ISSN 1392-1207. MECHANIKA. Vol. 32, No. 5, 2026: 369–376

Protocol- and Bearing-Allocation-Dependent Model Ranking in Vibration-Based Bearing Fault Diagnosis

Zeyun SHANG*, Jihang WANG*, Yiqian JIA*, Ruoqing WU*, Sumin GUO**
*Shandong University of Science and Technology, Qingdao, 266590, China, E-mails: 202583190091@sdust.edu.cn, 202583190031@sdust.edu.cn, 202483190038@sdust.edu.cn, 202483190059@sdust.edu.cn,
**Shandong University of Science and Technology, Qingdao, 266590, China, E-mail: skd994055@sdust.edu.cn (Corresponding Author)
https://doi.org/10.5755/j02.mech.45697

1. Introduction

Rolling bearings are central components of rotating machines, and their vibration measurements contain transient and spectral information associated with localized damage. Vibration analysis therefore remains a practical basis for condition monitoring and fault diagnosis [1, 2]. Data-driven methods can process raw waveforms, spectral transforms, or hybrid waveform–spectrum inputs. Model choice can depend on how vibration records are divided into development and evaluation sets. This becomes particularly relevant when a diagnostic conclusion is intended to guide condition monitoring under operating conditions that differ from those observed during model development.

In public bearing datasets, one physical bearing can supply many raw acquisitions, and each acquisition can supply many short windows. Window-level scores can therefore reflect shared file structure, bearing identity, and repeated condition information alongside diagnostic content [3, 4, 5]. We define a segmented window as the prediction unit and a verified physical bearing as the experimental unit. The windows are repeated observations nested within bearings. Bearing identity, group, raw-file, and hash checks establish which subsets contain independent specimens. Holding out an operating condition asks a different diagnostic question from testing on the recorded condition mixture. These distinctions motivate examining model ranking across both evaluation protocols and physical-bearing allocations.

This study examines whether a model selected after leakage control remains preferred when the evaluation protocol or physical-bearing allocation changes. We compare four compact waveform, spectrum, and fusion configurations using verified Paderborn records. The primary P1 allocation favors waveform–spectrum concatenation, whereas the spectrum model has the highest mean Macro-F1 in all three tasks that hold out a target operating condition. Within P1, reassigning physical bearings changes the highest-mean configuration despite unchanged leakage-control rules. This result shows that leakage control alone does not ensure that the same model will be selected across bearing allocations. The allocation audit uses split files and run manifests locked before their corresponding executions. A strict healthy/inner/outer (H/I/O) extension isolates both target conditions and test bearings. Training-budget, class-wise, and bearing-level analyses further examine the consistency and scope of the findings.

2. Related Work and Study Position

Waveform-domain, frequency-domain, and hybrid-fusion processing are established routes for extracting information from bearing vibration signals. Frequency-domain descriptions expose spectral composition, whereas waveform models retain local transient structure; fusion architectures seek to combine both [6, 7, 8, 9, 10]. Recent reviews document the breadth of deep-learning approaches for rotating-machine and bearing diagnosis [11, 12, 13, 14]. Cross-condition bearing diagnosis has also motivated transfer-learning and domain-adaptation methods [15, 16, 17, 18, 19]. Those studies show the importance of operating-condition variability, but their reported model orderings depend on the associated data-partition rule.

2.1. From leakage control to model-selection stability

Recent evaluation studies distinguish window, raw-file, group, bearing, and target-condition boundaries [5, 20]. Vieira et al. show that segment-wise and condition-wise partitions can introduce spurious correlations. They advocate bearing-wise evaluation, treat co-occurring faults as a multi-label problem, and examine training-bearing diversity across several public datasets. Their work examines how partition rules and bearing diversity affect performance estimates. Once the specified leakage controls are in place, does the model ranked first by mean Macro-F1 remain the same across protocols and bearing allocations?

We examine this question in two steps. Within P1, four additional bearing allocations vary the specimens assigned to each subset while preserving the same bearing, group, raw-file, hash, and training-only normalization controls. P2 then evaluates transfer to held-out operating conditions; the strict H/I/O extension additionally removes source–target bearing overlap. The ordering of the same candidate models is the primary outcome.

3. Materials and Experimental Design

Three evaluation designs address different uses. P1 tests on physical bearings absent from training and validation, using the recorded mixture of operating conditions. P2 tests on a held-out operating condition but retains bearing identities shared with the source data. The strict H/I/O extension holds out both the target condition and test-bearing identities for the healthy, inner-race, and outer-race classes. The same candidate panel is compared throughout.

3.1. Paderborn vibration records and preprocessing

The official Paderborn dataset provides the vibration records [21, 22]. The verified index contains 2,560 raw acquisitions from 32 physical bearings: six healthy, 11 inner-race, 12 outer-race, and three combined inner–outer bearings. We use the vibration_1 channel sampled at 64 kHz. Splits are formed at the raw-file or group level before segmentation. Bearing identity, group, raw-file, and SHA256 checks audit the separation between subsets. Each record is divided into non-overlapping 2,048-point windows with stride 2,048. Global z-score statistics are fitted on the training subset and applied unchanged to validation and test data. Main-task scores summarize window-level classification; physical bearings are the units for allocation sensitivity and cluster-level uncertainty analyses.

