Multiscale Ordinal Complexity Analysis for Wind Turbine Bearing Fault Diagnosis: Triaxial Accelerometer Feature Hierarchy and Specimen-Level Validation via GroupKFold Cross-Validation

Paulo R. L. Almeida[1,3], Thyago L. V. Lima[1], Alisson V. Brito[2], and Abel C. Lima-Filho[3]

[1] *Federal Institute of Paraíba (IFPB), Campus João Pessoa, João Pessoa-PB, Brazil*
paulo.almeida@ifpb.edu.br
thyagolvasconcelos@gmail.com

[2] *Center of Informatics (CI), Federal University of Paraíba (UFPB), Campus I, João Pessoa-PB, Brazil*
alisson@ci.ufpb.br

[3] *PPGEM, Federal University of Paraíba (UFPB), Campus I, João Pessoa-PB, Brazil*
abel@ct.ufpb.br

ABSTRACT

Permutation-entropy methods for bearing fault diagnosis in wind turbines are typically validated on laboratory test rigs without enforcing specimen-level separation in cross-validation or quantifying axis-specific diagnostic contributions. This study addresses these limitations using the Fraunhofer LBF operational wind turbine bearing dataset, applying multiscale permutation entropy (MPE) to triaxial front-bearing accelerometer signals across 44 bearing specimens (18 healthy, 10 inner race, 4 outer race, 12 roller element), with two corrupted files excluded following data quality screening. GroupKFold cross-validation with unique specimen-level group identifiers prevents the data leakage that arises when temporally correlated analysis windows are split without regard to bearing identity — a limitation present in all ten studies identified in a systematic literature search. MPE achieves $97.67 \pm 2.10\%$ window-level and 97.73% specimen-level accuracy using 12 features across three accelerometer axes, outperforming weighted permutation entropy (WPE, 91.83% window-level, 95.45% specimen-level) and matching a physically-augmented hybrid (MPE+Physical, 21 features) that contributes no additional specimen-level accuracy despite 33.0% feature importance within the combined set. The X-axis accelerometer (brng.f.x) accounts for 40.7% of classification importance, consistent with the primary radial load direction. Permutation entropy reveals a monotonic complexity hierarchy across fault types — Healthy (μ=0.702) < Roller Element (μ=0.889) < Inner Race (μ=0.981) < Outer Race (μ=0.992) — with Cohen's $d > 0.9$ for all pairwise comparisons. Roller element faults exhibit bimodal permutation entropy distributions, explained by load-zone-dependent impulsive generation. These results demonstrate that multiscale ordinal pattern analysis captures physically meaningful fault signatures in operational wind turbine data when evaluated under methodologically sound cross-validation protocols.

1. INTRODUCTION

Rolling element bearing faults in wind turbines represent a primary cause of unplanned downtime in energy generation assets, making reliable condition monitoring a prerequisite for cost-effective operation (Hameed, Hong, Cho, Ahn, & Song, 2009; Qiao & Lu, 2015). Vibration-based diagnostic methods have demonstrated strong performance in laboratory settings, yet their translation to operational wind turbines presents substantial challenges: variable wind loading, temperature-dependent signal characteristics, and limited labeled fault data from field installations place severe constraints on algorithm design and validation methodology. Previous investigation of optimal feature complexity for small-sample bearing fault detection demonstrates that carefully designed feature sets outperform larger feature collections under data constraints (P. R. L. Almeida, Lima, Brito, & Lima-Filho, 2025), motivating the 12-feature MPE representation evaluated in the present study.

Permutation entropy (PE) and its multiscale variants have emerged as computationally efficient complexity measures for bearing fault diagnosis, capturing ordinal pattern structure that remains robust under amplitude variations and opera-

https://doi.org/10.36001/IJPHM.2026.v17i2.4835

tional noise (Bandt & Pompe, 2002; Fu, Zhu, Zhu, & Yang, 2019). Reported classification accuracies range from 92.4% to 100% across laboratory studies (Sandoval, Leturiondo, Pozo, & Vidal, 2020; Wang, Yao, Cai, & Zhang, 2020; Ye, Yan, & Jia, 2021), with improved multiscale variants demonstrating 3–18% further improvement over standard PE-based methods (Minhas & Singh, 2021). However, three methodological limitations constrain the generalizability of these results to operational scenarios.

First, existing studies employ window-level cross-validation without specimen-level separation. When analysis windows derived from the same bearing specimen appear in both training and test partitions, the classifier learns specimen-specific signatures rather than generalizable fault features — a form of data leakage that inflates reported accuracy. A systematic search of the literature identifies this limitation across all eight entropy-based and ensemble studies retrieved (Table 1, rows 1–8).

Second, most studies use single-channel vibration signals without quantifying the diagnostic contribution of each measurement axis. Triaxial accelerometers are standard in industrial condition monitoring, yet the relative importance of X, Y, and Z axes for bearing fault discrimination remains uncharacterized for wind turbine applications.

Third, PE-based studies for wind turbine bearings predominantly use laboratory test rigs or pitch bearings under constant speed (Sandoval et al., 2020; Fu et al., 2019), not operational main bearings under realistic variable loading.

This study addresses these limitations through four primary contributions:

1. **Specimen-level GroupKFold validation**: unique group identifiers per bearing specimen prevent data leakage from temporally correlated analysis windows, yielding conservative and deployment-realistic performance estimates.
2. **Axis-specific diagnostic importance**: Shapley attribution across three accelerometer axes reveals that the X-axis accounts for 40.7% of classification importance, consistent with the primary bearing load direction.
3. **Physical interpretation of the PE hierarchy**: permutation entropy exhibits a monotonic complexity increase from healthy to faulty conditions, with fault-type-specific distributional structure — including bimodal distributions for roller element faults — explained by load-zone-dependent impulsive generation mechanisms.
4. **Comparative feature evaluation**: MPE, WPE, and MPE +Physical are evaluated under identical GroupKFold conditions, demonstrating that multiscale ordinal analysis captures the dominant diagnostic information with 12 features and that variance-based physical features add no specimen-level accuracy despite 33% feature importance within the hybrid set.

2. RELATED WORK

2.1. Permutation Entropy Variants for Bearing Fault Diagnosis

The use of multiscale permutation entropy for bearing fault diagnosis has a substantial prior record. Huo, Zhang, Shu, and Gallimore (2019) introduced fine-to-coarse multiscale permutation entropy (F2CMPE), reconstructing fine- and coarse-frequency components via wavelet packet decomposition before computing multiscale permutation entropy, combined with Laplacian score feature selection and SVM classification on laboratory test rig data. Subsequent variants address specific estimation weaknesses: composite multiscale weighted permutation entropy (CMWPE) targets stability on short windows (Zheng, Zheng, Du, et al., 2019), refined composite downsampling permutation entropy (rcDPE) improves estimation precision through anti-aliasing filtering at each scale (Dávalos, Jabloun, Ravier, & Buttelli, 2020), and multivariate interpolation-multiscale Chebyshev permutation entropy (MvIMCPE) extends multiscale analysis to multichannel settings (Li & Ding, 2026).

The present study differs from this body of work in three respects. First, prior MPE variants are validated under window-level or random splitting without specimen-level separation; Huo et al. (2019) and related studies do not report group-aware cross-validation, and the leakage this omission introduces has been quantified at over 40 percentage points of accuracy on standard benchmarks (Wheat, Mohrenschildt, Habibi, & Al-Ani, 2024). Second, prior work targets entropy estimator refinement — new coarse-graining schemes, decomposition front-ends, or multivariate formulations — whereas the present study uses standard MPE (Costa, Goldberger, & Peng, 2002) and instead contributes a validation protocol and a quantified sensor-axis attribution analysis. Third, no prior MPE study has been applied to the Fraunhofer LBF operational wind turbine dataset under realistic variable wind loading; existing MPE validations use controlled laboratory test rigs (Huo et al., 2019; Zheng et al., 2019; Sandoval et al., 2020; Wang et al., 2020; Ye et al., 2021). The contribution of the present work is therefore methodological and empirical rather than a new entropy estimator: it establishes what specimen-level GroupKFold validation and axis-resolved SHAP attribution reveal when applied to standard MPE on operational data, complementing rather than competing with the entropy estimator refinements summarized above.

The 100% accuracy reported by Huo et al. (2019) is not an isolated result; the same study tabulates five prior MPE-based methods spanning 2012–2018, all reporting 100% classification accuracy on laboratory test rig data with 4–11 fault states. This pattern indicates that headline accuracy on small, controlled laboratory benchmarks has been saturated for over a decade and is no longer a meaningful axis of differentiation for MPE-based bearing diagnostics. Notably, Huo et al.

