

Article

Shapelet-Based Bearing Fault Diagnosis Under Interpretability Constraints: A Recording-Level Evaluation

Lino González-García * , Luis Usero , Miguel-Angel Sicilia  and Elena García-Barriocanal 

Computer Science Department, University of Alcalá, 28805 Alcalá de Henares, Spain; luis.usero@uah.es (L.U.); msicilia@uah.es (M.-A.S.); elena.garciab@uah.es (E.G.-B.)

* Correspondence: lino.gonzalez@uah.es

Abstract

Shapelet-based classifiers offer structural interpretability: discriminative subsequences form an inspectable vibration-pattern vocabulary and a shallow decision-tree ensemble produces traceable fault-type predictions. We impose explicit interpretability constraints on the model and apply Bayesian optimisation within this bounded region; the primary experimental question is whether these constraints carry a performance penalty relative to an unconstrained baseline. Evaluation uses recording-level cross-validation on CWRU (Case Western Reserve University) and MFPT (Machinery Failure Prevention Technology) bearing datasets with Hilbert envelope demodulation. The central finding is that the constraints impose no systematic performance penalty: the shapelet classifier matches ROCKET, a non-interpretable baseline, on both datasets, with cross-validated mean F1 differences smaller than the fold-to-fold standard deviation. To further characterise the selected models under the controlled laboratory conditions studied here, we assess probability calibration and conformal prediction coverage as secondary analyses. Raw probability estimates are well-calibrated, but Platt scaling degrades under cross-severity distribution shift; split conformal prediction yields valid coverage on CWRU but fails on MFPT due to class-proportion mismatch across recording-level splits. Together, these results show that structural constraints supporting interpretability are compatible with competitive performance, and identify the conditions under which reliability tools succeed and fail in this setting.

Keywords: shapelets; bearing fault diagnosis; interpretability; conformal prediction; calibration; Bayesian optimisation; model selection; rotating machinery



Academic Editors: Xian-Bo Wang and Yunfeng Yan

Received: 12 June 2026

Revised: 6 July 2026

Accepted: 8 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Rolling-element bearings are among the most prevalent components in rotating industrial machinery; their fault diagnosis and health monitoring form a core task in condition-based maintenance, aimed at reducing unplanned downtime and improving operational reliability [1]. As a defective rolling element contacts a damaged race surface, it generates impulsive vibration events at characteristic fault frequencies; these amplitude-modulated signatures are captured by accelerometers and provide the basis for automated bearing fault classification [2]. The dominant approach today is deep learning: convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and Transformers achieve high classification accuracies on standard benchmarks such as the Case Western Reserve University (CWRU) dataset [2].

A key challenge for real-world deployment of these models is their opacity: a maintenance engineer who receives an alert cannot determine *which signal pattern* triggered it, cannot verify whether the model's reasoning matches established diagnostic knowledge, and cannot assess how confident to be in a specific prediction. Inherently interpretable models [3], whose internal structure can be inspected directly without post hoc explanation tools, are therefore attractive for safety-critical industrial applications.

Shapelets—discriminative subsequences of a time series—provide a natural interpretable feature vocabulary for vibration signals, where fault types correspond to characteristic waveform morphologies at bearing fault frequencies [4]. A downstream shallow decision-tree ensemble maps distances to those subsequences into fault-type predictions; the resulting classifier is structurally transparent, with each prediction traceable to a specific signal pattern and a short sequence of threshold comparisons.

Concretely, a shapelet-based rotation-forest classifier exposes four tunable parameters: the number of shapelet candidates evaluated during discovery, the number of retained shapelets (which bounds the feature space a maintenance engineer must reason about), the number of trees in the ensemble, and the maximum tree depth (which limits decision-path length). A critical design question is whether structural constraints that support interpretability—limiting complexity in ways that make the model amenable to inspection—carry a performance cost, or whether maximising performance inevitably drives the search toward deep, high-dimensional configurations that are structurally opaque. We operationalise these constraints as two structural proxies: the number of retained shapelets is capped at 15 (a vocabulary small enough for direct inspection) and tree depth at 8 (at most 256 leaf nodes per tree, visualisable in full). These bounds are necessary but not sufficient conditions for interpretability: a model satisfying them is amenable to inspection, but whether its shapelets correspond to physically meaningful fault patterns requires domain-expert validation beyond the scope of this study. Within this constrained subspace, Bayesian optimisation finds the best-performing configuration, so the performance results reported are achievable under structural conditions that support interpretability. Two further questions are then studied: (i) whether the resulting probability estimates are sufficiently well-calibrated to be acted upon; and (ii) whether the compact shapelet model remains competitive with a black-box baseline that carries no interpretability constraint. Neither question has been answered systematically for this class of model in the bearing fault diagnosis domain.

Main contribution. The main contribution of this paper is a systematic empirical study of the performance cost of interpretability constraints on a shapelet-based bearing fault classifier under recording-level cross-validation. We show that constraining the model to at most 15 retained shapelets and tree depth to at most 8 imposes no systematic F1 penalty relative to ROCKET, a non-interpretable baseline with 10,000 random convolutional kernels, across both CWRU and MFPT benchmark datasets. Bayesian optimisation with a fixed budget of $N = 200$ trials is used to find the best-performing configuration within this constrained space. To characterise the selected models beyond accuracy, we additionally assess probability calibration via Expected Calibration Error (ECE) [5] and evaluate split conformal prediction [6], reporting empirical coverage and prediction-set size at operationally meaningful confidence levels; in doing so, we identify a structural incompatibility between recording-level evaluation and the exchangeability assumption of split conformal prediction and diagnose its cause as class-proportion mismatch across recordings.

The remainder of this paper is structured as follows. Section 2 surveys related work on shapelet-based interpretable classifiers. Section 3 describes the data, models, and experimental design. Results and discussion are provided in Section 4, and Section 5 provides conclusions and outlook.

2. Background

2.1. Shapelets and Interpretability

Time-series shapelets are discriminative subsequences representative of a class [4]. Their similarity to a query series can be used for class identification via a distance transform that converts each series into a vector of minimum distances to the retained shapelets. Hills et al. [7] proposed the Shapelet Transform Classifier (STC), which separates shapelet discovery from classification, enabling the use of arbitrary downstream learners. Bostrom and Bagnall [8] extended STC to multiclass problems. In the fault-diagnosis domain, shapelet interpretability has a concrete operational meaning: each retained shapelet is a short vibration waveform segment, and the distance of an observed signal to that segment quantifies how closely the current vibration pattern resembles a known fault signature. A maintenance engineer can inspect the shapelet waveform, relate it to known fault-frequency content, and understand which pattern drove a specific prediction.

Shapelets have been applied to rotating machinery fault diagnosis in prior work [9], differing from the Shapelet Transform Classifier pipeline used here in feature extraction strategy, downstream classifier, and evaluation protocol. To our knowledge, no prior work has applied STC with Hilbert envelope demodulation as preprocessing—which converts amplitude-modulated impulsive fault signatures into geometrically stable envelope waveforms more directly amenable to shapelet matching—nor has any prior study evaluated calibration, conformal prediction, or recording-level cross-validation for shapelet-based bearing fault classifiers on standard benchmarks such as CWRU and MFPT.