The condition codes specify rotational speed, load torque, and radial force. N09_M07_F10 denotes 900 rpm, 0.7 Nm, and 1000 N; N15_M01_F10 denotes 1500 rpm, 0.1 Nm, and 1000 N; N15_M07_F10 denotes 1500 rpm, 0.7 Nm, and 1000 N; and N15_M07_F04 denotes 1500 rpm, 0.7 Nm, and 400 N [23]. The comparisons evaluate these operating-condition combinations. Fig. 1 summarizes this split-before-windowing workflow, which uses training-only normalization and the same fixed task split for all variants. P2 excludes the target condition from model development but retains source–target bearing overlap.

3.2. Operating-condition context in vibration signals

Fig. 2 compares inner- and outer-race waveforms and spectra across the source condition and three target conditions. These two localized fault categories have greater physical-bearing coverage than the combined-fault class. Source denotes N09_M07_F10; T1, T2, and T3 denote N15_M01_F10, N15_M07_F10, and N15_M07_F04, respectively. For each condition and fault class, we select the lower median-ranked bearing by full-record RMS and extract a centred 15-ms waveform segment. Each segment is standardized to emphasize temporal morphology rather than absolute RMS differences. The spectral curves show median log-magnitude rFFT spectra aggregated within and across bearings, with 0–10 kHz displayed from the full 0–32 kHz one-sided range for legibility. Source uses heavier lines, while targets use distinct line styles for grayscale readability. The Supplementary Material describes the selection and aggregation procedure.

3.3. Evaluation protocols

Table 1 summarizes the evaluation protocols and their isolation boundaries. P1 uses a fixed, class-stratified allocation of physical bearings across training, validation, and test subsets, with the recorded operating conditions represented within each subset. Bearing, group, and raw-file identities are disjoint across subsets, and the primary allocation is fixed before windowing. For the allocation audit, we generated only four additional splits, using split seeds 101, 202, 303, and 404, and evaluated all four. Within each class, sorted bearing IDs were shuffled using the split seed and fixed class offsets. Counts followed the rounded 0.70/0.15/remainder rule, subject to at least one bearing per subset.

The resulting train/validation/test quotas are 4/1/1 for healthy bearings, 8/2/1 for inner-race faults, 8/2/2 for outer-race faults, and 1/1/1 for combined faults. The split files and run manifests were locked before their corresponding executions. The completed alternative-P1 matrix contains 48 runs. Supplementary Section S11 and Table S20 provide the exact bearing IDs, class offsets, fallback rule, seeds, hashes, and locking records.

P2 uses N09_M07_F10 for training and validation. Its three target conditions are N15_M01_F10, N15_M07_F10, and N15_M07_F04. Each target is excluded from training, validation, normalization fitting, early stopping, and model selection. Source and target records share physical-bearing identities, but their raw files are disjoint. Thus, P2 evaluates transfer to an unobserved operating condition rather than to independent test bearings.

The strict extension holds out both target-condition records and test-bearing identities. It uses the healthy, inner-race, and outer-race classes because only three combined-fault bearings are available. The same nine held-out bearings–three per class–are evaluated under each of the three target conditions. Their identities are removed from the source data before the 17-bearing training and three-bearing validation sets are formed. Each target test set contains 180 raw files. All bearing-ID, raw-file-ID, and SHA256 overlaps between development and test subsets are zero.