(2019) achieves this 100% result on the CWRU dataset—the same benchmark for which Wheat et al. (2024) and Vieira, Bauler, Rosa, and Silva (2025) independently document substantial data leakage under conventional splitting protocols, suggesting that part of this saturation may itself be an artefact of validation design rather than genuine diagnostic ceiling. The present study therefore does not compete on benchmark accuracy; its contribution lies in demonstrating what methodologically rigorous validation (GroupKFold) and axis-resolved attribution (SHAP) reveal when standard MPE is applied to operational data where perfect accuracy is neither expected nor, as Section 6 shows, achieved (97.73% specimen-level, with the remaining error concentrated in the smallest-sample fault category).

2.2. Identified Methodological Limitations

A systematic literature search was conducted to identify representative studies on bearing fault diagnosis spanning permutation entropy methods, tree-ensemble classifiers, deep learning architectures, and cross-validation methodology. Screening prioritised five criteria: (i) focus on bearing fault diagnosis, (ii) reporting of validation protocol, (iii) vibration-based data, (iv) reporting of performance metrics, and (v) clarity of data source. Seventeen studies spanning 2014 to 2026 were retained for detailed analysis, summarised in Table 1. The search was not intended to be exhaustive but to identify representative methodological patterns across entropy-based, ensemble, and deep learning approaches to bearing diagnosis.

Three recurring limitations emerge from the entropy-based and ensemble studies in rows 1–8 of Table 1. First, all employ window-level cross-validation without specimen-level separation, permitting data leakage from temporally correlated windows. Second, only two studies address wind turbine-specific bearing configurations; most use general-purpose test rigs (Sandoval et al., 2020; Qi, Mauricio, & Gryllias, 2020). Third, PE exhibits non-monotonic patterns for certain decomposition levels under variable operating conditions (Fu et al., 2019), suggesting that the physical interpretation of PE-based discrimination remains incomplete.

The 2024–2026 entries in Table 1 illustrate two further patterns relevant to the present design choices. First, transformer and hybrid CNN architectures achieve near-ceiling accuracy on laboratory benchmarks such as CWRU and Paderborn (Gao, Wang, Li, & Yao, 2024; Y. Zhou, Chen, Zhang, et al., 2025), but these datasets provide hundreds to thousands of labeled samples per class; the Fraunhofer LBF dataset provides 44 specimens total, an operating regime where deep architectures are prone to overfitting and where tree-ensemble methods with engineered features remain competitive and more data-efficient (Zhang, Zhong, Zhang, & Jiang, 2023). This data efficiency argument is reinforced by recent applications of XGBoost and hybrid RNN-XGBoost architectures to bearing and induction-motor fault diagnosis (Bhende, 2025; Khan, Asad, Naseer, et al., 2026), which similarly report ensemble methods matching or exceeding deep architecture performance, motivating the inclusion of XGBoost as a direct comparison point in the present study (Section 6).

2.3. Positioning of the Present Study

The present study extends PE-based bearing diagnostics to the Fraunhofer LBF operational wind turbine dataset (Mostafavi & Friedmann, 2024a) by combining two elements not jointly evaluated in prior MPE-based work summarised in Table 1: GroupKFold cross-validation enforcing specimen-level separation, and Shapley-based quantification of axis-specific diagnostic importance for triaxial bearing accelerometers. Rather than proposing a new entropy estimator, the contribution is methodological — establishing a validation protocol and an interpretability analysis that together yield deployment-realistic performance estimates and actionable sensor-placement guidance for operational wind turbine condition monitoring.

3. THEORETICAL BACKGROUND

3.1. Permutation Entropy

For a time series $\{x_i\}_{i=1}^N$, PE captures the distribution of ordinal patterns in embedded vectors of length m (Bandt & Pompe, 2002):

$$\mathrm{PE}(m,\tau) = -\sum_{\pi} p(\pi)\log_2 p(\pi) \tag{1}$$

where the sum is over all $m!$ possible ordinal patterns π, and $p(\pi)$ is the relative frequency of pattern π in the embedding. Normalized PE divides by $\log_2(m!)$, bounding values in $[0,1]$.

3.2. Weighted Permutation Entropy

Weighted PE incorporates amplitude information through variance-based pattern weights (S. Zhou, Qian, Chang, Xiao, & Cheng, 2018):

$$\mathrm{WPE}(m,\tau) = -\sum_{\pi} \frac{w_\pi}{\sum_{\pi'} w_{\pi'}} \log_2 \frac{w_\pi}{\sum_{\pi'} w_{\pi'}} \tag{2}$$

where $w_\pi = \mathrm{Var}\big(\{x_i,\ldots,x_{i+(m-1)\tau}\}\big)$ is the variance of the embedding vector associated with pattern π.

3.3. Multiscale Permutation Entropy

MPE extends PE across temporal scales via coarse-graining (Costa et al., 2002; Wang et al., 2020). For scale factor s, the coarse-grained signal is:

$$y_j^{(s)} = \frac{1}{s}\sum_{i=(j-1)s+1}^{js} x_i, \quad j = 1,2,\ldots,\lfloor N/s \rfloor \tag{3}$$

Table 1. Systematic review of bearing fault diagnosis methods, extended to include 2024–2026 deep learning, ensemble, and entropy-based studies (n=17). None of the entropy-based or ensemble studies (rows 1–8) employ specimen-level cross-validation separation; the remaining rows represent recent deep learning, ensemble, and validation-methodology contributions.

Study	Method	Dataset	Classifier	Accuracy	Validation
Fu et al. (2019)	PE+VMD	Test rig	Multi-SVM	92.4–98.6%	Stratified split
Wang et al. (2020)	MPE+MSSM	Test rig	BAS-SVM	100%	N/S
Zhou et al. (2018)	WPE+EEMD	Cincinnati	SVM	97.78%	2/3 split
Ye et al. (2021)	MPE+VMD	CWRU	PSO-SVM	96.42–100%	Random split
Sandoval et al. (2020)	PE+SpEn	Pitch brng	Threshold	N/R	Fixed split
Minhas & Singh (2021)	SP2IMPE	Test rig	Feature-based	+3–18% vs wPE/IMPE	N/S
Huo et al. (2019)	F2CMPE+LS	CWRU	SVM	100% (10 cond.)	10-fold CV
Qi et al. (2020)	PE (blind)	Gearbox	Blind ind.	Best of entropy	N/R
2024–2026 deep learning and ensemble methods:					
Gao et al. (2024)	Twins Transformer	Lab rig	CNN+Transformer	Near-ceiling	N/S
Bhende (2025)	XGBoost	Vibration	XGBoost	95.48%	Stratified
Khan et al. (2026)	RNN-XGBoostNet	CWRU, IM	RNN+XGBoost	87.0→99.4%	N/S
Zhang et al. (2023)	12 tree models	XJTU-SY	gcForest/RF/XGB	up to 99.4%	3% training data
2024–2026 validation methodology and sensor attribution:					
Wheat et al. (2024)	Split comparison	CWRU, others	6 methods	Δ>40pp	3 split types
Vieira et al. (2025)	Bearing-wise split	CWRU	Multi-label	N/R	Component-aware
Nirmal et al. (2026)	Multi-domain features	Triaxial	XGBoost/RF/SHAP	96.65%	Train/test split
This study	MPE	Fraunhofer LBF	RF/XGB	97.73–100%	GroupKFold (specimen)

Notes: N/R = not reported; N/S = not specified; Imp. = improved; Brng = bearing; SpEn = spectral entropy; MSSM = Mahalanobis semi-supervised mapping; F2CMPE = fine-to-coarse multiscale permutation entropy; LS = Laplacian score; IM = induction motor.

PE is then computed at each scale, yielding the multiscale profile $\{\mathrm{PE}(s)\}_{s\in\mathcal{S}}$. Scale 1 recovers standard PE of the original signal; coarser scales reveal complexity at progressively lower frequencies relevant to bearing fault modulation.

3.4. Multi-Window Temporal Segmentation

To increase the number of independent analysis units per specimen, each signal of length N is segmented into overlapping windows of size W with step $S = W(1-\alpha)$:

$$M = \left\lfloor \frac{N-W}{S} \right\rfloor + 1 \tag{4}$$

where $\alpha \in [0,1)$ is the overlap fraction and M is the total number of windows extracted from the signal. Each window $w_k = x[kS : kS{+}W]$ is processed independently, yielding one feature vector per window. In this study, W=2000 samples (78.1 ms at 25.6 kHz) and α=0.50, giving S=1000 samples. Specimen-level classification is obtained by majority voting across all M windows of a given specimen.

3.5. GroupKFold Validation Rationale

Standard stratified K-fold cross-validation splits the $M \times |\mathcal{S}|$ analysis windows randomly, allowing windows from the same specimen to appear in both training and test partitions. Because windows derived from the same signal share bearing-specific characteristics (surface finish, mounting position, operating history), this constitutes data leakage that inflates reported accuracy. GroupKFold assigns a unique group identifier to each specimen (`condition_specimenID`) and guarantees:

$$\mathcal{G}_{\text{train}}^{(k)} \cap \mathcal{G}_{\text{test}}^{(k)} = \emptyset, \quad \forall\, k \in \{1,\ldots,K\} \tag{5}$$

where $\mathcal{G}_{\text{train}}^{(k)}$ and $\mathcal{G}_{\text{test}}^{(k)}$ indicate the sets of group identifiers present in the training and test partitions of fold k, respectively.

