The Eastasouth Journal of Information System and Computer Science
Vol. 1, No. 02, December, pp. 359-371

A Data-Driven Explainable AI Framework for Predicting Industrial Equipment Failures Using Machine Learning: A Bearing-Based Case Study

Md Emon Ahmed
Bernard M. Baruch College, City University of New York.

Article Info

Article history:
Received, December, 2023
Revised, December, 2023
Accepted, December, 2023

Keywords:
Actionable Alert Onset;
Condition Monitoring;
Distribution Shift;
Model Ablation;
Premature Episodes;
Shap Stability;
Temporal Validation;
Vibration Prognosis

ABSTRACT

Prediction accuracy does not necessarily indicate whether a machine-learning model warns of an impending equipment event at an actionable time. This study evaluates a data-driven, explainable framework using 15 accelerated XJTU-SY bearing tests. We audited 9,216 dual-axis vibration files, constructed 16 leakage-checked features from 9,201 preterminal observations, and compared an RMS score, logistic regression, and random forest using five bearing-disjoint folds. For the primary ten-minute recorded-end-of-test horizon, mean fold average precision was 0.6491, 0.4194, and 0.3098, respectively. Nevertheless, new actionable alarm episodes occurred in only 4, 4, and 6 of 15 held-out bearings, with 9, 32, and 27 premature episodes. An exploratory operating-condition holdout reduced actionable coverage to 0, 3, and 3 bearings. TreeSHAP explanations on 449 observations showed heterogeneous between-bearing feature rankings (median pairwise Spearman correlation 0.488). Feature-group ablation further demonstrated divergence between predictive ranking and alarm timing. These findings support decision-oriented evaluation alongside conventional metrics, but concern laboratory test termination rather than independently verified first damage or demonstrated industrial savings.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Name: Md Emon Ahmed
Institution: Bernard M. Baruch College, City University of New York.
Email: mdemon.ahmed@baruchmail.cuny.edu

1. INTRODUCTION

Vibration-based prognostics has increasingly used data-driven models to estimate bearing degradation and remaining useful life (RUL). The publicly released XJTU-SY bearing records support method development because measurements cover accelerated bearing run-to-termination experiments across three test conditions [1], [2]. However, a prediction score that separates preterminal observations from earlier snapshots does not directly establish whether a new alarm arrives within a useful intervention window. Alarm timing depends not only on score ranking but also on threshold selection, persistence and the decision window. This distinction motivates joint reporting of prediction metrics and bearing-level alarm outcomes.

The topic is not new in general. Zhu, Huang and Zhang [3] used XJTU-SY to investigate early rolling-bearing condition monitoring using dynamic asynchronous peak-factor ratios and the Honey Badger Algorithm, with envelope-spectrum confirmation for selected cases. Florian et al.

[4] presented a cost-oriented perspective on machine-learning predictive maintenance, while Hermansa et al. [5] studied industrial false-alarm reduction with explanatory methods. Accordingly, this study does not claim first use of XJTU-SY for early warning, first use of alarms in predictive maintenance or first combination of XAI with equipment monitoring. It examines a narrower empirical question: whether snapshot ranking and the onset timing of distinct alarm episodes lead to different interpretations under bearing-disjoint evaluation.

Explainable AI has also been established in maintenance research. Lundberg and Lee [6] introduced the SHAP framework, Baptista et al. [7] examined local SHAP values in prognostics, Brusa et al. [8] explored SHAP feature contributions for rotating machinery fault diagnosis, and Kundu and Hoque [9] discussed the need to assess explanation trust rather than simply reporting XAI outputs. We therefore treat SHAP feature attribution and its stability as exploratory diagnostics. Predictive attribution is not physical causal attribution, particularly when vibration features are correlated.

The research questions are: (RQ1) how well do three baseline methods produce actionable onset warnings on previously unseen test bearings in the laboratory end-of-test task; (RQ2) how do horizon, threshold strategy and persistence affect warning coverage and premature episode burden; (RQ3) do standard classification metrics and episode-oriented outcomes yield different model interpretations; and (RQ4) how consistent are random-forest SHAP rankings across held-out bearings, and what happens when predefined feature groups are removed? Our contribution is an auditable implementation and evidence for these evaluation distinctions, not an assertion that the component methods are unprecedented.

