock, and each vibration window was paired with the corresponding acoustic window. Only one fault condition was introduced in each recording run so that the class label remained unambiguous.

 ISSN NO: 2770-4491
May 2024

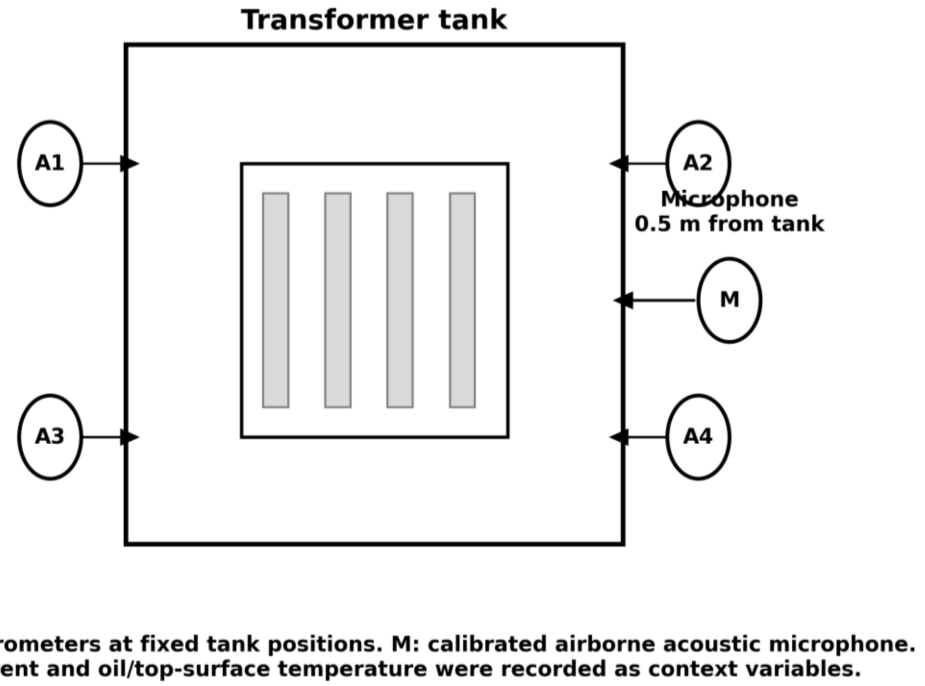

Fig. 2. Experimental sensor placement on the transformer tank.

C. Fault Classes

Five operating states were used: normal condition, winding looseness, axial displacement, radial deformation and incipient inter-turn fault. The fault-emulation procedure was non-destructive and controlled. Winding looseness was represented by reducing effective clamping pressure and by using a loosened mechanical fixture that changed the tank vibration. Axial displacement was represented by a controlled vertical shift of the winding fixture from its normal seated position. Radial deformation was represented by a controlled lateral displacement that reproduced the mechanical effect of radial bending or buckling. The incipient inter-turn condition was represented by a low-energy auxiliary shorted-turn emulator operated through 10-20 Ω current limitation, fuse protection, supervised switching and short test duration. The test was stopped if abnormal temperature rise, excessive current or unstable acoustic/vibration response was observed.

TABLE I
TRANSFORMER WINDING FAULT CLASSES USED FOR MODEL TRAINING

Class	Transformer condition	Description
C0	Normal	Healthy winding condition under the 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 emulator with 10-20 Ω current limitation and supervised short-duration switching

D. Sensor Specification and Acquisition Parameters

The acquisition settings were selected to preserve mechanical and acoustic fault information while remaining suitable for online monitoring. Both channels were sampled at 25.6 kHz using the same acquisition clock and a 24-bit data-acquisition unit with hardware anti-aliasing. The vibration channel used a laboratory accelerometer with nominal 100 mV/g sensitivity, checked before acquisition with a portable vibration calibrator. The acoustic channel used a calibrated condenser microphone checked with a 94 dB, 1 kHz acoustic calibrator. Load current and temperature were retained as context variables for interpretation and operating-condition checks; they were not used as primary diagnostic channels.