This ensures that performance estimates reflect genuine cross-specimen generalization rather than within-specimen memorization.

4. DATASET AND EXPERIMENTAL SETUP

4.1. Fraunhofer LBF Wind Turbine Bearing Dataset

Experimental validation employs the Fraunhofer LBF wind turbine bearing dataset (Mostafavi & Friedmann, 2024a), acquired from a physical wind turbine test bench under realistic variable wind loading and made publicly available via (Mostafavi & Friedmann, 2024b). Data were recorded continuously at 25.6 kHz and stored in 5-minute files; the final file of each session may be shorter due to early stoppage from below-cut-in wind speed. Bearing faults were induced via surface etching (IDEAL etching pencil), producing low-severity artificial defects that alter the ordinal complexity structure of vibration signals without necessarily increasing amplitude — motivating entropy-based analysis over amplitude-based features. Four fault conditions are evaluated, as summarized in

Table 2.

Table 2. Fraunhofer LBF dataset characteristics after data quality filtering. Two InnerRace recordings were excluded due to non-physical sensor values consistent with ADC overflow.

Condition	Specimens	Duration	Windows	Excluded
H	18	90 min	774	0
IR	10	60 min	430	2
OR	4	20 min	132	0
RE	12	60 min	1,320	0
Total	**44**	—	**2,656**	**2**

H = Healthy; IR = Inner Race; OR = Outer Race; RE = Roller Element.

The class distribution across analysis windows reflects the unequal measurement durations: Roller Element contributes 49.7% of windows (1,320/2,656) due to the longer recording duration (60 min, 12 specimens), while Outer Race contributes only 4.9% (132/2,656) owing to the 20-minute measurement protocol. This imbalance at window level does not propagate to specimen level, where the distribution is: Healthy (18 specimens, 40.9%), Inner Race (10, 22.7%), Outer Race (4, 9.1%), and Roller Element (12, 27.3%). Random Forest with default class weights is used without resampling; the GroupKFold folds are constructed to distribute specimens across conditions as evenly as possible given the specimen counts, resulting in fold sizes of 8–10 specimens per test partition (Table 3); the MPE per-fold accuracy sequence [94.3%, 100.0%, 99.8%, 97.3%, 96.9%] confirms balanced generalization across partitions despite fold 2 containing no outer race specimens.

The outer race condition warrants particular attention. With only 4 specimens and a 20-minute measurement protocol — the shortest of all conditions, reflecting the higher severity of outer race defects in this experimental setup — outer race classification is inherently high-variance. Any single misclassification represents a 25-percentage-point drop in specimen-level accuracy for this condition. This constraint is an intrinsic property of the dataset and not addressable through algorithmic improvements; future validation should include additional outer race specimens to obtain reliable performance estimates for this fault type.

4.2. Data Quality Screening and Sensor Selection

Two InnerRace recordings (signal_id = 8 and 10) were excluded following systematic amplitude screening. Both channels exhibited non-physical values exceeding 10^{32} g across the complete signal duration, consistent with ADC overflow or data acquisition failure in the final recording interval of a measurement session. The remaining 44 specimens are used in all analyses.

Rear bearing accelerometers (brng.r.x/y/z) were excluded due to baseline signal amplitudes $19.6\times$ lower than front bearing channels under healthy conditions (σ: 0.002–0.016 vs. 0.128–0.157 g). Under the low-severity etching-induced fault conditions of this dataset, surface defects are insufficient to generate detectable complexity changes at the rear bearing location. Anemometer (anm.mst, anm.roof) and wind vane (van) channels were excluded because their DC-coupled output (0–10 V, 0–50 m/s wind speed; 0–360° direction) is physically incompatible with entropy-based vibration complexity analysis. The final sensor set comprises three front bearing accelerometer axes: `brng.f.x`, `brng.f.y`, `brng.f.z`.

4.3. Signal Preprocessing and Windowing

Each recording is segmented into overlapping windows of W=2000 samples (78.1 ms at 25.6 kHz) with 50% overlap (S=1000 samples), yielding an average of 60 windows per specimen and 2,656 windows in total. Prior to feature extraction, each window is clipped to ± 250 g to remove any residual non-physical values arising from the excluded channels and then normalized to zero mean and unit variance: $w \leftarrow (w - \bar{w})/\hat{\sigma}_w$.

5. METHODOLOGY

5.1. Feature Sets

Three feature sets are evaluated under identical conditions:

MPE (proposed reference): Multiscale PE at scales $s \in \{1, 2, 4, 8\}$ with embedding dimension m=4, computed per accelerometer axis using coarse-grained signals (Eq. 3). This yields 12 features (4 scales $\times$ 3 axes).

WPE (baseline): Weighted PE with m=4, τ=1, amplitude-weighted ordinal pattern frequencies (Eq. 2), yielding 3 features (1 per axis). Implemented using vectorized variance -weighted bincount, achieving 6.3 ms per window versus 249 ms for the naïve loop implementation.

MPE+Physical (proposed hybrid): MPE augmented with three physically interpretable variance-scaling features per axis — variance ratio ($\mathrm{Var}_{s=1}/\mathrm{Var}_{s=8}$), scale slope (slope of $\log \mathrm{Var}$ vs. $\log s$), and scaling exponent (Hurst proxy via variance scaling across scales 2, 4, 8, 16) — yielding 21 features (12 MPE + 9 physical). Each feature targets a distinct aspect of bearing vibration dynamics. The variance ratio quantifies how vibration energy redistributes from high-frequency bearing-element-passage events (scale 1) to lower-frequency structural modulations (scale 8); a fault that introduces strong high-frequency impacts without proportional low-frequency modulation increases this ratio. The scale slope characterises the rate of energy decay across scales as a single coefficient, summarising whether the vibration spectrum is dominated by localised impulses (steep decay) or by broadband excitation (shallow decay). The scaling exponent (Hurst proxy) quantifies the self-similarity of the signal's variance structure across scales 2–16; values near 0.5 indicate uncorrelated (white-noise-like) dynamics, while deviations indicate persistent or anti-persistent fault-induced patterns consistent with the multifractal characteristics reported for bearing degradation (P. R. Almeida, Lima, Brito, & Lima Filho, 2026).

All features are extracted using a parallelized pipeline (7 CPU cores, Python joblib) with total processing time of 3.6 minutes for the complete dataset. Permutation entropy is computed via ordpy (order m=4); variance-based features use NumPy coarse-graining.

5.2. Feature Extraction Pipeline

Algorithm 1 formalises the complete feature extraction and validation procedure. The pipeline processes each bearing signal independently: after clipping and normalisation, each analysis window yields 12 MPE features (4 scales × 3 axes). Group identifiers are assigned at specimen level before cross-validation, guaranteeing that the leakage check in line 12 of Algorithm 1 is satisfied at every fold.

Algorithm 1 MPE Extraction and GroupKFold Validation Pipeline

Require: Bearing signals $\{x_i\}_{i=1}^{N_s}$, labels $\{y_i\}$, window W=2000, step S=1000, scales $\mathcal{S}=\{1,2,4,8\}$, m=4, folds K=5
Ensure: Window-level and specimen-level accuracy
1: **for** each signal x_i with specimen identifier g_i **do**
2: Clip: $x_i \leftarrow \text{clip}(x_i,\, -250\,\text{g},\, +250\,\text{g})$
3: Segment into M_i windows of size W, step S
4: **for** each window w_k, $k = 1, \ldots, M_i$ **do**
5: Normalise: $w_k \leftarrow (w_k - \bar{w}_k)/\hat{\sigma}_k$
6: **for** each axis $a \in \{\text{x}, \text{y}, \text{z}\}$ **do**
7: **for** each scale $s \in \mathcal{S}$ **do**
8: Coarse-grain: $y^{(s)} \leftarrow \text{CoarseGrain}(w_k^{(a)},\, s)$ (Eq. 3)
9: $f_{k,a,s} \leftarrow \text{PE}(y^{(s)},\, m)$ (Eq. 1)
10: **end for**
11: **end for**
12: **end for**
13: **end for**
14: Assign group identifiers: $g_i \leftarrow$ specimen ID
15: **for** fold $k = 1$ to K **do**
16: **Assert**: $\mathcal{G}_{\text{train}}^{(k)} \cap \mathcal{G}_{\text{test}}^{(k)} = \emptyset$
17: Fit StandardScaler on training windows only
18: Fit RF (300 trees) and XGBoost (300 trees) on scaled training features
19: Predict test windows for each classifier; majority vote per specimen
20: **end for**
21: **return** mean window accuracy ± std; specimen accuracy (per classifier)

5.3. Classification and Validation Protocol

Random Forest with 300 estimators (scikit-learn 1.3.2, random_state = 42, n_jobs = -1, all other hyperparameters at default values) is trained and evaluated under GroupKFold (K=5) with unique group identifiers per specimen (`condition_specimenID`). XGBoost (version 3.2.0, 300 estimators, `eval_metric=mlogloss`, `random_state=42`, default hyperparameters otherwise) is evaluated under identical conditions for comparison (Section 6). Permutation entropy is computed via ordpy (version 1.2.2); the analysis pipeline uses Python 3.11.15, NumPy 1.26.4, pandas 3.0.2, and statsmodels 0.14.6 for statistical testing. StandardScaler is fitted on training data within each fold and applied to test data, preventing information leakage through normalization. The absence of leakage is verified at each fold by ensuring that the training and test groups are disjoint, i.e., $\mathcal{G}_{\text{train}} \cap \mathcal{G}_{\text{test}} = \emptyset$.