The contribution is deliberately evaluative rather than algorithmic: (i) an audited bearing-disjoint alarm-onset protocol with a clearly defined minimum-lead window; (ii) joint reporting of snapshot ranking, per-bearing actionable episode coverage and premature-alarm rates; (iii) an additional operating-condition shift stress test with validation-only thresholds; and (iv) exploratory, unit-macro SHAP ranking stability and predefined feature-group refits. These analyses do not constitute the first early-warning use of XJTU-SY, the first application of SHAP to bearings, or a verified industrial intervention.

The XJTU-SY dataset has also supported attention-based remaining-useful-life prediction using a different task definition [10]. Thus, the present study is framed as an evaluation of prediction scores and event-onset decisions rather than a first demonstration of bearing prognosis.

2. LITERATURE REVIEW

2.1 Condition monitoring and prognostic maintenance

Condition-based maintenance links sensor acquisition, diagnostic interpretation, and maintenance decision-making [11]. A systematic machinery prognostics review distinguishes condition monitoring, health indicator construction, and remaining-life prediction [12]. A broader predictive-maintenance literature review highlights variation in data, models, and industrial contexts [13]. Multi-horizon classifiers for predictive maintenance and their implications for breakdowns and unused life were studied well before this work [14]. Deep-learning reviews likewise describe advances in machinery health monitoring [15]. These precedents preclude framing either predictive maintenance or combining data-driven scores with decision rules as novel in itself.

2.2 Validation, baseline models, and imbalanced outcomes

Evaluation of temporally correlated observations requires special care; ordinary cross-validation is not automatically appropriate for time-series tasks [16]. Random forests are a long-established supervised baseline [17]. For highly imbalanced binary outcomes, precision–recall reporting can be more

informative than sole reliance on receiver-operating-characteristic scores [18]. Accordingly, this study uses average precision and separates time-adjacent observations by their parent bearing during testing. Validation-bearing thresholds are learned without using test-bearing outcomes.

2.3 *Explainability versus deployment trust*

Model explanations should not be conflated with inherently interpretable or trustworthy systems in consequential applications [19]. TreeSHAP extends local explanations to tree ensembles and facilitates aggregation across cases [20], while other local model-agnostic tools such as LIME have long provided competing explanation strategies [21]. The logistic-regression and random-forest implementations used here follow established scikit-learn estimators [22]. In this study, SHAP describes contributions to predicted scores; it cannot certify physical failure causes or justify a maintenance action without independent validation.

2.4 *Unit-level uncertainty, performance metrics, and task scope*

Cross-validation can have substantial uncertainty when the number of independent experimental units is small [23]. Work comparing confusion-matrix measures also cautions against assuming one scalar metric adequately characterizes binary-classifier quality [24]. Deep neural approaches to remaining-useful-life estimation provide a distinct prognostic line of work, often with different endpoints and datasets [25]. Precision–recall and ROC representations capture different ranking characteristics, particularly under skew [26], while class imbalance has motivated specialized learning methods for decades [27]. These sources support an evaluation protocol that explicitly reports class prevalence, per-bearing alarm episodes, and constraints on cross-study comparisons.

Relative to these precedents and the XJTU-SY early-warning study of Zhu et al. [3], the defensible contribution is an auditable bearing-disjoint comparison of detection ranking and the timing of newly activated alarms, supplemented by explicitly exploratory explanation-stability and operating-condition-shift analyses. It is not the proposal of a new classifier or an independently demonstrated industrial failure-prediction service.

3. METHODS

3.1 *Dataset, experimental endpoint and quality control*

The XJTU-SY dataset contains accelerated bearing runs from 15 units, grouped into three operating conditions of five units each [1]. The conditions use 2,100 rpm/12 kN, 2,250 rpm/11 kN and 2,400 rpm/10 kN, respectively. Horizontal and vertical vibration signals were collected at 25.6 kHz for 32,768 points (1.28 seconds) per waveform snapshot, with successive snapshots recorded at approximately one-minute intervals. These are periodic short waveform acquisitions, not continuous 25.6-kHz measurement across each full bearing life. The experiment was concluded when a publisher-specified vibration-amplitude stop rule was reached.

All six original multipart archive pieces were obtained and extracted locally. A full numeric and schema audit processed 9,216 CSV waveform snapshots corresponding to all 15 units, checking row dimensions, numeric finiteness, flatline waveforms, per-bearing inventories and sequential filenames. The validated original counts were retained without excluding unusual vibration peaks. Fifteen terminal snapshots were removed from prediction positions, leaving 9,201 chronological preterminal records. Crucially, the last recorded test endpoint is a retrospective surrogate for an approaching event; the dataset does not independently annotate

the first onset of bearing damage for this classification task. Terminology such as recorded end-of-test or terminal event is used consistently to avoid conflating the two.

