Response to Reviewer

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

We sincerely thank the reviewer for the constructive and technically detailed comments. We respond point by point below, using the preserved experiments and a supplementary evidence package containing the recording manifests, paired model outputs and implementation records. We distinguish observations within the present public benchmark from confirmation on new acquisitions, and restrict the claims accordingly.

General assessment

The manuscript evaluates a fixed 30-node channel-frequency graph for cross-speed compound fault diagnosis of gearboxes using recording-disjoint operating-condition holdouts. The independently retrained self-loop control and additional topology comparisons provide a useful basis for assessing graph connectivity. The authors also appropriately acknowledge the stronger overall performance of the waveform CNN and the limitations of the evaluation. However, the independence of design selection, reproducibility of the comparison models, and reliability of the reported topology gains require further evidence. I therefore recommend major revision before the manuscript can be considered for publication. My specific comments are as follows.

Response

We appreciate the recognition of the recording-disjoint evaluation and the independently retrained self-loop control. We agree that development-stage choices, implementation transparency and acquisition-level uncertainty are central to interpreting the results. The most directly auditable topology comparison in this preprint is the fixed graph versus its matched identity-adjacency control; stronger waveform-CNN performance and the limits of other graph comparisons are explicitly acknowledged.

Comment 1

1. The discussion on p. 21 states that architecture and frequency-band selection followed earlier inspection of this dataset. Please describe the data, performance measures, and tuning decisions involved, and clarify whether the held-out conditions influenced design selection. Recording-disjoint splits alone do not exclude bias introduced during model development. Please freeze the design and confirm the topology gains using data not used for development, clearly distinguishing exploratory from confirmatory results.

Response

Thank you for highlighting the distinction between a recording-disjoint evaluation and a genuinely untouched development test. The method was developed iteratively using the same public collection of 64 acquisitions, comprising four operating conditions. Earlier trials included a CNN with a graph-residual branch (v2); their recording-level Macro-F1 results were inspected before the independently trained 30-node graph, full-window/local descriptors and matched self-loop control were specified for v3. The fixed-Hz band partition was also retained after previous inspection of this dataset. Thus, although v3 uses strict recording-disjoint training/validation/test partitions (32/16/16 acquisitions), its four held-out operating conditions cannot be described as completely unseen during method development.

The preserved records establish the transition from v2 to v3 but do not contain an exhaustive log of every rejected frequency-band boundary, architecture or hyperparameter choice. We therefore do not reconstruct an unsupported optimization history. The v3 definition and settings are frozen in the accompanying configuration and source archive. Its fixed-versus-self-loop gain is presented as an exploratory, within-benchmark observation—not confirmation on prospectively reserved acquisitions. Validation on a genuinely independent dataset or new machine/specimens would be necessary to establish a confirmatory gain, and is not claimed here.

Comment 2

2. Sections 3.4-3.6 do not sufficiently specify the random and adaptive graphs. Please provide the degree-preserving randomization algorithm, number of graph realizations, and equations for generating and normalizing the adaptive adjacency matrix. The residual-block operation order, loss weights, dropout rate, and threshold-search procedure also need clarification. Providing executable code, configurations, and recording-level split manifests would allow the central comparisons in Tables 6-8 to be reproduced.

Response

For the reproducible primary comparison, each three-channel window forms 30 channel–frequency nodes. Fixed edges connect consecutive frequency bands within each channel and all three channels at the same band, in addition to self-loops (57 non-self undirected edges). The graph operator uses row normalization, $\bar{A}_{ij} = A_{ij} / \sum_j A_{ij}$. Each of two residual blocks evaluates H = GELU(Wself X + Wneigh $\bar{A}$X + b), followed by Xnext = LayerNorm[X + Dropout(H)]; the dropout probability is 0.20. The identity control uses the same trainable structure, seeds and training protocol with $\bar{A}$ = I.

Training minimizes BCEWithLogitsLoss with a positive-label weight $w_k = \mathrm{clip}[(N - N_k^+)/N_k^+, 1, 10]$ calculated from the training split; negative-label weights are one. AdamW uses learning rate 0.001 and weight decay 10^{-4}, batch size 64 and at most 65 epochs, with ten-epoch early stopping on validation loss. A single decision threshold for all seven labels is selected by validation recording-level Macro-F1 from {0.3, 0.4, 0.5, 0.6, 0.7}; ties favor the candidate closest to 0.5. Fusion weights are chosen from {0, 0.25, 0.50, 0.75, 1.00}, accepting a nonzero graph contribution only for a validation gain greater than 0.01. The executable v3 code, configuration, recording-level split manifest and predictions are supplied with the response.

Regarding the additional random/adaptive graphs in Tables 6–8, the surviving projects preserve reported score summaries but not the original degree-preserving rewiring procedure, graph-realization manifest or adaptive-adjacency implementation. A precise realization count or adaptive equation cannot be authenticated retrospectively. We therefore treat these comparisons as descriptive context, rather than asserting that their exact construction is reproducible from the surviving files. The central auditable topology finding remains the independently retrained fixed-versus-self-loop comparison.