Table 3 reports the resulting partition composition. Each fold contains specimens from all four conditions where possible; fold 2 contains no outer race specimens, an unavoidable consequence of distributing four outer race specimens across five folds.

Table 3. GroupKFold partition composition (K=5). Each fold contains specimens from all four conditions where possible given the specimen counts; fold 2 contains no outer race specimens due to the limited total of four.

Fold	H	IR	OR	RE	Specimens	Windows
1	2	2	1	3	8	545
2	2	3	0	3	8	545
3	5	2	1	2	10	524
4	5	1	1	2	9	521
5	4	2	1	2	9	521
Total	**18**	**10**	**4**	**12**	**44**	**2,656**

Two accuracy metrics are reported throughout. *Window-level accuracy* is mean fold accuracy across individual 2000-sample windows. *Specimen-level accuracy* applies majority voting across all windows of each specimen, representing the operationally relevant metric for bearing-level diagnostic decisions.

5.4. Feature Importance Quantification

Two complementary importance measures are applied. Mean decrease in impurity (MDI), averaged across the five GroupKFold folds, quantifies aggregate feature importance with fold-to-fold standard deviation. TreeExplainer Shapley attribution (Lundberg & Lee, 2017) is computed on the first fold's test set ($n = 521$ windows), providing per-feature and per-sensor importance values. The SHAP array, of shape (n_s, n_f, n_c), where n_s is the number of samples, n_f the number of features, and n_c the number of classes, is aggregated as

$$\overline{|\phi|}_{ij} = \mathbb{E}_s[\mathbb{E}_c(|\phi_{ij}|)]\,, \tag{6}$$

yielding per-feature importance scores.

5.5. Permutation Entropy Distributional Analysis

Permutation entropy (PE) distributions are characterized per condition and per sensor axis to quantify both central tendency and dispersion, as well as inter-class separability. For each condition–axis pair, the following statistics are computed: mean (μ), median, standard deviation (σ), 5th and 95th percentiles, and the distribution mode estimated from a 100-bin histogram. These descriptors capture both the typical complexity level and the spread induced by intra-condition variability.

To assess class separability, Cohen's d effect size is computed for all pairwise condition comparisons:

$$d = \frac{|\mu_A - \mu_B|}{s_p}, \quad s_p = \sqrt{\frac{(n_A - 1)\sigma_A^2 + (n_B - 1)\sigma_B^2}{n_A + n_B - 2}} \tag{7}$$

where μ_A, μ_B, σ_A, σ_B are the condition means and standard deviations, n_A, n_B are the window counts, and s_p is the pooled standard deviation. Following Cohen's criterion, $d \geq 0.8$ denotes a large effect size (Cohen, 1988). This scale-invariant measure complements classification-based metrics by quantifying PE discriminability independently of the classifier.

In addition to unimodal statistics, distributional shape is examined. In particular, the roller element condition exhibits a bimodal PE distribution, which is characterized through visual inspection and interpreted in terms of load-zone-dependent dynamics. This bimodality reflects the intermittent nature of rolling element defects and is further analyzed in the Results section.

5.6. Statistical Significance Testing

To assess whether observed accuracy differences reflect genuine performance gaps rather than fold-to-fold variability, McNemar's test (McNemar, 1947) is applied to out-of-fold window-level predictions. For each pair of methods, predictions across all five GroupKFold folds are pooled into a single 2×2 contingency table of agreement/disagreement with ground truth, and the chi-squared statistic with continuity correction is computed. Significance is assessed at α=0.05.

6. RESULTS

6.1. Comparative Diagnostic Performance

Table 4 presents classification performance for five feature–classifier configurations under GroupKFold cross-validation. MPE-RF achieves 97.73% specimen-level accuracy (43/44) with only 12 features, matching PE+Complexity's window-level accuracy (98.43%) with half the features and improving specimen-level accuracy by 2.28 percentage points. WPE achieves 95.45% specimen-level accuracy with only 3 features — equivalent to PE+Complexity (24 features), demonstrating that amplitude-weighted ordinal patterns capture much of the dominant diagnostic information at minimal dimensionality. MPE+Physical exactly matches MPE-RF at specimen level, confirming that physical scaling features add no diagnostic value beyond what MPE already encodes.

Fold-level accuracy for MPE-RF is 94.3%, 100.0%, 99.8%, 97.3%, and 96.9%, demonstrating consistent performance across specimen groups, with no fold showing accuracy below 94%. Outer race (OR) specimen-level performance improves from 2/4 with WPE and PE+Complexity to 3/4 with MPE-RF and MPE+Physical, and to 4/4 with MPE-XGBoost. The remaining MPE-RF misclassification (`OuterRace_4`) results from a truncated 3.1-second recording yielding only 3 analysis windows, which is insufficient for reliable majority voting, rather than a limitation of the proposed framework. Figure 1 presents the window-level confusion matrices and the accuracy comparison across all five configurations.

The fifth configuration evaluates XGBoost (300 estimators, `eval_metric=mlogloss`, `random_state=42`) on the same 12-feature MPE representation under identical GroupKFold partitions. XGBoost achieves $98.09 \pm 1.93\%$ window-level accuracy and 100% specimen-level accuracy (44/44), correctly classifying all four outer race specimens where Random Forest misclassified one. Fold-level scores are 98.90%, 100.0%, 99.81%, 96.74%, and 95.01%, showing a similar range to RF but with the variance concentrated in different folds, suggesting that the two algorithms make complementary errors on the boundary cases. This result directly addresses calls for benchmarking against modern ensemble methods: with identical features and validation protocol, gradient boosting provides a measurable improvement over Random Forest on this dataset, particularly for the smallest-sample fault category.

McNemar's test confirms that the accuracy difference between MPE and WPE is statistically significant for both classifiers (MPE-RF vs. WPE-RF: χ^2=115.69, p<0.001; MPE-XGBoost vs. WPE-RF: χ^2=123.57, p<0.001), confirming that the 12-feature multiscale representation provides genuine discriminative advantage over the 3-feature weighted single-scale representation. In contrast, the window-level difference between MPE-RF and MPE-XGBoost is not statistically significant (χ^2=2.33, p=0.127): of 2,656 windows, both classifiers agree on 2,574 (96.9%), with MPE-XGBoost correcting 32 windows that MPE-RF misclassified and MPE-RF correcting 20 windows that MPE-XGBoost misclassified. The specimen-level improvement from XGBoost (100% vs. 97.73%) therefore reflects a favourable redistribution of a small number of window-level errors across specimen boundaries rather than a broad window-level superiority.

6.2. Axis-Specific Diagnostic Importance

Figure 2 presents feature importance and diagnostic performance for the PE+Complexity exploratory pipeline (24 features). Panel (a) shows RF feature importance decomposed by accelerometer axis: the X-axis (brng.f.x) accounts for 40.7% of total importance, the Z-axis (brng.f.z) for 30.1%, and the Y-axis (brng.f.y) for 29.3%, against a uniform baseline of 1/3. Shapley attribution (panel c) confirms this hierarchy indepen-

Table 4. Comparative diagnostic performance under GroupKFold cross-validation (K=5, 44 specimens, Fraunhofer LBF dataset). Window accuracy: mean $\pm$ std across folds. Specimen accuracy: majority voting. OR = outer race specimen accuracy (4 specimens total).

Method	Features	Window (%)	Specimen (%)	Correct	OR
WPE (S. Zhou et al., 2018)	3	91.83 ± 1.89	95.45	42/44	2/4
PE+Complexity (exploratory)	24	98.43 ± 1.56	95.45	42/44	2/4
MPE-RF (reference)	12	97.67 ± 2.10	97.73	43/44	3/4
MPE+Physical (proposed)	21	97.63 ± 1.88	97.73	43/44	3/4
MPE-XGBoost	**12**	$\mathbf{98.09 \pm 1.93}$	**100.00**	**44/44**	**4/4**

Healthy: 18/18 (100%), InnerRace: 10/10 (100%), RollerElement: 12/12 (100%) for all methods.

