



Fixed Channel Frequency Graphs for Cross Speed Compound Fault Diagnosis of Gearboxes

Junwei Shen^{1,*}

¹ Shanxi Institute of Mechanical and Electrical Engineering, Changzhi, 046000, China.

5 Correspondence to: Junwei Shen (shen_wangy@163.com)

Abstract. Compound bearing and gear faults require identification of multiple fault components under changes in operating speed. This study evaluates a fixed channel–frequency graph using 64 public recordings and recording-disjoint operating-condition holdouts. Each three-channel window is represented by 30 nodes containing full-window spectral statistics and eight local log-power descriptors. A two-layer graph network predicts seven binary fault labels. The fixed graph achieves mean recording-level Macro-F1 of 0.8284, compared with 0.7954 for a matched self-loop control and 0.8150 for a degree-matched random graph. The adaptive comparator achieves 0.8420. Five-seed condition results show higher adaptive than random graph means in all four holdouts, with differences of 0.0420, 0.0146, 0.0142, and 0.0370 at 1200, 1800, and 2400 r/min and variable speed, respectively. The fixed graph remains below the waveform CNN, which achieves 0.9708; validation-gated CNN–graph fusion reaches 0.9713. Validation rotation retains the low-speed deficit, and performance deteriorates at 0 and –5 dB test noise. The findings quantify the contribution of the prescribed topology within the evaluated graph representation, while showing that adaptive connectivity and waveform learning achieve higher overall scores on this benchmark.

Keywords: Compound fault diagnosis, cross-speed evaluation, channel–frequency graphs, multilabel classification, vibration analysis.

20 1 Introduction

Compound gearbox diagnosis involves both concurrent defects and changes in operating speed. Bearing and gear faults can coexist within an acquisition, while speed changes alter the vibration spectrum. A diagnostic model must therefore identify the constituent faults in recordings from a held-out operating condition. Feature-based and deep learning methods provide different representations for this task (Liu et al., 2018; Lei et al., 2020). Graph learning exchanges information between structured measurements (Li et al., 2022). Machinery diagnosis studies have connected scales, sensor streams, and samples (Zhao et al., 2023; C. Wang et al., 2025; J. Li et al., 2023; Yu et al., 2023). To determine the contribution of those connections, graph models require controls that retain the same features and trainable architecture while changing adjacency. Multilabel outputs also distinguish fault components within a compound state, as examined in compound-fault diagnosis studies (Xing et al., 2022; Z. Wang et al., 2025; Huang et al., 2021; W. Li et al., 2023).



30 We construct a fixed graph from three vibration channels and ten frequency bands per channel. Edges join adjacent bands
within a channel and corresponding bands across channels. These links define an aggregation structure; they are not
measurements of mechanical coupling. A separately trained self-loop model retains the same descriptors, initialization,
optimization, and parameter count. Frequency-chain, same-band channel, and degree-matched random graphs provide
additional topology comparisons. An adaptive graph, a raw-waveform CNN, and an SVM extend the comparison beyond the
35 prescribed fixed structure.

This study makes three contributions. First, it defines a 30-node channel–frequency representation containing full-window and
local spectral descriptors. Second, it evaluates edge families, random connectivity, adaptive connectivity, and descriptor
ablations, including condition-specific results from five training seeds for the random and adaptive graphs. Third, it reports
recording-level operating-condition holdouts, training–validation allocation sensitivity, compound-label recovery, validation-
40 gated fusion, and added-noise sensitivity. These experiments distinguish the contribution of graph connectivity from the
absolute performance of the final classifier.

The fixed graph improves mean recording-level Macro-F1 by 0.0330 over the self-loop model and by 0.0134 over the random
graph. The adaptive graph improves the fixed-graph mean by 0.0136, while the CNN has the highest mean among the
standalone models. The study therefore evaluates a specified graph representation and its limits rather than claiming that the
45 fixed topology is the best diagnostic architecture.

Long acquisitions yield many correlated windows. We assign complete recordings to partitions and aggregate window
predictions to one decision per recording. The 7488 windows are computational samples from 64 acquisitions, not 7488
independent experimental observations. Operating-condition holdout excludes test-condition recordings from both training
and validation.

50 2 Related work

Reviews of machinery diagnosis document the progression from hand-engineered vibration statistics to hierarchical learned
representations (Liu et al., 2018; Lei et al., 2020). Convolutional models extract local temporal patterns directly from raw
signals (Wen et al., 2018), whereas graph models add explicit relational inductive biases. A benchmark of graph methods
emphasizes joint evaluation of the node definition, adjacency construction, and readout (Li et al., 2022). Multiscale rotor-
55 bearing graphs (Zhao et al., 2023), multisensor spatiotemporal graphs (C. Wang et al., 2025), automated graph architecture
selection (J. Li et al., 2023), and graph-weighted diagnosis under noise (Yu et al., 2023) illustrate different ways to introduce
relationships. These systems address distinct tasks and datasets, so their published performance is not a head-to-head baseline
for the present data.

Normalized graph convolution (Kipf and Welling, 2017), inductive neighborhood aggregation (Hamilton et al., 2017),
60 attention-based messages (Veličković et al., 2018), and aggregation expressiveness (Xu et al., 2019) provide established graph-
learning formulations. The present network uses row-normalized fixed adjacency with separate self and neighbor



transformations. Its attention weights node representations at readout; it does not learn the edges of the fixed graph. Local spectral segments supply temporal descriptors without introducing time-indexed nodes or temporal edges.

Compound-fault studies examine label semantics, capsule representations, and multilabel transfer (Xing et al., 2022; Z. Wang et al., 2025; Huang et al., 2021; W. Li et al., 2023). The present training set contains the evaluated fault combinations at other speeds, so the task is cross-condition diagnosis of observed combinations. Studies of operating-condition transfer and domain generalization (Zhang et al., 2024; Gao et al., 2024; Chen et al., 2020; Y. Xiao et al., 2025; Liao et al., 2020) address related domain shifts. Our evaluation concerns the public experimental setting and does not measure transfer to a new machine or independent fault specimen. Multisensor fusion (X. Xiao et al., 2025) motivates cross-channel processing, and reported evaluation limitations (Matania et al., 2024) motivate recording-level rather than window-level assessment.

Zhao et al. (2024) examine held-out domains in an application-oriented diagnosis benchmark. Cao et al. (2025) investigate untrained planetary gearbox compound faults with a multitask dynamic graph attention model. Our test factor is operating condition, while the fault combinations remain represented in the source conditions. Their published scores concern different tasks and are not numerical baselines for the present experiment.

The graph unit also differs across studies. Graphs can connect samples, recordings, sensor locations, or feature scales (Li et al., 2022; Zhao et al., 2023; C. Wang et al., 2025; J. Li et al., 2023; Yu et al., 2023). Here, each window has its own 30-node graph, and the fixed adjacency depends only on channel and band identities. Constructing a test-window graph therefore requires no neighborhood statistics from other test recordings. This construction establishes an inductive input representation, while the experimental comparisons assess the chosen edge pattern.

3 Materials and methods

3.1 Data and diagnostic task

The public mixed bearing–gearbox vibration collection (Hou, 2025), documented in the companion data article (Hou et al., 2025), supplies the 64 CSV recordings used here. The released files used in the experiment have three measured channels sampled at 16 kHz under 1200, 1800, and 2400 r/min and a varying-speed condition (200–2400–200 r/min). Each condition contains 16 recordings: two normal, seven single-fault, and seven compound-fault states. The seven binary targets are bearing inner race, outer race, rolling element, and cage defects, and gear surface wear, tooth breakage, and eccentricity. Bearing–gear and inner–outer bearing combinations occur in the observed state set. A normal recording has an all-zero target vector.

The companion data article describes three-axis vibration acquisition at 16 kHz, 30-s recordings, and a constant 20-N·m load (Hou et al., 2025). The processed index uses 117 complete windows per recording. At 4096 samples per window, these windows cover 29.952 s, with the incomplete tail omitted. The evaluation addresses operating-condition changes in this collection; independent specimen and machine transfer are not evaluated.

