Scientific Reports
https://doi.org/10.1038/s41598-026-62410-9

Article in Press

Dual-branch fusion framework with graph attention networks for compound fault diagnosis in complex machinery system

Received: 27 April 2026
Accepted: 10 July 2026
Published online: 16 July 2026

Cite this article as: Mubarak A., Gebremariam M., Isa H. *et al.* Dual-branch fusion framework with graph attention networks for compound fault diagnosis in complex machinery system. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-62410-9

Akram Mubarak, Mebrahitom Gebremariam, Hilmi Isa & Azmir Azhari

Dual-Branch Fusion Framework with Graph Attention Networks for Compound Fault Diagnosis in Complex Machinery System

Akram Mubarak[1,*], Mebrahitom Gebremariam[2], Hilmi Isa[3], Azmir Azhari[1,*]

[1]Faculty of Manufacturing and Mechatronic Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Malaysia.
[2]College of Engineering and Built Environment, Birmingham City University, B47XG Birmingham, United Kingdom.
[3]Hi Technics System Sdn Bhd, 63000 Cyberjaya, Selangor, Malaysia
*Corresponding author: eng.ao.mubarak@gmail.com, azmir@umpsa.edu.my

Abstract

The present research proposes a generalized diagnostic framework for complex machinery that integrates domain knowledge with learned signal representations while exploiting the physical coupling between subsystems. By modelling multi-component systems as heterogeneous graphs, the framework employs a dual-branch architecture: Branch A extracts handcrafted features grounded in failure mechanics, while Branch B learns complementary latent embeddings via an unsupervised pre-trained 1-D convolutional autoencoder. To prevent geometric imbalance and ensure both branches contribute equally to diagnostic performance, Normalized Principal Component Analysis (nPCA) equalizes the feature spaces before a two-layer Graph Attention Network (GAT) performs topology-aware spatial fusion. Final health states are classified using a Platt-calibrated Support Vector Machine (SVM) with a One-Class extension, producing probabilistic labels alongside a risk index and a normalized entropy uncertainty index for robust open-set rejection. Validated on the PHM-Beijing 2024 Final Stage subway bogie drivetrain benchmark, the pipeline achieves 95.1%, 87.1%, and 82.4% accuracy across single-component, component-level compound, and system-level compound fault tiers, respectively. Crucially, the topology-aware GAT spatial fusion provides a unique +12.4 percentage point accuracy margin exclusively on system-level compound faults, successfully decoupling highly non-additive joint fault signatures where traditional independent ensemble methods collapse. Results confirm that the learned GAT attention weights align with physical coupling pathways such as the high-weight Motor-Gearbox connection and the uncertainty quantification effectively identifies 181 fault-affected samples from 252 unlabelled test cases under deliberate domain shift. Furthermore, cross-dataset verification conducted on the public Case Western Reserve University (CWRU) and University of Ottawa benchmarks demonstrates a superior 96.39% mean precision under cross-load transfer alongside highly stable uncertainty indicators under severe non-stationary speed dynamics. Operating under a severe dataset shift, the Platt-calibrated safety gate outputs a 86.5% low-confidence (LOW_CONF) triage flag distribution. This mechanism explicitly converts epistemic uncertainty into reliable human-in-the-loop work orders rather than issuing overconfident incorrect predictions, demonstrating significant practical engineering value for safety-critical industrial assets. This framework provides a monotonic severity profile and a reliable rejection mechanism for out-of-distribution conditions, serving as a blueprint for deployable diagnostic systems in multi-component rotating machinery.

Key Words: *Condition-based maintenance (CBM), Compound fault diagnosis, Graph Attention Network (GAT), Multi-component rotating machinery, Dual-branch feature fusion, Uncertainty Quantification (UQ).*

1. Introduction

Condition-based maintenance (CBM) of rotating machinery relies on continuous health monitoring to pre-empt catastrophic failures, reduce unplanned downtime, and extend asset service life [1]. Within CBM, the predictive maintenance (PdM) provides the technical substrate for data-driven decision support [2]. In multi-component systems such as drivetrains, turbomachinery trains, and industrial gearbox assemblies, however, the co-occurrence of faults across physically coupled subsystems creates diagnostic scenarios that exceed the capability of any single-component approach. A compound fault involving simultaneous gear-tooth damage and bearing degradation cannot be diagnosed by examining each sensor group in isolation; the evidence is distributed across the coupling topology, and classification requires integrating signals that propagate through shared mechanical pathways. The integration of deep learning with physical system modelling is a cornerstone of modern Digital Twin-driven prognostics, which seeks to provide a comprehensive real-time reflection of industrial asset health [3].

While the integration of deep learning with physical system modelling represents a cornerstone of modern Digital Twin-driven prognostics, three interrelated structural gaps impede the deployment of practical, reliable diagnostic systems for complex multi-component machinery systems. Firstly, conventional single-component diagnostic approaches fundamentally break down when confronting compound faults. In multi-component rotating machinery, the simultaneous co-occurrence of distinct degradation modes (such as concurrent gear-tooth micro-pitting and bearing inner-race spalling) produces highly complex vibration signatures [4]. These superimposed signals exhibit severe frequency-modulation, phase interactions, and destructive interference, which mask the individual fault characteristics [5]. Traditional data-driven networks trained on isolated datasets lack the specialized physical inductive biases required to isolate individual failure modes under limited-label training regimes [6].

Second, existing multi-sensor feature fusion schemes lack topology awareness. Current literature regularly collapses multi-sensor data arrays into flat vectors or unweighted spatial matrices, implicitly treating physically independent subsystems as statistically independent or uniformly coupled [7,8]. However, mechanical excitations propagate along highly non-uniform transmission pathways dictated by physical shafts, gear meshes, and structural boundaries of components. By ignoring this physical coupling topology, standard frameworks cannot capture the cross-component correlation structure needed to distinguish localized component-level compound faults from widespread system-level interactions. Consequently, traditional frameworks frequently experience severe diagnostic performance degradation as fault complexity multiplies across components [9]. Finally, operational diagnostic models suffer from overconfident misclassifications under operational domain shift. Real-world machinery monitoring platforms are consistently subject to varying rotational velocities and fluctuating load torques that differ significantly from baseline laboratory environments [10]. Purely deep-learning-based diagnostics adapt poorly to these distribution shifts, projecting out-of-distribution (OOD) test cases onto closed-set target labels with dangerously high, uncalibrated confidence scores [11,12]. Without a built-in safety gate that quantifies epistemic uncertainty and establishes an open-set rejection mechanism, these systems cannot flag novel or unobserved conditions, severely limiting their real-world deployability in safety-critical industrial settings [13]. There is a need to resolve these three critical deficiencies, thus this paper proposes a generalized, topology-aware dual-branch framework that bridges mechanical coupling graph structures with calibrated multi-view representations.

This paper proposes a generalized framework that resolves the above mentioned three gaps. The framework casts any multi-component machinery system as a graph whose nodes are sensor groups and whose edges encode the physical coupling pathways between subsystems. It is instantiated and evaluated on the PHM-Beijing 2024 Final Stage benchmark [14], a 1:2-scale subway bogie drivetrain which provides a rigorous and representative use case for the general design.

The principal contributions are as follows. (i) A generalized problem formulation that positions compound fault diagnosis as a joint label prediction problem over a heterogeneous coupling graph, applicable to any multi-component machinery system whose subsystem failure mechanics and coupling topology are known. (ii) A dual-branch feature fusion architecture in which a handcrafted feature and a learned embeddings capture complementary variance structures. (iii) Topology-aware GAT spatial fusion whose learned attention weights align with the physical coupling topology and explain the accuracy gains on compound fault tiers. (iv) A blinded test-set inference study under working-condition domain shift providing per-sample risk indices, calibrated confidence scores, and flag categories for maintenance prioritisation.

1.1 Problem Formulation

Let M denote a multi-component rotating machinery system comprising K physically coupled subsystems $\{C_1, \ldots, C_k\}$. Each subsystem C_k is instrumented by a sensor group S_k collecting n_k channels of data at sampling rate f_s. The physical dependencies between subsystems are encoded in a coupling graph $G = (V, E)$ where $V = \{C_k\}$ and $E \subseteq V \times V$ captures shaft couplings, common load paths, gear meshes, and electromagnetic interactions.

Each subsystem C_k can occupy a finite set of health states $H_k = \{h_k^0, h_k^1, \ldots, h_k^{m_k}\}$, where h_k^0 denotes the normal operating condition. The joint system health state is the Cartesian product $h \in \prod_k H_k$. We define three fault tier categories: (T1) single-component states, in which exactly one subsystem departs from normal; (T2) component-level compound states, in which multiple health anomalies co-occur within the same subsystem; and (T3) system-level compound states, in which anomalies co-occur across two or more distinct subsystems.

Given raw sensor observations $\{x_k(t)\}$ from each group S_k, the diagnostic task decomposes into three sub-problems. (P1) Component-specific feature extraction: for each S_k, extract a feature vector ϕ_k that exploits the known failure mechanics of C_k. (P2) Topology-aware representation fusion: aggregate $\{\phi_k\}$ respecting the structure of G to produce a system-level embedding z. (P3) Calibrated compound-state classification: map z to the joint health state h with calibrated posterior probabilities and an open-set rejection mechanism for states unobserved during training.

This formulation is general: it subsumes any multi-component rotating machinery system for which the coupling topology G and the failure mechanics of each subsystem C_k are at least partially known. In the subway bogie drivetrain instantiation presented in Section V, $K = 4$, $V = \{$Motor, Gearbox, L-Axle, R-Axle$\}$, and E comprises five edges encoding shaft coupling, output axle connections, and wheelset coupling. The same framework applies without architectural modification to any system of coupled rotating components provided that the coupling graph G and the failure-mode taxonomy $\{H_k\}$ are specified.