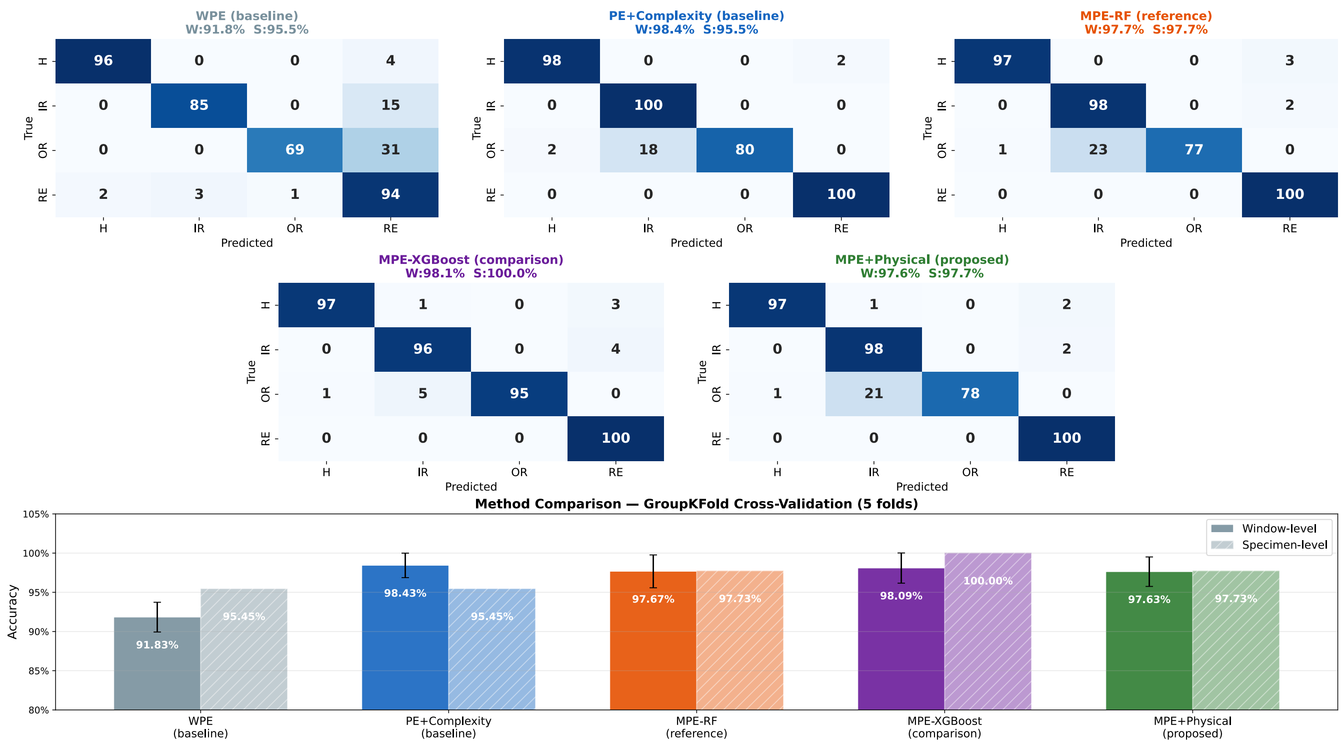

Figure 1. Comparative diagnostic performance across five feature/classifier configurations under GroupKFold cross-validation (K=5, 44 specimens, Fraunhofer LBF dataset). *Top two rows*: window-level confusion matrices (%) for WPE, PE+Complexity, MPE-RF, MPE-XGBoost, and MPE+Physical; method labels show window-level (W) and specimen-level (S) accuracy. *Bottom row*: accuracy comparison with window-level (solid bars, mean $\pm$ std across folds) and specimen-level (hatched bars) metrics. MPE-XGBoost achieves 100% specimen-level accuracy (44/44), correctly classifying all four outer race specimens where Random Forest misclassifies one.

dently: brng.f.x contributes 37.0% of mean absolute SHAP value, brng.f.z 33.1%, and brng.f.y 29.6%. This axis hierarchy is consistent with the MPE GroupKFold pipeline reported in Section 6.

The X-axis dominance is physically consistent with the primary radial load direction of the bearing under wind thrust forces. Bearing contact forces are predominantly transmitted in the main load direction, generating larger fault-induced impulsive events in that axis relative to transverse directions.

Panels (b) and (d) decompose importance by feature type (mean across axes) under the RF and SHAP metrics respectively; permutation entropy is the most discriminative measure under both, followed by spectral entropy and scaling exponent. Panel (e) shows the window-level confusion matrix (98.43%) for this pipeline under GroupKFold cross-validation with strict specimen-level separation, and panel (f) shows specimen-level accuracy by fault condition: Healthy (18/18), Inner Race (10/10), and Roller Element (12/12) achieve 100%, while Outer Race ($2/4$, 50%) reflects the limited specimen count for this condition (4 specimens total).

6.3. MPE Scale Importance and Physical Interpretation

Within the MPE feature set, scale 1 exhibits the highest mean importance (0.1342 per axis), followed by scale 8 (0.0952), scale 4 (0.0575), and scale 2 (0.0465), as shown in Table 5.

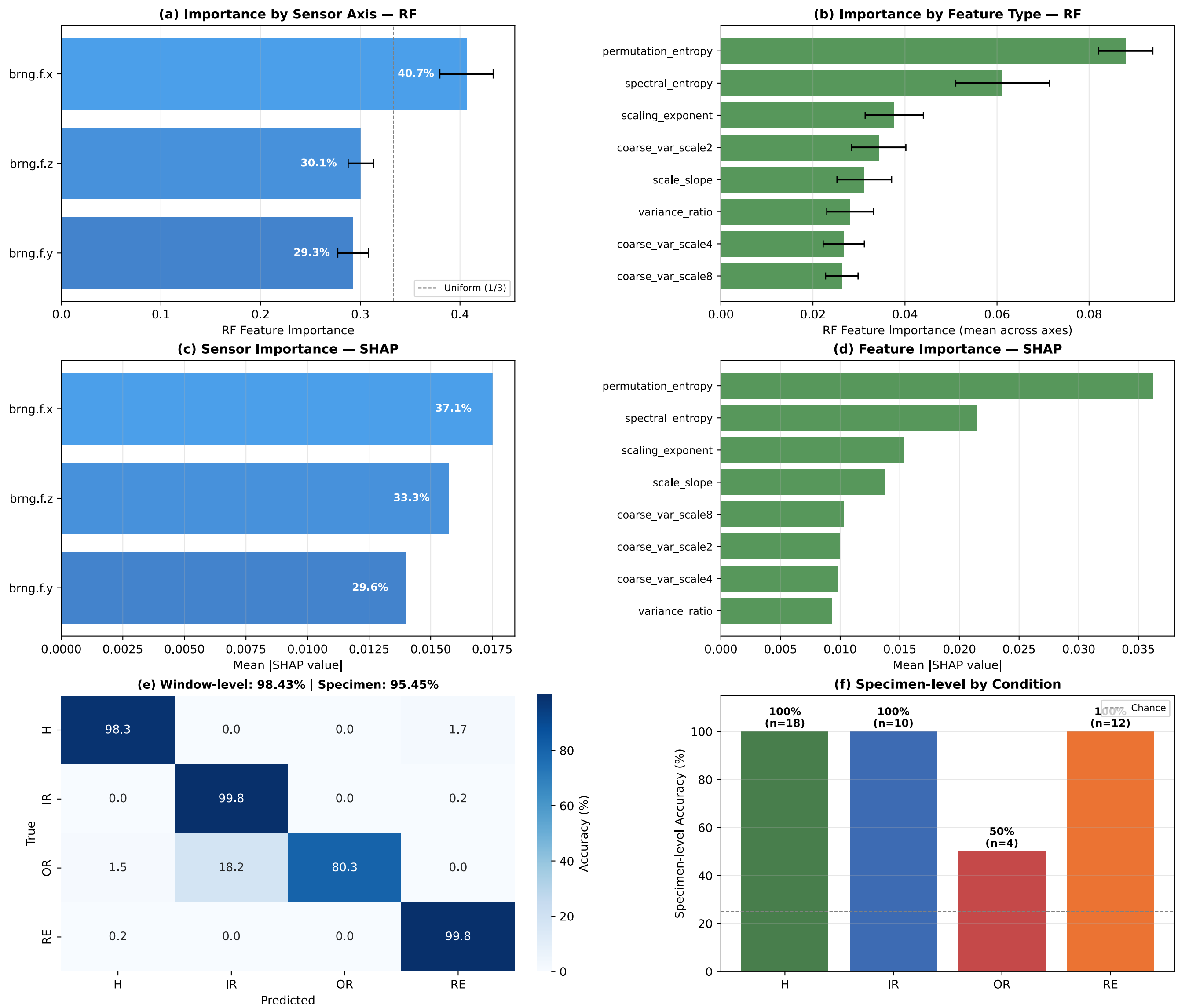

Figure 2. Feature importance and diagnostic performance for the PE+Complexity exploratory pipeline (24 features, brng.f.x/y/z, 44 specimens, Fraunhofer LBF dataset). (a) RF importance by sensor axis. (b) RF importance by feature type (mean across axes). (c) SHAP attribution by sensor axis. (d) SHAP attribution by feature type. (e) Window-level confusion matrix under GroupKFold cross-validation with specimen-level separation. (f) Specimen-level accuracy by fault condition.

