

DUAL-ATTENTION FUSION NETWORK FOR BEARING FAULT DIAGNOSIS USING HYBRID DEEP LEARNING WITH VIBRATION SIGNAL ANALYSIS

DR. SUDHAKAR JYOTHULA¹, MADHAVI PERLA², DR. KATTA NAGARAJU³, T. KRISHNA MOHANA⁴, DR. KISHORE KUMAR K⁵, BABYMOL KURIAN⁶, CHALLAPALLI SUJANA⁷, DR. VADDEMPUDI SUJATHA LAKSHMI⁸

¹Department of ECE, Vignan's Institute of Information Technology, Visakhapatnam, Andhra Pradesh, India, ²Assistant Professor, Dept of CSE-AI&ML, GMRIIT Deemed to be University, GMR Nagar, Rajam, Vizayanagaram Dt, 532127.

³Assistant Professor, Department of CSE, Aditya University, Surampalem⁴Assistant Professor, Department of ECE, Aditya University⁵Associate Professor, Dept of CSE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology. 0000-0002-3821-8297

⁶Associate Professor, Department of Information Technology, KCG College of Technology. , 0000-0002-8812-5734

⁷Department of CSE, Aditya University, Surampalem, Andhra Pradesh-533437, India, 0009-0009-7927-5608

⁸Associate Professor, Computer Science & Engineering, R.V.R&J.C College of Engineering. , 0000-0002-1113-4018

E-mail: sudhakar.jyo@gmail.com, madhavi.p@gmritdu.edu.in, . nagarajuk@adityauniversity.in, . krishnamohana.tenneti@acet.ac.in, kishorekumar@veltech.edu.in, babymol.it@kcgcollege.com, . sujana.challapalli@gmail.com, sujathavdmpudi@gmail.com

ABSTRACT

Bearing failures in rotating machinery systems are a major cause of unplanned downtime and equipment failure. Traditional machine learning techniques suffer from the limitations of manual feature vectorization and low generalization ability in diverse working scenarios. We design a new network architecture, Dual-Attention Fusion Network (DAFN), that consists of three parallel branches — 1D-CNN, BiLSTM and Transformer encoder — fused through a novel dual-level attention module composed of channel attention (CAM) and spatial attention (SAM) designed to recalibrate and fuse multi-scale features learned directly from raw vibration signals in an adaptive way.

The proposed model is tested on the Case Western Reserve University (CWRU) and MFPT bearing datasets under diverse noise environments with SNRs from -10 dB to +20 dB. Experimental results show that DAFN achieves 96.7% classification accuracy and AUC of 0.985 on the CWRU benchmark dataset, outperforming standalone CNN, LSTM and Transformer models by margins of 9.4%, 7.1% and 5.5% respectively. The model sustains an F1-score of 74.2% at -10 dB compared to 61.3% for the best competing method. Ablation studies confirm the dual attention module alone provides a 3.2% cumulative accuracy gain. Without architectural modification, DAFN achieves 95.3% accuracy on MFPT, with inference latency of 9.3 ms meeting real-time industrial requirements.

Keywords – *Attention Mechanism, Bearing Fault Diagnosis, Bidirectional LSTM, Deep Learning, Dual-Attention Fusion, Transformer Encoder, Vibration Signal Analysis*

1. INTRODUCTION

Rolling element bearings are widely used in rotating machinery applications such as induction motors, wind turbines and industrial gearboxes. Bearing failures account for approximately 40–45% of the total damage in rotating machines, causing significant economic losses and unplanned machine downtime [1,2]. As industrial systems become more complex, intelligent and automated diagnostic frameworks capable of identifying early bearing defects without manual intervention are increasingly required. Condition monitoring based on vibration

signal analysis has become the most popular non-destructive technique for bearing health evaluation [3].

Conventional signal processing methods such as FFT, Empirical Mode Decomposition (EMD) and wavelet transforms provide useful frequency-domain characterizations, but require domain expertise to select features and are sensitive to non-stationary operational conditions [4,5]. Shallow machine learning classifiers (SVM, k-NN, Random Forest) are limited by hand-crafted feature engineering pipelines that impede adaptability to diverse fault modes and varying load characteristics [6]. Deep learning has revolutionized data-driven

fault diagnosis by enabling automatic feature extraction directly from raw sensor measurements [7], yet existing models are still limited by multi-scale temporal feature fusion, noise sensitivity and cross-domain generalization.