3.2 *Causal feature engineering*

At each available preterminal snapshot, eight numeric features were constructed per vibration axis, for 16 predictors in total. Instantaneous features were root-mean-square amplitude (RMS), standard deviation, maximum absolute amplitude (peak) and crest factor, calculated from the independently checked waveform summary. Four causal historical RMS descriptors were a five-snapshot trailing mean, five-snapshot trailing standard deviation, a first difference and a trailing least-squares slope. For the earliest observations, only available current and preceding snapshots were used. Historical buffers were reset at the beginning of each bearing trajectory. A targeted future-perturbation test confirmed that changing later records did not alter earlier features in all 15 units.

Bearing identity, sample index, terminal lifetime, remaining minutes, label, failure mode and all other retrospective metadata were excluded from the 16-column predictor whitelist. Remaining life was used only to construct outcome labels and evaluate episode lead time. The Tier-A feature set did not include spectral-band measurements, kurtosis, skewness or features calculated from the complete future trajectory. These more elaborate candidate features must not be described as present in the experiments.

3.3 *Retrospective labels and bearing-disjoint folds*

For a bearing with ordered snapshots $j = 1,...,N$, each preterminal index $j < N$ received the target $y_H(j) = 1$ if $0 < N-j <= H$ and $y_H(j) = 0$ otherwise, where H is a horizon measured in one-minute recording intervals. The target indicates proximity to the recorded terminal snapshot. The primary horizon was H = 10 minutes, creating 150 positive observations among 9,201. Sensitivity horizons H = 5 and H = 20 had 75 and 300 positives, respectively. Since positives are concentrated in the last H snapshots of each bearing, the 9,201 observations are not independent samples.

The five original folds were assigned by bearing, not by snapshot. Every fold used nine bearing units for training, three distinct units for validation and three other units for testing, with one unit per operating condition represented in validation and testing. Each of the 15 bearing units was tested in exactly one fold. The same split was retained for all three prediction horizons. Feature scaling, model fitting and all threshold decisions were made without using that fold's test-bearing outcomes. Cross-horizon sensitivity used fresh training on the original training units under unchanged model settings.

3.4 *Models and validation-only threshold policies*

Three transparent baseline families were compared: (B0) the maximum of the two current RMS features as an uncalibrated risk score, (B1) a class-weighted logistic-regression classifier with StandardScaler fitted only on training units, and (B2) a random forest with 200 trees, balanced subsampling, one minimum leaf sample and random seed 42. The logistic regression used a maximum of 3,000 iterations and random seed 42. No test-driven hyperparameter search was conducted; random-forest outputs were treated as classifier scores, not independently calibrated failure probabilities.

Score thresholds were chosen per model and fold using a prespecified finite grid derived from validation scores. The prediction-oriented policy A maximized validation snapshot F1 with recall and threshold tie-breaks. The decision-oriented policy B prioritized the number of validation bearing units with

an actionable alarm episode onset and then reduced premature episodes per observed bearing-hour. Alarm persistence P1 required one above-threshold observation; P2 required two consecutive above-threshold observations. A new alarm episode began only upon activation from an inactive state and ended when the score fell below the threshold. An episode beginning prematurely and remaining continuously active through the actionable window was not recounted as actionable.

3.5 Primary episode-oriented evaluation

For the primary H10 analysis, the remaining recording intervals R at a newly activated episode were classified as premature when R > 10, actionable when 5 <= R <= 10 and late when 1 <= R < 5. A bearing was covered when at least one episode started in its actionable window. The numerator of actionable coverage was therefore an integer number of distinct test-bearing units out of 15, not the fraction of individually positive snapshots. The premature episode rate was divided by total observed preterminal minutes/60 across units. Additional descriptive summaries included missed actionable warnings, alarm-active time and first actionable lead. These quantities are laboratory decision proxies and do not measure real maintenance costs or downtime.

The sensitivity analysis reused the minimum lead of five minutes but adjusted the horizon-specific actionable window: exactly R = 5 for H5; R = 5,...,10 for H10; and R = 5,...,20 for H20. Consequently, actionable coverage does not have an identical geometric opportunity across horizons. Mean fold average precision and pooled F1 were computed separately for each horizon; average precision is also affected by changes in positive-class prevalence.