The prominence of scale 8 as the second most important scale suggests that fault-induced complexity changes propagate from high-frequency bearing element interactions (scale 1) to lower-frequency structural modulations (scale 8), with intermediate scales contributing less discriminative information.

The dominance of scale 1 has a direct physical interpretation. At the original 25.6 kHz sampling resolution, scale 1 captures complexity in the frequency range of bearing element passage frequencies: ball pass frequency outer race (BPFO), ball pass frequency inner race (BPFI), and ball spin frequency (BSF) for typical wind turbine operating speeds of 10–20 rpm. These characteristic frequencies generate impulsive events whose ordinal pattern structure is disrupted by surface defects, mak-

Table 5. MPE feature importance by scale and accelerometer axis (mean decrease in impurity, averaged across 5 GroupKFold folds). Scale 1 dominates in brng.f.x and brng.f.y; brng.f.z shows highest importance at scale 1 but distributes across scales differently from the other axes.

Axis	Scale 1	Scale 2	Scale 4	Scale 8
brng.f.x	0.1186	0.0454	0.0373	0.1051
brng.f.y	0.1199	0.0369	0.1034	0.1440
brng.f.z	0.1640	0.0573	0.0316	0.0365
Mean	**0.1342**	**0.0465**	**0.0575**	**0.0952**

ing scale-1 PE the most sensitive complexity indicator for the low-severity etching-induced faults in this dataset (mean importance 0.1342, Table 5).

Scale 8 ranks second in importance (mean 0.0952), capturing complexity at approximately $1/8$ of the original resolution, corresponding to modulation frequencies in the range of shaft rotation rate. The prominence of scale 8 as the second most important scale is consistent with the amplitude modulation mechanism of rolling element bearing faults, where impulsive events at bearing element passage frequencies are modulated at shaft rotation rate. Scales 2 and 4 contribute less (0.0465 and 0.0575, respectively), likely because they capture intermediate frequencies where neither bearing element passage nor rotational modulation is dominant.

6.4. Physical Feature Contribution

Within the hybrid MPE+Physical framework, MPE features account for 67.0% of total feature importance and physical features for 33.0%. The most significant physical feature is brng.f.x_variance_ratio (I=0.0679), followed by brng.f.z_scaling_exponent (I=0.0626). Despite their substantial individual importance, the ΔSpecimen $= 0.00$ pp result demonstrates that MPE implicitly encodes the self-similarity and energy redistribution characteristics that these variance-based measures quantify explicitly. Scale-1 MPE (I=0.1035) subsumes the diagnostic information in variance ratio (I=0.0679), and the MPE profile across scales captures the scaling exponent structure.

6.5. Permutation Entropy Hierarchy

Table 6 presents distributional statistics for PE (brng.f.x, scale 1) across fault conditions.

Table 6. Permutation entropy statistics by fault condition (brng.f.x, scale 1, m=4). All pairwise Cohen's d exceed 0.9 (large effect by Cohen's criterion $d \geq 0.8$).

Condition	μ	σ	Mode	n
Healthy	0.702	0.119	0.968	774
Roller Element	0.889	0.142	0.981	1,320
Inner Race	0.981	0.007	0.986	430
Outer Race	0.992	0.009	0.997	132

A monotonic hierarchy is established: Healthy $<$ Roller Element $<$ Inner Race $<$ Outer Race. Effect sizes are uniformly large: Healthy vs. Outer Race d=3.443, Healthy vs. Inner Race d=3.322, Healthy vs. Roller Element d=1.427, Inner Race vs. Outer Race d=1.298, Outer Race vs. Roller Element d=1.019, Inner Race vs. Roller Element d=0.916. This hierarchy is established for a single bearing model under the fault configurations present in the Fraunhofer LBF dataset; whether the same ordering holds for other bearing geometries, fault severities, or application domains has not been tested and is addressed as a limitation in Section 7.

6.6. Roller Element Bimodality

Roller element faults exhibit the largest within-condition variability ($\sigma = 0.142$), with a bimodal PE distribution visible across all three accelerometer axes. The lower mode (PE $\approx$ 0.56) corresponds to windows where the damaged roller traverses the load zone, generating periodic impulsive events that suppress ordinal variability. The upper mode (PE $\approx$ 0.98) corresponds to non-load-zone windows, where the vibration signature approaches the healthy regime. This intermittent fault signature is consistent with the rotating nature of roller element defects and explains the higher within-condition variance relative to inner and outer race faults, where the defect location is fixed with respect to the load zone.

7. DISCUSSION

7.1. GroupKFold as Methodological Contribution to PHM Benchmarking

The specimen-level GroupKFold protocol produces conservative accuracy estimates that more faithfully represent deployment performance than window-level splitting. The literature reports 92.4–100% accuracy under window-level splitting without specimen separation (Fu et al., 2019; Wang et al., 2020; Ye et al., 2021); these values are not directly comparable because they incorporate information leakage from same-specimen windows. The present results establish a reproducible methodological baseline for the Fraunhofer LBF dataset: 97.73% specimen-level accuracy under GroupKFold with MPE. Future studies on this dataset should adopt equivalent protocols to enable meaningful performance comparison.

7.2. Comparison with Prior PE-Based Results

Direct numerical comparison with prior studies requires caution because validation protocols differ substantially. Table 1 shows that the eight entropy-based and ensemble studies retrieved (rows 1–8) all use window-level splitting without specimen separation. Under equivalent window-level metrics, the present MPE framework achieves $97.67 \pm 2.10\%$, comparable to the 97.78% reported by S. Zhou et al. (2018) for WPE on the University of Cincinnati dataset and the 96.42%–100% reported by Ye et al. (2021) for VMD-MPE on CWRU.

Three methodological differences distinguish the present results. First, the Fraunhofer LBF dataset consists of operational wind turbine data under variable wind loading, whereas all comparison studies use laboratory test rigs under controlled conditions. Second, GroupKFold specimen-level separation produces conservative estimates that window-level splitting inflates; the 97.73% specimen- level accuracy reported here would be expected to decrease further under cross-turbine validation, representing a more honest estimate of deployment performance. Third, no signal decomposition preprocessing (VMD, EEMD) is applied, in contrast to the VMD-MPE pipelines of Ye et al. (2021) and Fu et al. (2019). The present results demonstrate that raw triaxial MPE without decomposition achieves competitive performance when evaluated under methodologically sound protocols, suggesting that signal decomposition may partially compensate for validation leakage rather than providing genuine feature improvement.

The 3-feature WPE baseline achieving 95.45% specimen-level accuracy merits particular attention. S. Zhou et al. (2018) required 8-feature WPE combined with EEMD decomposition and ensemble SVM to achieve 97.78% window-level accuracy on a different dataset. The present WPE result — 3 features, no decomposition, RF classifier, GroupKFold — demonstrates that amplitude-weighted ordinal patterns capture dominant diagnostic information with minimal feature engineering when the validation protocol prevents leakage-inflated estimates.

7.3. Tree Ensembles versus Deep Learning Architectures

Convolutional and transformer architectures achieve near-ceiling accuracy on CWRU and Paderborn (Gao et al., 2024; Y. Zhou et al., 2025), raising the question of why such architectures are not benchmarked here. Three considerations motivate the tree-ensemble choice instead. First, sample size: the Fraunhofer LBF dataset provides 44 specimens, an order of magnitude below typical deep learning benchmarks; Zhang et al. (2023) report that gcForest and Random Forest exceed 98% accuracy on the XJTU-SY dataset using only 3% of training data, outperforming XGBoost and neural-tree hybrids in this small-sample regime — directly supporting tree ensembles as the appropriate model class here. Second, the present results show that, with methodologically sound validation, tree ensembles already achieve 97.73–100% specimen-level accuracy (Table 4); the diagnostic ceiling for this dataset under GroupKFold has effectively been reached with engineered MPE features, leaving limited room for architecture-driven improvement. Third, interpretability: SHAP attribution on tree ensembles (Section 6) provides axis-resolved and feature-resolved explanations directly relevant to sensor placement decisions, whereas comparable attribution for deep architectures typically requires additional post-hoc machinery (Nirmal, Lakmal, Arachchige, & Rodrigo, 2026). The XGBoost result (Table 4) demonstrates that gradient boosting provides a measurable specimen-level improvement over Random Forest within the tree-ensemble family, particularly on the hardest-to-classify fault category, while preserving the interpretability and small-sample suitability discussed above. McNemar's test indicates that this improvement arises from a favourable redistribution of a small number of window-level errors rather than a broad accuracy gain (Section 6), consistent with both algorithms having converged toward the diagnostic ceiling achievable with MPE features on this dataset.