2.2. Reliability of Bearing Fault Classifiers

The bearing fault diagnosis literature is dominated by classification accuracy as the primary evaluation metric [2]. Reviews consistently report high accuracies on standard benchmarks, yet these results are routinely obtained with window-level random splits that allow overlapping windows from the same recording to appear in both train and test sets, artificially inflating reported performance and concealing generalisation difficulty. Beyond accuracy, the reliability of a classifier's probability estimates has received little systematic attention in the fault diagnosis domain, despite its direct operational relevance: a model that is confidently wrong about fault type can cause missed inspections or unnecessary shutdowns. Guo et al. [5] showed that high classification accuracy does not imply well-calibrated probability estimates, a finding with direct implications for deep learning-based fault detectors. Calibration of probabilistic predictors has been explored in predictive maintenance scenarios with tabular sensor features [10], but not for vibration-based bearing fault classification. Conformal prediction [6,11] offers a principled complement to calibration by providing prediction sets whose coverage is guaranteed by construction rather than inferred from model probabilities; class-conditional extensions [12] can enforce per-fault-class coverage guarantees. Its application to safety-critical transport systems has been demonstrated for railway signal detection [13], though conformal coverage guarantees have not been studied for bearing fault classification. To our knowledge, neither calibration analysis nor conformal prediction has been applied to vibration-based bearing fault classification, leaving a gap between the principled uncertainty-quantification literature and its application to vibration-based fault diagnosis.

2.3. Bayesian Optimisation for Model Selection

Hyperparameter optimisation for time-series classifiers involves expensive black-box evaluations, as each trial requires fitting the full model on a potentially large window dataset. Bayesian optimisation (BO) [14] addresses this by maintaining a probabilistic surrogate over the objective and selecting new trials via an acquisition function such

as Expected Improvement [15], achieving good configurations in far fewer evaluations than grid or random search. The sample-efficiency advantage is particularly pronounced for the shapelet family, where a single trial involves an exhaustive or stochastic search over candidate subsequences before the downstream classifier is even fitted. Modern BO platforms such as BoTorch [16] provide scalable Gaussian Process surrogates and Monte-Carlo acquisition functions that handle integer and mixed parameter spaces efficiently. Despite the computational cost of shapelet training, BO-based hyperparameter search for the STC pipeline has not been studied in the bearing fault diagnosis literature.

Recent work on interpretable and knowledge-driven fault diagnosis has explored complementary directions. Physics-informed domain adaptation methods embed bearing fault frequencies or physical constraints directly into deep network architectures [17], while knowledge-driven machine learning frameworks synthesise domain expertise with data-driven models to improve both accuracy and transparency in PHM applications [18]. Yin et al. [19] propose a fault evolution knowledge-driven adversarial meta-learning (FEK-AML) framework that embeds physical degradation trajectories into a two-stage meta-learning architecture for few-shot tool-state recognition under variable operating conditions. Li et al. [20] address cross-domain diagnosis under time-varying speeds using an interpretable dynamic weighted adaptation network, demonstrating that distribution shift across operating conditions is a central challenge for deployable fault classifiers. Three gaps remain unaddressed by these works: (i) the performance cost of *explicit structural constraints* (bounded feature vocabulary, limited decision depth) relative to unconstrained black-box baselines has not been quantified; (ii) probability calibration and conformal prediction have not been evaluated for interpretability-constrained classifiers under recording-level splits; and (iii) no prior work has characterised the interaction between recording-level evaluation and the exchangeability assumption of conformal prediction in this domain. The present paper addresses all three gaps.

3. Materials and Methods

3.1. Datasets

To prevent information leakage across temporal and operational boundaries, both datasets are partitioned at the *recording* level rather than by random window sampling, and performance is reported via three-fold recording-level cross-validation to avoid conclusions dependent on a single split.

CWRU. The Case Western Reserve University Bearing Dataset [21,22] comprises accelerometer recordings from a motor drive-end bearing. We use the 12 kHz drive-end channel at 0 HP load, covering four classes: Normal, Inner Race Fault (IRF), Ball Fault (BF), and Outer Race Fault (ORF), each at three severity levels (0.007, 0.014, and 0.021 inch diameter). The three severity levels define three rotating-severity folds; the severity assignment for each fold is shown in Table 1. Fold 2 is the most challenging, as it requires generalisation from mild to severe faults.

Table 1. CWRU three-fold severity assignment. Each fold leaves one severity level out for testing and rotates the remaining two between training and validation.

Fold	Train	Validation	Test
0	0.021 in	0.014 in	0.007 in
1	0.007 in	0.021 in	0.014 in
2	0.007 in	0.014 in	0.021 in

The single Normal recording is divided into three equal temporal thirds and assigned to folds in the same rotation; in the main analysis single split (Table 2) it is instead divided

70/10/20 % to preserve more training data, so window counts differ between the two experiments. This temporal partition is a pragmatic necessity—only one Normal recording exists—and represents a weaker form of recording-level independence than the severity-based fault split: consecutive temporal segments share operating conditions and may exhibit correlated signal characteristics not present across different severity levels. The potential influence on Normal-class classification performance is discussed in Section 4.2.

MFPT. The Machinery Failure Prevention Technology dataset [23] provides recordings at 97,656 Hz under variable load conditions, covering three classes: Normal, Inner Race Fault, and Outer Race Fault. Files within each class are sorted and partitioned into three balanced blocks; fold k uses block k for testing, block $(k + 1) \bmod 3$ for validation, and block $(k + 2) \bmod 3$ for training. Training windows are class-balanced by undersampling; validation and test windows are left unmodified to preserve natural class proportions. Because the number of recordings per class is not uniform across blocks, the resulting class proportions in the validation and test sets differ—a structural feature with implications for the conformal prediction analysis in Section 4.4.

Preprocessing. Raw accelerometer signals are processed through a four-step pipeline before being presented to any classifier.

1. *Recording-level z-score normalisation.* Each recording is normalised to zero mean and unit standard deviation. This removes amplitude differences between fault severity levels so the classifier sees waveform shape rather than signal scale.
2. *Sliding window segmentation.* The normalised signal is divided into fixed-length windows of 1024 samples with a stride of 512 samples (approximately 85 ms per window at 12 kHz; 10.5 ms at 97,656 Hz), yielding 50 % overlapping windows.
3. *Hilbert envelope demodulation.* The analytic signal is computed via the Hilbert transform and its instantaneous amplitude (the envelope) is extracted. This step removes the high-frequency carrier oscillation and retains the low-frequency amplitude modulation caused by recurring fault impacts, converting amplitude-modulated impulsive signals into geometrically stable waveforms amenable to shapelet matching.
4. *Per-window z-score normalisation.* Each envelope window is independently normalised to unit variance, removing residual amplitude differences between windows within the same recording.

The output is a set of 1024-dimensional vectors, one per window, representing the shape of the fault-frequency energy pattern. These vectors are the sole input to both the shapelet classifier and ROCKET; no hand-crafted spectral features are computed. Window counts for the main analysis split are reported in Table 2.

Table 2. Window counts per split after preprocessing, for the main analysis single split (CWRU: train = 0.007 in, val = 0.014 in, test = 0.021 in; MFPT: original single split). Both datasets use window = 1024, stride = 512. The CWRU Normal recording is divided into 70/10/20 % (main analysis only; the k-fold uses equal thirds, yielding different per-fold counts). MFPT training counts reflect undersampling to balance the three classes.