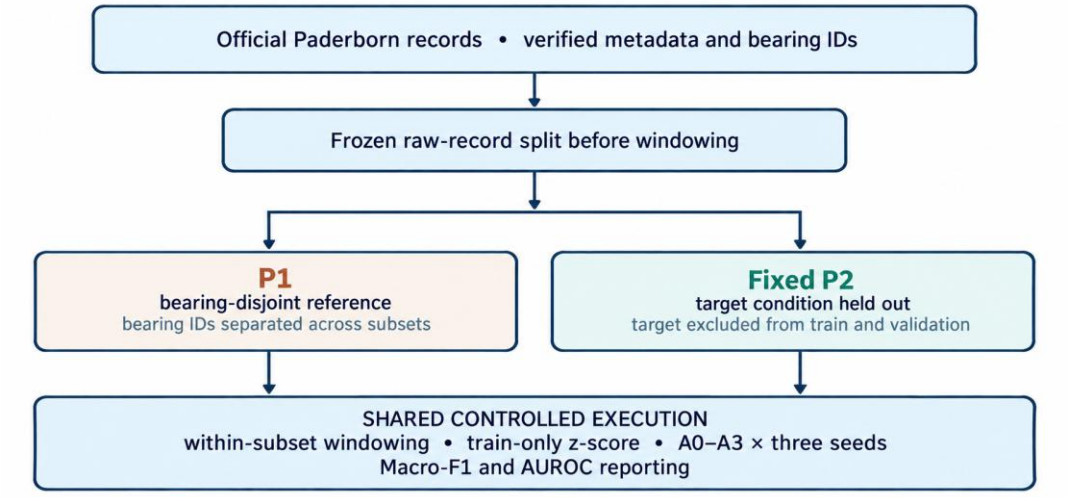

Fig. 1 Protocol-aware workflow for bearing fault diagnosis

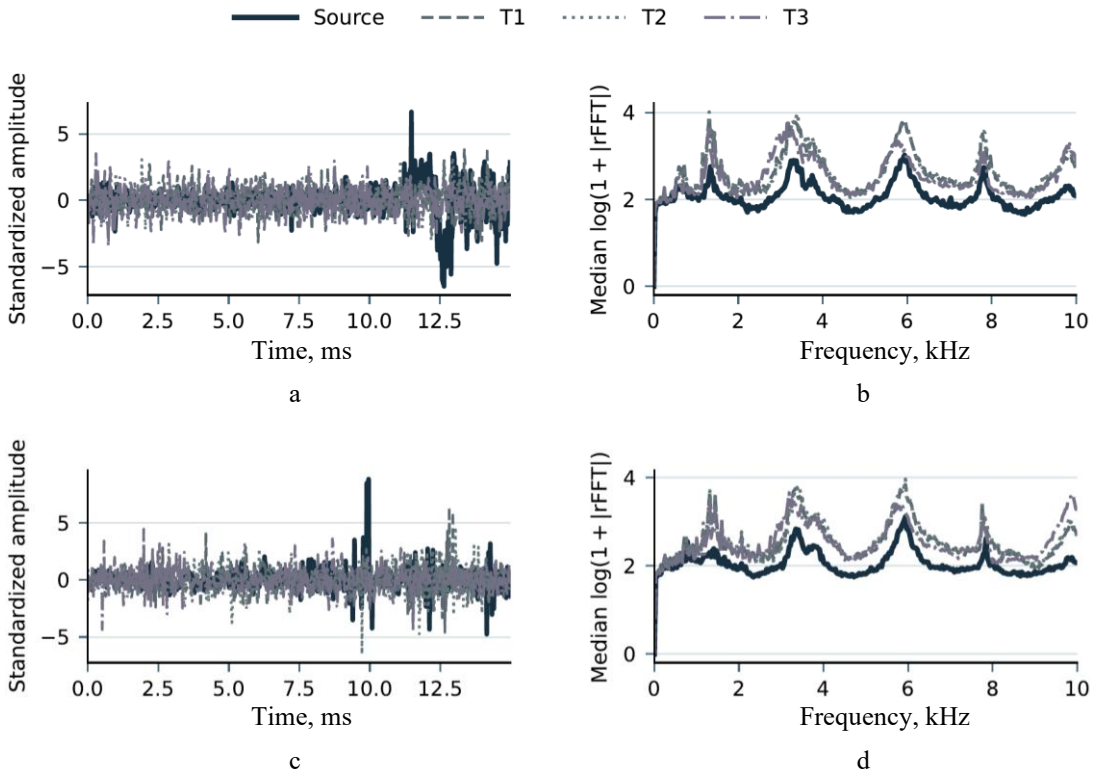

Fig. 2 Waveforms and spectra under different operating conditions: a – inner-race waveform, b – inner-race spectrum, c – outer-race waveform, d – outer-race spectrum

Table 1

Evaluated vibration-diagnosis protocols and isolation boundaries

Task	Source → target	Physical-bearing overlap	Raw-file overlap	Supported evaluation claim
P1	Mixed recorded conditions	0 / 0 (disjoint)	0	Bearing-disjoint reference; four classes in each subset
P2-0	N09_M07_F10 → N15_M01_F10	26 / 6	0	Target condition held out; bearing-overlapping
P2-1	N09_M07_F10 → N15_M07_F10	26 / 6	0	Target condition held out; bearing-overlapping
P2-2	N09_M07_F10 → N15_M07_F04	26 / 6	0	Target condition held out; bearing-overlapping
Strict H/I/O	N09_M07_F10 → three target subsets	0 / 0 (disjoint)	0	Joint condition-and-bearing isolation; three-class target subset

Note: Physical-bearing entries are verified train-test / validation-test intersections. P1 controls declared group, raw-file, verified-bearing, window-by-raw-file, and train-only-normalization boundaries. P2 uses bearing-overlapping source and target sets while excluding the target condition from training, validation, normalization fitting, early stopping, and model selection. The strict extension evaluates the same nine held-out H/I/O bearings under three target conditions; their identities are excluded from source training and validation.

3.4. Compact vibration-model variants and training

The panel represents two basic input choices and two ways of combining them. A0 uses waveforms, A1 uses spectra, A2 concatenates the two branch outputs, and A3 combines them through a learned gate. The fusion models reuse the same waveform and frequency branch types as the single-domain models. We keep this panel unchanged across protocols, bearing allocations, and training seeds to test whether the model ranked first by mean Macro-F1 changes with evaluation design. The compact architectures make the repeated allocation audit computationally feasible.