7.4. Physical Interpretation of the PE Hierarchy

The monotonic PE hierarchy — Healthy $<$ Roller Element $<$ Inner Race $<$ Outer Race — reflects the progressive disruption of ordinal pattern regularity by surface etching defects. Healthy bearings exhibit the lowest mean PE (μ=0.702) and highest variability (σ=0.119), reflecting diverse operational dynamics under variable wind loading. Fault conditions systematically increase PE toward the theoretical maximum, with outer race defects producing the broadest spectral excitation and thus the highest ordinal complexity (μ=0.992).

The dramatic reduction in standard deviation from healthy (σ=0.119) to faulty conditions ($\sigma<0.015$ for IR and OR) indicates that faults constrain bearing dynamics to a narrower, more reproducible complexity regime. This collapse of PE variability may precede the rise in mean PE during early fault progression, suggesting that distributional width may be a sensitive early warning indicator — a direction warranting future investigation with progressive fault datasets.

7.5. Cross-Axis Consistency of the PE Hierarchy

The monotonic PE hierarchy established for brng.f.x is reproduced across all three accelerometer axes, confirming that fault-induced ordinal complexity changes reflect a global alteration of bearing dynamics rather than axis-specific measurement artefacts. Table 7 presents the distributional statistics for all three axes.

Table 7. Permutation entropy mean (μ) and standard deviation (σ) by fault condition and accelerometer axis (scale 1, m=4). The monotonic hierarchy H $<$ RE $<$ IR $<$ OR is preserved across all three axes.

	brng.f.x		**brng.f.y**		**brng.f.z**	
Cond.	μ	σ	μ	σ	μ	σ
H	0.702	0.119	0.699	0.086	0.782	0.093
RE	0.889	0.142	0.838	0.158	0.938	0.074
IR	0.981	0.007	0.924	0.012	0.976	0.013
OR	0.992	0.009	0.965	0.006	0.993	0.009

H = Healthy; IR = Inner Race; OR = Outer Race; RE = Roller Element.

Two observations emerge from the cross-axis comparison. First, within-condition standard deviation is consistently lowest for inner race and outer race faults across all three axes ($\sigma<0.015$), confirming that fixed-location defects produce reproducible ordinal complexity regardless of measurement direction. Second, the healthy condition exhibits systematically higher variability across all axes (σ=0.086–0.119) relative to faulty conditions, yet maintains the lowest mean PE in all axes. This cross-axis robustness validates MPE as a physically grounded feature for this dataset and strengthens confidence that the GroupKFold results generalise across sensor orientations.

7.6. Roller Element Load-Zone Dependence

The bimodal PE distribution for roller element faults provides quantitative evidence of load-zone-dependent impulsive generation. This intermittent fault signature explains the higher within-condition variance (σ=0.142) relative to inner and outer race faults, and motivates distributional features — such as bimodality index or mode separation — as complementary diagnostic indicators for this specific fault type. From a PHM perspective, the fraction of windows in the low-PE mode may serve as a progression indicator as the defect severity increases.

7.7. Implicit Encoding of Physical Scaling Structure by MPE

The MPE+Physical result — 33% physical feature importance with 0% specimen-level accuracy gain — reveals that multiscale ordinal analysis implicitly encodes variance scaling structure. Scale-1 MPE (I=0.1035) subsumes the diagnostic information in variance ratio (I=0.0679), and the MPE profile across scales captures the self-similarity structure quantified by the scaling exponent. This finding has practical implications for sensor-efficient PHM system design: MPE with 12 features provides equivalent diagnostic power to a 21-feature hybrid without additional computational cost.

7.8. Axis Hierarchy and Sensor Placement Guidance

The X-axis dominance (40.7% RF importance, 37.0% SHAP) provides actionable guidance for triaxial accelerometer placement in wind turbine bearing monitoring. When sensor count or data bandwidth must be reduced — common in retrofit monitoring for legacy turbines or wireless SCADA-connected systems — the axis aligned with the primary radial load direction should be prioritized. The consistency between RF importance and SHAP attribution across two independent importance metrics strengthens the physical interpretation.

7.9. Limitations

The dataset contains only four outer race specimens; performance estimates for this condition are inherently high-variance. The RF outer race specimen-level accuracy of 3/4 (75%) should be interpreted in this context, noting that the single misclassification results from a truncated 3.1-second recording (`OuterRace_4`, 3 windows only); XGBoost correctly classifies this specimen (Table 4), but the 4-specimen sample size means this difference does not establish a reliable algorithm ranking for this condition. The GroupKFold partition itself reflects this constraint: with only four outer race specimens, one fold (fold 2, Table 3) necessarily contains zero outer race specimens in its test partition, while the remaining four folds each contain exactly one. This uneven distribution is an unavoidable consequence of K=5 folds applied to four specimens and does not indicate a flaw in the partitioning strategy, but it underscores why outer race performance estimates carry wider uncertainty than the other three conditions.

The low-severity etching-induced defects may not represent the full spectrum of naturally developing bearing damage. The monotonic PE hierarchy (Section 6) is established for a single bearing model; generalisation to other bearing geometries or fault severities has not been tested.

Single-dataset validation precludes assessment of generalizability across turbines. Cross-dataset transfer of entropy-based features has been demonstrated for stationary test-rig conditions (Minhas & Singh, 2021); however, this targets a different validation axis than the specimen-level leakage prevention addressed here, and the two approaches are complementary. Future work should extend GroupKFold validation to multi-turbine datasets and investigate prognostic applications where PE feature evolution tracks fault progression.

8. CONCLUSIONS

This study demonstrates that multiscale permutation entropy, evaluated under GroupKFold cross-validation with strict specimen-level separation, achieves $97.67 \pm 2.10\%$ window-level and 97.73% specimen-level accuracy for wind turbine bearing fault diagnosis using the Fraunhofer LBF operational dataset.

Four principal findings emerge. First, specimen-level GroupKFold is necessary: window-level cross-validation without specimen separation constitutes data leakage for multi-window temporal segmentation, and the protocol established here provides a reproducible methodological baseline for the Fraunhofer LBF dataset that enables valid comparison of future diagnostic studies. Second, the X-axis accelerometer accounts for 40.7% of diagnostic importance, consistent with the primary radial load direction of the bearing and providing actionable sensor placement guidance for triaxial accelerometer deployments in operational wind turbines. Third, PE reveals a monotonic fault-type complexity hierarchy — Healthy (μ=0.702) $<$ Roller Element (μ=0.889) $<$ Inner Race (μ=0.981) $<$ Outer Race (μ=0.992) — with Cohen's d>0.9 for all pairwise comparisons; roller element bimodality reflects load-zone-dependent impulsive generation, providing physical interpretability grounded in bearing mechanics. Fourth, MPE implicitly encodes physical scaling structure: variance-based physical features contribute 33% feature importance within MPE+Physical yet add no specimen-level accuracy, confirming that multiscale ordinal analysis captures self-similarity and energy redistribution characteristics without explicit scaling estimators.

These results establish MPE with GroupKFold as a methodologically sound and computationally efficient baseline for bearing fault diagnosis in operational wind turbine data, with physical interpretability grounded in the ordinal structure of fault-induced vibration complexity.

ACKNOWLEDGMENT

The authors acknowledge the Fraunhofer Institute for Structural Durability and System Reliability LBF for providing the wind turbine bearing dataset. The authors also acknowledge the support provided by CAPES and CNPq (Brazil) through the Graduate Program in Mechanical Engineering (PPGEM) at UFPB.

DATA AND CODE AVAILABILITY

The Fraunhofer LBF wind turbine bearing dataset is publicly available via Zenodo (DOI: 10.5281/zenodo.11820598) (Mostafavi & Friedmann, 2024b). The feature extraction, GroupKFold validation, and statistical analysis pipeline are available from the corresponding author upon reasonable request.