Dataset	Split	Normal	Inner Race	Ball	Outer Race	Total
CWRU	Train	332	235	238	237	1042
	Val	46	236	236	236	754
	Test	94	237	237	238	806
MFPT	Train	1143	1143	—	1143	3429
	Val	1143	285	—	570	1998
	Test	1143	285	—	285	1713

3.2. Interpretability-Constrained Model Selection

Each candidate model is a `ShapeletTransformClassifier` [7,8] that applies a learned shapelet transformation followed by a `RotationForest` classifier [24] whose base learner is a CART decision tree with a configurable maximum depth. The classifier operates in two stages. In the first stage, shapelet discovery, a pool of candidate subsequences of varying length is randomly sampled from the training windows. Each candidate is z-normalised and evaluated by the F-statistic between its minimum distance to each class, retaining those that best separate the fault classes. The retained shapelets define a distance transform: for each input window and each shapelet, the minimum z-normalised Euclidean distance between the shapelet and any position in the window is computed. A window of 1024 samples is thereby replaced by a compact vector of distances, one entry per retained shapelet. In the second stage, a `RotationForest` classifier is trained on these distance vectors. Rotation Forest builds an ensemble of decision trees, each trained on a randomly rotated subset of the features (rotation via PCA), which increases tree diversity while preserving the interpretable threshold-comparison structure of individual trees. Both stages take only the shapelet distance vectors as input; no raw signal enters the classifier directly. The four parameters that jointly determine model performance and interpretability are listed in Table 3.

Table 3. Optimisation parameter space. All parameters are integer-valued; the search space comprises every integer in [Min, Max] with step 1.

Parameter	Description	Min	Max
<code>n_shapelet_samples</code>	Candidate shapelets evaluated during discovery	100	500
<code>max_shapelets</code>	Shapelets retained as features	5	15
<code>n_estimators</code>	Trees in the rotation forest	2	50
<code>max_depth</code>	Maximum depth of each decision tree	3	8

`max_shapelets` is the primary interpretability driver on the feature-extraction side, fixing the number of vibration patterns the engineer must reason about. `max_depth` limits decision-path length to at most 2^d terminal nodes per tree. The upper bounds on these two parameters (15 and 8) are interpretability constraints chosen *a priori*: at most 15 shapelets constitute a vocabulary a maintenance engineer can inspect directly, and a tree with a depth of 8 has at most 256 leaf nodes. These constraints are structural proxies for interpretability: they are necessary but not sufficient conditions for a model to be genuinely interpretable, but they ensure the performance results reported are achievable within a structurally compact, inspectable model class.

Model selection is framed as black-box optimisation over this four-dimensional integer space, with macro-averaged F1 on the held-out validation set as the objective. We first evaluate a reference configuration (`n_shapelet_samples` = 200, `max_shapelets` = 15, `n_estimators` = 5, `max_depth` = 6) as trial 0 to anchor the surrogate from the start. Bayesian optimisation (BO) [14] navigates expensive black-box objective functions by maintaining a Gaussian Process surrogate over the objective and directing new evaluations towards promising regions via an acquisition function; the Expected Improvement (EI) criterion [15] balances exploration and exploitation. BO is particularly suited to shapelet model selection because each function evaluation requires training a full shapelet classifier, which is computationally costly at the window counts encountered here. BO is launched using BoTorch [16] via the Ax platform, preceded by 8 Sobol quasi-random draws [25] for uniform parameter-space coverage before the GP surrogate is warm enough to direct the search (Ax v0.3.7 default for a four-parameter experiment). The search runs for exactly $N = 200$ trials with no early stopping.

All experiments were run on an Apple M4 Max workstation (16-core CPU, 128 GB unified memory) under Python 3.13 with aeon 0.12 and BoTorch 0.17. Per-trial wall-clock time scales approximately linearly with `n_shapelet_samples`: on CWRU fold 0 ($n_{\text{train}} = 869$, $n_{\text{val}} = 865$) we measured a mean of 23.4 s per trial (range 9.8–41.9 s across trials with `n_shapelet_samples` $\in [100, 500]$), giving a projected BO runtime of approximately 78 min per fold on this dataset. MFPT folds are substantially larger (2565–5427 training windows versus 869 for CWRU), so per-trial cost scales accordingly; across the six folds (3 CWRU + 3 MFPT), total BO compute was approximately 13 hours. To assess BO seed stability, we re-ran 200-trial optimisation on CWRU fold 0 with five random seeds (0, 1, 2, 3, 42); results are reported in Table 4. Val F1 varies by only 0.010 across seeds (0.6500–0.6599, $\text{std} = 0.005$), and `n_estimators` = 2 is selected by every seed. The remaining hyperparameters (`n_shapelet_samples` in [408–463], `max_shapelets` in [8–10], `max_depth` in [6–8]) vary modestly but consistently favour the upper portion of the search space, confirming that the GP-EI surrogate converges reliably within 200 trials on this four-dimensional problem.

Table 4. BO seed stability on CWRU fold 0 (200 trials per seed). Val F1 varies by 0.010 across five seeds; `n_estimators` = 2 is selected unanimously.

Seed	Val F1	n_samples	max_shp	n_est	depth
0	0.6599	415	10	2	8
1	0.6500	463	10	2	8
2	0.6599	456	10	2	8
3	0.6599	408	8	2	6
42	0.6500	419	9	2	8
Mean $\pm$ std	0.656 $\pm$ 0.005				

3.3. Reliability Evaluation and Non-Interpretable Baseline

Calibration. Beyond classification accuracy, a model deployed in a maintenance workflow must produce probability estimates that can be acted upon: an engineer who receives a prediction together with a confidence of 90% needs that figure to be meaningful, not an artefact of the classifier’s internal scaling. This is particularly relevant for the interpretability-constrained model, whose compact structure raises an open question: does restricting the feature space to a small shapelet vocabulary and shallow trees produce well-calibrated probability estimates, or does the constraint distort the output distribution? A classifier is well-calibrated if its predicted confidence reflects the true posterior probability that its prediction is correct [5]. For fault diagnosis, miscalibration has direct operational consequences: an overconfident wrong prediction may lead to a missed inspection or an unnecessary shutdown. Expected Calibration Error (ECE) [26] quantifies miscalibration by partitioning predictions into M equal-width confidence bins and computing the weighted average gap between mean confidence and mean accuracy within each bin; we use $M = 10$ with top-label binning (each prediction is binned by its maximum predicted class probability). Dankowski and Ziegler [27] showed that tree ensembles tend to produce miscalibrated probability estimates that can be corrected post hoc. To assess whether post hoc correction helps or hurts under the recording-level distributional shift present in both datasets, we apply Platt scaling (sigmoid calibration) [28] fitted on the validation set and re-evaluate ECE on the test set. The same procedure is applied to ROCKET, allowing a direct comparison of the reliability profiles of the two models.

Conformal prediction. Even a well-calibrated point classifier cannot communicate the case where evidence is genuinely split between two fault types. Conformal prediction [11] addresses this by wrapping any point predictor to produce *prediction sets* with a finite-sample marginal coverage guarantee: for a user-chosen confidence level $1 - \alpha$,

the prediction set contains the true class with probability at least $1 - \alpha$ regardless of the underlying data distribution (under exchangeability). For the interpretability-constrained model, prediction sets have a concrete operational value that goes beyond the guarantee itself: when the set contains more than one fault type, an engineer can inspect the corresponding shapelets and understand why the model cannot discriminate—information that a black-box uncertainty score cannot provide. We apply the split conformal variant [6], which uses a held-out calibration set to compute nonconformity scores and thresholds them at the $(1 - \alpha)(1 + 1/n_{\text{cal}})$ quantile, to the best-found shapelet configuration. The non-conformity score for class k is $1 - \hat{p}_k(x)$, where $\hat{p}_k(x)$ is the model's predicted probability for class k . We report empirical coverage and mean prediction-set size at confidence levels $1 - \alpha \in \{0.80, 0.90, 0.95, 0.99\}$. Per-fault-class coverage at 95% is reported separately to surface any systematic gaps, since recording-level splits may produce class-proportion imbalances between the calibration and test sets that violate the exchangeability assumption.