The training-standardized waveform $\tilde{x}(n)$ is used directly as the A0 time-domain input. For A1, the one-sided rFFT coefficients of each standardized window of length $L = 2{,}048$ are computed as:

$$X(k) = \sum_{n=0}^{L-1} \tilde{x}(n) exp\left(-j\frac{2\pi kn}{L}\right). \quad (1)$$

The frequency-domain input is then obtained as the log-magnitude spectrum:

$$x^f(k) = log\left(1+|X(k)|\right). \quad (2)$$

In these equations, L is the window length, n indexes samples, k indexes frequency bins, and j is the imaginary unit. X denotes the complex Fourier coefficients, and the superscript f identifies the log-magnitude frequency-domain input.

The ranges are $n = 0, \ldots, L-1$ and $k = 0, \ldots, L/2$. This produces 1,025 one-sided spectral bins spanning 0–32 kHz. The DC bin is retained, and no per-window mean removal, tapering window, or separate frequency-domain normalization is applied.

We use the direct rFFT as a common spectral baseline so that input construction stays fixed across protocols and bearing allocations. Retaining DC preserves the magnitude of each window's mean relative to the training reference, and omitting tapering avoids additional sample weighting. The log transform compresses spectral magnitude. DC content and spectral leakage may therefore influence the observed ordering.

A2 concatenates the waveform and spectrum features before classification, whereas A3 uses a learned gate to interpolate between them. All models use compact one-dimensional convolutional branches with batch normalization, GELU activation, and adaptive average pooling. Table 2 summarizes the inputs, fusion strategies, and parameter counts; full layer specifications appear in the Supplementary Material. The candidates differ in representation, branch configuration, and fusion strategy. Their comparison therefore concerns the selection of complete model configurations.

All primary runs use unweighted cross-entropy loss, AdamW with learning rate 0.001 and weight decay 10^{-4}, batch size 64, a maximum of 10 epochs, and early stopping with patience 5. Validation Macro-F1 is the prespecified checkpoint-selection metric. Optimization seeds are 42, 2025, and 3407. The allocation audit applies the same training settings to four additional P1 allocations, giving $4\times4\times3=48$ runs.

Reported metrics are Accuracy, Macro-Precision, Macro-Recall, Macro-F1, and one-vs-rest Macro-AUROC. For each fixed split, metrics are reported as the mean ± sample standard deviation across the three optimization seeds. This variation describes training randomness, not resampling of physical bearings. Bearing-allocation effects are assessed separately through the additional P1 splits. Models are ranked by mean window-level Macro-F1. Where predictions are retained, paired bearing-cluster intervals assess uncertainty at the physical-bearing level.

The three P2 tasks use identical frozen source-side training and validation CSVs. Training was launched separately for each target task. These executions share the same source-bearing allocation and are treated as correlated optimization runs.

The budget extension covers primary P1 and representative P2-1. The strict H/I/O extension trains one three-class checkpoint per model and seed and evaluates it on the same nine held-out bearings under all three target conditions.

4. Results

4.1. Primary bearing-disjoint reference split

Table 3 summarizes the primary 48-run comparison. A2, which concatenates waveform and spectrum features, has the highest mean window-level Macro-F1 under the primary P1 allocation (0.5770±0.0189). It also has the highest score in each of the three optimization runs.

Table 2

Controlled model variants and parameter counts

Variant	Input representation	Fusion	Role	Parameters
A0	normalized waveform	none	time-domain baseline	4100
A1	log-magnitude one-sided rFFT	none	frequency baseline	2980
A2	waveform + log-magnitude rFFT	concatenation	waveform–spectrum concatenation	7076
A3	waveform + log-magnitude rFFT	learned gate	gated waveform–spectrum fusion	9028

Note: All variants share the optimizer, loss, maximum-epoch rule, split files, and optimization seeds. Full layer specifications are provided in the Supplementary Material.