NOMENCLATURE

PE	Permutation entropy
MPE	Multiscale permutation entropy
WPE	Weighted permutation entropy
RF	Random Forest
MDI	Mean decrease in impurity
SHAP	Shapley additive explanations
PHM	Prognostics and health management
m	Embedding dimension
s	Scale factor
τ	Time delay
W	Window size (samples)
S	Step size (samples)
K	Number of cross-validation folds
d	Cohen's effect size
μ	Mean permutation entropy
σ	Standard deviation
IR	Inner race fault
OR	Outer race fault
RE	Roller element fault
H	Healthy
brng.f.x/y/z	Front bearing accelerometer axes
BPFO	Ball pass frequency, outer race
BPFI	Ball pass frequency, inner race
BSF	Ball spin frequency

REFERENCES

Almeida, P. R., Lima, T. L., Brito, A. V., & Lima Filho, A. C. (2026). Multifractal characterization of bearing fault progression: Persistent-to-anti-persistent dynamics transition and cross-dataset validation of complexity-based features. *Chaos, Solitons & Fractals*, *206*, 117909. doi: 10.1016/j.chaos.2026.117909

Almeida, P. R. L., Lima, T. L. V., Brito, A. V., & Lima-Filho, A. C. (2025, December). Optimal feature complexity for small-sample bearing fault detection in manufacturing. *The International Journal of Advanced Manufacturing Technology*. doi: 10.1007/s00170-025-17296-8

Bandt, C., & Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. *Physical Review Letters*, *88*(17), 174102. doi: 10.1103/PhysRevLett.88.174102

Bhende, A. (2025). A robust xgboost approach for bearing fault prediction from vibration signals: Feature analysis and model evaluation. *Noise & Vibration Worldwide*. doi: 10.1177/09574565251390935

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates. doi: 10.4324/9780203771587

Costa, M., Goldberger, A. L., & Peng, C.-K. (2002). Multiscale entropy analysis of complex physiologic time series. *Physical Review Letters*, *89*(6), 068102. doi: 10.1103/PhysRevLett.89.068102

Dávalos, A., Jabloun, M., Ravier, P., & Buttelli, O. (2020). Improvement of statistical performance of ordinal multiscale entropy techniques using refined composite downsampling permutation entropy. *Entropy*. doi: 10.3390/e23010030

Fu, L., Zhu, T., Zhu, K., & Yang, Y. (2019). Condition monitoring for the roller bearings of wind turbines under variable working conditions based on the fisher score and permutation entropy. *Energies*, *12*(16), 3085. doi: 10.3390/EN12163085

Gao, Z., Wang, Y., Li, X., & Yao, J. (2024). Twins transformer: rolling bearing fault diagnosis based on cross-attention fusion of time and frequency domain features.. doi: 10.1088/1361-6501/ad53f1

Hameed, Z., Hong, Y. S., Cho, Y. M., Ahn, S. H., & Song, C. K. (2009). Condition monitoring and fault detection of wind turbines and related algorithms: A review. *Renewable and Sustainable Energy Reviews*, *13*(1), 1–39. doi: 10.1016/j.rser.2007.05.008

Huo, Z., Zhang, Y., Shu, L., & Gallimore, M. (2019). A new bearing fault diagnosis method based on fine-to-coarse multiscale permutation entropy, laplacian score and svm. *IEEE Access*. doi: 10.1109/ACCESS.2019.2893497

Khan, M., Asad, B., Naseer, M., et al. (2026). An improved hybrid lightweight approach for bearing fault detection and classification in three-phase squirrel cage induction motors. *Machines*. doi: 10.3390/machines14010068

Li, Y., & Ding, Q. (2026). Multivariate interpolation-multiscale chebyshev permutation entropy for bearing fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*. doi: 10.1109/TIM.2026.3657551

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*. doi: 10.48550/arXiv.1705.07874

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. doi: 10.1007/BF02295996

Minhas, A. S., & Singh, S. (2021). A new bearing fault diagnosis approach combining sensitive statistical features with improved multiscale permutation entropy method. *Knowledge-Based Systems*, *218*, 106883. doi: 10.1016/j.knosys.2021.106883

Mostafavi, A., & Friedmann, A. (2024a). Wind turbine condition monitoring dataset of fraunhofer lbf. *Scientific Data*, *11*, 1108. doi: 10.1038/s41597-024-03934-5

Mostafavi, A., & Friedmann, A. (2024b). *Wind turbine condition monitoring dataset of Fraunhofer LBF.* Zenodo. doi: 10.5281/zenodo.11820598

Nirmal, W., Lakmal, H., Arachchige, U., & Rodrigo, W. (2026). An interpretable machine learning framework for vibration-based fault detection using multi-domain features. In *International conference on soft computing and software engineering.* doi: 10.1109/SCSE70081.2026.11499945

Qi, J., Mauricio, A., & Gryllias, K. (2020). Comparison of blind diagnostic indicators for condition monitoring of wind turbine gearbox bearings. In *Proceedings of the asme turbo expo.* doi: 10.1115/GT2020-15278

Qiao, W., & Lu, D. (2015). A survey on wind turbine condition monitoring and fault diagnosis—part i: Components and subsystems. *IEEE Transactions on Industrial Electronics*, *62*(10), 6536–6545. doi: 10.1109/TIE.2015.2422112

Sandoval, D., Leturiondo, U., Pozo, F., & Vidal, Y. (2020). Low-speed bearing fault diagnosis based on permutation and spectral entropy measures. *Applied Sciences*, *10*(13), 4666. doi: 10.3390/app10134666

Vieira, J., Bauler, V., Rosa, R., & Silva, D. (2025). Towards a more realistic evaluation of machine learning models for bearing fault diagnosis. *arXiv preprint*. doi: 10.48550/arXiv.2509.22267

Wang, Z., Yao, L., Cai, Y., & Zhang, J. (2020). Mahalanobis semi-supervised mapping and beetle antennae search based support vector machine for wind turbine rolling bearings fault diagnosis. *Renewable Energy*, *155*, 1312–1327. doi: 10.1016/j.renene.2020.04.041

Wheat, L., Mohrenschildt, M., Habibi, S., & Al-Ani, D. (2024). Impact of data leakage in vibration signals used for bearing fault diagnosis. *IEEE Access*. doi: 10.1109/ACCESS.2024.3497716

Ye, M., Yan, X., & Jia, M. (2021). Rolling bearing fault diagnosis based on vmd-mpe and pso-svm. *Entropy*, *23*(6), 762. doi: 10.3390/e23060762

Zhang, H., Zhong, C., Zhang, Z., & Jiang, Y. (2023). A comparative study of rolling bearing fault diagnosis based on tree models. In *2nd international conference on digital society and intelligent systems.* doi: 10.1117/12.2673391

Zheng, J., Zheng, J., Du, Z.-l., et al. (2019). Composite multi-scale weighted permutation entropy and extreme learning machine based intelligent fault diagnosis for rolling bearing. *Measurement*. doi: 10.1016/J.MEASUREMENT.2019.05.002

Zhou, S., Qian, S., Chang, W., Xiao, Y., & Cheng, Y. (2018). A novel bearing multi-fault diagnosis approach based on weighted permutation entropy and an improved svm ensemble classifier. *Sensors*, *18*(6), 1934. doi: 10.3390/s18061934

Zhou, Y., Chen, J., Zhang, J., et al. (2025). Rolling bearing fault diagnosis based on emd-cnn-transformer. In *Youth academic annual conference of chinese association of automation.* doi: 10.1109/YAC66630.2025.11150136

BIOGRAPHIES

Paulo R. L. Almeida received his B.S. degree in Electrical Engineering from the Federal University of Paraíba (UFPB), Brazil, his M.S. degree in Electrical Engineering from the University of São Paulo (EESC/USP), Brazil, and his Ph.D. in Mechanical Engineering from UFPB in 2026. Since 2005, he has been a Professor at the Federal Institute of Education, Science and Technology of Paraíba (IFPB). His research interests include vibration signal analysis, entropy complexity measures, machine learning for fault diagnosis, signal processing, and prognostics and health management systems in the fields of condition monitoring and intelligent systems.

Thyago L. V. Lima received his B.S. degree in Industrial Automation Technology from IFPB, his M.S. degree in Mechanical Engineering from UFPB in 2012, and his Ph.D. in Mechanical Engineering from the same university in 2019. He is currently a Professor at IFPB. His research focuses on signal processing techniques for bearing fault detection and classification.

Alisson V. Brito received his Ph.D. in Electrical Engineering from the Federal University of Campina Grande (UFCG), Brazil, in cooperation with the Karlsruhe Institute of Technology (KIT). He is currently a Professor at UFPB and Coordinator of the Laboratory of Embedded Systems and Robotics (LaSER). His research interests include embedded systems, Internet of Things, and industry-academic collaborations.

Abel C. Lima-Filho received his Ph.D. in Mechanical Engineering from UFPB in 2009. He is currently an Associate Professor at the Department of Mechanical Engineering, UFPB, and coordinates the Graduate Program in Mechanical Engineering. His research focuses on mechatronics, industrial instrumentation, predictive maintenance, and condition monitoring of rotating machinery.