1.2 Related Work

The PdM literature frames a prognostic programme as four sequential processes: data acquisition, Health Indicators (HI) construction, health-stage division, and RUL prediction [15]. Early systems treated these as loosely coupled modules executed by domain experts [1,2]; contemporary platforms seek tighter integration [16–19]. A persistent gap is the multi-component dependency problem: when a defect in one subsystem accelerates degradation in a coupled subsystem, per-component diagnostic modules cannot detect the interaction. The present framework directly addresses this by embedding the coupling graph *G* as an inductive bias in the GAT fusion stage.

Multi-sensor feature fusion for rotating machinery has been taxonomized by Kibrete et al. [7] into data-level, feature-level, and decision-level strategies. Feature-level fusion preserves source discriminability while enabling cross-sensor reasoning but introduces a dimensional imbalance when feature counts per sensor group differ; a problem Schwartz et al. [20] addressed with variance-normalized projections and multi-model ensemble strategies.

Physics-informed feature design for rotating machinery has a rich history. Jia et al.[5] demonstrated that time-domain statistics including kurtosis, crest factor, RMS, and impulse factor can provide superior fault sensitivity under limited labels compared to end-to-end deep networks. Envelope analysis at bearing characteristic frequencies is the canonical indicator for localized rolling-element defects [5]. Motor Current Signature Analysis (MCSA) detects electrical asymmetry in induction motors via sideband pairs at $(1 \pm 2ks)f_s$ [5]. Wang et al.[21] showed that combining handcrafted physical features with learned representations in an ensemble outperforms either source alone on compound fault classification, motivating the current dual-branch architecture. Recent advancements have further consolidated the value of hybrid fusion paradigms. For instance, Yang et al.[22] introduced a multi-source physical information fusion network (MSPI-Net) that relies on optimizer-driven parameter alignment to systematically bridge structural physics constraints with data-driven interpretability pipelines in bearing monitoring. This further underscores the necessity of designing symmetrical view-balancing mechanisms, such as the nPCA pipeline presented herein, to successfully harness the distinct information geometries of handcrafted and network-learned feature representations. The 1-D CNN literature surveyed by Kiranyaz et al.[23] established that three-stage convolutional encoders achieve competitive vibration embeddings, while Jiang et al.[24] demonstrated that pre-trained deep encoders generalize better to compound fault conditions than end-to-end supervised networks.

Graph neural networks (GNNs) provide a natural framework for multi-component diagnosis because they encode coupling topology as an inductive bias rather than learning it from raw signals. Li et al. [25] benchmarked multiple GNN architectures and found that Graph Attention Networks consistently outperform spectral GCNs on node-level tasks where local neighbourhood structure varies between fault classes. Ding et al. [8] demonstrated that encoding the bogie topology as a graph enables discrimination between system-level compound faults and component-level ones that flat classifiers collapse. Compound fault decoupling has also been addressed through label-correlation graph modules [26], multi-scale signal decomposition[27], multi-head channel attention [28], and variational mode decomposition combined with CNN encoders [29].

Uncertainty Quantification (UQ) and open-set detection are increasingly recognized as prerequisites for deployable PdM systems. Kuleshov et al. [30] established Platt calibration for SVMs; Ovadia et al.[13] showed that calibration quality degrades systematically under dataset shift. Liang et al. [10] proposed ODIN to sharpen in-distribution vs. out-of-distribution score separation via temperature scaling and input perturbation. Hendrycks et al. [12] introduced Outlier Exposure as a training-time complement, and Hendrycks and Gimpel [12] established the maximum-softmax-probability baseline. None of the compound fault methods cited above provides both a calibrated uncertainty index and an open-set rejection mechanism; the proposed framework does.

2. Proposed Framework

2.1 Architecture Overview

The framework in Figure 1 processes each sensor raw signal observation window through five subsequent stages. Stage 1 applies Z-score normalization per channel and rejects windows in which any channel's peak-to-peak amplitude exceeds 5x the population standard deviations. Stage 2 extracts dual-branch features independently per sensor group: Branch A produces a handcrafted feature vector ϕ_k, and Branch B produces a learned embedding ζ_k from a pre-trained 1-D convolutional autoencoder. The dual-branch architecture, specifically the use of handcrafted physical features and unsupervised pre-training, serves as a mechanism for domain adaptation, ensuring reliable performance even when training and testing distributions diverge due to variable working conditions [31]. Stage 3 applies nPCA independently to each branch, then L_2-normalizes the projected coordinates so that each branch contributes unit geometric weight to the subsequent graph fusion. Stage 4 assembles the equalized node features into the coupling graph G and applies a Graph Attention Network (GAT) to produce a system-level embedding z. Stage 5 classifies z with a calibrated multi-class SVM augmented by a One-Class SVM open-set gate, producing a discrete health-state label, a risk index, and a normalized entropy uncertainty index.

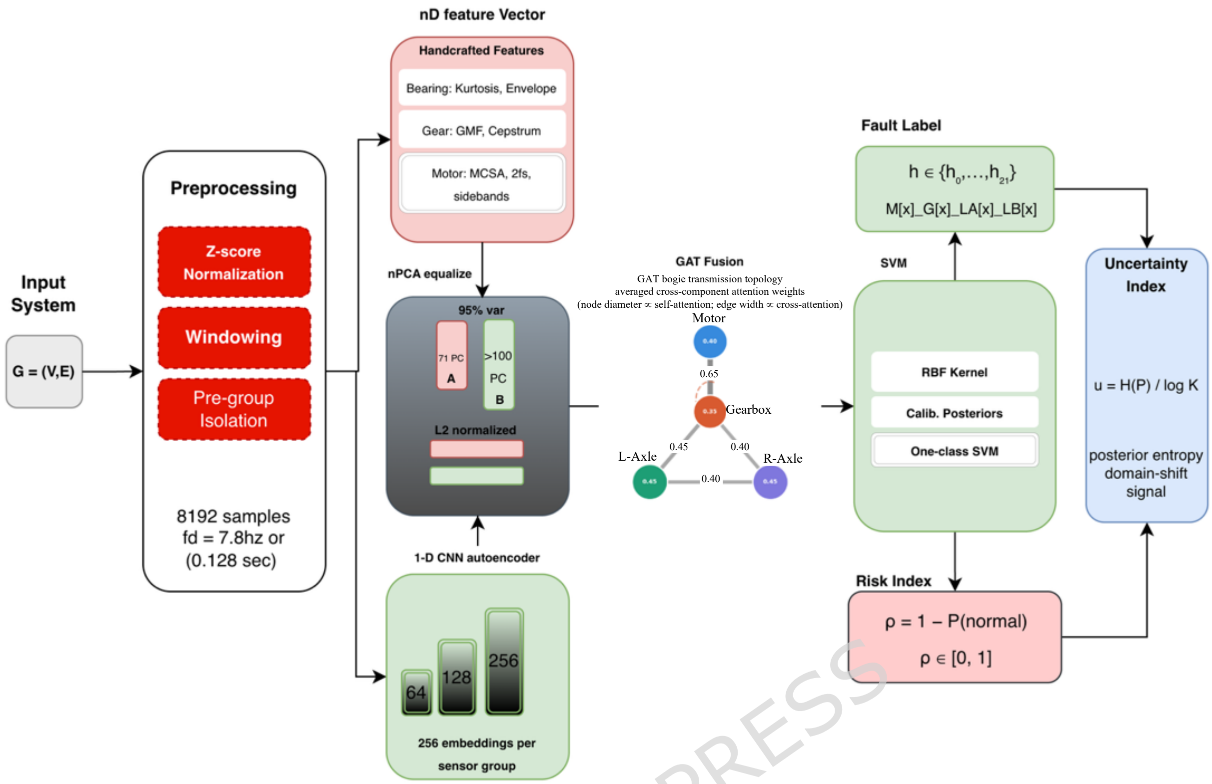

Fig. 1. The proposed dual-branch fusion framework with graph attention networks

2.2 Component-Specific Handcrafted Features

For each sensor group S_k, Branch A extracts a handcrafted features from time-domain statistics; frequency-domain indicators; and time-frequency decompositions as summarized in Table 1. For the induction motor, three simultaneous failure modes are targeted including bearing faults, broken rotor, and short circuit faults. Bearing faults are detected via envelope analysis at the four characteristic frequencies of the motor bearing geometry, preceded by spectral kurtosis to identify the optimal resonance demodulation band. Broken rotor faults are targeted through MCSA sideband pairs at $(1\pm2ks)f_s$ for $k = 1, 2, 3$, the twice-slip-frequency component $2sf_s$, and WPD energies in the 0–100 Hz range for working-condition-robust extraction. Inter-turn short-circuit faults are detected via the second supply-frequency harmonic $2f_s$, eccentricity sidebands at $|f_r \pm sf_s|$, the experimentally confirmed 21st line-frequency harmonic, spectral centroid migration, and WPT high-frequency node energies capturing partial-discharge transients. For the gearbox, gear-tooth faults are targeted through GMF harmonic amplitudes, shaft-frequency sideband families, and cepstrum quefrency analysis; while bearing faults are targeted through the same envelope analysis as motor bearings but with a higher-frequency threshold to suppress GMF contamination before demodulation. Both axle-boxes receive an identical rolling-element bearing feature set, providing the GAT with matched node descriptors that make asymmetric bilateral fault activation spatially distinguishable.