3.6 Exploratory operating-condition holdout stress test

To examine operating-condition shift, a post-hoc leave-one-operating-condition-out stress test was designed after reviewing the original H10 experiments. For each of three runs, the five bearings from one operating condition were held out as test units. From the two other conditions, Bearing*_5 in each condition served as validation (two units total), and Bearing*_1–4 in each served as training (eight units total). The original 16 predictors, 10-minute label, model hyperparameters, validation score-grid threshold selection and P1/P2 episode logic were reused without modifications. Thresholds were fixed on validation units before test scores were evaluated. The resulting three condition-disjoint test folds cover each of the 15 original bearings once. This test evaluates extrapolation to unseen rig operating conditions within the SAME dataset and was explicitly exploratory; it is not an independent external cohort, and its two validation bearings per fold create substantial estimation uncertainty.

3.7 Exploratory SHAP and feature-group ablation

Following inspection of the main H10 test results, the random forest was selected for explainability because its P1 validation-F1 setting showed the greatest actionable coverage (6/15 units), even though its mean AP ranked last. This choice is outcome-informed and makes subsequent explanation analyses exploratory. The original random forest was refit on training units in each fold. TreeSHAP with a tree-path-dependent reference explained deterministic samples from each held-out unit: up to 16 uniformly distributed preterminal snapshots and up to 16 terminal-proximate preterminal snapshots. The final sample comprised 449 observations from 15 test units. SHAP additivity and reproduction of the original held-out random-forest scores were verified.

Global feature importance was calculated first as each bearing's mean absolute SHAP contribution and then macro-averaged equally across the 15

independent units. Stability was quantified using Spearman's rho for the 105 unique pairs of per-bearing feature rankings, with median and interquartile range reported descriptively. The 105 comparisons share bearing units and therefore cannot be treated as independent hypothesis-test replicates.

Four predefined group removals were explored using additional random forests trained from scratch in the same bearing-disjoint folds: eight instantaneous amplitude features, eight historical RMS features, eight horizontal-channel features or eight vertical-channel features. Each variant retained eight predictors, underwent its own validation-only F1 threshold selection and was evaluated on the same already-used test bearings. Twenty model refits were completed. These retrospective ablations were not an independent confirmatory test and were not used to change the original primary model settings.

3.8 *Validation and reproducibility controls*

All original waveform, feature, label, and split artifacts were versioned and checked against recorded SHA-256 hashes. The five primary folds were evaluated once, and all pre-specified validation thresholds were saved. The later H5/H20 sensitivity models, TreeSHAP analysis, group ablations, and leave-one-operating-condition-out experiment were stored separately. Independent scripts recomputed saved average precision, confusion-matrix metrics, alarm episodes, and grouped aggregates. This numerical audit establishes internal consistency; it is not evidence of independent dataset replication.

4. RESULTS AND DISCUSSION

4.1 *Primary H10 baseline prediction and early-warning outcomes*

Table 1 summarizes the preregistered primary comparison at validation-F1 thresholds with P1 alarm persistence. The metrics are derived from 9,201 test prediction rows but actionable coverage denominators are the 15 independent test bearings.

Table 1. Primary H10 held-out baseline results, validation-F1 threshold and P1; actionable onset = 5–10 min before recorded test termination.

Model	Mean fold AP	Pooled F1	Actionable /15	Premature episodes	Premature/h
RMS	0.6491	0.3088	4	9	0.0587
Logistic regression	0.4194	0.2522	4	32	0.2087
Random forest	0.3098	0.2848	6	27	0.1761

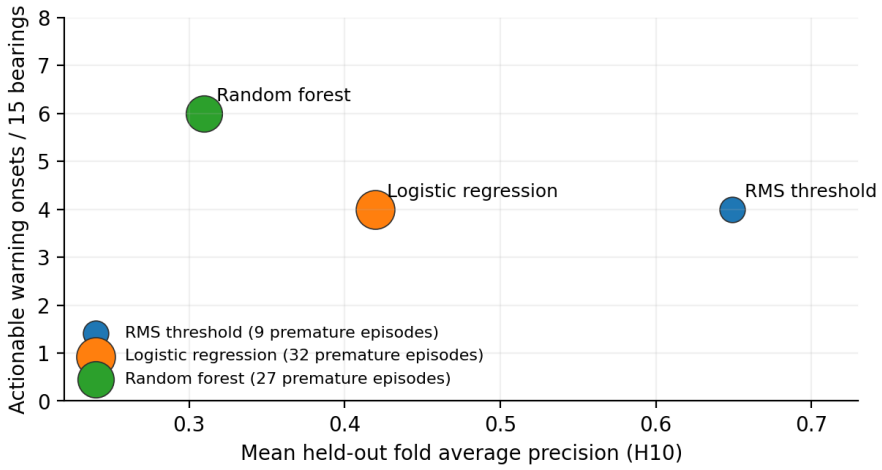


Figure 1. Original H10 prediction ranking versus actionable onset coverage; marker area indicates premature episode count.