After selecting CH 1–3 and excluding the time column, each recording is divided into nonoverlapping 4096-sample windows. This produces 7488 windows, with 1872 per condition. Each channel is centered by its window median and scaled by 1.4826



times its median absolute deviation (MAD). Standard deviation supplies the fallback scale when the MAD-based dispersion is
 95 near zero. Values are clipped to $[-8, 8]$, as summarized in Eq. (1).

$$\begin{aligned}\tilde{x}_{c,t} &= \text{clip}\left(\frac{x_{c,t} - m_c}{s_c}, -8, 8\right), \\ m_c &= \text{median}_t x_{c,t}, \quad s_c = 1.4826 \text{ MAD}_t x_{c,t}.\end{aligned}\quad (1)$$

The median and MAD in Eq. (1) are calculated independently for each channel and window. No statistics are pooled across
 held-out recordings for this normalization. Per-window scaling removes the original amplitude scale, so the normalized
 descriptors cannot be interpreted as calibrated vibration amplitudes.

100 **Table 1: Dataset composition**

Condition	Recordings	Windows	States
1200 r/min	16	1872	2 N / 7 S / 7 C
1800 r/min	16	1872	2 N / 7 S / 7 C
2400 r/min	16	1872	2 N / 7 S / 7 C
Variable	16	1872	2 N / 7 S / 7 C

N: normal; S: single fault; C: compound fault. There are 64 recordings and 7488 windows in total. The same 15 state definitions
 recur across conditions.

3.2 Spectral descriptors and graph construction

A Hann-windowed real FFT is computed for each complete window and its eight consecutive 512-sample segments. The ten
 105 frequency bands have boundaries 0, 64, 96, 128, 160, 256, 512, 1000, 2000, 4000, and 8000 Hz. The full-window and segment
 frequency spacings are given in Eq. (2).

$$\begin{aligned}\Delta f_{\text{full}} &= \frac{16000}{4096} = 3.90625 \text{ Hz}, \\ \Delta f_{\text{local}} &= \frac{16000}{512} = 31.25 \text{ Hz}.\end{aligned}\quad (2)$$

Each full window spans 0.256 s and each local segment spans 0.032 s. On the 31.25-Hz local frequency grid, the bands 64–
 96, 96–128, and 128–160 Hz contain one bin each, at 93.75, 125, and 156.25 Hz, respectively. Thus these local descriptors
 110 have limited within-band frequency resolution. The full-window grid has a spacing of 3.90625 Hz.

Let L denote the log-power spectrum and K_b the bins assigned to frequency band b on the relevant full-window or segment
 grid. Node (c, b) represents channel c and band b . Its descriptor is defined in Eq. (3).

$$\begin{aligned}\mathbf{z}_{c,b} &= [\text{mean } L_c(K_b), \text{max } L_c(K_b), \\ &\quad \text{std } L_c(K_b), \text{mean } L_{c,1}(K_b), \dots, \\ &\quad \text{mean } L_{c,8}(K_b)]^T \in \mathbb{R}^{11}.\end{aligned}\quad (3)$$



The descriptor contains three full-window log-power statistics and eight local mean log-power values, giving a 30×11 feature array. These are statistics of log power rather than total band energies. The frequency boundaries and fixed adjacency are identical across operating conditions.

The undirected binary adjacency includes self-loops, adjacent bands within a channel, and all channel pairs within the same band:

$$A_{(c,b),(c',b')} = \begin{cases} 1, & c = c', |b - b'| \leq 1, \\ 1, & c \neq c', b = b', \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

$$D_{ii} = \sum_j A_{ij}, \quad \hat{A} = D^{-1}A.$$

The ten cross-channel groups are three-node cliques. Together with the three frequency chains, this gives 27 adjacent-band edges and 30 cross-channel edges, or 57 undirected non-self edges. Including 30 self-loops, A has 144 nonzero entries. Endpoint nodes have degree 4 and interior nodes degree 5, including self-loops. Row normalization averages neighbor messages; it is distinct from symmetric normalization. Fig. 1 shows the complete cross-channel links and the resulting adjacency pattern.

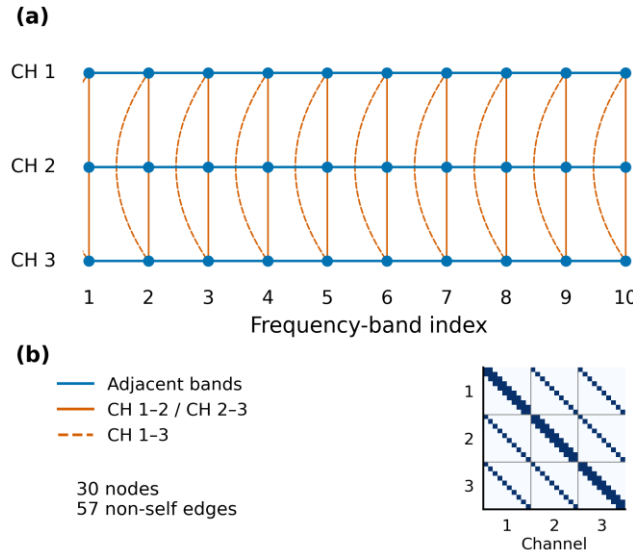


Figure 1: Fixed channel–frequency topology. (a) Three frequency chains and complete same-band channel connections; the dashed arcs explicitly show CH 1–CH 3. **(b)** Binary adjacency, including self-loops. The graph definition yields 30 nodes and 57 non-self edges.

3.3 Message passing and multilabel prediction

A width-48 feature embedding is augmented with learned channel and band embeddings. Two residual message blocks combine separate nodewise and neighbor transformations:



$$\begin{aligned} \mathbf{h}_{c,b}^{(0)} &= W_z \mathbf{z}_{c,b} + \mathbf{e}_c^{\text{ch}} + \mathbf{e}_b^{\text{band}}, \\ M^{(\ell)} &= \hat{A} H^{(\ell)}, \\ U^{(\ell)} &= H^{(\ell)} W_{\text{self}}^{(\ell)} + M^{(\ell)} W_{\text{nbr}}^{(\ell)}, \\ H^{(\ell+1)} &= \mathcal{B}_\ell(H^{(\ell)}, U^{(\ell)}). \end{aligned} \quad (5)$$

Biases are omitted from Eq. (5). The residual block B comprises GELU, dropout, layer normalization, and a residual connection. The same block and feature dimensions are used in the independently trained self-loop model. Learned channel and band embeddings encode node identity; shaft speed is not supplied as an input embedding.
 The final node representations are pooled by the mean, elementwise maximum, and learned attention. Their concatenation has dimension 144 and feeds a seven-logit head:

$$\begin{aligned} a_i &= \frac{\exp s(\mathbf{h}_i)}{\sum_{j=1}^{30} \exp s(\mathbf{h}_j)}, \\ \mathbf{g} &= \left[\frac{1}{30} \sum_i \mathbf{h}_i \parallel \max_i \mathbf{h}_i \parallel \sum_i a_i \mathbf{h}_i \right] \\ &\in \mathbb{R}^{144}, \quad \ell = f_{\text{head}}(\mathbf{g}) \in \mathbb{R}^7. \end{aligned} \quad (6)$$

Here, s is the learned node-scoring function and f_{head} the classification head. A sigmoid maps logits to seven component-fault probabilities. The weighted binary cross-entropy can be written using positive and negative coefficients as

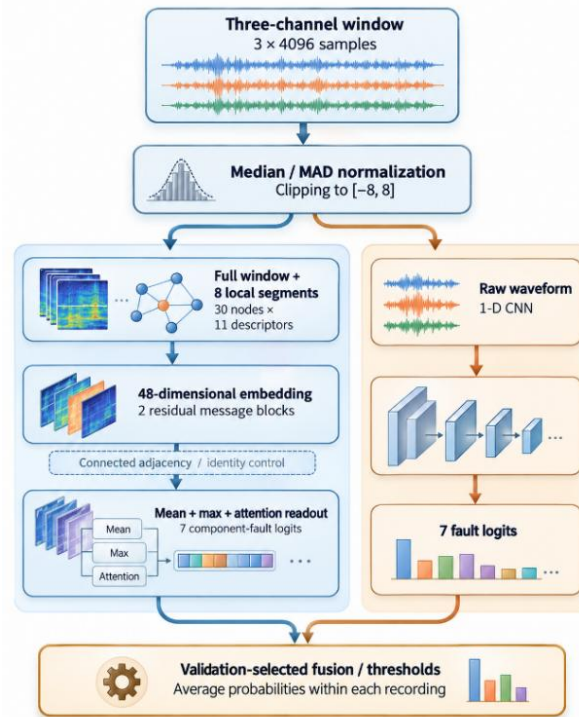
$$\mathcal{L}_{\text{BCE}} = -\frac{1}{7N} \sum_{n=1}^N \sum_{k=1}^7 [w_k^+ y_{nk} \log p_{nk} + w_k^- (1 - y_{nk}) \log (1 - p_{nk})]. \quad (7)$$

The coefficients w^+ and w^- in Eq. (7) denote positive- and negative-target weights; a coefficient of one gives an unweighted term. Training uses window-level targets. Checkpoint selection and diagnostic evaluation use predictions aggregated at the recording level.