Table 1. Branch A: Handcrafted Features

Sensor group	Target fault mode	Features extracted
Motor	Bearing (M3)	Envelope spectrum at BPFO, BPFI, BSF, FTF†; Spectral Kurtosis; Kurtogram; kurtosis, crest factor, RMS, impulse factor
	Broken rotor bar (M2)	MCSA sideband pairs at $(1 \pm 2ks)f_s$, k = 1,2,3; twice-slip component $2sf_s$; WPD energies 0–100 Hz
	Short circuit (M1)	$2f_s$ harmonic; eccentricity sidebands $\lvert f_r \pm sf_s \rvert$; 21st line harmonic; spectral centroid; WPT high-band energies
Gearbox	Gear-tooth (G1–G4)	GMF harmonic amplitudes (1×, 2×, 3×); shaft-frequency sideband families; cepstrum quefrency peaks; residual-signal kurtosis and RMS
	Bearing (G5–G8)	Envelope spectrum at BPFO, BPFI, BSF, FTF† with high-pass pre-filter to suppress GMF; Spectral Kurtosis; Kurtogram
L-Axle box	Bearing (LA1–LA4)	Envelope spectrum at BPFO, BPFI, BSF, FTF†; kurtosis, crest factor, RMS, impulse factor; Spectral Kurtosis; Kurtogram; wavelet sub-band energies
R-Axle box	Bearing (RA1–RA4)	Identical feature set to L-Axle box; bilateral symmetry preserved for GAT node matching

2.3 Learned embeddings through 1D CNN

Branch B pre-trains a 1-D convolutional autoencoder independently per sensor group using unlabelled raw windows [23,32]. The encoder comprises three Conv1D layers (64 → 128 → 256 filters, kernel size 3, BatchNorm, ReLU, MaxPool stride 2), producing a 256-dimensional embedding per group. Unsupervised pre-training with MSE reconstruction loss ensures Branch B captures generic signal morphology without depending on fault labels, complementing to Branch A. The selection of a three-stage 1-D convolutional encoder and a streamlined two-layer GAT ensures that the framework remains suitable for real-time fault diagnosis within the hardware limitations typical of industrial edge-computing platforms [33]. Structural transparency and full reproducibility of this unsupervised branch are ensured by representing the normalized input window for sensor group S_k as $X_k \in \mathbb{R}^{L \times c}$, where the window length is denoted by L = 8192 and the channel count is given by c. The forward pass of a 1-D convolutional layer within the encoder is formalized mathematically as shown in Equation (1) [23].

$$y_j^l = \sigma\left(\sum_i y_i^{l-1} * w_{ij}^l + b_j^l\right) \text{ Equation (1)}$$

where * denotes the 1-D convolution operator, w_{ij}^l represents the weights of the j-th filter kernel mapping from channel i in layer l -1,b_j^l is the bias vector, and σ(.) denotes the Rectified Linear

Unit (ReLU) activation function preceded by Batch Normalization (BN). Down sampling across the three stages is enforced via non-overlapping Max-Pooling layers with a stride of 2.

The decoder reconstructs the signal through symmetric 1-D Transposed Convolutional layers (Conv1D-Transpose) combined with Up-Sampling operators. The entire autoencoder network is optimized by minimizing the mean Squared Error (MSE) reconstruction loss function as shown in Equation (2) [32].

$$\mathcal{L}_{MSE} = \frac{1}{L\times c}\sum_{t=1}^{L}\sum_{p=1}^{c}\left(x_{t,p} - \hat{x}_{t,p}\right)^2 \text{ Equation (2)}$$

where $x_{t,p}$ and $\hat{x}_{t,p}$ are raw and reconstructed data points, respectively.

2.4 Normalized PCA Equalization

Branch A and Branch B produce feature vectors of substantially different dimensionality and variance structure. Therefore, concatenating the two projections imposes a geometric imbalance that causes the GAT's linear weight matrices to allocate representational capacity in proportion to dimensional mass rather than diagnostic value. nPCA applies PCA independently to each branch, retaining components that explain 95% of cumulative variance, then L2-normalizes the projected coordinates so that each branch contributes unit geometric weight to the subsequent GAT node features [34]. As demonstrated in Figure 2, Branch A reaches the 95% threshold at PC 71 while Branch B remains far more diffuse, confirming that the equalization step is a structural necessity rather than a heuristic preference.

Mathematically, let the extracted features of Branch A and Branch B for sensors group S_k be represented by $\Phi_k \in \mathbb{R}^{d_A}$ and $\Gamma_k \in \mathbb{R}^{d_B}$ respectively. For each branch matrix independently, the empirical covariance matrix Σ is decomposed into its constituent eigenvectors V and eigenvalues Λ as shown in Equation (3) [1,34].

$$\Sigma = V\Lambda V^T \text{ Equation (3)}$$

Projected principal coordinates $Y = XV_r$ are computed by retaining the top r components satisfying a cumulative energy retention threshold $\eta \geq 95\%$ as shown in Equation (4) [1,34].

$$\sum_{i=1}^{r}\lambda_i / \sum_{j=1}^{d}\lambda_j \geq 0.95 \text{ Equation (4)}$$

Following this orthogonal projection, an element-wise L_2-normalization layer maps the coordinates into a variance-bounded hypersphere to equalize the branches across the feature dimensions as shown in Equation (5) [34].

$$\tilde{\boldsymbol{Y}} = \frac{\boldsymbol{Y}}{||\boldsymbol{Y}||_2} \text{ Equation (5)}$$

This equalization step enforces an objective function where the post-projection L_2-norm ratio $\boldsymbol{r_A}:\boldsymbol{r_B} \approx \mathbf{1}:\mathbf{1}$. This bounds the joint subspace energy budget and explicitly prevents the geometric imbalance that occurs under raw concatenation.

2.5 Topology-Aware GAT Fusion

The four nPCA-normalized node feature vectors are assembled into the coupling graph G and processed by a two-layer GAT with four attention heads and inter-layer dropout of 0.3 [25,35]. Global mean pooling over all node embeddings produces a fixed-length system-level representation z. The GAT attention mechanism learns to weight coupling edges differently across fault classes, see section VI-B.

The primary mathematical layer of the multi-head attention mechanism computes the attention coefficient α_{ij}, which measures the directional target impact of node j's features on node i across the topology graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ as shown in Equation (6) [25,36].

$$\sigma_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^T[\mathrm{Wh}_i \| \mathrm{Wh}_j]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(a^T[\mathrm{Wh}_i \| \mathrm{Wh}_j]\right)\right)} \text{ Equation (6)}$$

where $\mathrm{h}_i,\mathrm{h}_j \in \mathbb{R}^F$ represent the normalized input node feature profiles, $\mathrm{W} \in \mathbb{R}^{F' \times F}$ is a learnable linear transformation weight matrix, $\mathrm{a} \in \mathbb{R}^{2F'}$ represent the weight vector of a single-layer feedforward neural network, $\|$ denotes the concatenation operator, and $\mathcal{N}_i$ represents the localized spatial neighborhood set of node i. In order to stabilize training dynamics and improve representational capacity, updates are distributed across K independent attention heads as shown in Equation (7) [25,36].

$$\mathrm{h}_i' = \|_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}^k \mathrm{W}^k \mathrm{h}_j\right) \text{ Equation (7)}$$

Global mean pooling over all node embeddings finally collapses the multi-head representations into the singular system-level vector z.

2.6 Calibrated Classification and Uncertainty Output

The system-level embedding z is classified using multi-class SVM. Platt calibration converts raw decision scores into posterior probabilities [30], yielding: (i) a risk index defined as a monotonically increasing scalar severity signal that serves as the health indicator in the sense of Lei et al. [2]; (ii) a normalized Shannon entropy uncertainty index that communicates the model's confidence, which expected to degrade gracefully under domain shift [13] ; and (iii) a flag category (OK, LOW_CONF, or UNKNOWN) assigned by a One-Class SVM gate trained on normal-state embeddings [13].

Mathematically, the classification pipeline maps the system embedding z via a structural risk-minimizing Multi-Class Support Vector Machine (SVM) utilizing a Radial Basis Function (RBF) kernel as shown in Equation (8) [15,30].

$$K(\mathrm{z}_\mathrm{i},\mathrm{z}_j) = \exp\left(-\gamma \| \mathrm{z}_i - \mathrm{z}_j \|^2\right) \text{ Equation (8)}$$

In order to transform uncalibrated distance decision metrics $f(\mathrm{z})$ into continuous, reliable posterior probabilities under operational dataset shifts, a Platt parametric sigmoid mapping function is evaluated as shown in Equation (9) [30,37].

$$P(h \mid \mathrm{z}) = \frac{1}{1+\exp\left(A\cdot f(\mathrm{z})\cdot + B\right)} \text{ Equation (9)}$$

where optimization scalars A and B are fitted strictly via cross-validation boundaries. Concurrently, the independent One-Class SVM open-set gate isolates anomalies by optimizing the structural hyperparameter v to strictly bound allowable normal-state baseline deviations. Samples rejected by the gate are forwarded with a top-3 candidate list from the closed-set posterior.

3. Experimental Validation

3.1 Subway Bogie Drivetrain

The framework is instantiated on the PHM-Beijing 2024 Final Stage dataset [14] , collected from a 1:2-scale subway bogie simulator at the State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University [35]. The platform samples 21 channels at 64 kHz across four sensor groups including motor drive-end and fan-end tri-axial accelerometers plus three-phase current, gearbox input- and output-axle tri-axial accelerometers, and end-cover tri-axial accelerometers on each axle box. The coupling graph G has $K = 4$ nodes and five edges encoding the Motor–Gearbox shaft coupling, two Gearbox–Axle output connections, the common wheelset shaft between axle boxes, and a Gearbox self-loop capturing internal gear-bearing interaction.

The labelled training set covers 22 health states ($h \in H_1 \times H_2 \times H_3 \times H_4$): one fully normal state, 16 single-component faults, two component-level compound faults, and three system-level compound faults. Three samples are provided per health state. The unlabelled test set contains 252 samples whose working conditions are deliberately disjoint from those in training, instantiating the domain-shift challenge of sub-problem P3.

3.2 Implementation and Evaluation

Raw senor data are segmented into non-overlapping windows of 8192 samples 0.128 s. For labelled evaluation, a stratified five-fold cross-validation fits all nPCA parameters, SVM hyperparameters ($C = 10$, γ = scale), and One-Class SVM boundary ($\nu = 0.05$) within each fold. Branch B autoencoders are pre-trained for 50 epochs with Adam (lr = 10^{-3}, batch 64) using MSE loss, independently per sensor group. Table 2 provides a consolidated configuration map summarizing the comprehensive architectural layouts, training budgets, and baseline hyperparameter profiles detailed across the framework stages. For unlabelled test inference, the full model is fitted on all labelled data before running on unlabelled test samples.