ISSN NO: 2770-4491
May 2024

TABLE II
SENSOR SPECIFICATION AND DATA ACQUISITION SETTINGS

Parameter	Value used in this study	Purpose
Vibration sensor	Calibrated accelerometer	Captures tank and winding vibration response
Acoustic sensor	Calibrated condenser microphone	Captures airborne acoustic activity near the tank
Accelerometer sensitivity	Nominal 100 mV/g; checked before acquisition	Supports repeatable vibration scaling
DAQ resolution	24-bit synchronized acquisition	Maintains channel timing and amplitude resolution
Sampling rate	25.6 kHz	Preserves vibration and acoustic fault information
Window length	1 s	Provides frequency resolution for early detection
Overlap	50%	Improves sample continuity without crossing run-level split boundaries
Input image size	128 x 128 pixels	Compatible with compact CNN processing
Transformer rating	15 kVA, 11/0.415 kV, 50 Hz	Defines the laboratory test-bench condition
Loading range	25%-100% of rated load	Captures load-dependent mechanical and acoustic response
Vibration position	Fixed tank-wall positions close to winding region	Reduces sensor-placement variation
Acoustic position	0.5 m from tank wall at winding height	Keeps microphone distance constant

E. Signal Preprocessing and Time-Frequency Conversion

The raw vibration and acoustic signals were synchronized, corrected for DC offset and filtered. A fourth-order Butterworth band-pass filter was applied to each channel. The vibration channel was filtered from 10 Hz to 5 kHz, and the acoustic channel was filtered from 20 Hz to 8 kHz. Each signal was divided into 1 s windows with 50% overlap. Vibration windows were converted to CWT scalograms using a complex Morlet wavelet with 64 scale levels, following the time-frequency representation principles described by Mallat [14]. Acoustic windows were converted to Mel-spectrograms using a 1024-point FFT, 512-sample hop length, Hann window and 64 Mel bands. The resulting maps were resized to 128 x 128 pixels. Normalization statistics were computed from the training set only and then applied to validation and test sets. Data augmentation was limited to low-level noise injection, small time shifts and random gain scaling during training.

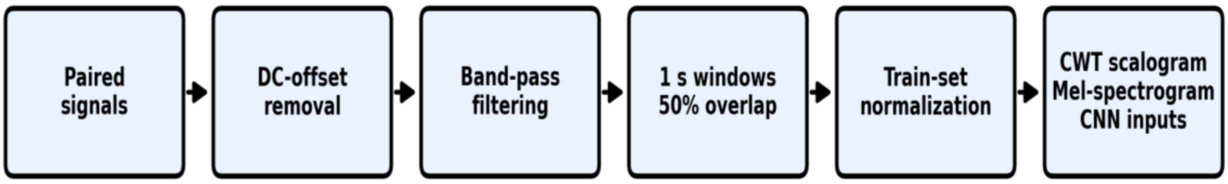

Fig. 3. Signal preprocessing and time-frequency transformation pipeline. The raw-run split is completed before window generation so that overlapping windows from one recording run cannot appear in different subsets.

Let x_v[n] and x_a[n] denote the synchronized vibration and acoustic signals. For channel m, a windowed segment is extracted with length L and stride S as follows:

$$x_m^{(k)}[n] = x_m[n + kS], 0 \leq n < L, m \in \{v, a\} \quad (1)$$

The digital band-pass filtering operation is expressed as:

$$\tilde{x}_m[n] = (h_m * x_m)[n] \quad (2)$$

 ISSN NO: 2770-4491
May 2024

The vibration branch uses the magnitude of the CWT coefficient matrix:

$$C_v(s,\tau) = \left| \frac{1}{\sqrt{s}} \sum_n x_v[n]\psi \left(\frac{n-\tau}{s} \right) \right| \tag{3}$$

The acoustic branch uses Mel-scaled spectral energy:

$$P_a(f,t) = |\mathrm{STFT}\{x_a[n]\}|^2 \tag{4}$$

$$M_a(b,t) = \sum_f H_b(f) P_a(f,t) \tag{5}$$

Each feature map is normalized with statistics estimated from the training set:

$$\hat{X}_m = \frac{X_m - \mu_{train,m}}{\sigma_{train,m} + \epsilon} \tag{6}$$

IV. Proposed Deep Learning Framework

A. Model Overview

The model contains two parallel input branches. The vibration branch receives the CWT scalogram, while the acoustic branch receives the Mel-spectrogram from the same time window. Each branch applies two convolutional blocks to learn local time-frequency features. The resulting feature maps are reshaped along the time dimension and passed through a BiLSTM layer. Additive attention produces a weighted feature vector for each modality. The two vectors are concatenated, regularized with dropout and classified through a dense softmax layer into the five winding states.