3.4 Topology and descriptor comparisons

The topology comparison includes self-loops, within-channel frequency chains, same-band cross-channel links, a degree-matched random graph, the fixed graph, and an adaptive graph. The two edge-family variants isolate the prescribed frequency and channel links. The random graph provides a comparison with alternative edge organization at matched node degrees. The adaptive graph is an additional comparator; equal parameter count and computational cost with the fixed graph have not been established.

removal of the eight local descriptors. The fixed graph is also referred to as the connected graph in the figures. Here, fixed graph means the prescribed combination of frequency-chain and same-band channel links, not an all-to-all graph. Table 6 summarizes all variants. Tables 7 and 8 give five-seed random and adaptive results and the corresponding condition means.



155 **Figure 2: A compact model schematic. The graph branch uses 30 nodes and the same seven-label output capacity in both fixed-**
 160 **adjacency and identity-adjacency fits. The raw-signal CNN is a separate comparator; validation controls optional fusion.**

3.5 Controls and validation gate

The identity control has the same 30 nodes, learned parameters, initialization, minibatch ordering, augmentation, loss, and early-stopping rule as the connected graph; its adjacency is the identity matrix. It is retrained independently rather than
 160 produced by masking edges in a trained graph. We separately report a fixed-weight edge intervention, in which the learned connected model is evaluated with its adjacency replaced by the identity. This intervention probes reliance on trained neighbor messages, but it is not a substitute for the retrained matched ablation.

The raw-waveform CNN has three convolutional stages with kernel sizes 9, 7, and 5, batch normalization, GELU, pooling, and a seven-logit head. The SVM uses training-fold-standardized time and spectral statistics with radial-basis one-versus-rest
 165 classifiers. A validation gate optionally blends CNN and graph logits. Its weight and decision thresholds are selected on validation recordings and frozen before test inference, as specified in Eq. (8).

The SVM features include channelwise central tendency, dispersion, root-mean-square amplitude, peak and impulsiveness statistics, spectral centroid, bandwidth, dominant frequency, entropy, and eight relative spectral-energy bins. Seven binary classifiers provide the multilabel output. This is a feature-based baseline, whereas the graph controls examine adjacency
 170 changes within the graph representation.

CNN and graph logits are combined before sigmoid conversion and recording-level aggregation:



$$\begin{aligned}\ell_w^{\text{fuse}} &= (1 - \alpha)\ell_w^{\text{CNN}} + \alpha\ell_w^{\text{graph}}, \\ \bar{\mathbf{p}}_r &= \frac{1}{|\mathcal{W}_r|} \sum_{w \in \mathcal{W}_r} \sigma(\ell_w^{\text{fuse}}), \\ \hat{y}_{rk} &= \mathbb{1}\{\bar{p}_{rk} \geq \tau_k\}.\end{aligned}\tag{8}$$

The graph weight α is selected from $\{0, 0.25, 0.50, 0.75, 1.00\}$; the validation threshold search spans 0.3–0.7. A nonzero graph weight is accepted only when validation recording-level Macro-F1 exceeds that of the pure CNN by more than 0.01. Otherwise α is set to zero. The same gate is applied to the self-loop model. Selected weights and thresholds remain unchanged during clean and noisy test inference.

3.6 Splits training and evaluation

Four reference folds hold out one operating condition for testing. The preceding condition in the cyclic order 1200, 1800, 2400 r/min, and variable speed supplies validation data; the remaining two conditions supply training data (Table 2). Each fold contains 32 training, 16 validation, and 16 test recordings, with all windows of a recording assigned together. The neural comparisons use seeds 13, 29, 42, 71, and 97. The reported training settings are AdamW with an initial learning rate of 0.001, at most 65 epochs, batch size 64, and early stopping after ten epochs without validation improvement. Training augmentation rolls windows by up to 128 samples and adds Gaussian noise with probability 0.5 at an SNR sampled from 5–25 dB. The SVM is fitted once per fold.

Training–validation allocation sensitivity is evaluated for the fixed graph. For each test condition, each of the other three conditions is used in turn for validation, with the remaining two used for training. The 12 allocations retain 32 training, 16 validation, and 16 test recordings each. Table 2 defines the reference folds and Table 11 reports the full rotation. Because both source training and validation change, this experiment measures their joint allocation sensitivity.

Table 2: Four operating condition holdouts

Test	Validation	Training
1200	Variable	1800, 2400
1800	1200	2400, Variable
2400	1800	1200, Variable
Variable	2400	1200, 1800

Speeds are in r/min. Each row contains 32 training, 16 validation, and 16 test recordings. The preceding condition is defined cyclically in the order 1200, 1800, 2400, Variable.

Table 3: Reported network and training settings

Item	Setting
Node features / embedding	11 / 48 dimensions



Item	Setting
Graph blocks / pooled vector	2 / 144 dimensions
Outputs / objective	7 / weighted binary cross-entropy
CNN kernel sizes	9, 7, 5
Optimizer / initial learning rate	AdamW / 0.001
Batch size / maximum epochs	64 / 65
Early stopping patience	10 validation epochs
Training seeds	13, 29, 42, 71, 97
Training noise	Probability 0.5; 5–25 dB
Roll augmentation	Up to 128 samples

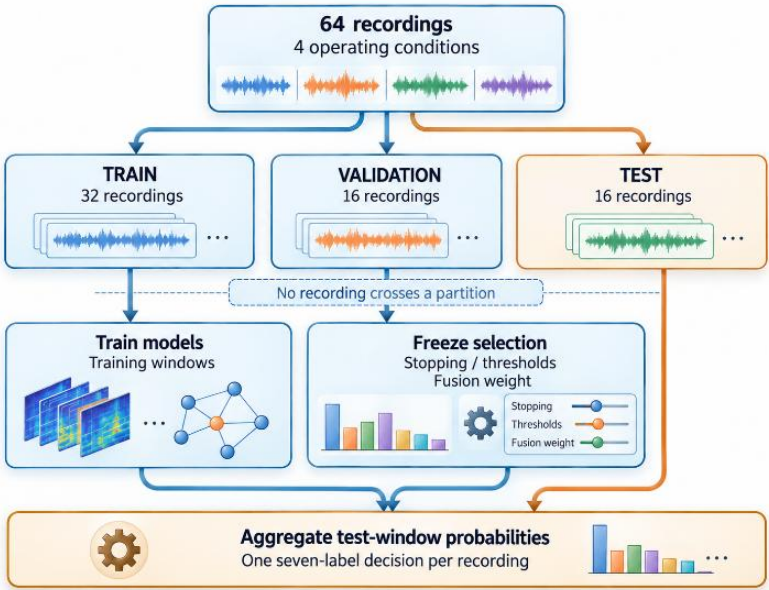
Window probabilities are averaged within each recording before applying validation-selected thresholds. Macro-F1 is the arithmetic mean of seven recording-level label F1 scores. Overall neural scores average the four conditions and five seeds; full-vector exact match measures recovery of all seven labels. Added-noise tests apply fixed Gaussian-noise realizations at 10, 5, 0, and −5 dB after normalization. Clean-validation thresholds and fusion weights remain frozen.

$$\begin{aligned}
 F_{1,k} &= \frac{2 \text{TP}_k}{2 \text{TP}_k + \text{FP}_k + \text{FN}_k}, \\
 \text{MacroF1} &= \frac{1}{7} \sum_{k=1}^7 F_{1,k}, \\
 \text{ExactMatch} &= \frac{1}{R} \sum_{r=1}^R \mathbb{1}\{\hat{\mathbf{y}}_r = \mathbf{y}_r\}.
 \end{aligned} \tag{9}$$

Exact match requires agreement of all seven binary predictions with a recording’s target vector. Macro-F1 gives each fault label equal weight. Each held-out condition has one positive recording for each of the rolling-element and cage labels, three for each gear label, and five for each inner- and outer-race label. These small supports make label scores sensitive to individual recording errors. The results are descriptive; no significance test treating the 20 fold–seed scores as independent observations is used.