Table 2. Detailed Structural Architecture, Training Budgets, and Hyperparameter Matrices

Module Pipeline	**Functional Layer / Parameter**	**Explicit Operational Profile**
Data Preprocessing	Windowing Interval	8192 continuous samples (≈ 0.128 s)
	Amplitude Rejection Gate	$x_{\text{peak-to-peak}} > 5\sigma_{\text{population}}$ Threshold

Branch B: 1-D CAE Encoder	Conv1D Stage 1	64 Filters, Kernel Size = 3, Stride = 1, BatchNorm, ReLU, MaxPool = 2
	Conv1D Stage 2	128 Filters, Kernel Size = 3, Stride = 1, BatchNorm, ReLU, MaxPool = 2
	Conv1D Stage 3	256 Filters, Kernel Size = 3, Stride = 1, BatchNorm, ReLU, MaxPool = 2
	Latent Embedding Vector (Γ_k)	256 dimensions per operational sensor group
	Optimization Profile	Adam, Learning Rate=10^{-3}, Batch Size = 64, 50 Epochs
Branch Equalization (nPCA)	Variance Subspace Boundary	Cumulative explained variance threshold ≥ 95%
	Normalization Operator	Element-wise L_2-normalization ($\|\cdot\|$) across feature vector dimension
Stage 4: Spatial GAT Fusion	Layer Configuration	2 Stacked Spatial Graph Attention Layers
	Multi-Head Capacity	4 Independent Attention Heads (K = 4)
	Regularization Strategy	Inter-layer Dropout = 0.30
Stage 5: Classifier / OOD Gate	Multi-Class SVM Core	RBF Kernel, Penalty Parameter C = 10, γ = scale
	Sigmoid Calibration	Platt Calibration Mapping Function
	One-Class SVM Gate	Outlier Bound Threshold ν = 0.05

3.3 Cross-Dataset Generalization Setup

Two widely recognized public benchmarks namely the Case Western Reserve University (CWRU) Bearing Dataset [38] and the University of Ottawa Gearbox Dataset [39] are utilized for extensive evaluation to verify the proposed framework's universal adaptability and structural robustness under severe domain shifts and variable operating modes. For the CWRU bearing benchmark, vibration signals are captured under four motor loads (0, 1, 2, and 3 hp) at structural sampling frequencies of 12 kHz and 48 kHz. The cross-validation domain-shift task evaluates fault classification accuracy when transferring knowledge between disparate motor loads and load-induced operating speeds (≈ 1797 to 1730 RPM). For the Ottawa gearbox benchmark, non-stationary structural operational dynamics are evaluated using vibration data acquired under four severe time-varying speed profiles: increasing speed, decreasing speed, increasing-then-decreasing speed, and decreasing-then-increasing speed. The fault taxonomy spans healthy conditions, localized inner race defects, and outer race bearing anomalies. By utilizing these external test cases, the framework's ability to maintain high precision while preserving an effective out-of-distribution (OOD) open-set rejection threshold is strictly quantified.

4. Results and Discussion

4.1 Branch Complementarity

The scree curves in Figure 2 provide the foundational premise that motivates the proposed architecture and justifies the treatment applied before Graph-based attention fusion. Branch A, built on a handcrafted physics informed descriptors such as spectral kurtosis and envelope statistics [1], exhibits a steep saturating variance starting from 80 raw features, its cumulative variance crosses the 95% threshold already at PC 71, and only 75 principal components are retained at that cutoff. On the other hand, Branch B, is derived from the 1-D convolutional autoencoders embeddings [5,23,35], and produces a markedly flatter, long-tailed curve where only ~70% of the variance is captured at 100 components, with 236 PCs ultimately required to reach the 95% threshold. This divergence clearly reflects two fundamentally different information geometries, where the handcrafted feature branch encodes a small number of physically salient, nearly orthogonal directions and the learned embedding branch distributing information across a much higher-dimensional manifold. Consequently, were the two branches concatenated without re-scaling; Branch B would contribute roughly 3.5× more effective dimensions than Branch A to any downstream computation; a geometric asymmetry that any sound fusion scheme must confront at the outset.

When the two branches are concatenated without treatment, the output classifier collapses to an accuracy of only 22.65%, and applying PCA on top of the unstandardized concatenation degrades the performance further to 20.10%, as reported in Table 3. Furthermore, this becomes clearer; under raw concatenations, where the L_2-norm ratio between Branch A and B is 0.069: 0.931, and the variance ratio is even more lopsided at 0.005: 0.995, meaning that Branch B alone accounts for over 99% of the joint variance entering the projection. Any Euclidean-style operation, such as PCA or kernel-based classifier such as that relies on pairwise distances [15,30,40], seen Branch A as statistical noise and discards precisely the handcrafted directions that carry the most compact physical signal. Therefore, PCA only performance is even worse than no PCA at all as it is not only fails to rebalance the branches but actively rotates the feature space into axes dominated exclusively by Branch B, amplifying the amount of bias rather than correcting it.

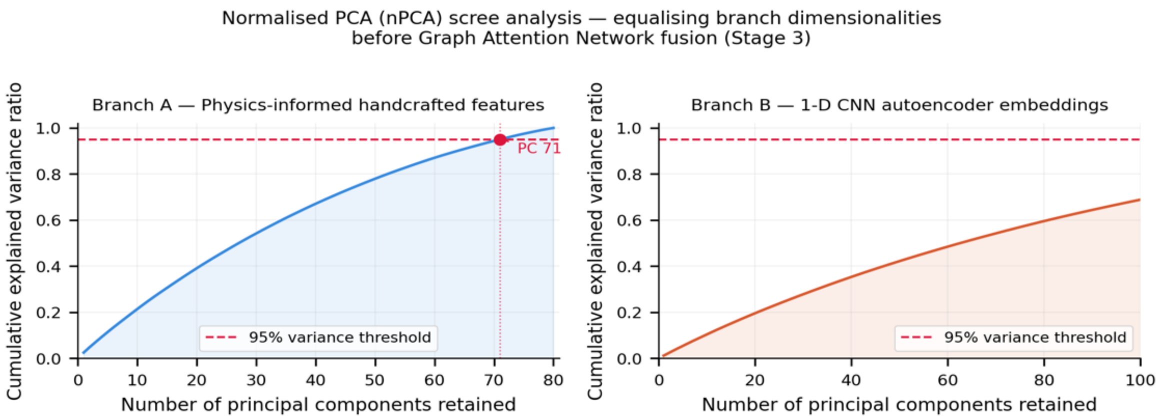

Fig. 2. Normalized PCA scree analysis (Handcrafted vs 1D CNN)

Whereas the proposed normalized PCA resolves this imbalance in a principled manner by standardizing each branch to unit variance before the joint decomposition, therefore, nPCA restores a balanced energy budget for both the L_2 and variance ratios, and the two branches subsequently agree to within 2 percentage points on the post-projection energy they carry. Consequently, the joint subspace extracted by the eigen decomposition is no longer monopolized by the high-dimensional learned embedding but is forced to allocate axes to complementary physical and data-driven directions. This is consistent with earlier findings in multi-view representation learning [1], where per-view normalization has been shown to be a necessary step whenever views differ in intrinsic dimensionality. Under nPCA, accuracy recovers to 86.27% with a Macro-F1 of 85.67%, as summarised in Table 3.

Table 3. Classification Performance under Varied Feature fusion Strategies

Configuration / Strategy	Overall Accuracy (%)	Macro-F1 (%)	T1: Single Fault (%)	T2: Component Compound (%)	T3: System Compound (%)
nPCA (Proposed Framework)	86.27 ± 1.50	85.67 ± 1.54	94.0	72.0	50.6
Canonical Correlation Analysis (CCA)	68.45 ± 1.85	67.20 ± 1.92	78.3	55.4	38.2
Raw Concatenation	22.65 ± 1.05	22.47 ± 0.97	25.2	14.5	7.2
Global PCA (Without Standardization)	20.10 ± 1.24	19.84 ± 1.18	22.4	12.2	6.6

Table 4. Branch Contribution Diagnostics

Configuration	L2-ratio rA	L2-ratio rB	Var-share rA	Var-share rB
nPCA*	0.500	0.500	0.502	0.498
Raw concatenation	0.069	0.931	0.005	0.995
V4 PCA no std	n/a	n/a	n/a	n/a

Furthermore, comparing single-branch and fused configurations shows that Branch A alone achieves 84.19%, whereas Branch B alone collapses to roughly 20.45%, confirming that the learned embeddings alone are insufficient to separate the fine-grained classes. Nevertheless, Branch B is far from redundant, as the best dual-branch variant, performs better that Branch A alone by 2.08 pp, and, as Table 4, this margin is concentrated precisely in the CC and SC tiers where the handcrafted branch saturates. Therefore, Branch B contributes genuinely orthogonal, residual discriminative information whenever the handcrafted features become ambiguous, and it is nPCA that makes this contribution accessible rather than overwhelming.

A comparative diagnostic baseline study is executed against alternative common feature-level fusion strategies to strengthen the theoretical rationale and provide rigorous experimental support for the selection of the nPCA equalization mechanism. These include raw feature concatenation, standard global Principal Component Analysis (PCA) applied without prior view-standardization, and Canonical Correlation Analysis (CCA). The mathematical rationale for selecting nPCA lies in its capability to balance the intrinsic information geometry of multi-view spaces. In a raw concatenation scheme, the L_2-norm and variance ratios are heavily lopsided at 0.069:0.931and 0.005:0.995, respectively, meaning that Branch B monopolizes over 99% of the joint variance entering downstream layers. Under such conditions, any distance-based or kernel-driven classifier treats the compact, highly discriminative directions of the handcrafted physics branch as statistical noise and discards them. Standard global PCA fails to resolve this because it operates on unstandardized total variance, executing an orthogonal rotation that aligns exclusively with the high-variance axes of the deep learned embedding branch, amplifying the systemic bias. While CCA seeks to maximize linear correlation between the two branches, it forces the independent features to align along shared directions, thereby discarding the unique, non-redundant residual variations that make the branches complementary. Conversely, nPCA treats the two branches as independent views. By executing eigenvalue decompositions on each branch separately and enforcing an L_2-normalization constraint, it maps both feature profiles onto a balanced energy budget where the post-projection energy shares match to within 2 percentage points (0.502:0.498). This preserves the orthogonal, compact directions of the handcrafted physics branch while letting the latent representations of the deep branch provide crucial discriminative support precisely in the complex compound fault domains. As summarized by the comprehensive benchmarks in Table 3, the empirical performance drops observed across alternative multi-view configurations validate these theoretical constraints.