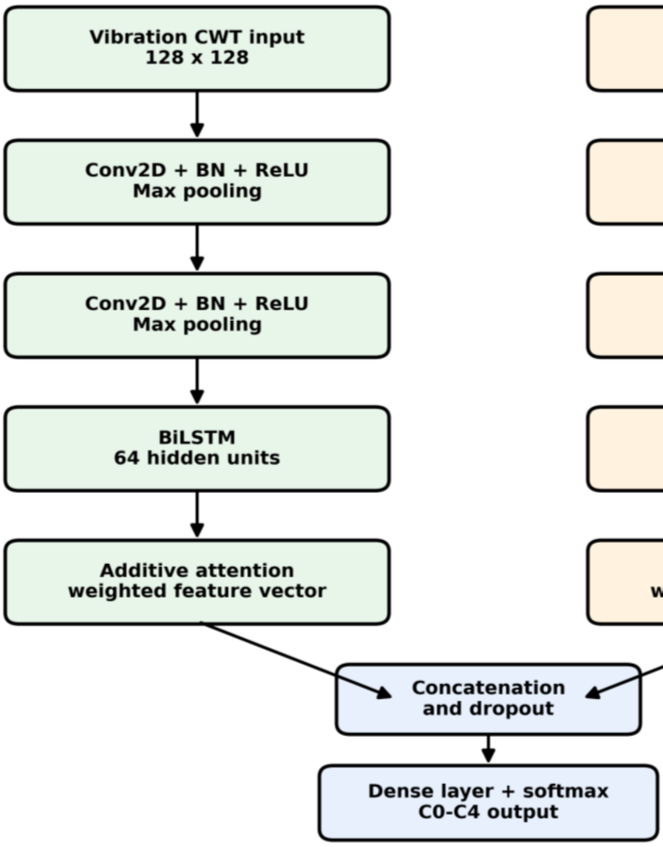

Fig. 4. Dual-branch CNN-BiLSTM attention fusion architecture.

B. CNN Feature Extraction Blocks

Each branch uses two compact convolutional blocks. Each block contains a two-dimensional convolutional layer, batch normalization, ReLU activation and max pooling. The first block extracts local spectral patterns, while the second block learns deeper time-frequency textures associated with looseness, displacement, deformation and short burst-like acoustic signatures. Equation (7) represents one generic convolutional block applied successively in each branch. The trained fusion model contained about 1.87 million trainable parameters and occupied about 7.5 MB with 32-bit floating-point weights. The measured inference time on the GPU workstation was 3.8 ms per paired vibro-acoustic window.

ISSN NO: 2770-4491
May 2024

$$F_m^{(\ell)} = \mathrm{Pool}\left(\mathrm{ReLU}\left(\mathrm{BN}\left(W_m^{(\ell)} * F_m^{(\ell-1)} + b_m^{(\ell)}\right)\right)\right) \quad (7)$$

C. BiLSTM Temporal Learning

Although the model receives time-frequency images, the convolutional feature maps preserve a time order. A BiLSTM layer was therefore used to capture forward and backward dependencies in the extracted feature sequence. This is useful when a fault-related pattern is short, intermittent or distributed across neighbouring time regions.

$$h_t = \left[\overrightarrow{\mathrm{LSTM}}(q_t); \mathrm{LST\acute{M}}(q_t)\right) \quad (8)$$

D. Attention-Based Feature Weighting

The attention layer assigns larger weights to hidden states that contain stronger fault evidence. This prevents the classifier from treating all time-frequency regions as equally informative.

$$e_t = u^T tanh(W_h h_t + b_h), \alpha_t = \frac{exp(e_t)}{\sum_j exp\left(e_j\right)}, z_m = \sum_t \alpha_t\, h_t \quad (9)$$

E. Multimodal Fusion and Classification

The vibration feature vector and acoustic feature vector are concatenated before classification. The fused vector is passed to a dense layer and a softmax output layer. Categorical cross-entropy is used as the training loss.