TP, FP, and FN in Eq. (9) count recording-level outcomes within each 16-recording test fold. The overall standard deviation combines differences between operating conditions with training-seed variation. In Tables 7 and 8, within-condition standard deviations are computed from five reported seed scores using a divisor of four. Five fits reuse the same test recordings and do not add independent acquisitions.

The contribution of a label to one fold Macro-F1 is bounded by $\frac{1}{7}$; its contribution to the equally weighted four-condition mean is bounded by $\frac{1}{28}$. These bounds describe the averaging scheme, not the change caused by any single recording error. Small aggregate differences require interpretation alongside the per-label support and condition-specific results.



210

Figure 3: One full operating condition is held out for testing in each fold. Training, validation, and test partitions contain distinct recordings; all window predictions are returned to one decision per acquisition.

4 Results

4.1 Overall performance and condition effects

215 Table 4 reports overall recording-level performance. The fixed graph achieves Macro-F1 of 0.8284 and exact match of 0.6375, compared with 0.7954 and 0.5688 for self-loops. The Macro-F1 difference is 0.0330. The paired results show positive fixed-minus-self-loop differences in 14 of 20 fold–seed runs and positive means in all four conditions. Those repeated fits share test recordings, so the count is a descriptive comparison rather than 20 independent experimental replications.

Table 4: Overall recording level performance

Model	Macro-F1	Exact match
SVM	0.8849	0.8125
CNN	0.9708 ± 0.0206	0.9188
Self-loop graph	0.7954 ± 0.1837	0.5688
Fixed graph	0.8284 ± 0.1804	0.6375
CNN + graph gate	0.9713 ± 0.0261	0.9250
CNN + self-loop gate	0.9737 ± 0.0239	0.9281



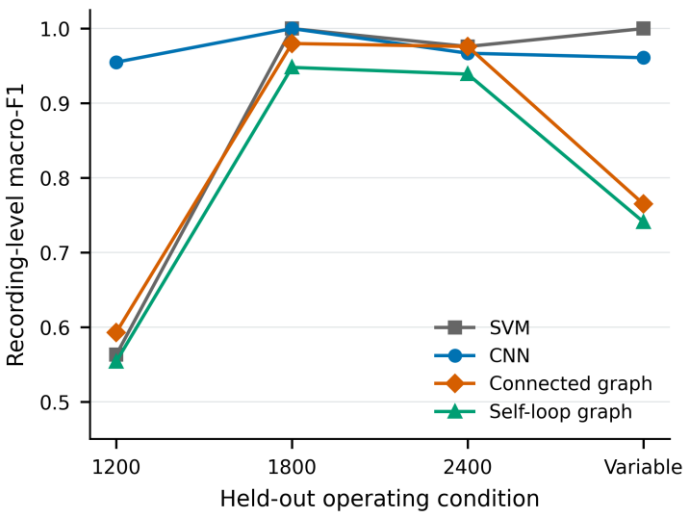
220 Neural results: mean and standard deviation across four test conditions $\times$ five seeds. SVM: four fold fits; no seed standard deviation is reported. Exact match is the fraction of fully correct seven-label vectors.

The CNN achieves the highest standalone mean Macro-F1, 0.9708. CNN–fixed-graph fusion reaches 0.9713, a difference of 0.0005, while CNN–self-loop fusion reaches 0.9737. Thus the fixed graph improves its matched graph comparator, but the fusion results do not show a topology-specific advantage in the highest-scoring overall classifier.

225 **Table 5: Macro F1 by held out condition**

Test	SVM	CNN	Fixed graph	Self-loop
1200	0.563	0.955	0.593	0.554
1800	1.000	1.000	0.980	0.948
2400	0.976	0.967	0.976	0.939
Variable	1.000	0.961	0.765	0.741

Neural entries average five seeds and are rounded to three decimals. Small differences reconstructed from this table may differ from differences calculated before rounding.



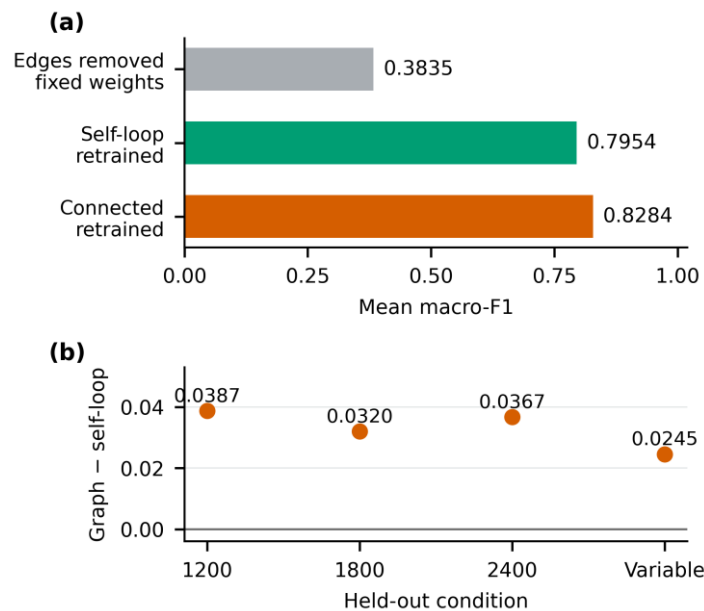
230 **Figure 4: Recording-level Macro-F1 across held-out operating conditions. The connected graph exceeds the matched self-loop mean in each condition, but the CNN is stronger at 1200 r/min and under variable speed. Lines connect categorical conditions only to aid reading.**

The fixed graph achieves 0.976 at 2400 r/min, compared with 0.967 for the CNN, using the rounded condition means in Table 5. The CNN scores higher at 1200 r/min and under variable speed. The SVM reaches 1.000 at 1800 r/min and variable speed but scores 0.563 at 1200 r/min. Overall means therefore conceal substantial differences between conditions.



235 **4.2 Retrained topology control and edge intervention**

A fixed-weight adjacency intervention and independently retrained controls answer different questions. Replacing the trained fixed graph’s adjacency with the identity reduces Macro-F1 from 0.8284 to 0.3835. Independently training the self-loop model gives 0.7954. The intervention measures the trained model’s sensitivity to an adjacency change, while the retrained contrast compares classifiers learned with different message neighborhoods. The reported fixed-minus-self-loop means are 0.0387, 0.0320, 0.0367, and 0.0245 for 1200, 1800, and 2400 r/min and variable speed, respectively; six seed-level differences are negative.



245 **Figure 5: Two distinct topology comparisons. (a) Removing edges with connected-model weights fixed versus independently retrained self-loop and connected models. (b) Connected-minus-self-loop fold means calculated before table rounding. Markers show fold-mean contrasts.**

The retrained Macro-F1 difference of 0.0330 corresponds to 3.30 percentage points on the 0–100 score scale, or 4.15% relative to the self-loop mean of 0.7954. This comparison quantifies the observed benefit of the fixed adjacency under the reported training protocol. The larger fixed-weight intervention effect is not an estimate of the gain over a retrained classifier. Table 6 and Fig. 6 compare topology and descriptor variants. The fixed graph exceeds the frequency chain (0.8110), same-band channel graph (0.8070), and random graph (0.8150) by 0.0174, 0.0214, and 0.0134, respectively. It has the highest aggregate score among the listed fixed-topology variants. The random comparison supports an observed advantage for the prescribed edge organization, subject to the implemented degree-matching procedure.