Comment 3

3. The reported difference in mean Macro-F1 between the fixed and random graphs is 0.0134, but Table 8 does not provide the fixed graph's scores for individual training seeds. Please supply paired results and recording-level predictions for the fixed graph and its controls. Assess training variability separately from variability across random graph realizations, and quantify uncertainty in the performance differences while preserving the recording-level dependence structure; windows and training seeds are not additional independent acquisitions.

Response

We recovered the complete four-condition × five-seed paired results for the fixed and self-loop models, with every test recording's seven target labels, predicted probabilities and selected threshold (Supplementary Files S3–S4). Their mean recording-level Macro-F1 values are 0.8284 and 0.7954, respectively; the mean paired difference is +0.0330. This difference is positive in all four condition averages and in 14/20 fold–seed comparisons, but negative in six seedwise comparisons. The full paired results are provided rather than inferred from rounded table means.

Test condition	Fixed graph	Self-loop	Paired difference
1200 r/min	0.5929	0.5542	+0.0387
1800 r/min	0.9798	0.9479	+0.0320
2400 r/min	0.9756	0.9389	+0.0367
Variable speed	0.7653	0.7408	+0.0245
Overall	0.8284	0.7954	+0.0330

A descriptive 10,000-replicate paired bootstrap, resampling the 16 recordings within each held-out condition while keeping the same selected recordings across models and seeds, gives a conditional 95% percentile interval of approximately [0.0007, 0.0510] for this fixed-minus-self-loop mean difference (S9). Because rare labels have only one positive recording in some conditions, and the same acquisitions are reused across seeds, this interval is interpreted cautiously and is not evidence of transfer to new machines or specimens.

For the fixed-versus-random difference of 0.0134, the original random-graph per-realization predictions and realization identifiers are unavailable. Consequently, its training-seed and graph-realization variability cannot be separated or assigned a justified paired interval. We do not treat this small difference as a statistically verified advantage of the prescribed edges.

Comment 4

4. The 12 training-validation allocations in Table 11 and Fig. 9 cover only the fixed graph. They therefore do not establish whether its advantage over the controls depends on the reference split. Please evaluate the self-loop and random graph controls under the same allocations and report paired performance differences. Also report the decision thresholds and fusion weights selected in each fold to assess selection stability with the small validation sets.

Response

We agree with the interpretation of Table 11: rotating the two source conditions used for training and validation measures allocation sensitivity of the fixed graph, not allocation-independent superiority over the controls. Across the three allocations for each test condition, fixed-graph Macro-F1 ranges are 0.582–0.608 (1200 r/min), 0.963–0.980 (1800 r/min), 0.962–0.976 (2400 r/min) and 0.742–0.773 (variable speed). These data support the persistence of the low-speed deficit; they do not establish that the fixed graph outranks self-loop or random connectivity in all twelve configurations. Matched control runs for those twelve allocations were not preserved, so such a claim is not made.

The archived v3 outputs do contain all validation-selected thresholds and fusion weights for the twenty reference-fold seed runs (S3). The fixed-graph fusion assigns zero graph weight in 15/20 runs, a weight of 0.25 in 3/20 and 0.50 in 2/20. The self-loop fusion similarly assigns zero weight in 15/20 runs. The model thresholds span 0.3–0.7 across folds and seeds. These reported choices make the limited stability of selection visible rather than implying that the 16-recording validation sets yield fixed decisions.

Comment 5

5. The frequency-band boundaries in Section 3.2 need stronger justification, particularly because three narrow bands contain only one local FFT bin each. Table 5 reports fixed-graph Macro-F1 values of 0.593 at 1200 r/min and 0.765 under variable speed. Please examine representative spectra and misclassified recordings to explain these deficits. An order-aligned comparison or a band-boundary perturbation study would help establish the operating range over which this representation is effective.

Response

The boundaries are 0, 64, 96, 128, 160, 256, 512, 1000, 2000, 4000 and 8000 Hz. Their purpose is a reproducible, nonuniform fixed-frequency partition with finer low-frequency coverage, not a claim that the bands track gear or bearing orders under changing speed. At 16 kHz, the full 4096-sample FFT has 3.90625-Hz spacing, whereas each 512-sample local FFT has 31.25-Hz spacing; consequently, the three narrow bands mentioned by the reviewer each contain only one local bin. The full-window descriptors retain finer resolution, but the local descriptors have no within-band frequency resolution in those intervals.

We examined the recovered recording-level errors and representative spectra from the same inner-race-plus-broken-tooth state at three operating conditions (S8 and S10). The fixed graph correctly recovered the full label vector in 0/5 fits at 1200 r/min, 5/5 at 2400 r/min and 0/5 under variable speed. Across all labels in the 1200-r/min fold, outer-race false positives were especially frequent (29 of 80 repeated recording–seed decisions); the variable-speed fold likewise showed outer-race and gear-label confusions. The illustrative spectra differ across conditions after identical normalization, supporting a plausible sensitivity of fixed-Hz representations to operating-condition shifts, but not proving a unique cause.