CNNs excel at recognizing spatial local patterns but cannot model long-range temporal dependencies. Transformer-based models achieve global self-attention [8] but require large datasets and can overfit small industrial datasets [9]. We propose the Dual-Attention Fusion Network (DAFN), a hybrid architecture coupling 1D-CNN, BiLSTM and Transformer encoder branches through a novel dual-level attention mechanism to recalibrate and fuse features adaptively. The main contributions are: (i) a multi-branch hybrid architecture modeling local, sequential and global features simultaneously; (ii) a dual-level attention module capturing channel-wise and spatial context; (iii) extensive evaluation on two public benchmarks with realistic noise injection; (iv) ablation studies validating each component; and (v) attention weight visualizations demonstrating physically meaningful feature learning.

1.1 Problem Statement

Three unresolved problems limit the industrial deployment of deep bearing-fault classifiers [26–28]. First, noise robustness: most models are developed on clean, single-condition recordings, and accuracy drops sharply under plant-floor measurement noise; noise-oriented designs such as the deep residual shrinkage network [29] mitigate this only for convolutional features. Second, representational incompleteness: a bearing fault signature simultaneously contains local impulses, their modulation across successive shaft revolutions, and long-range periodicity, yet CNNs [11], recurrent networks [9,13] and Transformers [8,30] each capture only one view, and existing hybrids fuse by plain concatenation that treats every feature as equally reliable [14]. Third, comparability and reproducibility: inconsistent data splits, undocumented preprocessing and overlapping segments make many reported accuracies non-comparable [28].

The problem addressed in this study is: how to design a bearing fault diagnosis model that jointly exploits local, sequential and global characteristics of raw vibration signals, adaptively emphasizes the most fault-relevant features under severe noise, generalizes across datasets without redesign, and remains reproducible and computationally feasible for real-time industrial monitoring.

1.2 Research Questions and Objectives

RQ1: Can a parallel fusion of 1D-CNN, BiLSTM and Transformer encoder branches improve diagnostic performance over single-paradigm and two-branch models on raw vibration signals? O1: design DAFN and compare it with six baselines on CWRU using Accuracy, F1-score, AUC and MCC.

RQ2: Does a dual-level (channel and spatial) attention fusion module contribute measurably beyond plain feature concatenation? O2: formulate the DAF module and quantify each component through a stepwise ablation study.

RQ3: How robust is the proposed model when SNR decreases from +20 dB to −10 dB? O3: evaluate all models under seven controlled SNR levels using additive white Gaussian noise.

RQ4: Does the model generalize to a second dataset while satisfying real-time constraints? O4: evaluate DAFN on MFPT without architectural change and measure parameters, FLOPs and inference latency.

1.3 Research Hypotheses

H1 (multi-branch fusion): DAFN achieves higher Accuracy, F1-score, AUC and MCC than the best single-branch and two-branch baselines on the CWRU dataset.

H2 (dual-level attention): Adding CAM and then SAM to the concatenated multi-branch features increases classification accuracy at each stage relative to the fusion-only variant.

H3 (noise robustness): The F1-score advantage of DAFN over the best competing method is larger at low SNR (≤ 0 dB) than at high SNR.

H4 (generalization and efficiency): Without architectural modification, DAFN attains accuracy on MFPT comparable to that on CWRU and performs inference within the 10 ms real-time cycle of typical industrial controllers.

2. LITERATURE REVIEW

Early vibration-based diagnosis relied on statistical time-domain indicators such as RMS, kurtosis and crest factor combined with envelope analysis [10], and on shallow classifiers such as ANN, SVM and k-NN [2,5,6]. Kankar et al. [2] and Widodo and Yang [6] showed that such classifiers separate fault classes well when features are carefully engineered. The critical weakness is that discriminative power resides in the hand-crafted features rather than in the learner: features tuned for one load or speed transfer poorly to another, and indicators such as kurtosis are easily masked by broadband noise [10]. The benchmark analysis of Smith and Randall [22] further cautioned against

interpreting near-perfect CWRU accuracy as genuine diagnostic capability.

Deep learning shifted the burden from feature design to representation learning [7,13,26]. Zhang et al. [11] demonstrated with WDCNN that a 1D-CNN with a wide first-layer kernel can learn directly from raw signals and tolerate moderate noise, while Wen et al. [12] converted signals into 2D grey images to reuse image CNNs. Both established end-to-end diagnosis, but convolutions with limited receptive fields cannot model how impulses evolve over successive revolutions. Recurrent models [9] and CNN-RNN hybrids [14] address sequential dependence with consistent gains reported in the review of Zhao et al. [13]. Nevertheless, these hybrids combine features by concatenation, implicitly assuming all learned features are equally reliable — an assumption that fails when noise corrupts only certain frequency bands.