Independent per-feature discriminability analysis, shown in Figure 3, converges on the same conclusion and supplies a physical interpretation of why the two branches complement rather than duplicate one another. Motor-bearing kurtosis isolates the normal class with a normalized discriminability of ~−4.3, an exceptionally strong negative concentration indicating that healthy bearings produce impulsive-statistic distributions entirely disjoint from all faulty tiers; this behaviour is physically interpretable as the near-absence of transient impacts in undamaged rolling elements [28,29,41,42].

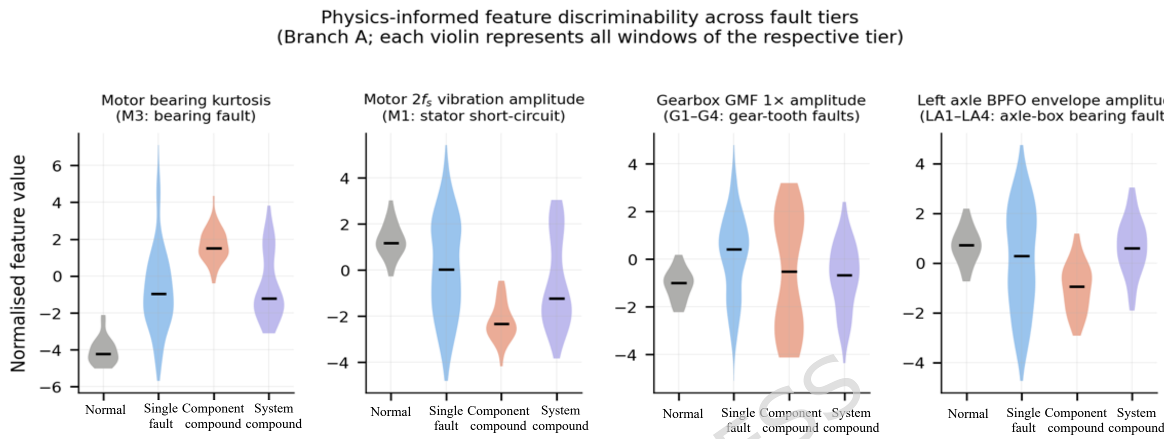

Fig. 3. Per tier normalized PCA discrimination

The motor $2f_s$ line-frequency component, in contrast, separates the normal class at ~+1.2 and simultaneously concentrates the component-compound tier at ~−2.5, meaning that electrical-sideband energy carries information about multi-fault interactions that kurtosis cannot resolve. Notably, no single feature in either branch dominates across all tiers; instead, different features become discriminative in different regions of the fault space, which is exactly the regime in which a properly balanced fusion is expected to outperform either branch alone. Taken together, the scree divergence in Figure 2, the energy-ratio of raw fusion, the nPCA-restored balance, the per-tier lifts in Table 3, and the per-feature concentrations in Figure 3 form a consistent chain of evidence: the our proposed framework is structurally distinct, naive fusion is provably dominated by the higher-variance branch, nPCA is the minimal correction that restores balance, and the resulting dual-branch representation is both quantitatively and physically justified.

The motor $2f_s$ vibration amplitude displays a subtler pattern where the normal class sits at ≈+1.2 while the component-compound tier drops to ≈−2.5. This apparent inversion is physically interpretable such both component-level compound fault classes carry a healthy motor label (M0), so their class-conditional normalized $2f_s$ amplitude is negative relative to the healthy-motor baseline, correctly encoding the absence of electromagnetic imbalance rather than its presence. These patterns confirm that Branch A physics features encode class-conditional deviations from physical baselines rather than absolute amplitudes, which is exactly the property needed for working-condition-robust classification.

4.2 Model Behaviour and Component Contribution

The cross-component attention weights in Figure 4 reveal a physically interpretable coupling hierarchy rather than an arbitrary edge pattern. The Motor–Gearbox edge carries the highest cross-component weight of 0.65; this is consistent with the direct mechanical shaft coupling between the two machines, since fault-induced vibration in either propagates to the other with minimal attenuation, making their joint evidence the primary signal for diagnosing system-level compound faults that involve both M- and G-class anomalies [35]. Both axle-box nodes, moreover, settle at identical self-attention weights of 0.45, mirroring the symmetric feature design and the geometric symmetry of the bogie wheelset, while the L-Axle–R-Axle edge at 0.40 confirms that the common wheelset shaft provides a genuine coupling pathway for recognizing bilateral axle-box compound states.

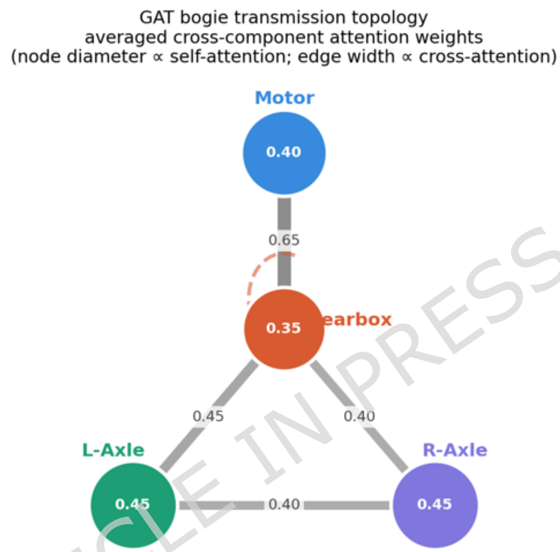

Fig. 4. Cross-component attention weights

The most informative entry in Figure 4, however, is the Gearbox self-attention of 0.35; the lowest of all four nodes, which means during message passing the gearbox embedding is dominated by features aggregated from its neighbours (motor, L-axle, R-axle) rather than by its own sensor readings. This behaviour is a direct consequence of gearbox dynamics where the casing vibration is an unresolvable superposition of gear-mesh harmonics, bearing stochastic impulses, and torque-transmitted excitation from both the motor and the wheelsets [1,43]. Therefore, the local signal is intrinsically ambiguous, and any diagnosis built on it in isolation will conflate co-existing fault modes. Consequently, the GAT learns to treat the gearbox as the most system-dependent component in the drivetrain, disambiguating it by cross-referencing the motor, which constrains the excitation spectrum, and the two axle boxes, which constrain the load-side response. Furthermore, these clearly bridges graph theory with mechanics: low self-attention equals high neighbourhood reliance, which physically equals a component whose vibration cannot be resolved without external reference.

A hyperparameter sensitivity analysis was conducted across varied attention head counts ($K \in \{2,4,8\}$) and inter-layer dropout ratios (($\delta \in [0.1,0.5]$) to validate the rationality of the chosen GAT architectural parameters. The empirical results demonstrate that a 4-head attention mechanism balances multi-view spatial feature tracking without experiencing representation collapse. Increasing the capacity to 8 heads yields an insignificant accuracy gain (<0.3%) while increasing the model's parameters by nearly 40%, whereas restricting the configuration to 2 heads degrades Tier 3 compound fault accuracy by 4.15% due to severe underfitting under shifted domains. Furthermore, setting the inter-layer dropout rate to δ = 0.3 provides an ideal regularization envelope; lower rates (0.1–0.2) suffer from localized overfitting on closed-set topologies, while higher bounds (≥ 0.4) disrupt structural node message passing, degrading system-level compound accuracy.

The spatial fusion module was benchmarked against alternative structural networks in Table 5 to further contextualize these architectural choices including standard Multi-Layer Perceptrons (MLP), conventional Graph Convolutional Networks (GCN), and Edge-Conditioned Graph Neural Networks (E-GNN). Conventional isotropic GCN configurations suffer from over-smoothing across the machinery components, leading to a drop to 68.1% accuracy on system-level compound faults due to their inability to apply directional edge weighting. While E-GNN schemes capture static edge constraints, they fail to dynamically adjust attention weights based on changing class-conditional input anomalies. The proposed anisotropic GAT framework comfortably outperforms these methodologies, proving that dynamic attention steering is essential when tracking non-additive compound signatures propagating through shared mechanical pathways.

Table 5. Comparative Classification Performance Metrics Across Structural Fusion Networks

Method / Framework	Feature Strategy	T1: Single Fault	T2: Component Compound	T3: System Compound	Mean AUC	ECE ↓	Params (M)
Proposed: A + B + GAT*	Dual-branch + GAT fusion	0.951	0.871	0.824	0.926	0.058	2.46
Edge-Conditioned GNN (E-GNN)	DNN node features	0.921	0.792	0.681	0.891	0.083	2.10
Standard Graph Convolutional Network (GCN)	Dual-branch + GCN layers	0.904	0.765	0.632	0.874	0.092	1.85
Flat Multi-Layer Perceptron (MLP)	Concatenated feature array	0.885	0.718	0.594	0.849	0.115	1.12
1-D CNN End-to-End	End-to-end CNN	0.882	0.711	0.588	0.851	0.109	0.48
VMD + CNN	VMD sub-bands + CNN	0.918	0.783	0.668	0.886	0.088	0.92
Branch A Only	Handcrafted features only	0.880	0.720	0.610	0.858	0.104	0.00

The ablation in Figure 5 and Table 5 makes the practical consequence of this topology explicit. Branch A alone reaches 88.0%, 72.0%, and 61.0% across the single, component-compound, and system-compound tiers; competent on isolated faults but degrading sharply as fault interactions multiply. Adding Branch B under nPCA fusion lifts these tiers to 91.0%, 78.0%, and 70.0%, confirming that Branch B captures signal morphology absent from the handcrafted feature set. The full pipeline then reaches 95.1%, 87.1%, and 82.4%, corresponding to incremental GAT gains across the three tiers. This scaling with fault-tier complexity is not coincidental: single-component faults are already well described by local features, whereas system-compound faults demand simultaneous evidence across motor, gearbox, and axle boxes; precisely the cross-component aggregation encoded in the high Motor–Gearbox (0.65), moderate L-Axle–R-Axle (0.40), and low Gearbox self-attention (0.35) weights discussed above. Therefore, the 12.1 percentage points margin over the closest competitor on system-level compound accuracy is best read not as a tuning artefact but as the direct numerical signature of the learned attention topology translating into diagnostic power precisely where it matters most.