Table 6: Topology and descriptor comparison

Variant	Macro-F1
Self-loop control	0.7954
Frequency chain	0.8110
Same-band channel	0.8070
Degree-matched random	0.8150 ± 0.1721
Complete fixed graph	0.8284
Uniform bands	0.8080
No local segments	0.8160
Adaptive graph	0.8420 ± 0.1598

Random and adaptive entries report mean ± sample SD across 20 condition–seed scores; these SDs combine operating-
condition differences and training variability. Other variants report aggregate means without standard deviations.
Uniform bands reduce Macro-F1 from 0.8284 to 0.8080, and removal of local descriptors reduces it to 0.8160. These
differences describe the evaluated frequency partition and descriptor choices without identifying a mechanical cause. The
adaptive graph achieves 0.8420, exceeding the fixed graph by 0.0136. The condition-specific results in Tables 7 and 8 show
the corresponding holdout scores; aggregate performance alone does not establish significance or equivalent model cost.

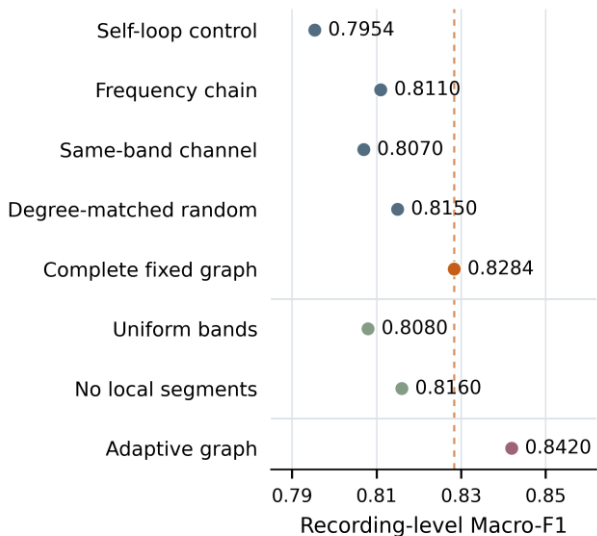


Figure 6: Recording-level Macro-F1 for topology and descriptor variants. The dashed line marks the complete fixed graph. The horizontal scale is restricted to show score differences; points represent aggregate scores without uncertainty estimates.



4.3 Condition specific random and adaptive graph results

265 Table 7 gives the five seed scores for each random and adaptive graph holdout. Table 8 summarizes their within-condition means and sample standard deviations alongside the fixed-graph condition means from Table 5. Recalculation from the reported four-decimal seed values gives overall means of 0.81503 for the random graph and 0.841975 for the adaptive graph, reproducing the four-decimal aggregate entries in Table 6. These additional data resolve condition and training variability without increasing the number of test recordings.

Table 7: Per seed recording level Macro F1

Graph	Test	Seed 13	Seed 29	Seed 42	Seed 71	Seed 97	Mean	SD
Random	1200	0.5410	0.6010	0.5650	0.5800	0.5880	0.5750	0.0231
Random	1800	0.9637	1.0000	0.9714	0.9670	0.9632	0.9731	0.0154
Random	2400	0.9524	0.9841	0.9740	0.9558	0.9740	0.9681	0.0134
Random	Variable	0.7030	0.7630	0.7410	0.7800	0.7330	0.7440	0.0294
Adaptive	1200	0.5890	0.6550	0.6000	0.6400	0.6010	0.6170	0.0287
Adaptive	1800	1.0000	1.0000	0.9870	0.9643	0.9870	0.9877	0.0146
Adaptive	2400	1.0000	1.0000	0.9762	0.9610	0.9740	0.9822	0.0172
Adaptive	Variable	0.7380	0.8180	0.7710	0.8020	0.7760	0.7810	0.0308

270 Speeds are in r/min. SD is the sample standard deviation of the five reported seed scores. Means and SDs are recalculated from the reported four-decimal scores. Each set of five scores evaluates the same 16 test recordings.

Table 8: Graph comparison by held out condition

Test	Random mean $\pm$ SD	Fixed mean	Adaptive mean $\pm$ SD
1200	0.5750 $\pm$ 0.0231	0.593	0.6170 $\pm$ 0.0287
1800	0.9731 $\pm$ 0.0154	0.980	0.9877 $\pm$ 0.0146
2400	0.9681 $\pm$ 0.0134	0.976	0.9822 $\pm$ 0.0172
Variable	0.7440 $\pm$ 0.0294	0.765	0.7810 $\pm$ 0.0308

Fixed means are retained at the three-decimal precision in Table 5. Fixed-graph seed scores were not supplied, so no fixed-graph seed SD or paired seedwise comparison is reconstructed.

275 Adaptive-minus-random mean differences are 0.0420 at 1200 r/min, 0.0146 at 1800 r/min, 0.0142 at 2400 r/min, and 0.0370 under variable speed. Against the displayed fixed means, adaptive improvements are approximately 0.024, 0.008, 0.006, and



0.016, respectively; their precision is limited by the rounded fixed values. All three graphs have their lowest condition mean at 1200 r/min. The graph construction comparison therefore retains the low-speed deficit.

Adaptive within-condition SDs are 0.0287, 0.0146, 0.0172, and 0.0308, compared with 0.0231, 0.0154, 0.0134, and 0.0294 for random connectivity. The adaptive graph has a higher mean in each condition, but a smaller SD only at 1800 r/min. These results do not establish improved seed stability. No significance test or claim of equal model complexity is inferred from the seed table.

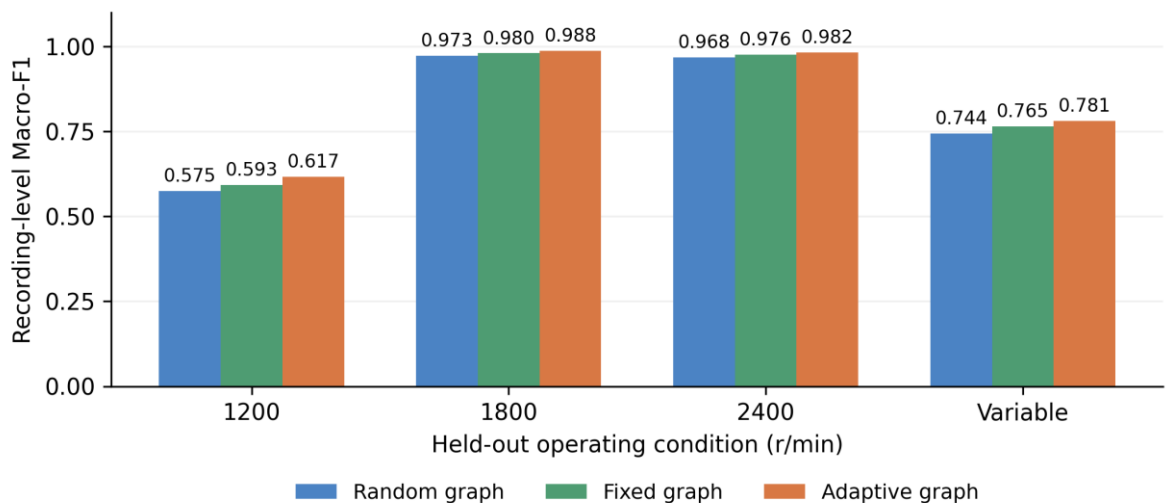


Figure 7: Mean recording-level Macro-F1 for random, fixed, and adaptive graphs. Random and adaptive bars use recalculated seed means from Table 7; fixed bars use the rounded means in Table 5. The ordinate starts at zero. Bars show means only; within-condition SDs are reported in Table 8.