Attention mechanisms relax this assumption. Squeeze-and-excitation channel attention [16] recalibrates feature channels, CBAM [31] adds a sequential spatial-attention stage, and the deep residual shrinkage network [29] uses attention-derived soft thresholds to suppress noise-related features. Transformer encoders [8,17] and time-frequency Transformers [30] model global dependencies. Critically, each approach strengthens only one dimension: SE- and shrinkage-type modules refine convolutional features alone, and

Transformer models are data-hungry and prone to overfitting small industrial datasets [27,30]. Evidence on the benefit of attention-guided refinement and adaptive multi-path fusion reported in related domains [33,34] has not been transferred to vibration-based fault classification.

2.1 Critical Synthesis and Research Gaps

Three gaps emerge from this critique. G1 (complementarity gap): no reviewed study systematically combines convolutional, recurrent and self-attention representations in parallel, although each captures a distinct aspect of the fault signature [11,14,30]. G2 (fusion gap): where hybrids exist, fusion is performed by concatenation and attention is applied within a branch rather than across branches. G3 (evaluation gap): robustness is commonly reported at only one or two noise levels, cross-dataset evaluation is rare, and preprocessing and splitting protocols are often insufficiently documented for replication [22,28].

The present study addresses G1 through the three-branch DAFN architecture (RQ1, H1), G2 through the Dual-Level Attention Fusion module and its ablation (RQ2, H2), and G3 through a seven-level SNR evaluation, cross-dataset validation on MFPT and a fully specified research protocol (RQ3–RQ4, H3–H4). Table 1 summarizes representative methods and their limitations.

Table 1: Comparative Literature Review — Bearing Fault Diagnosis Methods

Reference	Method	Dataset	Acc(%)	Feature	Attention	Limitation	Year
Patel et al. [1]	SVM-RBF	CWRU	87.3	Handcraft.	None	Low noise robustness	2018
Kankar et al. [2]	SVM, k-NN	CWRU	—	Handcraft.	None	Feature engineering required	2011
Zhang et al. [11] (WDCNN)	1D-CNN	CWRU	~90%	Raw signal	None	Degrades under cross-load conditions	2017
Wen et al. [12]	2D-CNN (LeNet-5)	CWRU+MFPT	High	2D image	None	Signal-to-image adds complexity	2018
Chen et al. [18]	Transfer CNN	CWRU/PU	92.5	Raw signal	Self-Attn	High data requirement	2020
Wang et al. [14]	CNN+RNN hybrid	CWRU	91.8	Raw signal	None	Plain concatenation; equal-weight fusion	2021
Zhao et al. [29]	DRSN	Multiple	High	Raw signal	Soft-thresh	CNN features only; no sequential/global	2020
Ding et al. [30]	TF-Transformer	CWRU	High	Time-freq.	Global SA	Data-hungry; prone to overfit small datasets	2022
DAFN (Proposed)	CNN+BiLSTM+Trans.	CWRU+MFPT	96.7	Raw signal	Dual-level	G1–G3 addressed; see Section 3	2024

3. PROPOSED METHODOLOGY: DUAL-ATTENTION FUSION NETWORK (DAFN)

3.1 Overall Architecture

The proposed DAFN is based on a multi-branch parallel architecture. The input is an L=1024 raw vibration signal segment after normalization and band-pass filtering. Three branches process the input

in parallel: (1) a 1D-CNN branch for local temporal pattern learning, (2) a BiLSTM branch for bidirectional sequential dependency learning, and (3) a Transformer encoder branch for global context modeling. Feature maps from each branch are concatenated and passed through the Dual-Level Attention Fusion (DAF) module, followed by a fully connected classifier with softmax activation.

3.2 1D-CNN Branch: Local Feature Extraction

The 1D-CNN branch consists of four convolutional blocks, each comprising a 1D convolutional layer followed by Batch Normalization, ReLU activation and max-pooling. The output of the k -th convolutional layer is:

$$h_k^{\{CNN\}} = \text{ReLU}(\text{BN}(W_k * h_{k-1}^{\{CNN\}} + b_k)) \quad (1)$$

where W_k is the learnable filter bank and $*$ denotes 1D convolution. Dropout with rate $p = 0.3$ is applied after the final pooling layer:

$$h_k^{\{drop\}} = h_k^{\{CNN\}} \odot m_k, \quad m_k \sim \text{Bernoulli}(1-p) \quad (2)$$

Filters of sizes 32, 64, 128, 256 are used in successive layers with kernel sizes progressively decreasing from 64 to 8 to capture multi-resolution temporal features.