An order-aligned comparison or a systematic perturbation of the band boundaries is not part of the archived experiments; accordingly, we make no claim of demonstrated speed/order invariance. The supported operating range is confined to the four evaluated conditions, with explicitly weaker results at 1200 r/min and variable speed.

Comment 6

6. Tables 9-10 mainly report aggregate exact-match and per-label F1 scores, while rolling-element and cage faults each have only one positive recording per condition. Please provide the mapping from the 64 recordings to the seven-label targets, missed and spurious labels for each compound combination, and the false-positive rate on normal recordings. For labels represented in both single and compound faults, report their performance separately to clarify the effect of fault coexistence.

Response

The original 64-recording-to-seven-label mapping is supplied as S1, and S4 contains the corresponding recorded test probabilities, thresholds and predicted labels for every model/seed. S5 reports missed and spurious component labels separately for each of the seven observed compound combinations. In particular, the fixed graph achieves 78/140 exact compound-vector decisions (0.5571), compared with 70/140 (0.5000) for self-loops and 127/140 (0.9071) for the waveform CNN. The denominator represents 28 distinct compound acquisitions evaluated under five seeds, not 140 independent samples.

Fault combination	Exact	Missed labels	Spurious labels
Inner + outer race	19/20	None	Ball 1; Cage 1
Inner race + broken tooth	9/20	Inner 2; Broken 1	Outer 10; Ecc. 9
Inner race + eccentricity	9/20	Inner 2	Outer 10; Broken 10
Inner race + gear wear	10/20	Inner 2; Wear 2	Outer 9; Broken 4; Ecc. 1
Outer race + broken tooth	7/20	Outer 2; Broken 5	Inner 7; Wear 5; Ecc. 6
Outer race + eccentricity	11/20	Ecc. 3	Inner 5; Wear 3; Broken 5
Outer race + gear wear	13/20	Wear 2	Inner 6; Broken 1; Ecc. 1

On the eight distinct normal recordings, the fixed graph produced at least one false-positive label in 20/40 repeated recording–seed decisions (50.0%); the corresponding values are 27/40 (67.5%) for self-loops and 13/40 (32.5%) for the CNN. The fixed graph's label-level false-positive frequency on normal inputs is 30/280 (10.7%). These are repeated-seed descriptive rates, not estimates from 40 independent normal acquisitions.

For labels represented in both single and compound states, the fixed-graph pooled-decision F1 scores are: inner race, 0.833 versus 0.860; outer race, 0.755 versus 0.834; gear wear, 0.780 versus 0.857; broken tooth, 0.750 versus 0.723; and eccentric gear, 0.976 versus 0.787 (single versus compound, respectively; S6). Subgroup supports differ, so these values describe the observed mixtures rather than isolate a causal coexistence effect. Rolling-element and cage defects have no positive compound examples in this dataset, precluding that comparison for those labels.

Comment 7

7. Please specify the FFT normalization and treatment of zero power in the log-power descriptors of Eq. (3). For the noise experiments, define the power reference used for SNR calculation, whether noise is independent across channels, and the number of noise realizations. Please also repair the sentence fragment "removal of the eight local descriptors." in Section 3.4 and use consistent model names, particularly "fixed graph" and "connected graph", throughout the manuscript.

Response

The FFT features use a Hann window and the expression $L(f) = \log[1 + |\mathrm{RFFT}(w \odot x)(f)|^2/N]$, with $N = 4096$ for each full window and $N = 512$ for each local segment. The factor $1/N$ is applied to squared FFT magnitude; this is not a calibrated one-sided power spectral density. The log1p operation maps zero power to zero, so no separate zero-power replacement is required. For each frequency band, the full-window features are mean, maximum and population standard deviation of $L(f)$, together with the eight local segment means.

Test noise is added after per-channel, per-window median/MAD normalization. For each channel the reference power is $P = \mathrm{mean}(x^2)$, and independent, zero-mean Gaussian samples are scaled to variance $\max[P/10^{(SNR/10)}, 10^{-12}]$. Thus the noise is independent across channels and time samples. One fixed pseudo-random realization is used per test window at each evaluated SNR (10, 5, 0 and −5 dB), and reused across models and training seeds; the underlying draw is rescaled between SNR levels. The test therefore evaluates one matched noise realization per window/SNR, not repeated independent noise trials.

The incomplete sentence in Section 3.4 is corrected in wording as follows: “The descriptor ablation removes the eight local segmentwise log-power features, retaining the three full-window log-power statistics per node.” We also agree to use “fixed graph” consistently for the specified topology, including figure labels; “connected graph” refers to that same model and should not be used as a separate name in the manuscript.

Supporting material

The accompanying Reviewer1_Supplementary_Analyses.zip includes the recording-label manifest (S1), reference split manifest (S2), paired training-seed results and selected parameters (S3), recording-level predictions (S4), compound and subgroup error tables (S5–S8), the paired recording-bootstrap calculation (S9), representative spectra (S10), and the recovered v3 source/configuration. The supplementary audit does not substitute unperformed independent-data validation or reconstruct undocumented random/adaptive graph runs.