Non-interpretable baseline. As a non-interpretable performance ceiling, ROCKET [29] applies 10,000 random convolutional kernels to the same envelope-transformed windows as the shapelet classifier and fits a RidgeClassifierCV on the resulting features. Each kernel has a randomly chosen length, dilation, and bias; it is convolved with the input window to produce a feature map from which two scalar summary statistics are extracted: the maximum value and the proportion of positive values. The 10,000 kernels therefore produce 20,000 features per window, which are passed directly to a ridge regression classifier (RidgeClassifierCV) with no further feature engineering or hyperparameter tuning. ROCKET is thus a black-box method: the random kernels have no interpretable meaning and individual predictions cannot be traced back to recognisable signal patterns. For the F1 comparison, ROCKET is trained on the combined training and validation set. For the reliability comparison, a separate ROCKET model is trained on the training set only so that Platt scaling can be fitted on the held-out validation set, yielding calibrated probability estimates on the test set; this calibration model therefore uses strictly less training data than the F1-comparison model, a necessary asymmetry given that Platt scaling requires a held-out validation split.

ROCKET is used as the sole non-interpretable baseline because it provides a well-established and reproducible performance ceiling for time-series classification. On the UCR and UEA benchmark archives, ROCKET matches or exceeds the accuracy of 1D-CNN, LSTM, and InceptionTime architectures on the majority of datasets while requiring no architecture search or GPU resources [29]. Deep learning baselines would moreover require substantially larger window counts to generalise reliably: under the recording-level evaluation protocol used here, each fold has at most a few hundred training windows, a regime where random-kernel methods are known to be more data-efficient than gradient-trained networks. MiniROCKET is a faster variant of the same family; its accuracy on most datasets is within one percentage point of full ROCKET, so its inclusion would not alter the interpretable/black-box comparison. The comparison is therefore intentionally narrow: the shapelet classifier is evaluated against the strongest available general-purpose time-series classifier that does not rely on architecture tuning or domain-specific feature engineering, rather than against a broad suite of baselines whose relative ordering under recording-level splits has not been established.

4. Results and Discussion

The primary question—whether interpretability constraints impose a performance cost relative to an unconstrained black-box baseline—is addressed in Sections 4.1 and 4.2. Sections 4.3 and 4.4 then characterise the reliability of the selected model's outputs as secondary analyses under the same controlled laboratory conditions.

4.1. Bayesian Optimisation Results and Selected Configurations

Table 5 summarises the best configurations found by Bayesian optimisation for each dataset, together with the corresponding baseline and test macro-F1 scores.

Table 5. Bayesian optimisation summary. **F1_{base}**: validation macro-F1 of the baseline configuration ($\{200, 15, 5, 6\}$) trained on the training set only; **Trials**: total BO trials evaluated (fixed budget $N = 200$, including the baseline as trial 0); **F1_{best}**: best validation macro-F1 found by BO; **Best params**: Ax-selected configuration ($n_{\text{samp}}/\text{max_sh}/n_{\text{est}}/\text{max_d}$); **F1_{test}**: test macro-F1 of the best configuration retrained on train+val; **Δ (%)**: relative improvement in validation F1 computed from unrounded values (apparent discrepancies with the 3 dp displayed values are due to rounding).

Dataset	F1 _{base}	Trials	F1 _{best}	Best Params	F1 _{test}	Δ (%)
CWRU	0.375	200	0.655	418/8/2/5	0.641	+74.5
MFPT	0.820	200	0.865	457/12/33/4	0.894	+5.6

On CWRU, the expert-chosen baseline achieves a validation macro-F1 of only 0.375, reflecting the genuine difficulty of cross-severity generalisation (training on 0.007 inch faults, testing on 0.021 inch faults). The best validation F1 reached across all 200 trials is 0.655, well below the near-perfect scores routinely reported with window-level random splits in the literature [2], confirming that recording-level evaluation exposes genuine cross-severity generalisation difficulty rather than interpolation within the same recording. After 200 trials the GP surrogate selects a compact configuration (eight retained shapelets, two trees, depth 5), improving validation F1 by +74.5% to 0.655 (Figure 1, left). With at most 8 shapelets, each tree has at most $2^5 = 32$ terminal nodes, making the ensemble inspectable by design; the small number of trees contributes rotation-forest diversity while preserving transparency. Full cross-validation results are reported in Section 4.2.

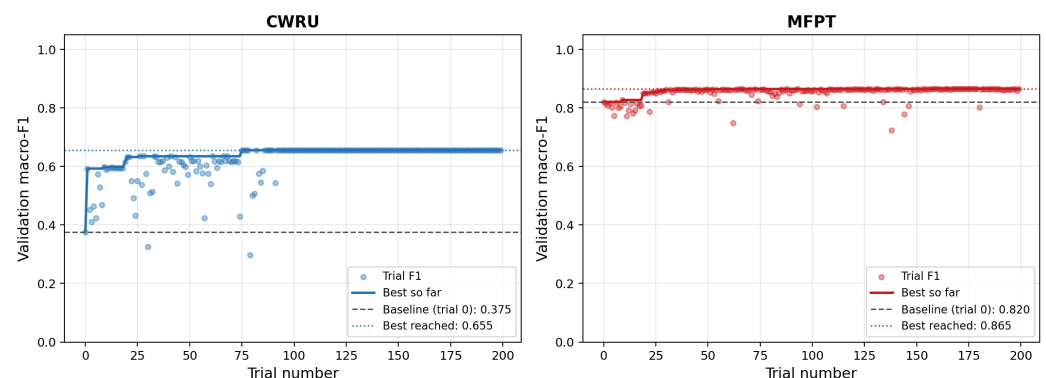


Figure 1. Bayesian optimisation convergence over 200 trials for CWRU (left) and MFPT (right). Dots show per-trial validation macro-F1; the solid line tracks the running best. Trial 0 (starred) is the expert baseline. On CWRU, performance improves steeply in the first 25 trials before plateauing; on MFPT the baseline already scores high and gains are incremental throughout.

On MFPT, the baseline starts considerably higher ($F1 = 0.820$), consistent with this particular single split proving favourable—as confirmed by the cross-validation results in Section 4.2, which show substantial block-to-block variance. BO selects a larger configuration—12 retained shapelets, 33 trees, depth 4—from a wider candidate vocabulary (457 shapelet samples evaluated), consistent with the heterogeneous per-class signal-to-noise characteristics of the MFPT recordings. The best validation F1 (0.865) is reached near trial 160 (Figure 1, right), indicating that the GP surrogate continues to refine the model well beyond the initial Sobol phase, and retraining on the combined train-plus-validation set yields a test F1 of 0.894 (+5.6% over baseline). Inspection of the full trial

history (Figure 2) reveals that the F1 surface is flat in the neighbourhood of the optimum: 144 of the 200 evaluated configurations achieve validation F1 within 1% relative of the best (i.e., ≥ 0.857), including configurations with as few as five trees (trial 149, $F1 = 0.861$) or eight trees (trials 104 and 108, $F1 = 0.863$). A practitioner who prioritises transparency over the last fraction of a percentage point can therefore adopt a substantially more compact ensemble (5–8 trees rather than 33) at a negligible performance cost.