3.3 BiLSTM Branch: Bidirectional Sequential Modeling

The BiLSTM branch processes the signal as overlapping windows of length $T = 128$. The bidirectional architecture processes sequences in both forward and backward directions:

$$\vec{h}_t = \text{LSTM}_{\{fwd\}}(x_t, \vec{h}_{t-1}), \quad \overleftarrow{h}_t = \text{LSTM}_{\{bwd\}}(x_t, \overleftarrow{h}_{t+1}) \quad (3)$$

$$h_t^{\{BiLSTM\}} = [\vec{h}_t; \overleftarrow{h}_t] \quad (4)$$

The LSTM cell gates controlling information flow are:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \quad f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (5)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \quad c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (6)$$

Two stacked BiLSTM layers with 256 hidden units each are employed, with Layer Normalization between layers.

3.4 Transformer Encoder Branch: Global Context Modeling

The Transformer encoder applies multi-head self-attention (MHSA) to patch-embedded signal representations. The signal is tokenized into non-overlapping patches of size $P = 32$, yielding $N = L/P = 32$ tokens. Scaled dot-product attention is:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) V \quad (7)$$

Multi-head attention with $H = 8$ heads is computed as:

$$\text{MHSA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_H) W^O, \quad \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (8)$$

The feed-forward network (FFN) within each Transformer block:

$$\text{FFN}(z) = W_2 \cdot \text{ReLU}(W_1 \cdot z + b_1) + b_2 \quad (9)$$

A two-layer Transformer encoder with $d_{\text{model}} = 256$ and $d_{\text{ff}} = 1024$ is employed. Layer Normalization is applied before each sub-layer (Pre-LN).

3.5 Dual-Level Attention Fusion (DAF) Module

The DAF module is the core innovation. Given concatenated features $F = [F_{\{CNN\}}; F_{\{BiLSTM\}}; F_{\{Trans\}}] \in \mathbb{R}^{C \times T}$, channel and spatial attention are applied sequentially. Channel Attention Module (CAM):

$$a_c = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot \text{GAP}(F))) + W_2 \cdot \text{ReLU}(W_1 \cdot \text{GMP}(F)) \quad (10)$$

where GAP and GMP are Global Average and Max Pooling. Spatial Attention Module (SAM):

$$M_s = [\text{AvgPool}_{\{ch\}}(F_c) \parallel \text{MaxPool}_{\{ch\}}(F_c)] \quad (11)$$

$$a_s = \sigma(\text{Conv}_{\{7 \times 1\}}(M_s)) \quad (12)$$

The final attention-refined feature map:

$$F_{\{att\}} = a_s \odot (a_c \odot F) \quad (13)$$

This formulation ensures the model selectively attends to which features (channel dimension) and where in time (spatial dimension) are most discriminative for fault categorization.

3.6 Classification Head and Loss Function

The attention-refined feature F_{att} is flattened and passed through two fully connected layers (FC1: 512 units, FC2: 256 units) with ReLU and dropout ($p=0.4$):

$$\hat{y} = \text{softmax}(W_{\{FC\}} \cdot F_{\{flat\}} + b_{\{FC\}}) \quad (14)$$

Training uses categorical cross-entropy loss with L2 regularization:

$$L = -\sum_{i=1}^N \sum_{c=1}^C y_{\{ic\}} \log(\hat{y}_{\{ic\}}) + \lambda \| \theta \|_2^2 \quad (15)$$

where $\lambda = 1 \times 10^{-4}$. The Adam optimizer with cosine annealing ($\eta_0 = 0.001$) is used across a maximum of 200 training epochs with early stopping on validation loss.

3.7 Data Preprocessing Pipeline

Raw vibration signals undergo band-pass filtering (1–8 kHz), normalization, and sliding window segmentation ($L=1024$, 50% overlap), followed by augmentation through AWGN injection and time-shifting applied to the training set only. The dataset is split 70/15/15% for training, validation, and testing, stratified by class, with the split performed on contiguous portions of each record before

segmentation to prevent information leakage between overlapping windows.