Table 3

Three-seed primary task results

Task	Model	Accuracy	Macro-F1	Macro-AUROC	F1 rank
P1	A2 concatenation	0.6745 ± 0.0146	0.5770 ± 0.0189	0.7165 ± 0.0048	1
P1	A3 gated fusion	0.5257 ± 0.0367	0.4411 ± 0.0254	0.6273 ± 0.0238	2
P1	A0 time-only	0.5090 ± 0.0555	0.4316 ± 0.0666	0.6682 ± 0.0152	3
P1	A1 frequency-only	0.3636 ± 0.0372	0.3372 ± 0.0239	0.5079 ± 0.0089	4
P2-0	A1 frequency-only	0.4242 ± 0.0086	0.3870 ± 0.0449	0.7260 ± 0.0064	1
P2-0	A3 gated fusion	0.4235 ± 0.0435	0.2616 ± 0.0310	0.6602 ± 0.0534	2
P2-0	A2 concatenation	0.3872 ± 0.0057	0.1978 ± 0.0101	0.6728 ± 0.0113	3
P2-0	A0 time-only	0.3496 ± 0.0078	0.1594 ± 0.0060	0.6402 ± 0.0115	4
P2-1	A1 frequency-only	0.4412 ± 0.0077	0.4097 ± 0.0501	0.7423 ± 0.0092	1
P2-1	A3 gated fusion	0.4343 ± 0.0522	0.2707 ± 0.0292	0.6695 ± 0.0546	2
P2-1	A2 concatenation	0.3975 ± 0.0094	0.2309 ± 0.0213	0.7117 ± 0.0144	3
P2-1	A0 time-only	0.3456 ± 0.0073	0.1582 ± 0.0107	0.6572 ± 0.0142	4
P2-2	A1 frequency-only	0.4550 ± 0.0124	0.4220 ± 0.0678	0.7374 ± 0.0051	1
P2-2	A2 concatenation	0.3945 ± 0.0126	0.3279 ± 0.0097	0.7290 ± 0.0042	2
P2-2	A0 time-only	0.4063 ± 0.0040	0.3002 ± 0.0117	0.7406 ± 0.0109	3
P2-2	A3 gated fusion	0.4007 ± 0.0534	0.2594 ± 0.0295	0.6593 ± 0.0450	4

Note: Values are mean ± sample standard deviation across seeds 42, 2025, and 3407. Models are ranked within each task by mean Macro-F1, using the frozen 48-run registry.

Section 4.4 tests whether this ordering persists when the physical bearings are reassigned under the same P1 controls.

4.2. Target-condition shifts

The P2 tasks yield a different model ordering from primary P1. A1 has the highest mean window-level Macro-F1 in P2-0 (0.3870±0.0449), P2-1 (0.4097±0.0501), and P2-2 (0.4220±0.0678). Its score is also highest in each of the three optimization runs for each task. These tasks share source-bearing identities with their target sets, as specified in Section 3.3.

4.3. Protocol-dependent contrast

The primary P1 ordering is A2 > A3 > A0 > A1, whereas A1 has the highest mean Macro-F1 in all three P2 tasks. The two evaluation designs therefore select different candidates from the same panel. They differ in condition distribution, training-set composition, and source–target bearing overlap, so the contrast reflects the overall evaluation design. The next analysis varies bearing allocation within P1 while holding its leakage controls constant.

4.4. Physical-bearing allocation sensitivity

The four additional P1 allocations select A3, A3, A1, and A2 by mean Macro-F1 under unchanged leakage controls (Table 4; Fig. 3). Supplementary Table S10 reports paired bearing-resampling intervals for the observed leader–runner-up differences, conditional on each fixed five-bearing test set. Only the two outer-race bearings contribute within-class resampling variation, so these intervals do not establish population-level superiority. Supplementary Figure S2 shows the corresponding per-bearing window-accuracy profiles.

4.5. Budget check

With maximum training budgets of 10, 30, and 50 epochs, A2 retains the highest Macro-F1 in each primary-P1 optimization run, and A1 does so in each representative-P2-1 run. Supplementary Tables S6–S8 report the complete results and stopping records.

A four-class audit of retained predictions from the 10-epoch primary-P1 budget-control runs also places A2 highest by mean Macro-F1 (Supplementary Tables S18–S19). These runs use the same split but are distinct from the historical primary registry. The audit summarizes all five test bearings, including the combined-fault bearing; its paired resampling interval is conditional on this fixed class composition.

4.6. Combined-fault coverage and recall

The combined inner–outer class has only one physical bearing in each primary-P1 subset, and all four models have zero mean recall for this class. Macro-F1 gives this class equal weight, and incorrect combined-class predictions also affect the other classes. The four-class ranking

Table 4

Physical-bearing allocation sensitivity under P1

Allocation	Highest mean	Top F1	Runner-up	Runner-up F1	Gap
P1-Alt-01	A3 gated	0.563±0.014	A2 concat	0.532±0.099	0.0306
P1-Alt-02	A3 gated	0.816±0.024	A2 concat	0.727±0.021	0.0888
P1-Alt-03	A1 spectrum	0.316±0.031	A2 concat	0.271±0.014	0.0455
P1-Alt-04	A2 concat	0.391±0.040	A0 waveform	0.379±0.017	0.0126

Note: Each alternative retains bearing, group, raw-file, and SHA256 disjointness. Values are three-seed mean ± sample SD; the final column reports the leader-runner-up gap.