$$z_f = [z_v; z_a) \quad (10)$$

$$\hat{y} = \mathrm{softmax}\left(W_f z_f + b_f\right) \quad (11)$$

$$L = -\sum_{c=1}^{C} y_c\, log(\hat{y}_c) \quad (12)$$

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

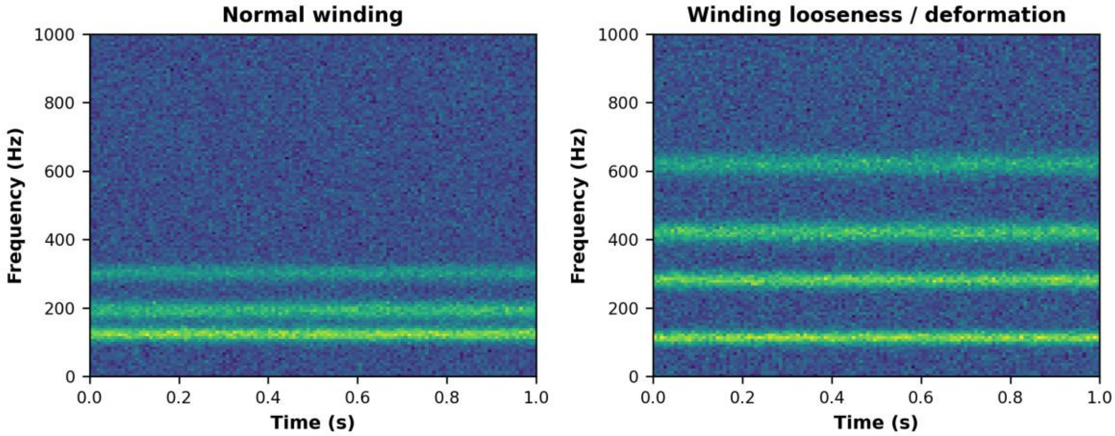

Fig. 5. Representative vibration time-frequency maps for normal and faulty conditions.

The representative maps in Figs. 5 and 6 were generated from synchronized windows in the prepared dataset and are used to show typical time-frequency differences between healthy and fault-related conditions. The faulty vibration map in Fig. 5 shows stronger localized energy bands and a shift in dominant frequency regions. Such changes are consistent with looseness, displacement or deformation that alters the tank and winding response. The acoustic map in Fig. 6 shows broader harmonic activity and short burst-like regions under the fault-related condition. These acoustic patterns complement the vibration features and help the fusion model when a fault signature is weak in one sensor channel.

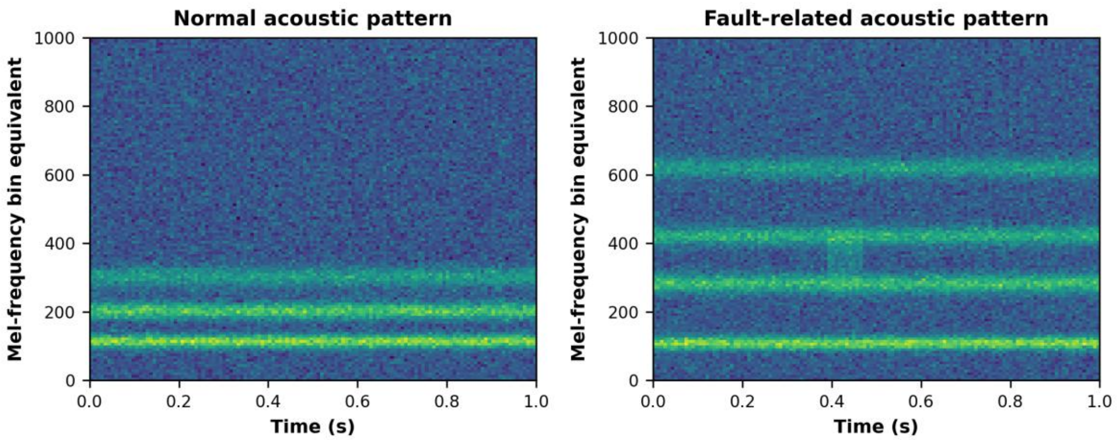

Fig. 6. Representative acoustic Mel-spectrograms for normal and faulty conditions.

V. Experimental Setup

A. Dataset Preparation

The dataset was prepared from synchronized vibration and acoustic recordings collected under the five winding-condition classes listed in Table I. It 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 split was made by raw recording run before window-based learning. 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. In window counts, the split produced 5,040 training windows, 1,080 validation windows and 1,080 test windows. Each sample contained one vibration window, one acoustic window and one class label.

TABLE IV
DATASET STRUCTURE USED FOR TRAINING AND EVALUATION

Item	Description
Transformer type	15 kVA, 11/0.415 kV, 50 Hz oil-immersed laboratory distribution transformer
Operating condition	25%-100% controlled load range using a resistive-inductive load bank
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
Raw run duration	6.5 s retained waveform per run, giving 12 paired windows per run
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
Leakage prevention	Split by raw recording run; no windows from one run appeared in more than one subset
Validation procedure	Five independent training runs with seeds 21, 42, 84, 126 and 168
Environment	Python 3.11, TensorFlow/Keras 2.15, NumPy/SciPy preprocessing scripts and CUDA-enabled GPU workstation

B. Training Configuration

The network was trained using the Adam optimizer with an initial learning rate of 0.001 [15]. The batch size was 32, and the maximum number of epochs was 80. Early stopping with validation-loss patience of 10 epochs was applied to limit overfitting. Dropout was fixed at 0.4 after the fusion layer. 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
Trainable parameters	Approximately 1.87 million
Average inference time	3.8 ms per paired vibro-acoustic window; effective update interval governed by 0.5 s window stride

C. Baseline Models

The fusion-attention model was compared with traditional machine-learning classifiers, single-modality deep learning models and fusion variants. Traditional models used statistical and spectral handcrafted features extracted from the same train-test partitions. Single-modality models used only vibration or only acoustic input. Fusion variants tested the effect of feature fusion, attention and controlled augmentation.