4.4 Compound fault and component analysis

Table 9: Exact match on compound fault recordings

Model	Compound exact match
CNN	0.9071
Self-loop graph	0.5000
Fixed graph	0.5571
CNN + graph gate	0.9143
CNN + self-loop gate	0.9143

Seven compound recordings per condition give 28 distinct acquisitions, evaluated repeatedly across five seeds. The 140 fold–seed decisions do not represent 140 independent compound specimens.



Compound exact match increases from 0.5000 for self-loops to 0.5571 for the fixed graph (Table 9). The CNN achieves 0.9071 and both fusion variants reach 0.9143. These results quantify recovery of every constituent fault label within compound recordings. The CNN has substantially higher complete-label recovery than either standalone graph.

Table 10: Per label F1 and positive support

Fault	CNN	Fixed graph	Self-loop	n ⁺
Inner race	0.943	0.863	0.802	5
Outer race	0.931	0.823	0.838	5
Rolling element	1.000	0.903	0.778	1
Cage	1.000	0.833	0.853	1
Gear wear	0.993	0.836	0.841	3
Broken tooth	0.950	0.706	0.644	3
Eccentric gear	0.979	0.834	0.810	3

295 F1 values average the fold–seed evaluations. n⁺ is positive recording support in each held-out condition, not the number of windows.

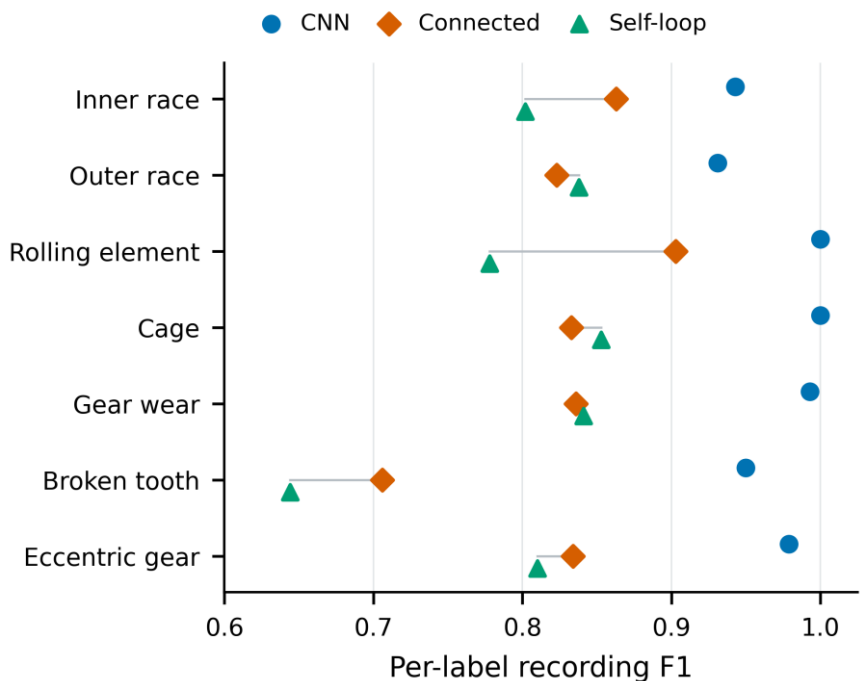


Figure 8: Per-label mean recording F1. Gray segments connect connected and self-loop values. Both positive and negative topology effects occur; rolling-element and cage labels each have one positive recording per test condition.



300 Table 10 shows that the topology contrast varies by fault label. Fixed-minus-self-loop mean F1 differences are +0.125 for rolling element, +0.062 for broken tooth, +0.061 for inner race, and +0.024 for eccentric gear. Differences are −0.015 for outer race, −0.020 for cage, and −0.005 for gear wear. These are differences between rounded table entries. Rolling-element and cage scores each depend on one positive recording per test condition, so they cannot establish a general label-specific advantage across independent specimens.

305 At 1200 r/min, the fixed graph has a condition mean of 0.593, compared with 0.554 for self-loops and 0.955 for the CNN. At 2400 r/min it reaches 0.976, and under variable speed it reaches 0.765. The positive condition means against self-loops therefore coexist with weak absolute performance in two holdouts. Fig. 8 presents the component-level comparison.

4.5 Sensitivity to training and validation allocation

Table 11 and Fig. 9 report all 12 training–validation allocations for the fixed graph. With 1200 r/min held out, Macro-F1 ranges from 0.582 to 0.608 and averages 0.5943 across the three allocation scores. The corresponding allocation means are 0.9717 at 1800 r/min, 0.9690 at 2400 r/min, and 0.7600 under variable speed. Maximum–minimum ranges are 0.026, 0.017, 0.014, and 0.031, respectively. These summaries are calculated from the displayed three-decimal values.

310

Table 11: Complete fixed graph under validation rotation

Test	Val.	Training	Macro-F1
1200	1800	2400, Var.	0.608
1200	2400	1800, Var.	0.582
1200	Var.	1800, 2400	0.593
1800	1200	2400, Var.	0.980
1800	2400	1200, Var.	0.972
1800	Var.	1200, 2400	0.963
2400	1200	1800, Var.	0.969
2400	1800	1200, Var.	0.976
2400	Var.	1200, 1800	0.962
Var.	1200	1800, 2400	0.742
Var.	1800	1200, 2400	0.773
Var.	2400	1200, 1800	0.765



Speeds are in r/min; Var. denotes variable speed. Each test condition is excluded from training and validation. The reference
 315 allocations are 1200/Var., 1800/1200, 2400/1800, and Var./2400 for test/validation, respectively. Their rounded scores
 reproduce the graph column of Table 5.

Every 1800- and 2400-r/min allocation achieves at least 0.962, while the highest 1200-r/min score is 0.608. Variable-speed
 scores range from 0.742 to 0.773. The low-speed deficit persists across all source allocations. Variable speed has the largest
 observed allocation range. This experiment cannot attribute score changes solely to validation-based model selection because
 320 the training pair changes at the same time.

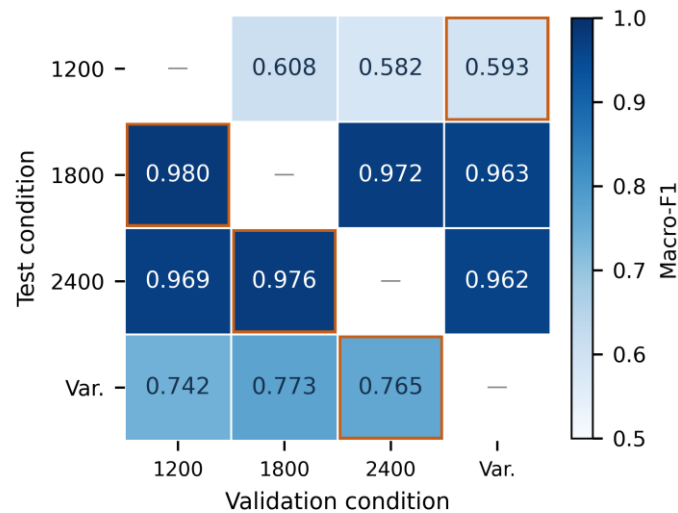
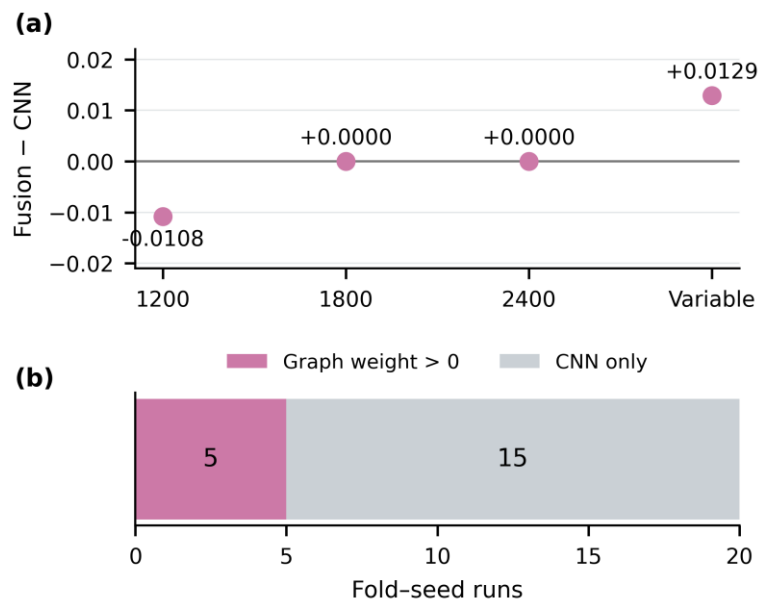


Figure 9: Complete fixed graph Macro-F1 under validation rotation. Rows identify the test condition and columns the validation condition; the other two conditions are used for training. Diagonal cells are excluded. Orange outlines mark the reference allocations in Table 2. All three entries in a row use the same held-out test recordings.