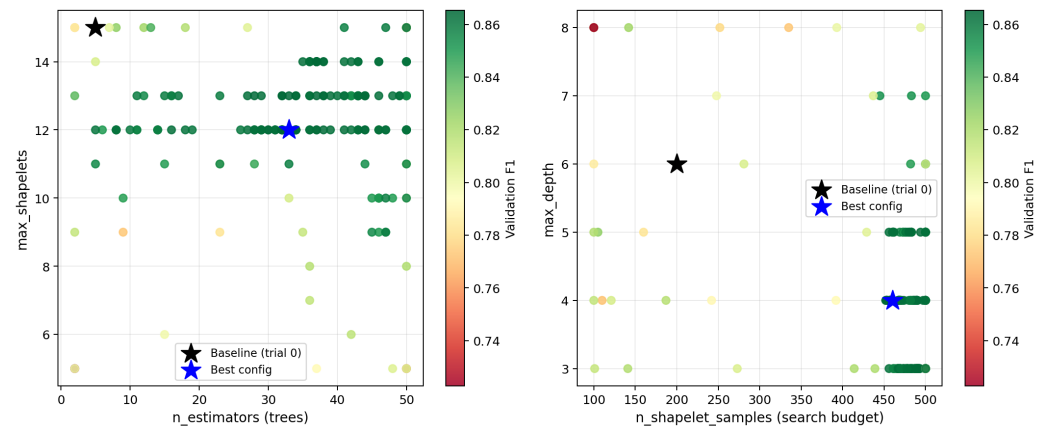


Figure 2. Parameter-space exploration for MFPT (200 trials), coloured by validation macro-F1 (green = high, red = low). The black star marks the expert baseline (trial 0); the blue star marks the best configuration found. The broad green region across a wide range of $n_estimators$ and $max_shapelets$ values confirms the flat F1 surface near the optimum: many configurations achieve near-identical performance, including substantially more compact ensembles.

Physical interpretability of the retained shapelets. The structural constraints (at most eight shapelets, $depth \leq 8$) ensure model compactness, but do the retained shapelets correspond to physically meaningful vibration patterns? To address this question, we extracted the four shapelets retained by the best CWRU configuration (eight shapelets requested; four selected after self-similar removal), computed the FFT magnitude spectrum of each, and compared dominant frequencies with the known bearing fault characteristic frequencies for the CWRU 6205-2RS deep groove ball bearing at 1797 RPM ($BPFI = 158.4$ Hz, $BPFO = 111.2$ Hz, $BSF = 82.8$ Hz, $FTF = 12.4$ Hz).

Figure 3 shows the results. Shapelet 1 (Inner Race, length = 181 samples) has a dominant FFT peak at 175.8 Hz, an 11% deviation from BPFI (158.4 Hz); Shapelet 2 (Inner Race, length = 549 samples) has a dominant peak at 164.1 Hz, a 3.6% deviation from BPFI. Both Inner Race shapelets therefore correspond to impulsive patterns whose repetition rate is consistent with the ball pass frequency on the inner race, supporting the claim that these structural patterns carry physical meaning beyond arbitrary waveform segments.

The two Normal shapelets (Shapelets 3–4, lengths 121 and 134 samples) have dominant spectral content at 900–930 Hz, well above any of the fault characteristic frequencies. These patterns capture the high-frequency vibration texture of a healthy bearing under load, consistent with the absence of periodic impulsive bursts in a normal operating condition.

Ball and Outer Race fault classes did not obtain dedicated shapelets in this run; their classification relies on distance from the retained Normal and Inner Race patterns. This is a structural limitation of the compactness constraint: with only four retained patterns across four classes, some fault types are implicitly represented through their dissimilarity from explicitly modelled patterns rather than through their own characteristic waveforms. Domain-expert validation—e.g., comparing shapelet spectra against acoustic emission signatures for each fault mode—is therefore recommended before operational deployment.



Figure 3. Retained shapelets from the best CWRU configuration (train+val, $n_{\text{samp}} = 418$, max_shapelets = 8; four selected after self-similar removal). **Left:** shapelet waveform. **Centre:** FFT magnitude (dashed lines mark fault characteristic frequencies: BPFI = 158 Hz red, BPFO = 111 Hz green, BSF = 83 Hz orange, FTF = 12 Hz purple; solid black vertical line is the spectral peak). **Right:** normalised autocorrelation (dotted lines at expected inter-impact periods for each fault frequency). Shapelets 1–2 (Inner Race) have dominant spectral energy near BPFI; Shapelets 3–4 (Normal) capture high-frequency background texture at 900–930 Hz, inconsistent with any fault characteristic frequency.

4.2. Comparison with Non-Interpretable Baseline

Table 6 compares the best shapelet BO configuration against the ROCKET non-interpretable baseline on F1; calibration results are reported in Section 4.3.

On CWRU, the shapelet classifier achieves test macro-F1 of 0.641 versus ROCKET's 0.438, a gap of +0.203 in favour of the structurally constrained model. A compact model satisfying the interpretability constraints (eight retained shapelets, two trees, depth 5) thus outperforms a black-box baseline (10,000 random kernels, ridge regression) under a challenging cross-severity transfer setting. On MFPT, ROCKET leads (0.989 vs. 0.894, gap −0.095 F1), reflecting the advantage of a dense feature space in the matched-condition split used here.

Table 6. Comparison of best shapelet BO configuration against ROCKET baseline on the single fixed test split (CWRU fold 2; original MFPT split). **Gap F1:** Shapelet F1 – ROCKET F1 (positive = shapelet higher); **Gap ECE:** ROCKET Platt ECE – Shapelet Platt ECE (positive = shapelet better calibrated).

Dataset	Shapelet F1	ROCKET F1	Gap F1	Shapelet ECE	ROCKET ECE	Gap ECE
CWRU	0.641	0.438	+0.203	0.269	0.281	+0.012
MFPT	0.894	0.989	−0.095	0.046	0.057	+0.011

To assess whether these single-split results are representative, Table 7 reports per-fold test macro-F1 under three-fold recording-level cross-validation with independent Bayesian optimisation per fold.

Table 7. Three-fold recording-level cross-validation: test macro-F1 per fold. CWRU: leave-one-severity-out (fold k tests on severity k). MFPT: leave-one-block-out (fold k tests on block k). Independent Bayesian optimisation is performed per fold with a fixed budget of $N = 200$ trials. Bold values indicate the fold in which the shapelet classifier outperforms ROCKET.

Dataset	Fold	Shapelet F1	ROCKET F1
CWRU	0 (test = 0.007 in)	0.763	0.722
	1 (test = 0.014 in)	0.470	0.513
	2 (test = 0.021 in)	0.481	0.440
	Mean $\pm$ std	0.571 $\pm$ 0.136	0.558 $\pm$ 0.119
MFPT	0 (test = block 0)	0.485	0.510
	1 (test = block 1)	0.984	0.998
	2 (test = block 2)	0.802	0.867
	Mean $\pm$ std	0.757 $\pm$ 0.206	0.792 $\pm$ 0.206

On CWRU, the single-split advantage holds: shapelet leads ROCKET on two of three folds (fold 0: 0.763 vs. 0.722; fold 2: 0.481 vs. 0.440) and on average (0.571 vs. 0.558). The fold 0/fold 2 asymmetry is physically meaningful—severe-fault training transfers better to mild-fault testing than the reverse, and shapelet features capture this transferable structure—while fold 1 is the sole fold favouring ROCKET (0.513 vs. 0.470). The fold-to-fold standard deviation (0.12–0.14) far exceeds the mean method gap (0.013 F1), confirming that cross-severity distributional shift dominates both methods equally; structural transparency does not entail a performance penalty on this dataset.

On MFPT, the single-split picture is less representative. Fold-to-fold variance (std = 0.206) far exceeds the mean method gap (0.035 F1): fold 1 yields near-perfect F1 for both methods (shapelet 0.984, ROCKET 0.998), while fold 0 falls to 0.485 and 0.510; fold 2 is intermediate (0.802 vs. 0.867). Notably, in fold 0, the BO search achieved validation F1 = 1.000 yet generalised to only 0.485 on the test block, the sharpest illustration across both datasets of optimism under val/test distribution mismatch. The earlier single-split result (shapelet 0.894, ROCKET 0.989) corresponds to a particularly favourable split and should not be taken as representative of overall MFPT difficulty; block-to-block distributional shift is the dominant performance driver, not model capacity.