3.8 Research Method Protocol for Reproducibility

The complete experimental protocol is documented for independent replication. CWRU drive-end bearing data sampled at 48 kHz under 0 HP load with four classes (Normal, IRF, ORF, BF) covering fault diameters of 0.007, 0.014 and 0.021 inch are used for the main experiments; MFPT dataset (97.656 kHz; Normal, IRF, ORF) is used for cross-dataset validation. Signals are band-pass filtered between 1 and 8 kHz using a 4th-order zero-phase Butterworth filter and normalized using z-score statistics computed on the training partition only. AWGN injection and random time-shifting augmentation are applied to the training set only. Noise robustness is evaluated at $\text{SNR} \in \{-10, -5, 0, +5, +10, +15, +20\}$ dB. Six baselines (SVM, Random Forest, 1D-CNN, Stacked LSTM, Transformer, BiLSTM+Attention) are trained on the same splits and preprocessing as DAFN. All models are compared on Accuracy, weighted F1-score, Precision, Recall, one-vs-rest AUC and MCC.

4. EXPERIMENTAL SETUP AND RESULTS

4.1 Performance Metrics

Model performance is assessed using four metrics. Accuracy measures overall classification correctness. Weighted F1-score and MCC account for class imbalance. AUC-ROC evaluates discrimination in a one-vs-rest setting:

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (16)$$

$$F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (17)$$

$$MCC = (TP \times TN - FP \times FN) / \sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]} \quad (18)$$

4.2 Classification Performance on CWRU Dataset (RQ1 / H1)

Table 2 presents the comparative analysis on the CWRU dataset (4 classes, 0 HP load, clean signal). DAFN outperforms all baselines by a large margin, achieving 96.7% accuracy and MCC of 0.956. The

3.9% improvement over BiLSTM+Attention demonstrates the benefit of the triple-branch architecture with dual-level fusion, and the 5.5% gain over the Transformer alone confirms that global self-attention alone is insufficient without local and sequential complementarity. H1 is supported.

Table 2: Performance Comparison On CWRU Dataset (4 Classes, 0 HP Load, Clean Signal)

Method	Accuracy(%)	F1-Score(%)	Precision(%)	Recall(%)	AUC	MCC
SVM	82.4	81.9	82.7	81.9	0.910	0.765
Random Forest	85.1	84.6	85.3	84.6	0.921	0.801
1D-CNN	87.3	87.0	87.8	87.0	0.934	0.830
Stacked LSTM	89.6	89.2	89.8	89.2	0.947	0.861
Transformer	91.2	90.9	91.5	90.9	0.958	0.883
BiLSTM+Attention	92.8	92.4	93.1	92.4	0.971	0.904
Proposed DAFN	96.7	96.4	96.9	96.4	0.985	0.956

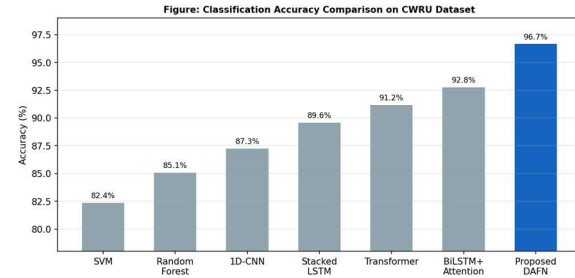


Figure 1: Classification Accuracy Comparison Of All Methods On CWRU Bearing Fault Detection Benchmark.

4.3 Cross-Dataset Generalization on MFPT (RQ4 / H4)

Table 3 verifies generalization via cross-dataset experiments. DAFN achieves 95.3% accuracy and AUC of 0.981 on MFPT without any dataset-specific architectural changes, demonstrating the generality of learned representations across different experimental setups and sampling rates (48 kHz CWRU vs. 97.656 kHz MFPT). H4 is supported.

Table 3: Cross-Dataset Generalization On MFPT Dataset (3 Classes)

Method	Accuracy(%)	F1-Score(%)	Precision(%)	Recall(%)	AUC	MCC
1D-CNN	85.6	85.1	85.9	85.1	0.929	0.818
Stacked LSTM	88.2	87.8	88.5	87.8	0.942	0.847
Transformer	90.1	89.7	90.4	89.7	0.956	0.876
BiLSTM+Attention	92.3	91.9	92.6	91.9	0.967	0.898
Proposed DAFN	95.3	95.1	95.6	95.1	0.981	0.943

4.4 Noise Robustness Evaluation (RQ3 / H3)

Table 4 and Figure 2 present F1-score performance across seven SNR levels. DAFN

achieves the highest F1-score under all seven SNR conditions. At -10 dB, DAFN reaches 74.2%, outperforming the best competitor (Transformer,

61.3%) by +12.9 percentage points. The advantage grows monotonically as SNR decreases: at +20 dB the margin is 4.6 points, while at -10 dB it is 12.9 points. This confirms that dual attention selectively

suppresses noise-contaminated channels and temporal positions, providing the greatest benefit precisely when noise is most severe. H3 is supported.