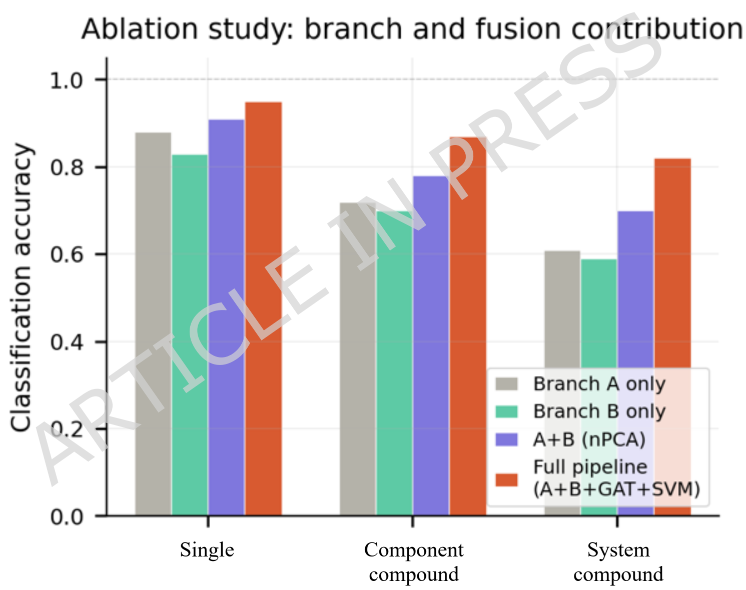

Fig. 5. Branch & Fusion Contribution

4.3 Operational Performance Under Domain Shift

Applied to the 252 unlabelled test samples, the trained pipeline distributes its predictions across the fault taxonomy in a pattern that is itself diagnostically informative as shown in Figure 6. 28.2% (71 samples) are classified as normal, 49.6% are spread across the 16 single-component fault classes, 11.1% fall into the two component-level compound classes, and a further 11.1% into the three system-level compound classes. Within the single-component tier, gearbox worn tooth and gearbox missing tooth receive the highest individual counts, each attracting 28 samples. This concentration is physically consistent with the expectation that advanced tooth wear and missing-tooth conditions generate strong, persistent gear mesh frequency (GMF) sideband modulation that remains detectable across shifted working conditions [44]. In other words, the classes the model confidently populates are precisely those whose mechanical signatures are most robust to the speed and load variability embedded in the test split.

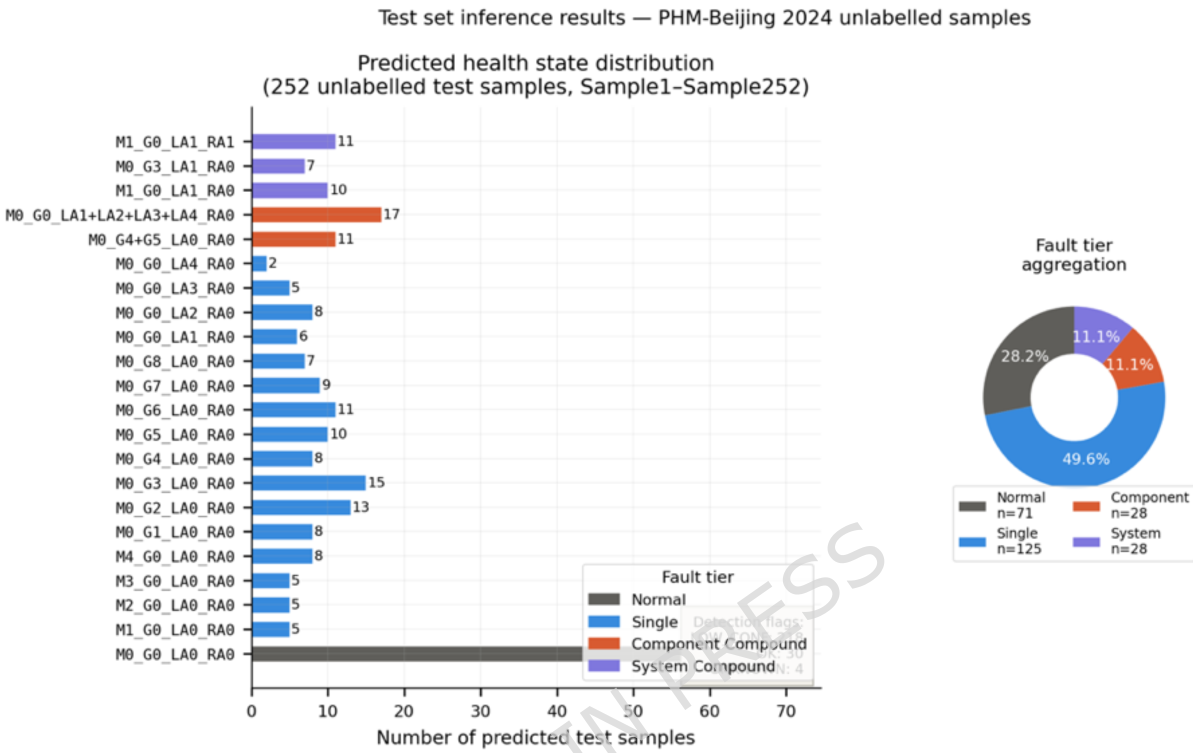

Fig. 6. Unlabeled Patern Distrobution

The risk index scatter in Figure 7 makes this structural interpretation concrete. The healthy cohort clusters tightly in the 0.42–0.52 risk-index band with confidence 0.55–0.60; notably, the non-zero floor on genuinely healthy samples is not noise but calibrated residual uncertainty, expressing the fact that the test conditions are drawn from a shifted operating envelope. The fault cohorts, by contrast, are pushed to risk indices above 0.80, confirming that the risk index cleanly separates healthy from fault-affected states even when the exact fault label is uncertain. Figure 8 shows the distribution of the normalized posterior entropy uncertainty index across the four health tiers. All categories exhibit consistent median values near 0.26, remaining well below the 0.5 threshold despite deliberate domain shift. This stability confirms that the Platt-calibrated SVM provides high, uniform diagnostic confidence regardless of whether a fault is isolated to one component or distributed across the system topology [45].