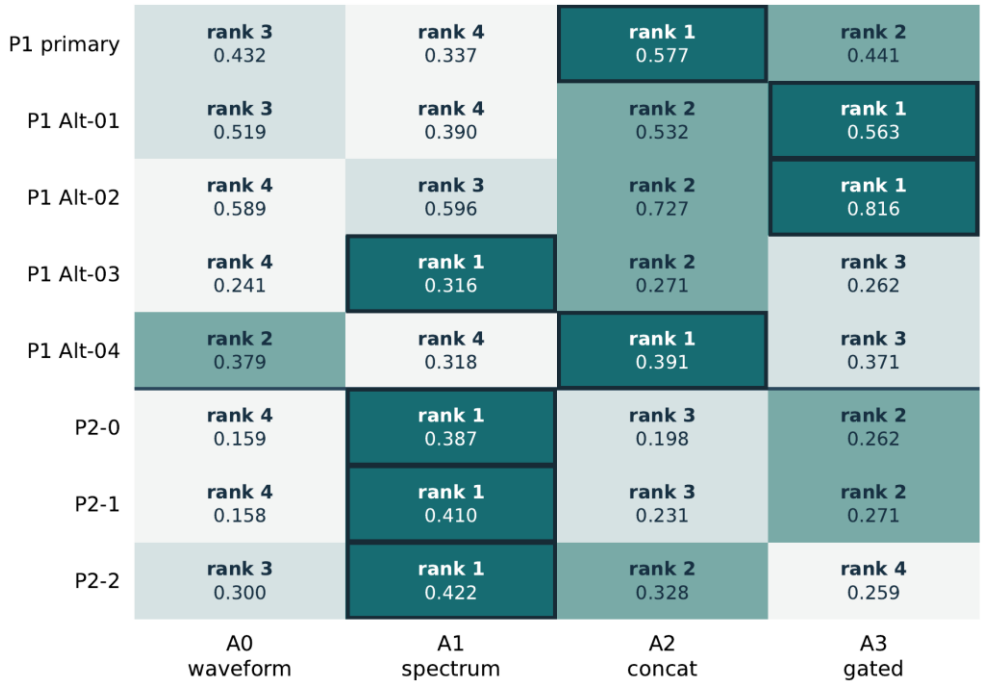

Fig. 3 Model rankings across evaluation protocols and bearing allocations

therefore, partly reflects performance on a class for which generalization remains uncertain. The alternative allocations yield both zero and nonzero combined-fault recall. Their classwise results are included in the accompanying reproducibility materials.

4.7. H/I/O-restricted sensitivity

In the H/I/O sensitivity analysis, we remove test windows whose true label is combined, while retaining the original four-class predictions for the remaining windows. A prediction of the combined class therefore remains an error for a retained H/I/O window. Restricted Macro-F1 is the unweighted mean of the healthy, inner-race, and outer-race F1 scores. A2 still ranks first by mean Macro-F1 in primary P1, and A1 ranks first in all three P2 tasks. P1-Alt-01 changes from A3 to A2, but the leading model still varies across allocations (Supplementary Tables S13–S14 and Figure S1).

4.8. Strict condition-and-bearing isolation

The strict H/I/O extension excludes test-bearing identities from the source data while holding out the target condition. A1 has the highest three-seed mean Macro-F1 for all three targets (0.386, 0.377, and 0.374). The same nine physical bearings supply the test observations under these three conditions. The paired bearing-cluster intervals for the difference between A1 and the next-highest model all include zero (Supplementary Table S17), so these data do not establish statistical separation between them.

5. Discussion and Engineering Implications

Fig. 3 shows how model selection changes across the audited protocols and allocations. Each cell reports the within-row Macro-F1 rank and three-seed mean; outlined cells mark the highest mean. A2 ranks first by mean Macro-F1 in primary P1. Under the same leakage controls, the four additional P1 allocations select A3, A3, A1, and A2, respectively. Below the horizontal rule separating P1 allocations from P2 tasks, A1 ranks first in all three P2 tasks. These results show that the selected model depends on both the evaluation protocol and the physical bearings assigned to its subsets. A1 also has the highest mean score in the strict H/I/O extension after source–target bearing overlap is removed, although the paired intervals do not distinguish it from the next-highest model.

Before selecting a model, engineers should decide whether the evaluation needs to represent new bearings, new operating conditions, or both. P1 tests independent bearings under the recorded mix of conditions. P2 tests new operating conditions with shared bearing identities, while the strict extension tests new conditions on independent bearings. The study should match its evaluation design to the intended use, report the bearing and class coverage of each subset, and compare candidates across several feasible bearing allocations. If the leading model changes with allocation or paired intervals do not distinguish the leading candidates, computational cost and deployment constraints can inform a provisional choice. The unresolved performance comparison should be reported and reassessed using additional independent bearings.

The number of independent bearings and operating conditions needed for model selection depends on class coverage, specimen diversity, the size of model differences, and the precision required for the engineering decision. The present data do not support a universal numerical threshold. Many windows from one bearing cannot substitute for independent specimens. An evaluation should include independent bearings for each fault class, conditions relevant to the intended transfer, and an uncertainty analysis based on bearings. Stability across allocations and adequately precise paired comparisons are more informative than a single sample-count threshold. Here, the wide strict-H/I/O intervals show that nine test bearings per target do not resolve the leading-model comparison.

The H/I/O analyses clarify how sparse combined-fault coverage affects the four-class results. Removing true combined-fault windows preserves the primary-P1 and P2 selections, but changes P1-Alt-01 from A3 to A2. Allocation sensitivity remains, so the main finding is not confined to that sparse class. The four-class ranking, however, does not establish a reliable model preference for combined faults. Fig. 2 provides the signal context for these comparisons, showing how waveform morphology and spectral profiles vary across the evaluated conditions.