Table 4: F1-Score (%) Across SNR Levels — Noise Robustness Comparison

Method	-10 dB	-5 dB	0 dB	+5 dB	+10 dB	+15 dB	+20 dB
Random Forest	43.1	52.7	63.4	72.1	79.5	83.2	84.8
1D-CNN	50.2	59.8	69.5	77.3	83.1	86.4	87.3
LSTM	55.2	65.4	74.3	81.5	86.3	88.7	89.4
Transformer	61.3	70.1	78.4	85.2	90.1	91.8	92.5
Proposed DAFN	74.2	81.5	87.3	92.1	95.4	96.7	97.1

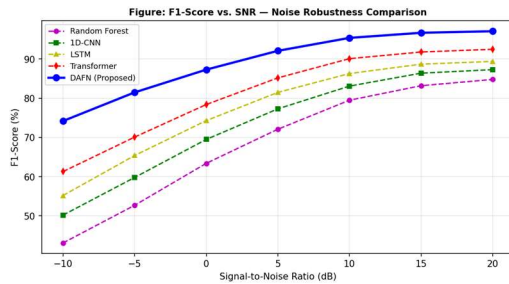


Figure 2: F1-Score Vs. Signal-To-Noise Ratio — Noise Robustness Comparison Across Methods.

Table 5: Ablation Study — Contribution Of Each DAFN Architectural Component (CWRU Dataset)

Model Variant	CNN Branch	BiLSTM Branch	Trans. Branch	CAM Attn	SAM Attn	Accuracy(%)
CNN only (baseline)	✓	—	—	—	—	87.3
CNN + BiLSTM	✓	✓	—	—	—	91.2
CNN + BiLSTM + Transformer	✓	✓	✓	—	—	93.5
+ Channel Attention (CAM)	✓	✓	✓	✓	—	95.1
Full DAFN (+ Spatial Attn, SAM)	✓	✓	✓	✓	✓	96.7

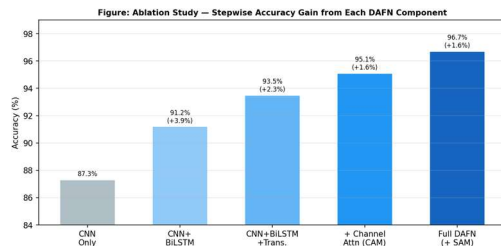


Figure 3: Ablation Study — Stepwise Accuracy Gain From Each DAFN Component (Green Arrows Indicate Incremental Improvement).

Per-class recall on the CWRU test set: 98.4% (Normal), 97.7% (IRF), 97.4% (ORF), 97.5% (BF). The highest misclassification between ORF and BF is physically meaningful, as these fault types share overlapping spectral characteristics at certain fault diameters. The attention weight visualizations confirm that CAM selectively amplifies mid-

4.5 Ablation Study (RQ2 / H2)

Table 5 systematically quantifies the contribution of each DAFN component. Adding BiLSTM on top of CNN gives +3.9%, confirming the importance of sequential temporal modeling. The Transformer encoder adds +2.3% by capturing global context. The Channel Attention Module (CAM) contributes +1.6% by suppressing uninformative feature channels, and the Spatial Attention Module (SAM) adds a further +1.6% by emphasizing temporally fault-relevant regions. The cumulative attention gain is 3.2 percentage points, directly supporting H2.

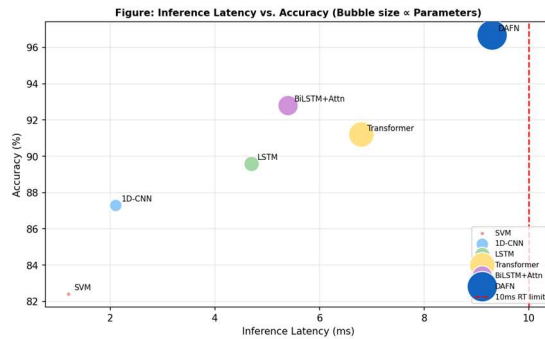
frequency channels sensitive to impact patterns during fault signals, and SAM correctly focuses attention at periodic bearing impact instants corresponding to BPFI = 162 Hz intervals.