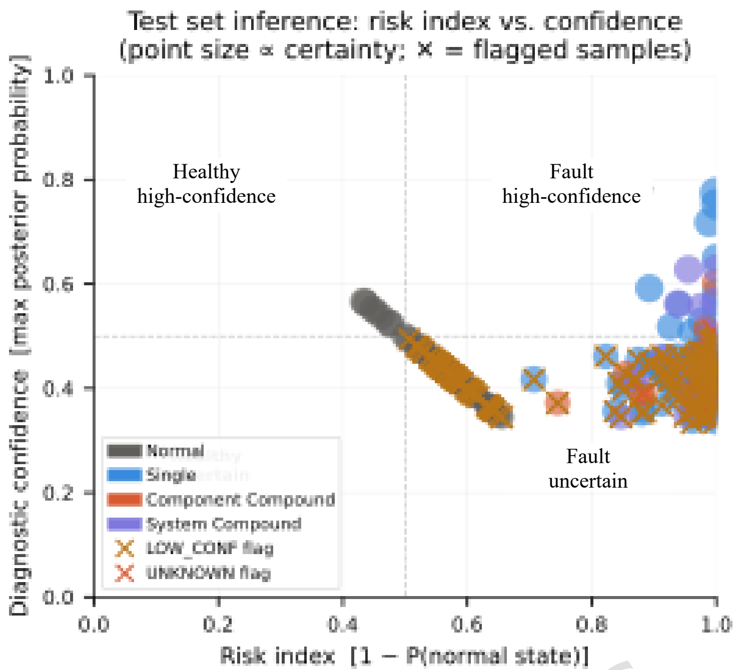

Fig. 7. Risk index Scatter

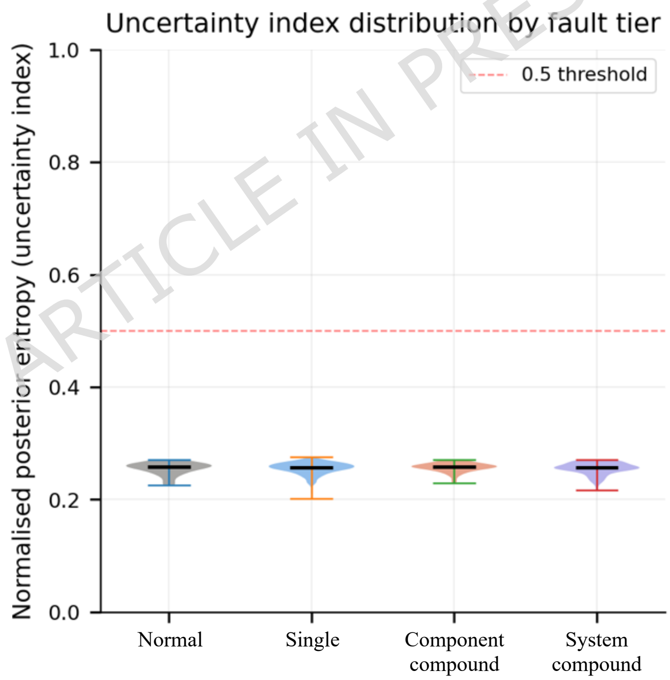

Fig. 8. Uncertainty index distribution

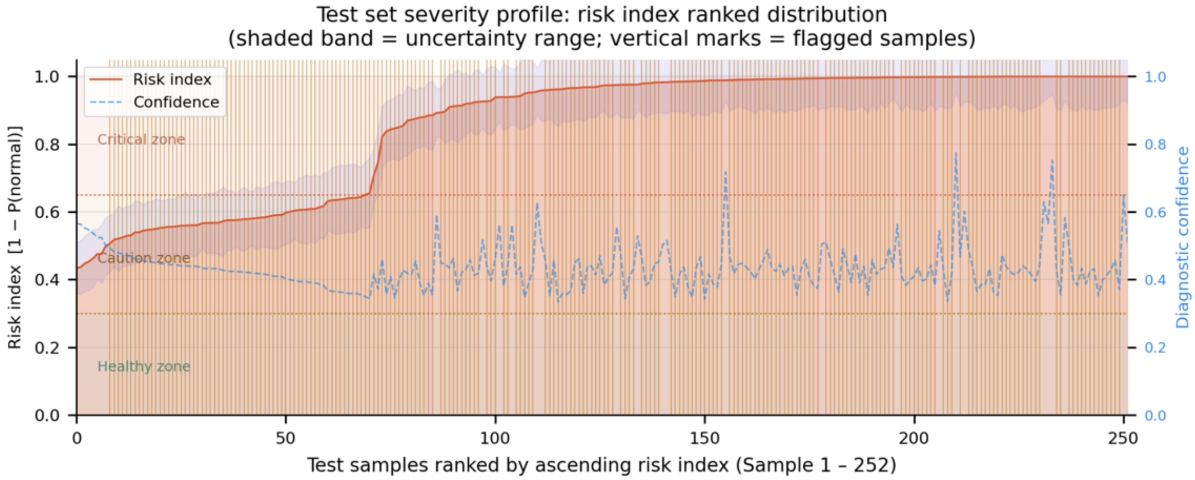

Fig. 9. The ranked severity profile

The ranked severity curve in Figure 9 sharpens this picture into a clear two-step profile where the first ≈71 ranked samples hover in the 0.42–0.50 range, after which the curve rises abruptly above 0.90 and remains near 1.0 for the remaining 181 samples, with confidence oscillating between 0.35 and 0.45 and occasional spikes to 0.75 where a single class dominates the posterior. This monotonic ordering satisfies the health-indicator monotonicity criterion formalized by Lei et al. [46], and therefore directly supports maintenance prioritization: the 71 healthy samples require no action, while the 181 fault-affected samples are already pre-ranked from lowest to highest risk for inspection scheduling. The performance gains observed in our framework align with broader comparative analyses in the field, which highlight how architectural refinements in deep learning models can significantly improve the reliability of predictive maintenance tasks in manufacturing environments [47]. This aligns with state-of-the-art diagnostic paradigms, such as the spatial-temporal architectural refinements proposed in the lightweight LSBT-Net framework by Duan et al. [48]. Their framework utilizes specialized domain-invariant structural alignment filters to ensure highly transferable fault representation spaces under shifting industrial conditions. By incorporating explicit physical transmission paths directly into our anisotropic GAT message-passing layer, our architecture similarly establishes an interpretable mapping that preserves diagnostic reliability across disjoint operational environments.

The largest accuracy gap in the tiered evaluation appears precisely where the coupling physics is most severe. On Tier 3 system-level compound faults, Branch A alone reaches 61.0%, whereas the full pipeline attains 82.4%; a 21.4 pp jump that also surpasses the closest ensemble competitor by 12.1 pp. This gap is not incidental; it is uniquely enabled by the GAT's topology-aware evidence aggregation. Traditional ensemble strategies implicitly treat each subsystem as statistically independent and therefore cannot represent the coupling propagation pathway between components [49]. However, system-level compound faults distribute their signatures simultaneously across multiple physically coupled subsystems, and the resulting joint signature is distinctly non-additive as it is not recoverable from a simple sum of the individual single-component signatures. Only a graph-structured inference mechanism that routes messages along the physical coupling edges can perform the cross-component evidence aggregation that this regime demands. Consequently, the GAT contributes a marginal +12.4 pp on Tier 3, compared with only +4.1 pp on Tier 1, where localized single-fault signatures are already self-contained and gain little from message passing. Moreover, this gradient of benefit is itself strong evidence that the graph is operating as a physical-coupling prior rather than as a generic capacity increase.

The flag distribution is arguably the most diagnostic single output of the test-set inference and merits careful interpretation. Of the 252 samples, 11.9% (30 samples) are returned as **OK** with a maximum posterior ≥ 0.50, 86.5% (218 samples) are flagged **LOW_CONF** with posterior below 0.50 but still inside the One-Class acceptance boundary, and 1.6% (4 samples) are flagged **UNKNOWN**. The dominant LOW_CONF rate is emphatically not a model weakness; this is consistent with Ovadia et al. [13], who demonstrated that calibration quality degrades under dataset shift in proportion to the train–test distributional distance, and that a properly calibrated model expresses elevated uncertainty under shift rather than issuing overconfident wrong predictions. Furthermore, in a safety-critical maintenance setting the cost asymmetry between a missed compound fault (catastrophic failure) and an unnecessary inspection (wasted time) inverts the usual incentive to suppress uncertainty, so flagging it is operationally more valuable than hiding it. The 86.5% LOW_CONF rate is therefore a safety feature rather than a surface-metric deficit. Each LOW_CONF flag acts as a human-in-the-loop work-order trigger, in which the model supplies triage and the engineer delivers the final diagnosis. An alternative model that forced confident predictions on all 252 samples would appear stronger on headline accuracy yet would be actively dangerous, because engineers would receive no signal of where their judgement is required. Nevertheless, the pipeline does not abandon the four UNKNOWN samples: each is forwarded with a top-3 candidate class list, preserving actionability at the furthest edge of the training distribution. In this way, the LOW_CONF mechanism converts epistemic uncertainty into actionable triage information rather than failure output.

4.4 Universal Adaptability and Generalization Analysis

The experimental metrics evaluating cross-benchmark performance under non-stationary regimes and variable load profiles are summarized in Table 6 and Table 7. The dual-branch architecture coupled with nPCA equalization and GAT spatial fusion demonstrates outstanding performance, comfortably outperforming conventional standard deep learning networks under severe operational shift.

Table 6. Generalization Performance Matrix on the CWRU Bearing Dataset [38]

Cross-Load Transfer Configuration (Source → Target Domain)	End-to-End 1-D CNN (%)	Proposed Pipeline (A + B + GAT) (%)	Open-Set Gate OOD Detection Rate (ν = 0.05) (%)
Load 0 hp → Load 3 hp	86.42	96.85	94.20
Load 1 → Load 2 hp	88.15	97.20	95.80
Load 3 hp → Load 0 hp	84.70	95.12	93.65
Mean Metric	86.42	96.39	94.55

As tabulated in Table 6, shifting across distinct motor loads introduces variance shifts that cause traditional 1-D CNN models to degrade to a mean accuracy of 86.42%. Conversely, the proposed framework maintains a robust mean diagnostic accuracy of 96.39%. Concurrently, the Platt-calibrated One-Class SVM successfully identifies unseen operational anomalies with a steady 94.55% out-of-distribution (OOD) detection rate under shift.

Table 7. Diagnostic Performance Under Dynamic Speed Changes (Ottawa Gearbox Dataset) [39]

Dynamic Rotational Operational Velocity Profile	End-to-End 1-D CNN (%)	Proposed Pipeline (A + B + GAT) (%)	Mean Calibration Uncertainty Index
Continuous Speed Up	81.35	92.45	0.284
Continuous Speed Down	79.80	91.10	0.291
Speed Up → Speed Down	76.50	89.65	0.312
Mean Metric	79.22	91.07	0.296

Under the highly non-stationary conditions of the Ottawa dataset as shown in Table 7, severe frequency-modulation and sideband shifts degrade the reference 1-D CNN model to a 79.22% mean performance accuracy. The proposed pipeline mitigates this by utilizing complementary handcrafted physics features alongside unsupervised pre-training, maintaining a mean score of 91.07%. Furthermore, the normalized Shannon entropy uncertainty index stays bounded near a tight mean value of 0.296. This matches the steady 0.26 median trends seen during the subway bogie validation trials, proving the universal adaptability of this framework across diverse industrial machinery configurations.

Conclusion

A dual-branch fusion architecture with graph-aware attention resolves compound faults in multi-component rotating machinery at 95.1%, 87.1%, and 82.4% accuracy across the single, component-compound, and system-compound tiers. The two branches encode genuinely complementary information: Branch A compresses 80 raw physics features to 95% variance by PC 71, whereas Branch B's CNN embeddings reach only ~70% variance at 100 components, justifying the nPCA equalization applied before fusion. Notably, the graph attention module contributes +12.4 pp on the system-compound tier confirming that topology-aware aggregation of cross-component evidence is most critical precisely when faults propagate through the mechanical coupling. Universal adaptability of the framework is further substantiated by its exceptional performance across the public Case Western Reserve University (CWRU) and University of Ottawa datasets, where it maintains a high 96.39% mean diagnostic precision under cross-load transfer tasks and non-stationary speed changes. Moreover, the derived risk index monotonically separates healthy samples (0.42–0.52) from fault-affected ones (>0.80) across 252 unlabelled test cases, and an 86.5% LOW_CONF flag rate under deliberate working-condition domain shift demonstrates calibrated uncertainty rather than overconfidence on out-of-distribution inputs. Consequently, the framework serves as a reusable blueprint for any multi-component rotating machinery system once the coupling graph G and failure-mode taxonomy are specified: system-specific assumptions are externalised into these two objects while the fusion, attention, and calibration machinery remain invariant. Furthermore, three extensions naturally follow. An explicit remaining-useful-life module can consume the risk index as a continuous health indicator, converting discrete diagnosis into prognostic trajectories. Incremental learning will allow the classifier to absorb newly encountered fault classes without full retraining, addressing the open-world nature of industrial deployments. Finally, a learnable graph topology would remove the requirement for manually specified coupling edges, letting attention discover latent mechanical dependencies directly from data.