Table 8 reports the per-fold best configurations and the validation–test F1 gap. The gap varies substantially: positive values (test > val) arise when the test recording block is easier to classify than the validation block; negative values indicate optimism in the BO search. The extreme case is MFPT fold 0 (gap = −0.516), where the validation block is perfectly classified but the test block is not, confirming that BO overfits to the specific distributional characteristics of the validation recording block rather than to the overall task. Across the remaining five folds, the gap is moderate ($|\text{gap}| \leq 0.37$); the instability is specific to MFPT fold 0 and reflects the unusually large between-block distributional mismatch

for that partition. A more robust parameter selection strategy—such as cross-validation within the training+validation pool—would reduce this optimism, but is computationally expensive given the cost of each shapelet model training.

Table 8. Per-fold BO configurations and validation–test F1 gap (Gap = Test F1 – Val F1; positive indicates better generalisation than the BO objective suggested).

Dataset	Fold	n_shp	max_shp	n_est	depth	Val F1	Gap
CWRU	0 (test = 0.007 in)	135	8	30	5	0.475	+0.288
	1 (test = 0.014 in)	337	8	22	3	0.842	−0.372
	2 (test = 0.021 in)	404	6	11	3	0.659	−0.179
MFPT	0 (test = block 0)	308	12	20	4	1.000	−0.516
	1 (test = block 1)	220	8	21	4	0.840	+0.144
	2 (test = block 2)	100	5	2	3	0.682	+0.121

These results suggest a practical guideline for method selection. The shapelet classifier is preferable when (i) training and deployment conditions differ in fault severity or operating regime, since shapelet distance features capture structurally stable waveform patterns that transfer across amplitude changes; (ii) interpretability is an operational requirement, since the retained shapelets and tree thresholds remain auditable; or (iii) labelled data per class is scarce, since the compact feature space (at most 15 distances) avoids overfitting under small training sets. ROCKET is preferable when (i) training and test conditions are drawn from the same operational regime, so its dense random feature space can exploit fine-grained within-condition variation; and (ii) raw classification accuracy is the sole criterion and computational cost of the 10,000-kernel transform is acceptable. Neither method reliably dominates under the recording-level evaluation protocol used here; the fold-to-fold variance far exceeds the method gap on both datasets.

Table 9 reports the per-class F1 for the shapelet classifier across all six folds, addressing two concerns raised by reviewers: whether the temporal split of the single Normal recording introduces bias (R1.2/R3.3), and which fault classes drive the fold-to-fold performance variation. For the Normal class, F1 is consistently high across all folds (CWRU: 0.99–1.00; MFPT: 0.46–0.98), indicating that the temporal partition does not introduce systematic bias for the healthy-class examples; the single low value (MFPT fold 0, Normal F1 = 0.46) reflects poor overall classification on that fold for all classes due to the distributional mismatch between block 2 (train) and block 0 (test). The main source of CWRU fold-to-fold variance is fault-class misclassification under cross-severity transfer: Outer Race F1 drops to 0.12 in fold 2, and Inner Race to 0.21 in fold 1, confirming that severity-level shift is the dominant challenge rather than any artefact of the Normal-class split.

Table 9. Per-class F1 for the shapelet classifier across all cross-validation folds (retrained on train+val with fold-specific BO parameters). N = Normal, IR = Inner Race, Ba = Ball, OR = Outer Race.

Dataset	Fold	N	IR	Ba	OR
CWRU	0 (test = 0.007 in)	1.00	0.82	0.55	0.74
	1 (test = 0.014 in)	1.00	0.21	0.36	0.30
	2 (test = 0.021 in)	1.00	0.29	0.51	0.12
MFPT	0 (test = block 0)	0.46	0.72	—	0.28
	1 (test = block 1)	0.98	1.00	—	0.97
	2 (test = block 2)	0.88	0.84	—	0.68

4.3. Calibration Analysis

Having established the primary performance comparison, we now characterise the reliability of the selected model's probability estimates as a secondary analysis relevant to deployment under matched laboratory conditions. Table 10 reports ECE before and after Platt scaling for the best shapelet configuration and the ROCKET baseline.

Table 10. Expected Calibration Error (ECE) on the test set. **Shapelet raw:** ECE of the best BO configuration without post hoc calibration; **Shapelet Platt:** ECE after Platt scaling fitted on the validation set; **ROCKET Platt:** ECE of the ROCKET baseline after Platt scaling. CWRU Platt ECE increases because Platt scaling was fitted on a different severity level than the test set (distribution shift).

Dataset	Shapelet Raw	Shapelet Platt	ROCKET Platt
CWRU	0.034	0.269	0.281
MFPT	0.073	0.046	0.057

On CWRU, the raw ECE of 0.034 shows that the interpretability-constrained shapelet model produces well-calibrated probability estimates even under cross-severity transfer: the model is uncertain about its predictions, but its confidence estimates are not systematically overconfident or underconfident. Platt scaling, fitted on the validation set (0.014 inch faults), degrades calibration substantially on the test set ($0.034 \rightarrow 0.269$): when the Platt sigmoid is fitted on a different severity distribution than the test set, the estimated scaling parameters push predicted probabilities away from the decision boundary, a known failure mode of post hoc calibration under covariate shift. Crucially, this failure is not a property of the interpretable model but of post hoc calibration under distribution shift; ROCKET's Platt ECE (0.281) suffers the same degradation.

On MFPT, where validation and test recordings share the same operational regime, Platt scaling behaves as expected ($0.073 \rightarrow 0.046$). The calibrated shapelet ECE (0.046) is lower than the ROCKET baseline (0.057), indicating that the interpretability-constrained model can be *better* calibrated than its non-interpretable counterpart when the evaluation conditions are matched.

4.4. Conformal Prediction Sets

As a further secondary analysis, we evaluate split conformal prediction to assess whether the model's uncertainty can be communicated as prediction sets with formal coverage guarantees. Table 11 reports empirical coverage and mean prediction-set size at four confidence levels for each dataset, and Table 12 shows per-class coverage at the 95% level.

On CWRU, empirical coverage meets or exceeds the nominal level at every tested confidence threshold. Strictly, the CWRU calibration (0.014 inch) and test (0.021 inch) sets are drawn from different severity levels and are therefore not exactly exchangeable; nonetheless, coverage holds in practice, suggesting the cross-severity distributional shift is mild enough not to violate the quantile threshold meaningfully. At $1 - \alpha = 0.95$, empirical coverage is 0.979—nearly three percentage points above nominal—indicating a conservative predictor. Mean prediction-set size at 95% confidence is 2.93 out of four possible classes, consistent with the CWRU model's limited certainty in the cross-severity setting: the conformal predictor correctly signals that fault classification under severity transfer is inherently ambiguous.

On MFPT, empirical coverage falls *below* the nominal level at all four confidence thresholds (Table 11). The marginal conformal guarantee requires that calibration and test data be exchangeable. In the recording-level split, the MFPT validation and test sets are drawn from different recordings with unequal class proportions (Table 2: val has

1143/285/570 windows for Normal/IR/OR, while test has 1143/285/285). This class-distribution mismatch violates exchangeability: the Outer Race class contributes twice as many windows to the calibration set as to the test set, yet the two sets come from different recording blocks whose signal characteristics differ. Outer Race test windows therefore have systematically higher nonconformity scores than the calibration windows that set the global quantile threshold, so the threshold is too low to include Outer Race in the prediction set at the required frequency. Per-class coverage at 95% confirms this: Outer Race achieves only 0.344 empirical coverage (Table 12), far below the nominal level. This outcome illustrates that recording-level splits, while necessary to prevent data leakage, can violate the exchangeability assumption of split conformal prediction when class proportions differ between the calibration and test recordings. A conformal predictor with a larger or class-stratified calibration set would be required to recover valid marginal coverage on MFPT.