TABLE VI
BASELINE MODEL CATEGORIES

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

D. Evaluation Metrics

The evaluation metrics were accuracy, precision, recall, F1-score, false alarm rate, inference latency and noise robustness. For the five-class task, precision, recall and F1-score were computed per class and reported as macro-averaged values across the five winding states. Accuracy measures the overall proportion of correct predictions. Precision measures the reliability of predicted fault labels. Recall measures the ability to detect actual fault samples. F1-score balances precision and recall. False alarm rate is the proportion of healthy windows incorrectly classified as any fault class. Inference latency was measured as the average prediction time for one paired vibration-acoustic window on the GPU workstation. Noise robustness was evaluated only on the held-out test set after model training.

For the false-alarm-rate expression, FP_H denotes healthy windows incorrectly classified as fault windows, while TN_H denotes healthy windows correctly classified as healthy.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{13}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

$$\text{FAR} = \frac{FP_H}{FP_H + TN_H} \tag{17}$$

VI. Results and Discussion

A. Basis of the Reported Results

All performance values in this section were computed on the held-out test set after five independent training runs. Because the comparison studies used different transformers, sensors, fault conditions and datasets, Table X is used to position the proposed method rather than to claim direct cross-dataset superiority.

B. Classification Performance

The fusion-attention model achieved 98.92 ± 0.24% accuracy across five independent training runs and outperformed the vibration-only, acoustic-only and fusion-without-attention baselines. The gain from fusion is physically plausible because winding looseness and displacement modify the tank vibration path, while the microphone captures airborne components associated with rubbing, impacts and short burst-like mechanical events. The median-run confusion matrix matched the stated 1,080-window test set and showed that most residual errors occurred between axial displacement and radial deformation, where the mechanical signatures are close. Table VIII reports the run-wise test accuracies for the five random seeds. The fusion-attention model improved the mean accuracy by about 1.10 percentage points over the fusion-without-attention baseline. The paired comparison used the same run-level split for both internal models.

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
Fusion without attention	97.82 ± 0.31	97.48 ± 0.33	97.31 ± 0.36	97.39 ± 0.34
Vibro-acoustic fusion-attention model	98.92 ± 0.24	98.61 ± 0.27	98.47 ± 0.29	98.54 ± 0.26

TABLE VIII
RUN-WISE ACCURACY FOR THE MAIN INTERNAL COMPARISON

Run seed	Fusion without attention (%)	Fusion-attention model (%)
21	97.43	98.60
42	97.61	98.78
84	97.82	98.92
126	98.03	99.08
168	98.21	99.22

TABLE IX
OPERATIONAL METRICS OF THE FINAL FUSION-ATTENTION MODEL

ISSN NO: 2770-4491
May 2024

Metric	Value	Interpretation
False alarm rate	0.46% on the held-out test set	Healthy windows were rarely classified as fault windows
Macro ROC-AUC	0.998	Classes were well separated under the controlled laboratory conditions
Average inference latency	3.8 ms per paired window	Computational delay is small relative to the 1 s window length
Model footprint	Approximately 7.5 MB using 32-bit floating-point weights	Compact enough for edge-assisted monitoring