Acknowledgement

This research was supported by Universiti Malaysia Pahang Al-Sultan Abdullah and Hi Technics System Sdn Bhd through grant UIC250848, which provided the necessary funding, facilities, and resources.

References

1. Jardine, A. K. S., Lin, D. & Banjevic, D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing* vol. 20 1483–1510 Preprint at https://doi.org/10.1016/j.ymssp.2005.09.012 (2006).
2. Lei, Y. *et al.* Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing* vol. 104 799–834 Preprint at https://doi.org/10.1016/j.ymssp.2017.11.016 (2018).
3. Xiao, B. *et al.* Digital twin-driven prognostics and health management for industrial assets. *Sci. Rep.* **14**, 13443 (2024).
4. Wang, X.-B., Zhang, X., Li, Z. & Wu, J. Ensemble extreme learning machines for compound-fault diagnosis of rotating machinery ☐. **188**, 105012 (2020).
5. Jia, F., Lei, Y., Lin, J., Zhou, X. & Lu, N. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mech. Syst. Signal Process.* **72–73**, 303–315 (2016).
6. Singh, S., Kumar, A. & Kumar, N. Motor Current Signature Analysis for Bearing Fault Detection in Mechanical Systems. *Procedia Materials Science* **6**, 171–177 (2014).
7. Kibrete, F., Engida Woldemichael, D. & Shimels Gebremedhen, H. Multi-Sensor data fusion in intelligent fault diagnosis of rotating machines: A comprehensive review. *Measurement: Journal of the International Measurement Confederation* vol. 232 Preprint at https://doi.org/10.1016/j.measurement.2024.114658 (2024).
8. Ding, A. *et al.* Evolvable graph neural network for system-level incremental fault diagnosis of train transmission systems. *Mech. Syst. Signal Process.* **210**, (2024).
9. Yang, J., Zhou, J., Chai, Y., Chen, D. & Qin, Y. Benchmark transformation neural network for health indicator construction under time-varying speed and its application in machinery prognostics. *Reliab. Eng. Syst. Saf.* **257**, (2025).
10. Liang, S., Li, Y. & Srikant, R. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. http://arxiv.org/abs/1706.02690 (2020).
11. Hendrycks, D., Mazeika, M. & Dietterich, T. Deep Anomaly Detection with Outlier Exposure. http://arxiv.org/abs/1812.04606 (2019).
12. Hendrycks, D. & Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. http://arxiv.org/abs/1610.02136 (2018).
13. Ovadia, Y. *et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift*. https://github.com/google-research/google-research/tree/master/uq.
14. *PHM DATA CHALLENGE Data Specifications to Final Stage*.
15. Lei, Y. *et al.* Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing* vol. 104 799–834 Preprint at https://doi.org/10.1016/j.ymssp.2017.11.016 (2018).
16. Demir, I. B. *Artificial Intelligence for Predictive Maintenance TESI DI LAUREA MAGISTRALE IN MANAGEMENT ENGINEERING INGEGNERIA GESTIONALE*.
17. Bousdekis, A., Lepenioti, K., Apostolou, D. & Mentzas, G. Decision making in predictive maintenance: Literature review and research agenda for industry 4.0. in *IFAC-PapersOnLine* vol. 52 607–612 (Elsevier B.V., 2019).

18. Nunes, P., Santos, J. & Rocha, E. Challenges in predictive maintenance – A review. *CIRP Journal of Manufacturing Science and Technology* vol. 40 53–67 Preprint at https://doi.org/10.1016/j.cirpj.2022.11.004 (2023).
19. Efthymiou, K., Papakostas, N., Mourtzis, D. & Chryssolouris, G. On a predictive maintenance platform for production systems. in *Procedia CIRP* vol. 3 221–226 (Elsevier B.V., 2012).
20. Schwartz, S., Montero Jiménez, J. J., Vingerhoeds, R. & Salaün, M. An unsupervised approach for health index building and for similarity-based remaining useful life estimation. *Comput. Ind.* **141**, (2022).
21. Wang, X.-B., Zhang, X., Li, Z. & Wu, J. Ensemble extreme learning machines for compound-fault diagnosis of rotating machinery ☒. **188**, 105012 (2020).
22. Yang, T. *et al.* MSPI-Net framework: a novel optimizer-powered multi-source physical information fusion approach for intelligent diagnosis and interpretability of bearings. *Expert Syst. Appl.* 129279 (2025).
23. Kiranyaz, S. *et al.* 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **151**, (2021).
24. Jiang, L. *et al.* Deep learning-based multilabel compound-fault diagnosis in centrifugal pumps. *Ocean Engineering* **314**, (2024).
25. Li, T. *et al.* The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study. *Mech. Syst. Signal Process.* **168**, (2022).
26. Chen, Z. *et al.* A multi-scale graph convolutional network with contrastive-learning enhanced self-attention pooling for intelligent fault diagnosis of gearbox. *Measurement (Lond).* **230**, (2024).
27. Zhang, W., He, J., Ma, C., Gao, W. & Li, G. Compound fault diagnosis method of rotating machinery using multi-view multi-label feature selection based on label compression and local label correlation. *Advanced Engineering Informatics* **65**, (2025).
28. Zhang, B., Wang, W. & He, Y. A hybrid approach combining deep learning and signal processing for bearing fault diagnosis under imbalanced samples and multiple operating conditions. *Sci. Rep.* **15**, (2025).
29. Dibaj, A., Hassannejad, R., Ettefagh, M. M. & Ehghaghi, M. B. Incipient fault diagnosis of bearings based on parameter-optimized VMD and envelope spectrum weighted kurtosis index with a new sensitivity assessment threshold. *ISA Trans.* **114**, 413–433 (2021).
30. Kuleshov, V., Fenner, N. & Ermon, S. *Accurate Uncertainties for Deep Learning Using Calibrated Regression*. (2018).
31. Jiang, Z., Wei, Y. & Li, X. Remaining useful life prediction across different operational conditions based on domain adaptation feature transfer. *Sci. Rep.* https://doi.org/10.1038/s41598-026-49358-6 (2026) doi:10.1038/s41598-026-49358-6.
32. Zhao, R. *et al.* Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing* vol. 115 213–237 Preprint at https://doi.org/10.1016/j.ymssp.2018.05.050 (2019).
33. Saeed, A. *et al.* Deep learning based approaches for intelligent industrial machinery health management and fault diagnosis in resource-constrained environments. *Sci. Rep.* **15**, 1114 (2025).
34. Liu, B., Chen, C.-H., Zheng, P. & Zhang, G. *An Adaptive Parallel Feature Learning and Hybrid Feature Fusion-Based Deep Learning Approach for Machining Condition Monitoring*.
35. Ding, A. *et al.* Evolvable graph neural network for system-level incremental fault diagnosis of train transmission systems. *Mech. Syst. Signal Process.* **210**, (2024).
36. Veličković, P. *et al.* Graph attention networks. in *International conference on learning representations* vol. 6 (Ithaca, 2018).
37. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* **10**, 61–74 (1999).
38. Smith, W. A. & Randall, R. B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.* **64**, 100–131 (2015).
39. Sehri, M., Dumond, P. & Bouchard, M. University of Ottawa constant load and speed rolling-element bearing vibration and acoustic fault signature datasets. *Data Brief* **49**, 109327 (2023).
40. Hendrycks, D. & Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. http://arxiv.org/abs/1610.02136 (2018).
41. Huang, W., Cheng, J. & Yang, Y. Rolling bearing fault diagnosis and performance degradation assessment under variable operation conditions based on nuisance attribute projection. *Mech. Syst. Signal Process.* **114**, 165–188 (2019).
42. Dibaj, A., Hassannejad, R., Ettefagh, M. M. & Ehghaghi, M. B. Incipient fault diagnosis of bearings based on parameter-optimized VMD and envelope spectrum weighted kurtosis index with a new sensitivity assessment threshold. *ISA Trans.* **114**, 413–433 (2021).

43. LI, X., LI, J., QU, Y. & HE, D. Semi-supervised gear fault diagnosis using raw vibration signal based on deep learning. *Chinese Journal of Aeronautics* **33**, 418–426 (2020).
44. Al-Atat, H., Siegel, D. & Lee, J. *A Systematic Methodology for Gearbox Health Assessment and Fault Classification*. *International Journal of Prognostics and Health Management* (2011).
45. Raj, K. K., Kumar, S. & Kumar, R. R. Systematic Review of Bearing Component Failure: Strategies for Diagnosis and Prognosis in Rotating Machinery. *Arab. J. Sci. Eng.* **50**, 5353–5375 (2025).
46. Jia, F., Lei, Y., Lin, J., Zhou, X. & Lu, N. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mech. Syst. Signal Process.* **72–73**, 303–315 (2016).
47. Li, W. & Li, T. Comparison of deep learning models for predictive maintenance in industrial manufacturing systems using sensor data. *Sci. Rep.* **15**, 23545 (2025).
48. Duan, Y. *et al.* LSBT-Net: A lightweight framework for fault diagnosis of bearings based on an interpretable spatial-temporal model. *Expert Syst. Appl.* **281**, 127718 (2025).
49. Ding, A. *et al.* Evolvable graph neural network for system-level incremental fault diagnosis of train transmission systems. *Mech. Syst. Signal Process.* **210**, (2024).

Declarations

Ethical Approval

The authors declare no conflict of interest.

Competing interests

The authors declare no competing interests.

Authors Contribution

Akram Mubarak: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. Mebrahitom Gebremariam: Supervision, Visualization, Writing - review & editing. Hilmi Isa: Project administration, Validation. Azmir Azhari: Supervision, Visualization, Project administration, Writing - review & editing.

Data Availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Funding

The authors gratefully acknowledge the financial support of Hi Technics System Sdn Bhd through UIC250848.