6. Scope and Limitations

The evidence covers one Paderborn subset, one source condition, three target conditions, four additional P1 allocations, and four compact model configurations. The configurations are not capacity matched, so the results concern their complete designs rather than an isolated effect of input representation. Three optimization seeds give only a limited view of training variability. Bearing-cluster uncertainty is conditional on the observed specimens, and the strict extension covers nine test bearings per target while omitting combined faults.

Absolute performance is modest: the best mean Macro-F1 is about 0.58 in primary P1 and 0.39–0.42 in P2. These scores describe the tested configurations and training settings, not a performance ceiling for the dataset. Increasing the maximum budget to 30 and 50 epochs preserves the observed selections in primary P1 and representative P2-1, but does not establish that the models are fully optimized. Whether more advanced architectures or different optimization strategies show similar ranking changes remains an open question. The present evidence supports protocol-aware selection among the tested candidates; it does not demonstrate deployment-ready diagnosis.

7. Conclusions

Within the tested model panel, rankings change across evaluation protocols and physical-bearing allocations. A2 ranks first by mean Macro-F1 in primary P1, whereas A1 ranks first in the three P2 tasks. The additional P1 allocations select different leading models under unchanged leakage controls. Leakage control is therefore necessary for credible evaluation but does not by itself ensure a stable model choice. For engineering use, models should be evaluated under the intended condition shift and their rankings checked across feasible bearing allocations.

Data Availability

The Paderborn bearing dataset is available from Paderborn University [22, 24]. The anonymized reproducibility archive accompanying this submission provides fixed split CSV files, physical-bearing mappings, per-run metrics, model configurations, retained primary-P1 budget-control prediction records, strict-extension manifests, and traceability records. The supplementary material provides the primary, allocation-sensitivity, and strict H/I/O audit summaries, split registries, and result-hash records needed to inspect the evidence trail.

Conflict of Interest

The authors declare no conflict of interest.

References

1. **Randall, R. B.; Antoni, J.** 2011. Rolling element bearing diagnostics – A tutorial, Mechanical Systems and Signal Processing 25(2): 485-520. https://doi.org/10.1016/j.ymssp.2010.07.017.
2. **Vakharia, V.; Gupta, V. K.; Kankar, P. K.** 2015. Ball bearing fault diagnosis using supervised and unsupervised machine learning methods, International Journal of Acoustics and Vibration 20(4): 244-250. https://doi.org/10.20855/ijav.2015.20.4387.
3. **Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O.** 2012. Leakage in data mining: Formulation, detection, and avoidance, ACM Transactions on Knowledge Discovery from Data 6(4): 15. https://doi.org/10.1145/2382577.2382579.
4. **Kapoor, S.; Narayanan, A.** 2023. Leakage and the reproducibility crisis in machine-learning-based science, Patterns 4(9): 100804. https://doi.org/10.1016/j.patter.2023.100804.
5. **Wheat, L.; von Mohrenschildt, M.; Habibi, S.; Al-Ani, D.** 2024. Impact of Data Leakage in Vibration Signals Used for Bearing Fault Diagnosis, IEEE Access 12: 169 879-169 895. https://doi.org/10.1109/ACCESS.2024.3497716.
6. **Zhang, W.; Peng, G.; Li, C.; Chen, Y.; Zhang, Z.** 2017. A New Deep Learning Model for Fault Diagnosis with Good Anti-Noise and Domain Adaptation Ability on Raw Vibration Signals, Sensors 17(2): 425. https://doi.org/10.3390/s17020425.
7. **Ince, T.; Kiranyaz, S.; Eren, L.; Askar, M.; Gabbouj, M.** 2016. Real-Time Motor Fault Detection by 1-D Convolutional Neural Networks, IEEE Transactions on Industrial Electronics 63(11): 7067-7075. https://doi.org/10.1109/TIE.2016.2582729.
8. **Li, W.; Qiu, M.; Zhu, Z.; Wu, B.; Zhou, G.** 2016. Bearing fault diagnosis based on spectrum images of vibration signals, Measurement Science and Technology 27(3): 035005. https://doi.org/10.1088/0957-0233/27/3/035005.
9. **Kim, K.; Kim, Y. S.** 2024. Vibration spectrogram analysis for bearing fault diagnosis based on Grad-CAM for feature selection and statistical approach, Journal of Mechanical Science and Technology 38(11): 5885-5898. https://doi.org/10.1007/s12206-024-1010-3.
10. **Shao, H.; Jiang, H.; Wang, F.; Zhao, H.** 2017. An enhancement deep feature fusion method for rotating machinery fault diagnosis, Knowledge-Based Systems 119: 200-220. https://doi.org/10.1016/j.knosys.2016.12.012.
11. **Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R. X.** 2019. Deep learning and its applications to machine health monitoring, Mechanical Systems and Signal Processing 115: 213-237. https://doi.org/10.1016/j.ymssp.2018.05.050.
12. **Liu, R.; Yang, B.; Zio, E.; Chen, X.** 2018. Artificial intelligence for fault diagnosis of rotating machinery: A review, Mechanical Systems and Signal Processing 108: 33-47. https://doi.org/10.1016/j.ymssp.2018.02.016.
13. **Hoang, D.-T.; Kang, H.-J.** 2019. A survey on Deep Learning based bearing fault diagnosis, Neurocomputing 335: 327-335. https://doi.org/10.1016/j.neucom.2018.06.078.
14. **Lei, Y.; Yang, B.; Jiang, X.; Jia, F.; Li, N.; Nandi, A. K.** 2020. Applications of machine learning to machine fault diagnosis: A review and roadmap, Mechanical Systems and Signal Processing 138: 106587. https://doi.org/10.1016/j.ymssp.2019.106587.
15. **Zhang, X.; Zhao, J.-m.; Li, H.-p.; Yang, R.-f.; Teng, H.-z.** 2020. Rolling Bearings Fault Diagnosis under Variable Conditions Using RCMFE and Improved Support Vector Machine, International Journal of Acoustics and Vibration 25(3): 304-317. https://doi.org/10.20855/ijav.2020.25.31596.
16. **Tong, Z.; Li, W.; Zhang, B.; Zhang, M.** 2018. Bearing Fault Diagnosis Based on Domain Adaptation Using Transferable Features under Different Working Conditions, Shock and Vibration 2018: 6714520. https://doi.org/10.1155/2018/6714520.
17. **Zhu, J.; Chen, N.; Shen, C.** 2020. A New Deep Transfer Learning Method for Bearing Fault Diagnosis Under Different Working Conditions, IEEE Sensors Journal 20(15): 8394-8402. https://doi.org/10.1109/JSEN.2019.2936932.
18. **Liu, Z.; Zhang, P.; Yu, Y.; Li, M.; Zeng, Z.** 2024. A novel fault diagnosis model of rolling bearing under variable working conditions based on attention mechanism and domain adversarial neural network, Journal of Mechanical Science and Technology 38(3): 1101-1111. https://doi.org/10.1007/s12206-024-1208-8.
19. **Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V.** 2016. Domain-Adversarial Training of Neural Networks, Journal of Machine Learning Research 17(59): 1-35. Available at: https://jmlr.org/papers/v17/15-239.html.
20. **Vieira, J. P.; Bauler, V. A.; Rosa, R. K.; Silva, D.** 2026. Towards a more realistic evaluation of machine learning models for bearing fault diagnosis, Mechanical Systems and Signal Processing 258: 114640. https://doi.org/10.1016/j.ymssp.2026.114640.
21. **Lessmeier, C.; Kimotho, J. K.; Zimmer, D.; Sextro, W.** 2016. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification, PHM Society European