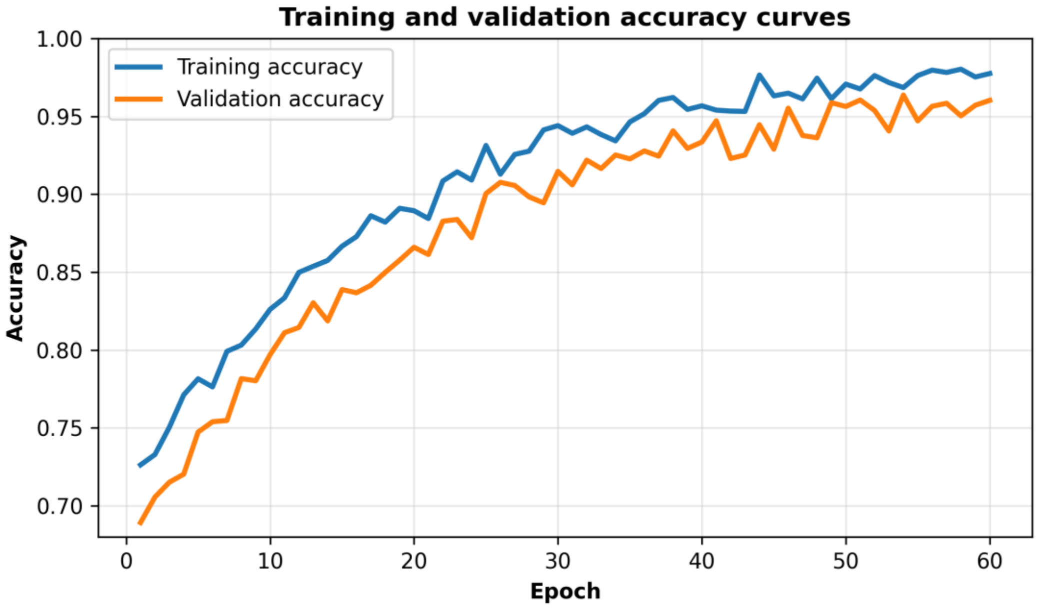

Fig. 7. Training and validation accuracy curves.

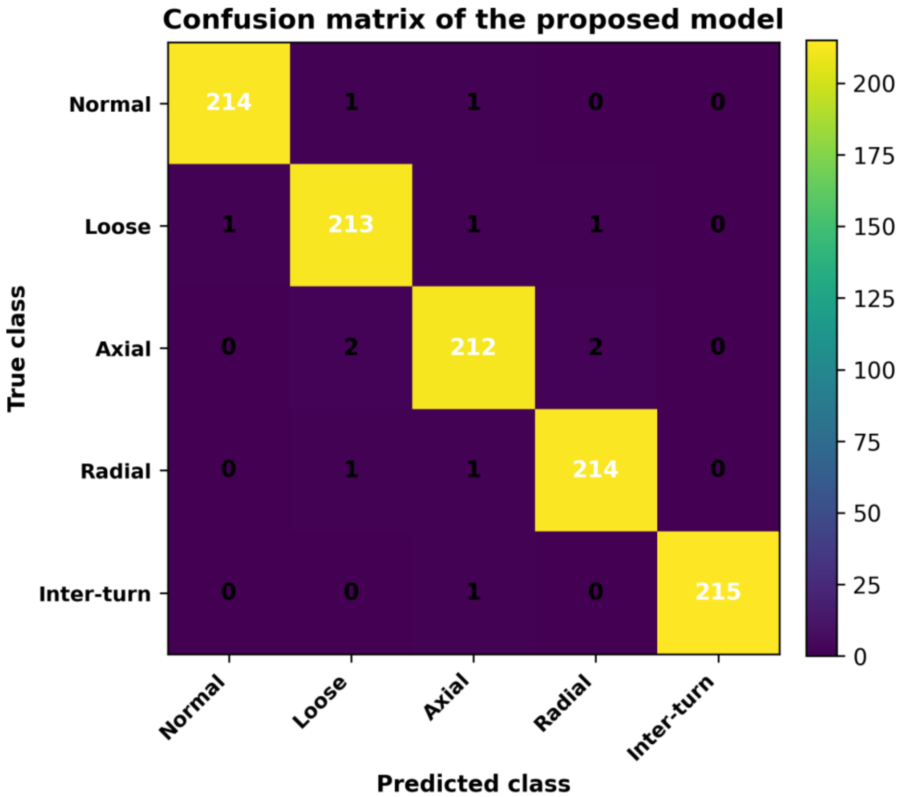

Fig. 8. Confusion matrix for the median-performing run on the 1,080-window test set.