Table 11. Conformal prediction results at selected confidence levels. **Cov.:** empirical coverage (italicised cells indicate coverage below the nominal level); **Size:** mean prediction-set size; **Sing.:** fraction of singleton prediction sets. MFPT coverage falls below nominal due to class-distribution mismatch between the recording-level calibration and test splits (see Section 4.4).

Dataset	$1 - \alpha$	Cov.	Size	Sing.
CWRU	0.80	0.877	2.03	23.7%
	0.90	0.944	2.14	13.5%
	0.95	0.979	2.93	3.8%
	0.99	0.993	3.78	0.0%
MFPT	0.80	0.713	0.85	84.5%
	0.90	0.831	1.00	99.6%
	0.95	0.883	1.71	28.7%
	0.99	0.967	2.56	3.9%

Table 12. Per-fault-class empirical coverage at 95% confidence level. Split conformal prediction provides only marginal (class-averaged) coverage guarantees; class-conditional coverage can deviate. MFPT Outer Race coverage (0.344) is far below nominal due to class-proportion mismatch between the calibration (2 OR recordings) and test (1 OR recording) splits.

Dataset	Normal	Inner Race	Ball	Outer Race
CWRU	1.000	0.949	0.975	0.996
MFPT	0.988	1.000	—	0.344

Class-conditional conformal prediction as a corrective experiment. Standard (marginal) split conformal prediction computes a single global nonconformity threshold and guarantees coverage only averaged across all classes. When calibration and test sets have different class proportions—as occurs in the recording-level MFPT split—this global threshold is driven by the majority class (Inner Race and Normal) and falls too low to cover the Outer Race class at the required frequency. Mondrian (class-conditional) conformal prediction [11] addresses this by computing a separate nonconformity threshold per class, so each class achieves its own coverage guarantee.

We implemented class-conditional conformal prediction using the `class_cond=True` option of the `crepes` library and compared it with the standard global predictor on both datasets, using the same models and calibration data as in Tables 11 and 12. Table 13 reports per-class coverage at the 95% level.

On CWRU, where coverage was already near-nominal under the global predictor, class-conditional prediction reduces mean set size and increases Ball class coverage to 0.903

(from 0.975), with minor changes for other classes; the marginal improvement confirms that the problem diagnosed for MFPT is specifically attributable to its between-recording class-proportion mismatch.

On MFPT, class-conditional conformal prediction substantially increases Outer Race coverage from 0.344 to **0.677** at the 95% nominal level. This represents a factor-of-two improvement: the class-specific threshold now accommodates the higher nonconformity scores of Outer Race test windows without being dominated by the calibration distribution of the larger classes. However, an Outer Race coverage of 0.677 still falls below the 0.95 nominal target. The residual gap reflects the *fundamental* exchangeability violation: even with a class-specific threshold, the calibration Outer Race windows (from recording block 2, 2 OR files) and the test Outer Race windows (from recording block 0, 1 OR file) are drawn from recordings with different distributional characteristics, so the nonconformity quantile remains miscalibrated. At the 99% nominal level, class-conditional prediction achieves an Outer Race coverage of **0.933**, approaching the target. These results illustrate a practical guideline: under recording-level splits with class-proportion imbalance, class-conditional conformal prediction substantially mitigates but does not fully eliminate coverage failure; collecting additional Outer Race calibration recordings, or using a class-stratified calibration set, would be required to recover the formal guarantee.

Table 13. Per-class empirical coverage at 95% confidence: standard (global) vs. class-conditional (Mon-drian) conformal prediction. Values below the nominal level (0.95) are italicised. Class-conditional prediction substantially improves MFPT Outer Race coverage (0.344 → 0.677) but does not fully recover the guarantee due to the fundamental exchangeability violation between calibration and test recording blocks.

Dataset	Predictor	Normal	IR	Ball	OR
CWRU	Global	1.000	<i>0.949</i>	0.975	0.996
	Class-cond.	<i>0.989</i>	0.992	<i>0.903</i>	<i>0.950</i>
MFPT	Global	0.988	1.000	—	<i>0.344</i>
	Class-cond.	<i>0.935</i>	<i>0.986</i>	—	0.677

5. Conclusions

We have studied the performance–interpretability–reliability trade-off for shapelet-based bearing fault classification under recording-level evaluation, combining Bayesian optimisation within an interpretability-constrained parameter space with calibration analysis and conformal prediction on two standard vibration datasets.

The central finding is that interpretability constraints do not prevent competitive performance: the shapelet classifier leads ROCKET on average on CWRU and remains within fold-to-fold standard deviation on MFPT, with both datasets showing that distributional shift across recordings is the dominant uncertainty, not model capacity.

The selected configurations are structurally compact by design. On CWRU, Bayesian optimisation selects eight retained shapelets and two trees with depth 5: a maintenance engineer needs to inspect only eight vibration-pattern templates and can trace any prediction through at most 32 decision paths. On MFPT, a larger configuration (12 shapelets, 33 trees, depth 4) is selected, yet the flat F1 surface means that a practitioner who prefers a more compact ensemble—five to eight trees—loses less than one percentage point of F1. These structural properties are the operational value of the interpretability constraints: they bound the cognitive load of model inspection and make it feasible for a domain expert to relate each shapelet to known fault-frequency content, cross-checking whether the model’s reasoning is physically grounded. This contrasts with ROCKET’s 10,000 random convolutional kernels, whose aggregate feature representation offers no such pathway to physical interpretation.

Two further findings concern reliability: post hoc calibration via Platt scaling is unreliable under cross-severity shift and should not be applied without a matched calibration set, and recording-level splits can violate the exchangeability assumption of split conformal prediction when class proportions differ across recordings. Class-conditional (Mondrian) conformal prediction substantially mitigates the resulting coverage failure—raising MFPT Outer Race coverage from 0.344 to 0.677 at the 95% nominal level—but does not fully recover the formal guarantee, as the residual gap reflects the fundamental between-recording distributional mismatch rather than the choice of threshold mechanism. Together these results establish a baseline for interpretability-constrained bearing fault diagnosis and identify the conditions under which reliability tools can and cannot be trusted.

The scope of this study is intentionally narrow: both CWRU and MFPT were recorded under fixed rotational speed, single-fault conditions, and controlled laboratory noise levels. The conclusions on classification performance, calibration, and conformal coverage therefore apply specifically to these controlled conditions and cannot be transferred directly to real industrial environments, which typically involve variable speed, compound or evolving faults, and high ambient noise. The main technical limitations are: (i) the fixed window length of 1024 samples, applied uniformly across sampling rates without matching the temporal scale of the fault frequencies of interest; (ii) the restriction to single-fault recordings, excluding compound faults and early-stage continuous degradation; and (iii) the absence of variable-speed or noise-augmented conditions, which limits the generalisability of the reliability findings to non-stationary industrial settings. Extending the evaluation to variable-speed datasets (e.g., FEMTO-ST) and compound-fault scenarios is reserved for future work.