325 4.6 Validation gated fusion

The validation gate selects nonzero fixed-graph weight in 5 of 20 runs and $\alpha = 0$ in the remaining 15. When $\alpha = 0$,
 inference uses CNN logits alone. Mean fusion Macro-F1 is 0.0005 above the CNN and 0.0024 below CNN–self-loop fusion.
 Both performance and gate-selection frequency are therefore needed to interpret the fused system.

CNN–graph fusion and CNN have identical reported means at 1800 and 2400 r/min. Under variable speed, fusion averages
 330 0.9738 versus 0.9609 for CNN; at 1200 r/min it averages 0.9443 versus 0.9551. Fig. 10 shows these condition contrasts and
 the aggregate gate-selection count. Seed-level covariance is not available for these fusion summaries. Moreover, the higher
 overall CNN–self-loop fusion score prevents attribution of the variable-speed improvement specifically to the fixed graph’s
 inter-node edges.



335 **Figure 10: Validation-gated fusion. (a) Fold-mean Macro-F1 difference from CNN, using the reported four-decimal values. (b)**
340 **Nonzero graph weight is selected in 5 of 20 runs; the other 15 use CNN alone. The count is not a per-condition breakdown.**

4.7 Sensitivity to added test noise

At 10 and 5 dB, the fixed graph achieves Macro-F1 of 0.8169 and 0.7409, compared with 0.8012 and 0.7222 for self-loops and 0.9719 and 0.9637 for the CNN. At 0 dB the graph scores 0.2632 and CNN 0.6749; at -5 dB they score 0.1286 and 0.1604.
340 CNN-graph fusion scores 0.6185 at 0 dB, below the CNN. Table 12 reports the rounded results. These tests show substantial degradation at the two lowest SNRs and do not support a claim of strong-noise robustness.

Table 12: Macro F1 under added Gaussian test noise

Model	10 dB	5 dB	0 dB	-5 dB
CNN	0.972	0.964	0.675	0.160
Fixed graph	0.817	0.741	0.263	0.129
Self-loop	0.801	0.722	0.281	0.149
CNN + graph gate	0.973	0.959	0.619	0.156

Scores average four conditions × five seeds for the learned clean checkpoints. Thresholds and fusion weights remain fixed. Table values are rounded to three decimals.

345 The noise results reuse learned checkpoints and average four test conditions and five seeds. Training augmentation spans 5–25 dB, while the two lowest test SNRs lie outside that range. Fig. 11 includes the clean scores for reference.

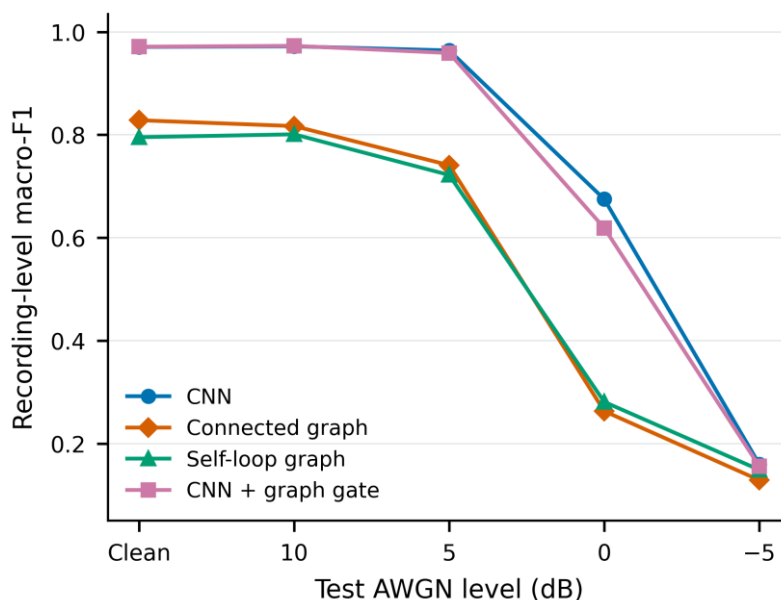


Figure 11: Noise sensitivity with clean checkpoints and fixed decision settings. Clean scores are added as a reference; noisy points reproduce Table 12. The noise axis is categorical and does not treat Clean as a numerical SNR.

350 5 Discussion

The independently trained fixed graph improves Macro-F1 by 0.0330 over self-loops. Frequency-only and channel-only scores are lower than the combined fixed topology, and the random graph is lower by 0.0134. These comparisons quantify the contribution of the declared adjacency in this representation. The fixed-weight intervention additionally demonstrates sensitivity to changing adjacency after training, but its larger score reduction cannot be treated as a retrained ablation effect.

355 The condition-specific five-seed results qualify the topology comparison. The adaptive graph exceeds the fixed graph's displayed mean in each condition and its overall mean by 0.0136. It also exceeds the random graph in 17 of the 20 reported condition–seed comparisons, with two ties and one negative difference. Its within-condition standard deviations are higher than the random graph's in three of the four conditions. Thus the data show higher average adaptive scores, not uniformly lower training variability. Fixed-graph seed scores are unavailable for a matched seedwise comparison with these two graph
 360 variants.

Descriptor ablations reduce Macro-F1 by 0.0204 with uniform bands and by 0.0124 without local descriptors. These results support the evaluated descriptor design at the aggregate level. They do not isolate the cause of the 1200-r/min deficit or measure the representation of a specific transient fault frequency. Allocation rotation retains the low-speed deficit, and changing graph construction does not remove the condition gap.

365 Allocation sensitivity is smaller than the observed low-speed versus higher-speed contrast: within-test ranges are 0.014–0.031, while all three 1200-r/min allocations remain below all 1800- and 2400-r/min allocations. Because training and validation



change jointly, these ranges describe source allocation rather than validation alone. The rotation experiment covers the fixed graph only and does not establish baseline rankings or fusion stability across allocations.

370 The conclusions are limited to 64 acquisitions from one public experimental setting. Evaluated fault combinations are represented in training at other speeds, and five training seeds reuse the same test acquisitions. Architecture and band selection followed earlier inspection of this dataset, so the study is exploratory rather than a prospective evaluation of a frozen design. The waveform CNN remains stronger overall, and the 0.0005 fusion difference does not establish a practical topology-specific gain.

375 The variable-speed trajectories span speeds contained in the stationary conditions. The present tests nevertheless hold out the variable-speed recordings as a distinct condition. Recording-level separation excludes window reuse between partitions, but the protocol does not test independent machine or specimen transfer. Mechanical coupling, edge-level causality, and deployment performance are outside the reported measurements.

6 Conclusions

380 A fixed 30-node channel–frequency graph was evaluated for seven-label compound gearbox diagnosis with recording-disjoint operating-condition holdouts. It achieved mean Macro-F1 of 0.8284, exceeding self-loops by 0.0330 and the random graph by 0.0134. Five-seed random and adaptive data reproduce aggregate means of 0.8150 and 0.8420. Adaptive means exceed random and fixed means in each condition, but do not show uniformly lower seed variability. Validation rotation retains the 1200-r/min deficit, and all tested neural systems deteriorate at severe added-noise levels. The waveform CNN achieves 0.9708 and CNN–graph fusion 0.9713. The fixed topology therefore improves its matched graph representation on this dataset, while the 385 adaptive comparator and waveform CNN achieve higher overall scores. Independent machine and specimen evaluations are required to assess broader transfer.

Data availability

The vibration recordings are publicly available in the Mendeley Data archive (Hou, 2025), with acquisition details in (Hou et al., 2025).