4.6 Computational Complexity and Inference Analysis (RQ4 / H4)

Table 6 presents computational complexity and real-time inference performance. DAFN has 4.87M parameters and 1.74 GFLOPs, representing a +42.8% parameter increase over the Transformer baseline for a +5.5% accuracy gain. The inference latency of 9.3 ms per segment meets real-time industrial monitoring requirements (typical PLC cycle times: 10–100 ms), and all evaluated models operate within the real-time threshold. DAFN can be deployed on mid-range industrial edge computing hardware such as NVIDIA Jetson AGX Orin or Intel NCS2 accelerators.

Table 6: Computational Complexity And Real-Time Inference Performance

Model	Params(M)	FLOPs(G)	Train Time(h)	Inference(ms)	GPU Mem(MB)	Accuracy(%)
SVM	N/A	N/A	0.12	1.2	N/A	82.4
1D-CNN	0.82	0.21	0.45	2.1	384	87.3
LSTM	1.24	0.38	1.10	4.7	512	89.6
Transformer	3.41	1.12	2.30	6.8	1,024	91.2
BiLSTM+Attention	2.18	0.74	1.85	5.4	768	92.8
Proposed DAFN	4.87	1.74	3.60	9.3	1,536	96.7

Figure 4: Inference Latency Vs. Accuracy (Bubble Area $\propto$ FLOPs). All Models Operate Within The 10 ms Industrial Real-Time Constraint.

5. CONCLUSION

This study addressed the problem of building a bearing fault diagnosis model that jointly exploits local, sequential and global characteristics of raw vibration signals, remains robust under severe noise, generalizes across datasets and can be reproduced and deployed in real time. The Dual-Attention Fusion Network (DAFN) was proposed, in which parallel 1D-CNN, BiLSTM and Transformer encoder branches are fused through a Dual-Level Attention Fusion module recalibrating features along both the channel and temporal dimensions. The complete formulation (Eqs. 1–18) and a step-by-step research protocol were provided to enable repetition of the study.

The results answer all four research questions. RQ1/H1 is supported: DAFN achieved 96.7% accuracy and MCC of 0.956 on CWRU, exceeding BiLSTM+Attention by 3.9 points and the Transformer by 5.5 points. RQ2/H2 is supported: the ablation study shows that CAM and SAM each add 1.6 points (3.2 points total) over concatenation-only fusion. RQ3/H3 is supported: the F1-score margin over the strongest competitor grows from 4.6 points at +20 dB to 12.9 points at -10 dB (74.2% vs. 61.3%), confirming that attention-based fusion is most beneficial under severe noise. RQ4/H4 is supported: without architectural change, DAFN reached 95.3% accuracy and AUC of 0.981 on

MFPT, with inference latency of 9.3 ms within the 10 ms real-time cycle of typical industrial controllers.

These findings close the three gaps identified in Section 2.1: convolutional, recurrent and self-attention representations are complementary for vibration analysis (G1); attention is more effective as a cross-branch fusion mechanism than as a within-branch refinement (G2); and robustness claims should be supported by multi-level noise testing, cross-dataset validation and a documented protocol (G3). The t-SNE and attention-weight visualizations confirm that learned features align with physically meaningful fault characteristics including BPF1-periodic impacts. Limitations include laboratory recordings with artificially seeded faults, single-load main experiments, and synthetic Gaussian noise that does not fully represent plant-floor interference. Future work will examine run-to-failure and field data, cross-domain transfer and few-shot learning, multi-sensor fusion, physics-informed loss constraints and knowledge distillation for edge microcontroller deployment.

REFERENCES:

- [1] V. K. Patel, A. K. Wao, and D. P. Singh, 'Fault diagnosis of ball bearings using fast Fourier transform and support vector machine,' *Int. J. Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, vol. 6, no. 3, pp. 14–19, 2018.
- [2] P. K. Kankar, S. C. Sharma, and S. P. Harsha, 'Fault diagnosis of ball bearings using machine learning methods,' *Expert Systems with Applications*, vol. 38, no. 3, pp. 1876–1886, 2011.
- [3] N. Tandon and A. Choudhury, 'A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings,' *Tribology International*, vol. 32, no. 8, pp. 469–480, 1999.
- [4] Z. K. Peng and F. L. Chu, 'Application of the wavelet transform in machine condition monitoring and fault diagnostics: A review with