Conference 3(1). https://doi.org/10.36001/phme.2016.v3i1.1577.
22. Paderborn University; Konstruktions- und Antriebstechnik. 2026. Bearing DataCenter [online; accessed 2026-07-02]. Available at: https://mb.uni-paderborn.de/kat/forschung/bearing-datacenter.
23. Paderborn University. 2026. Operating Conditions – bearing datacenter [online; accessed 2026-07-22]. Available at: https://mb.uni-paderborn.de/en/kat/research/bearing-datacenter/operating-conditions.
24. Paderborn University. 2026. Data sets and download [online; accessed 2026-07-02]. Available at: https://mb.uni-paderborn.de/kat/forschung/bearing-datacenter/data-sets-and-download.

Z. Shang, J. Wang, Y. Jia, R. Wu, S. Guo

PROTOCOL- AND BEARING-ALLOCATION-DEPENDENT MODEL RANKING IN VIBRATION-BASED BEARING FAULT DIAGNOSIS

S u m m a r y

Model selection for bearing diagnosis can depend on the evaluation design and the physical bearings assigned to each subset. We investigate this dependence using 2,560 Paderborn vibration records and four compact waveform, spectrum, and fusion configurations. P1 separates physical bearings across training, validation, and test sets. Three P2 tasks hold out target operating conditions while retaining source–target bearing overlap, and a strict healthy/inner/outer extension isolates both conditions and test bearings. A2, the waveform–spectrum concatenation model, has the highest mean Macro-F1 in primary P1 (0.577), whereas A1, the spectrum model, has the highest mean in all three P2 tasks (0.387–0.422). Four additional P1 allocations select A3, A3, A1, and A2 under unchanged leakage controls. A1 also has the highest mean in the strict extension, but the paired bearing-cluster intervals for its differences from the next-highest models include zero. These results show that model selection within the tested panel can change even after leakage control. Evaluation should therefore reflect the intended operating-condition shift and test whether reassigning physical bearings changes the rankings.

Keywords: rolling bearing, vibration diagnosis, evaluation protocol, model ranking, operating-condition shift, physical-bearing split, data leakage.

Received July 31, 2026
Accepted October 15, 2026



This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).