390 Author contributions

Junwei Shen: Writing–original draft, Writing–review & editing, Methodology, Software.

Competing interests

The authors declare that they have no conflict of interest.



The authors have no relevant financial or non-financial interests to disclose.

395 **Ethical approval**

This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent

This article does not contain any studies with human participants. So informed consent is not applicable here.

Financial support

400 No funding.

References

- Cao, L., Wang, Y., Li, J., Qian, Z., Meng, Z., and Liu, M.: MT-CDGAT: A multi-label diagnosis model for untrained planetary gearbox compound faults based on multi-task cross dynamic graph attention networks, *Neurocomputing*, 639, 130307, <https://doi.org/10.1016/j.neucom.2025.130307>, 2025.
- 405 Chen, Z., He, G., Li, J., Liao, Y., Gryllias, K., and Li, W.: Domain adversarial transfer network for cross-domain fault diagnosis of rotary machinery, *IEEE Trans. Instrum. Meas.*, 69, 8702–8712, <https://doi.org/10.1109/TIM.2020.2995441>, 2020.
- Gao, T., Yang, J., Wang, W., and Fan, X.: A domain feature decoupling network for rotating machinery fault diagnosis under unseen operating conditions, *Reliab. Eng. Syst. Saf.*, 252, 110449, <https://doi.org/10.1016/j.ress.2024.110449>, 2024.
- Hamilton, W. L., Ying, R., and Leskovec, J.: Inductive representation learning on large graphs, *Advances in Neural Information Processing Systems*, 30, <https://arxiv.org/abs/1706.02216>, 2017.
- 410 Hou, Y., Tang, J., Zhang, J., Pang, P., Zhang, H., and Li, X.: Mixed fault dataset of bearings and gearboxes under variable operating conditions, *Data Brief*, 63, 112187, <https://doi.org/10.1016/j.dib.2025.112187>, 2025.
- Hou, Y.: Mixed fault dataset of bearings and gearboxes under variable operating conditions, *Mendeley Data*, V1 [data set], <https://doi.org/10.17632/yrms2v89p7.1>, 2025.
- 415 Huang, R., Li, J., Liao, Y., Chen, J., Wang, Z., and Li, W.: Deep adversarial capsule network for compound fault diagnosis of machinery toward multidomain generalization task, *IEEE Trans. Instrum. Meas.*, 70, 1–11, <https://doi.org/10.1109/TIM.2020.3042300>, 2021.
- Kipf, T. N. and Welling, M.: Semi-supervised classification with graph convolutional networks, *Proc. Int. Conf. Learn. Represent. (ICLR)*, <https://arxiv.org/abs/1609.02907>, 2017.



- 420 Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., and Nandi, A. K.: Applications of machine learning to machine fault diagnosis: A review and roadmap, *Mech. Syst. Signal Process.*, 138, 106587, <https://doi.org/10.1016/j.ymssp.2019.106587>, 2020.
- Li, J., Cao, X., Chen, R., Zhang, X., Huang, X., and Qu, Y.: Graph neural network architecture search for rotating machinery fault diagnosis based on reinforcement learning, *Mech. Syst. Signal Process.*, 202, 110701, <https://doi.org/10.1016/j.ymssp.2023.110701>, 2023.
- 425 Li, T., Zhou, Z., Li, S., Sun, C., Yan, R., and Chen, X.: The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study, *Mech. Syst. Signal Process.*, 168, 108653, <https://doi.org/10.1016/j.ymssp.2021.108653>, 2022.
- Li, W., Lan, H., Chen, J., Feng, K., and Huang, R.: WavCapsNet: An interpretable intelligent compound fault diagnosis method by backward tracking, *IEEE Trans. Instrum. Meas.*, 72, 1–11, <https://doi.org/10.1109/TIM.2023.3282664>, 2023.
- 430 Liao, Y., Huang, R., Li, J., Chen, Z., and Li, W.: Deep semisupervised domain generalization network for rotary machinery fault diagnosis under variable speed, *IEEE Trans. Instrum. Meas.*, 69, 8064–8075, <https://doi.org/10.1109/TIM.2020.2992829>, 2020.
- Liu, R., Yang, B., Zio, E., and Chen, X.: Artificial intelligence for fault diagnosis of rotating machinery: A review, *Mech. Syst. Signal Process.*, 108, 33–47, <https://doi.org/10.1016/j.ymssp.2018.02.016>, 2018.
- 435 Matania, O., Dattner, I., Bortman, J., Kenett, R. S., and Parmet, Y.: A systematic literature review of deep learning for vibration-based fault diagnosis of critical rotating machinery: Limitations and challenges, *J. Sound Vib.*, 590, 118562, <https://doi.org/10.1016/j.jsv.2024.118562>, 2024.
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y.: Graph attention networks, *Proc. Int. Conf. Learn. Represent. (ICLR)*, <https://arxiv.org/abs/1710.10903>, 2018.
- 440 Wang, C., Wang, Y., Wang, Y., Li, X., and Chen, Z.: Richly connected spatial–temporal graph neural network for rotating machinery fault diagnosis with multi-sensor information fusion, *Mech. Syst. Signal Process.*, 225, 112230, <https://doi.org/10.1016/j.ymssp.2024.112230>, 2025.
- Wang, Z., Xuan, J., Shi, T., and Lin, Y.: Multi-label domain adversarial reinforcement learning for unsupervised compound fault recognition, *Reliab. Eng. Syst. Saf.*, 254, 110638, <https://doi.org/10.1016/j.ress.2024.110638>, 2025.
- 445 Wen, L., Li, X., Gao, L., and Zhang, Y.: A new convolutional neural network-based data-driven fault diagnosis method, *IEEE Trans. Ind. Electron.*, 65, 5990–5998, <https://doi.org/10.1109/TIE.2017.2774777>, 2018.
- Xiao, X., Li, C., He, H., Huang, J., and Yu, T.: Rotating machinery fault diagnosis method based on multi-level fusion framework of multi-sensor information, *Inf. Fusion*, 113, 102621, <https://doi.org/10.1016/j.inffus.2024.102621>, 2025.
- Xiao, Y., Shao, H., Yan, S., Wang, J., Peng, Y., and Liu, B.: Domain generalization for rotating machinery fault diagnosis: A survey, *Adv. Eng. Informat.*, 64, 103063, <https://doi.org/10.1016/j.aei.2024.103063>, 2025.
- 450 Xing, S., Lei, Y., Wang, S., Lu, N., and Li, N.: A label description space embedded model for zero-shot intelligent diagnosis of mechanical compound faults, *Mech. Syst. Signal Process.*, 162, 108036, <https://doi.org/10.1016/j.ymssp.2021.108036>, 2022.



- Xu, K., Hu, W., Leskovec, J., and Jegelka, S.: How powerful are graph neural networks?, Proc. Int. Conf. Learn. Represent. (ICLR), <https://arxiv.org/abs/1810.00826>, 2019.
- 455 Yu, X., Tang, B., and Deng, L.: Fault diagnosis of rotating machinery based on graph weighted reinforcement networks under small samples and strong noise, *Mech. Syst. Signal Process.*, 186, 109848, <https://doi.org/10.1016/j.ymssp.2022.109848>, 2023.
- Zhang, G., Kong, X., Wang, Q., Du, J., Wang, J., and Ma, H.: Single domain generalization method based on anti-causal learning for rotating machinery fault diagnosis, *Reliab. Eng. Syst. Saf.*, 250, 110252, <https://doi.org/10.1016/j.ress.2024.110252>, 2024.
- 460 Zhao, C., Zio, E., and Shen, W.: Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study, *Reliab. Eng. Syst. Saf.*, 245, 109964, <https://doi.org/10.1016/j.ress.2024.109964>, 2024.
- Zhao, X., Yao, J., Deng, W., Ding, P., Zhuang, J., and Liu, Z.: Multiscale deep graph convolutional networks for intelligent fault diagnosis of rotor-bearing system under fluctuating working conditions, *IEEE Trans. Ind. Informat.*, 19, 166–176, <https://doi.org/10.1109/TII.2022.3161674>, 2023.