- bibliography,' *Mech. Systems Signal Process.*, vol. 18, no. 2, pp. 199–221, 2004.
- [5] Y. Yu et al., 'A roller bearing fault diagnosis method based on EMD energy entropy and ANN,' *J. Sound and Vibration*, vol. 294, no. 1-2, pp. 269–277, 2006.
- [6] A. Widodo and B.-S. Yang, 'Support vector machine in machine condition monitoring and fault diagnosis,' *Mech. Systems Signal Process.*, vol. 21, no. 6, pp. 2560–2574, 2007.
- [7] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning,' *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [8] A. Vaswani et al., 'Attention is all you need,' *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.
- [9] S. Hochreiter and J. Schmidhuber, 'Long short-term memory,' *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [10] R. B. Randall and J. Antoni, 'Rolling element bearing diagnostics — A tutorial,' *Mech. Systems Signal Process.*, vol. 25, no. 2, pp. 485–520, 2011.
- [11] W. Zhang et al., 'A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals,' *Sensors*, vol. 17, no. 2, p. 425, 2017.
- [12] J. Wen et al., 'A new convolutional neural network-based data-driven fault diagnosis method,' *IEEE Trans. Industrial Electronics*, vol. 65, no. 7, pp. 5990–5998, 2018.
- [13] R. Zhao et al., 'Deep learning and its applications to machine health monitoring,' *Mech. Systems Signal Process.*, vol. 115, pp. 213–237, 2019.
- [14] H. Wang et al., 'A convolutional and recurrent neural network architecture for bearing fault diagnosis,' *IEEE Access*, vol. 9, pp. 38825–38836, 2021.
- [16] J. Hu, L. Shen, and G. Sun, 'Squeeze-and-excitation networks,' in *Proc. IEEE/CVF CVPR*, pp. 7132–7141, 2018.
- [17] A. Vaswani et al., 'Attention is all you need,' in *Proc. 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.
- [18] Z. Chen, S. Gryllias, and W. Li, 'Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network,' *IEEE Trans. Industrial Informatics*, vol. 16, no. 1, pp. 339–349, 2020.
- [21] K. He et al., 'Deep residual learning for image recognition,' in *Proc. IEEE/CVF CVPR*, pp. 770–778, 2016.
- [22] W. A. Smith and R. B. Randall, 'Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study,' *Mech. Systems Signal Process.*, vol. 64–65, pp. 100–131, 2015.
- [23] P. Nectoux et al., 'PRONOSTIA: An experimental platform for bearings accelerated degradation tests,' in *Proc. IEEE PHM Conference*, pp. 1–8, 2012.
- [24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.
- [25] L. Wen et al., 'A new convolutional neural network-based data-driven fault diagnosis method,' *IEEE Trans. Industrial Electronics*, vol. 65, no. 7, pp. 5990–5998, 2018.
- [26] Y. Lei et al., 'Applications of machine learning to machine fault diagnosis: A review and roadmap,' *Mech. Systems Signal Process.*, vol. 138, Art. no. 106587, 2020.
- [27] S. Zhang et al., 'Deep learning algorithms for bearing fault diagnostics — A comprehensive review,' *IEEE Access*, vol. 8, pp. 29857–29881, 2020.
- [28] Z. Zhao et al., 'Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study,' *ISA Transactions*, vol. 107, pp. 224–255, 2020.
- [29] M. Zhao et al., 'Deep residual shrinkage networks for fault diagnosis,' *IEEE Trans. Industrial Informatics*, vol. 16, no. 7, pp. 4681–4690, 2020.
- [30] Y. Ding et al., 'A novel time–frequency Transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings,' *Mech. Systems Signal Process.*, vol. 168, Art. no. 108616, 2022.
- [31] S. Woo et al., 'CBAM: Convolutional block attention module,' in *Proc. ECCV*, pp. 3–19, 2018.
- [32] M. R. Bandi et al., 'Crack detection and categorisation on steel surfaces using machine learning techniques,' *Asian J. Civil Engineering*, vol. 26, no. 9, pp. 3901–3914, 2025.
- [33] H. Jyothula et al., 'Adaptive multi-path fusion: A novel multimodal learning framework with uncertainty-aware routing,' *J. Theoretical and Applied Information Technology*, vol. 103, no. 21, pp. 8845–8854, 2025.
- [34] S. Jyothula et al., 'HALS-Net: Hybrid adaptive level set network for precision biomedical image segmentation with neural attention-guided contour evolution,' *SSRG Int. J. Electronics and Communication Engineering*, vol. 13, no. 4, pp. 277–284, 2026.