 ISSN NO: 2770-4491
May 2024

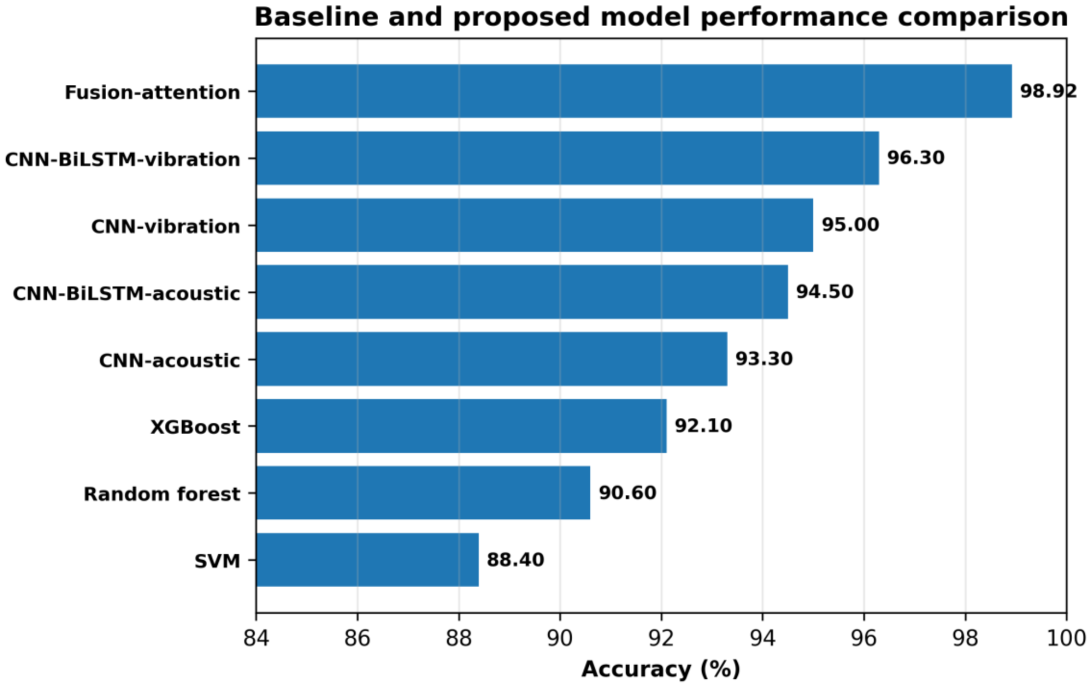

Fig. 9. Performance comparison with baseline models.

C. Confusion Matrix Analysis

The median-run confusion matrix in Fig. 8 contains 1,080 test windows, with 216 windows per class. The matrix gives 1,068 correct predictions and 12 misclassifications, corresponding to 98.89% accuracy for that run. Most remaining errors occur between axial displacement and radial deformation because both faults can shift tank vibration patterns in nearby frequency regions. Acoustic evidence reduces this confusion but does not remove it completely. The normal class and incipient inter-turn class were detected most consistently in the median-performing run.

D. Comparison with Published Studies

The method builds on earlier transformer vibration-diagnosis studies by adding synchronized acoustic information and attention-based feature fusion. Because the studies in Table X used different datasets, transformer ratings, sensors, operating conditions and fault-emulation procedures, the table is used to position the proposed method rather than to claim direct numerical equivalence. The comparison shows the value of moving from single-source diagnosis toward fused diagnostic evidence for transformer condition monitoring.

TABLE X
COMPARISON WITH PUBLISHED TRANSFORMER FAULT-DIAGNOSIS STUDIES

Study	Signal source	Main method	Reported result	DOI
Hong et al. [2]	Transformer vibration	Vibration image with CNN	High-accuracy vibration-image diagnosis under their experimental setting	10.1109/TPWRD.2020.2988820
Li et al. [4]	Dry-type transformer vibration	CWT image with CNN	High-accuracy CWT/CNN vibration diagnosis under controlled setup	10.3390/s23104781
Ma et al. [3]	Acoustic emission / airborne acoustic signal	Non-contact acoustic fault detection	Robust acoustic detection under controlled conditions	10.3390/app12052759
Chen et al. [5]	Transformer diagnostic features	Multiscale entropy and stochastic configuration network	Automated power-transformer fault diagnosis	10.3390/e24081135

ISSN NO: 2770-4491
May 2024

Study	Signal source	Main method	Reported result	DOI
This study	Vibration and acoustic signals	Dual-branch CNN-BiLSTM attention fusion	98.92 ± 0.24% accuracy from five training runs on the laboratory dataset	—

E. Ablation Study

The ablation study separates the contribution of each modelling component. The vibration branch is strong because winding deformation changes the mechanical transfer path from winding and core to the tank wall. The acoustic branch is informative but more exposed to ambient noise and microphone placement. Fusion without attention improves the result by combining these two physical responses, while attention gives an additional gain by assigning more weight to time-frequency regions that contain clearer fault evidence. Controlled augmentation improves noise tolerance, although the improvement is modest because the dataset is balanced and the split is controlled at recording-run level.

TABLE XI
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 balance of accuracy and robustness

F. Noise Robustness Analysis

Noise robustness is important because acoustic signals in substations may be affected by fans, switching operations, nearby equipment, wind and human activity. Vibration signals may also be affected by mounting condition and transformer loading. In this study, noise was added only to the held-out test set after training. Gaussian perturbations were used for vibration data, while Gaussian and recorded ambient acoustic perturbations were used for acoustic data. Table XII and Fig. 10 show that the fusion model maintained higher accuracy than either single-sensor model as SNR decreased. The accuracy still dropped at very low SNR, confirming the need for sensor calibration, stable mounting, local noise profiling and periodic recalibration during deployment.