A deeper limitation concerns the meaning of interpretability. This paper operationalises interpretability through two structural proxies: number of retained shapelets and maximum tree depth. Spectral analysis of the retained CWRU shapelets (Figure 3) shows that both Inner Race shapelets have dominant FFT energy within 11% of the ball pass frequency on the inner race (BPFI = 158.4 Hz), and the Normal shapelets capture high-frequency background texture inconsistent with any fault characteristic frequency. This partial alignment suggests the shapelet transform can discover physically grounded patterns under interpretability constraints, but Ball and Outer Race fault classes did not obtain dedicated shapelets in this run, so their classification relies on dissimilarity from the Normal and Inner Race templates rather than on explicitly modelled fault-specific waveforms. Full validation that all discovered shapelets correspond to physically meaningful patterns for each fault class requires further domain-expert evaluation and larger datasets where all fault classes can obtain dedicated shapelet templates.

Three-fold recording-level cross-validation (Table 9) further shows that the Normal class is consistently well-classified (CWRU F1 0.99–1.00 across all folds), confirming that the pragmatic temporal split of the single Normal recording does not introduce systematic bias for the healthy class. The main source of fold-to-fold variance is fault-class misclassification under cross-severity transfer, not Normal-class artefacts.

The class-conditional conformal experiment (Table 13) moves the conformal analysis from diagnosis to prescription: Mondrian conformal prediction raises MFPT Outer Race coverage from 0.344 to 0.677 at the 95% nominal level by computing class-specific nonconformity thresholds. The residual gap (0.677 vs. 0.95) reflects a fundamental exchangeability violation that threshold adjustment alone cannot resolve; collecting additional Outer Race calibration recordings remains the recommended remedy.

Future work directions include: Platt scaling adapted to distribution shift (e.g., temperature scaling fitted jointly on source and target distributions); Pareto-front optimisation treating F1, ECE, and interpretability as simultaneous objectives to map the full three-way

trade-off frontier; extension to compound faults and variable operating speed (FEMTO-ST dataset); and domain-expert validation of shapelet-frequency alignment across all fault classes and bearing geometries.

Author Contributions: Conceptualisation, L.G.-G., L.U., M.-A.S. and E.G.-B.; Methodology, L.G.-G., L.U., M.-A.S. and E.G.-B.; Software, L.G.-G.; Formal Analysis, L.G.-G.; Data Curation, L.G.-G.; Writing—Original Draft Preparation, L.G.-G.; Writing—Review and Editing, L.U., M.-A.S. and E.G.-B.; Visualisation, L.G.-G.; Supervision, L.U., M.-A.S. and E.G.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All datasets used are open access; references are provided in the main text. Experiment code is available from the authors upon request.

Acknowledgments: During the preparation of this manuscript the authors used Claude (Anthropic, claude-sonnet-4-6) to assist with drafting, editing, and consistency checking of the manuscript text. The authors have reviewed and edited the output and take full responsibility for the content of this publication. The AI tool was not used to generate research data or experimental results.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lei, Y.; Li, N.; Guo, L.; Li, N.; Yan, T.; Lin, J. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mech. Syst. Signal Process.* **2018**, *104*, 799–834. [\[CrossRef\]](#)
2. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [\[CrossRef\]](#)
3. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A review of machine learning interpretability methods. *Entropy* **2021**, *23*, 18.
4. Ye, L.; Keogh, E. Time series shapelets: A novel technique that allows accurate, interpretable and fast classification. *Data Min. Knowl. Discov.* **2011**, *22*, 149–182.
5. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
6. Angelopoulos, A.N.; Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [\[CrossRef\]](#)
7. Hills, J.; Lines, J.; Baranauskas, E.; Mapp, J.; Bagnall, A. Classification of time series by shapelet transformation. *Data Min. Knowl. Discov.* **2014**, *28*, 851–881.
8. Bostrom, A.; Bagnall, A. Binary shapelet transform for multiclass time series classification. *Trans. Large-Scale Data Knowl.-Centered Syst.* **2017**, *32*, 24–46. [\[CrossRef\]](#)
9. Chen, G.; Lu, Y.; Su, R. Interpretable fault diagnosis with shapelet temporal logic: Theory and application. *Automatica* **2022**, *142*, 110350. [\[CrossRef\]](#)
10. Johansson, U.; Löfström, T.; Sönströd, C. Well-Calibrated Probabilistic Predictive Maintenance using Venn-Abers. *arXiv* **2023**, arXiv:2306.06642.
11. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: Berlin/Heidelberg, Germany, 2005.
12. Romano, Y.; Sesia, M.; Candès, E. Classification with valid and adaptive coverage. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3581–3591.
13. Andéol, L.; Fel, T.; de Grancey, F.; Mossina, L. Conformal prediction for trustworthy detection of railway signals. *AI Ethics* **2024**, *4*, 157–161. [\[CrossRef\]](#)
14. Mockus, J. On the Bayes methods for seeking the extremal point. *IFAC Proc. Vol.* **1975**, *8*, 428–431. [\[CrossRef\]](#)
15. Jones, D.R.; Schonlau, M.; Welch, W.J. Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **1998**, *13*, 455–492. [\[CrossRef\]](#)
16. Balandat, M.; Karrer, B.; Jiang, D.R.; Daulton, S.; Letham, B.; Wilson, A.G.; Bakshy, E. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21524–21538.
17. He, C.; Shi, H.; Liu, X.; Li, J. Interpretable physics-informed domain adaptation paradigm for cross-machine transfer diagnosis. *Knowl.-Based Syst.* **2024**, *288*, 111499. [\[CrossRef\]](#)
18. Yan, R.; Zhou, Z.; Shang, Z.; Wang, Z.; Hu, C.; Li, Y.; Yang, Y.; Chen, X.; Gao, R.X. Knowledge Driven Machine Learning Towards Interpretable Intelligent Prognostics and Health Management: Review and Case Study. *Chin. J. Mech. Eng.* **2025**, *38*, 5. [\[CrossRef\]](#)

19. Yin, C.; Dong, Y.; He, J.; Wang, Y. A fault evolution knowledge-driven adversarial meta-learning method for few-shot tool state recognition under variable working conditions. *Eng. Appl. Artif. Intell.* **2026**, *167*, 113806. [CrossRef]
20. Li, X.; Li, S.; Zhang, G.; Xie, Y.; Wang, T.; Chu, F. A novel interpretable dynamic weighted domain adaptation network for cross-domain fault diagnosis of bearings under time-varying speeds. *Eng. Appl. Artif. Intell.* **2026**, *170*, 114173. [CrossRef]
21. Case Western Reserve University. Bearing Data Center. 2000. Available online: <https://engineering.case.edu/bearingdatacenter> (accessed on 1 June 2026).
22. Smith, W.A.; Randall, R.B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.* **2015**, *64*, 100–131. [CrossRef]
23. Machinery Failure Prevention Technology Society. MFPT Bearing Fault Dataset. 2013. Available online: <https://www.mfpt.org/fault-data-sets/> (accessed on 1 June 2026).
24. Rodriguez, J.J.; Kuncheva, L.I.; Alonso, C.J. Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1619–1630. [CrossRef] [PubMed]
25. Sobol, I.M. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 86–112. [CrossRef]
26. Naeini, M.P.; Cooper, G.; Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian binning. In Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 2901–2907.
27. Dankowski, T.; Ziegler, A. Calibrating random forests for probability estimation. *Stat. Med.* **2016**, *35*, 3949–3960. [CrossRef] [PubMed]
28. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*; Smola, A.J., Bartlett, P., Schölkopf, B., Schuurmans, D., Eds.; MIT Press: Cambridge, MA, USA, 1999; pp. 61–74.
29. Dempster, A.; Petitjean, F.; Webb, G.I. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Min. Knowl. Discov.* **2020**, *34*, 1454–1495. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.