TABLE XII
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

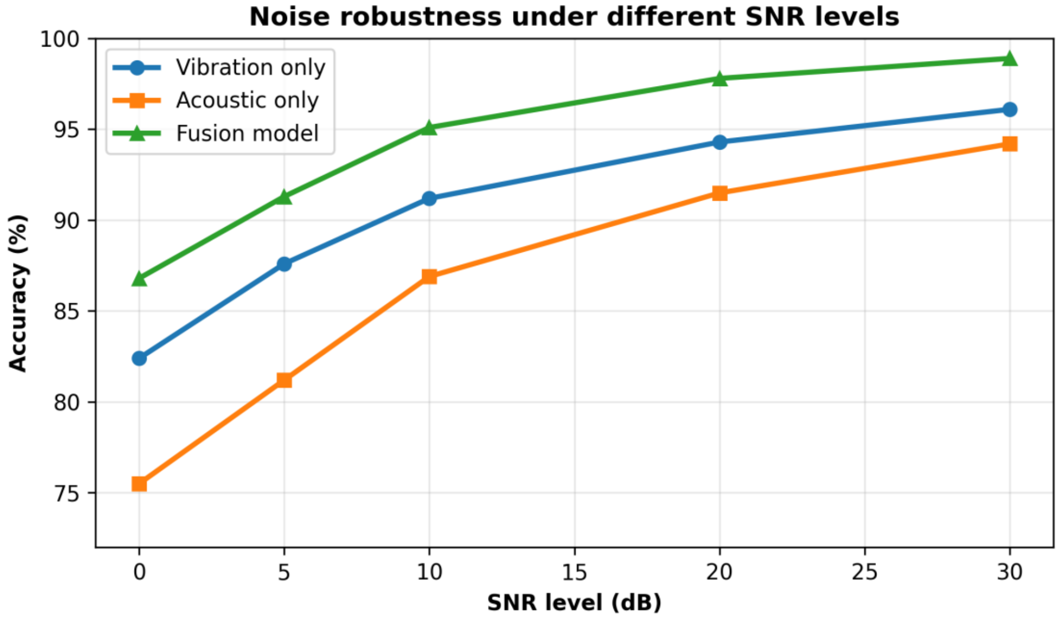

Fig. 10. Noise robustness comparison under different SNR levels.

G. Practical Deployment Considerations

The framework can support online transformer condition monitoring when the sensor layout is reproduced carefully and calibration is maintained. Mounting pressure, tank position, microphone distance and local noise sources can shift the signal distribution, so these factors must be controlled during deployment. The 3.8 ms inference time indicates a small computational delay per paired 1 s window; the effective alarm update interval is governed by the 0.5 s segmentation stride and preprocessing time. For substation use, the model output should be interpreted as an alarm signal and checked against established tests such as frequency response analysis, dissolved gas analysis, thermal inspection and expert review before major maintenance decisions are made.

VII. Conclusion

This study developed a vibro-acoustic deep learning framework for early detection of transformer winding faults. The system combines vibration measurements from accelerometers with acoustic measurements from a calibrated microphone. The recordings were obtained from a controlled 15 kVA, 11/0.415 kV, 50 Hz oil-immersed laboratory transformer operated under a 25%-100% load range. Healthy winding, winding looseness, axial displacement, radial deformation and incipient inter-turn conditions were represented one at a time through controlled fault emulation. The vibration and acoustic signals were transformed into time-frequency images and processed by a dual-branch CNN-BiLSTM attention fusion model.

The fusion-attention model achieved 98.92 ± 0.24% accuracy across five independent training runs and outperformed the vibration-only, acoustic-only and fusion-without-attention baselines. The gain from fusion is physically plausible because winding looseness and displacement modify the tank vibration path, while the microphone captures airborne components associated with rubbing, impacts and short burst-like mechanical events. The median-run confusion matrix matched the stated 1,080-window test set and showed that most residual errors occurred between axial displacement and radial deformation, where the mechanical signatures are close.

References

[1] IEEE Guide for the Application and Interpretation of Frequency Response Analysis for Oil-Immersed Transformers, IEEE Std C57.149-2012, 2013.

[2] 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.

[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.

ISSN NO: 2770-4491
May 2024

[4] 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.
[5] 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.
[6] S. Alsuhaibani, Y. Khan, A. Beroual, and Z. Malik, "A review of frequency response analysis methods for power transformer diagnostics," Energies, vol. 9, no. 11, Art. no. 879, Nov. 2016, doi: 10.3390/en9110879.
[7] 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.
[8] 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.
[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. International Conference on Learning Representations, 2015.
[10] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[11] 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.
[12] 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.
[13] 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.
[14] S. Mallat, A Wavelet Tour of Signal Processing: The Sparse Way, 3rd ed. Burlington, MA, USA: Academic Press, 2009.
[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. International Conference on Learning Representations, 2015.