The RMS baseline had the highest mean test-fold average precision. The random forest, despite substantially lower AP, generated actionable episode onsets for the greatest number of distinct bearing units. Relative to RMS, its additional two covered units coincided with three times as many premature episodes. These descriptive results challenge reliance on snapshot ranking alone, but the 15-unit sample and wide unit-level uncertainty do not support a statistical superiority claim. The study's original Wilson 95% intervals for actionable coverage were 10.9%-52.0% for RMS (4/15) and 19.8%-64.3% for random forest (6/15).

4.2 Threshold and persistence trade-offs

For H10 random forest under prediction-F1-selected thresholds, requiring two consecutive above-threshold recordings (P2) reduced premature episodes from 27 to 11 and actionable unit coverage from 6/15 to 3/15. This illustrates an operational trade-off: persistence can suppress short-lived alarms but can also shift an episode onset outside the actionable interval. Under the alternative validation decision-onset policy, H10 changes were not consistently favorable across models and persistence settings; all 12 original H10 configurations are preserved in the results archive. The analysis does not justify selecting a new test-set-optimal rule after observing these outcomes.

4.3 H5 and H20 sensitivity analyses

Table 2 lists validation-F1-selected P1 results for the two additional horizons. For H5, the actionable onset is exactly five minutes before the recorded terminal snapshot; for H20, it may occur anywhere from five to 20 minutes before that endpoint.

Table 2. Horizon sensitivity: warning windows differ (H5: R=5 only; H10: R=5–10; H20: R=5–20), so rows across horizons are not directly comparable.

Horizon	Model	Mean fold AP	Actionable /15	Premature/h
H5	RMS	0.5595	0	0.0652
H5	Logistic	0.3075	2	0.1435
H5	RF	0.2998	1	0.0978
H10	RMS	0.6491	4	0.0587
H10	Logistic	0.4194	4	0.2087
H10	RF	0.3098	6	0.1761
H20	RMS	0.735	5	0.228
H20	Logistic	0.573	10	0.365
H20	RF	0.375	5	0.163

At H20, validation decision-onset thresholding for logistic regression with P1 maintained 10/15 actionable unit coverage while reducing its premature episode rate from about 0.365 to 0.176 per bearing-hour. The RMS H20 decision-onset P1 setting reported 6/15 actionable units and about 0.033 premature episodes per bearing-hour. These findings remain horizon-specific descriptive comparisons: the H20 window offers 16 possible actionable one-minute positions versus only one for H5 and six for H10. The positive class prevalence also differs between horizons. No unexecuted uniform-window post-hoc comparison is reported.

4.4 SHAP importance and across-bearing stability

Across 449 deterministically selected H10 test observations explained by fold-trained random forests, maximum SHAP additivity error was 3.11 x 10^-14. Macro-averaged mean absolute SHAP ranked vertical standard deviation (std_v), horizontal standard deviation (std_h), vertical trailing RMS mean (rms_ma5_v), horizontal trailing RMS mean (rms_ma5_h) and current horizontal RMS (rms_h) as the five

leading attributes. The per-bearing rank Spearman correlation across 105 bearing pairs had a median of 0.488 and interquartile range 0.350-0.679. Thus, explanatory feature order varied meaningfully between the units. This is an observation about model attribution, not a proven physical interpretation of fault evolution.

4.5 Feature-group ablation results

Table 3 summarizes the exploratory RF H10 group-drop results. Each comparison used the original five unit-disjoint folds, fresh training and validation-only threshold selection. The full 16-feature model is the unchanged Step 4 reference.

Table 3. Exploratory H10 random-forest group-drop ablations on the same previously used bearing units.

Feature set	Mean fold AP	Δ AP vs full	Actionable /15	Premature episodes
Full RF	0.3098	reference	6	27
Drop amplitude	0.3095	−0.0003	7	37
Drop temporal	0.4001	+0.0904	2	26
Drop horizontal	0.3064	−0.0034	8	69
Drop vertical	0.2983	−0.0114	6	93

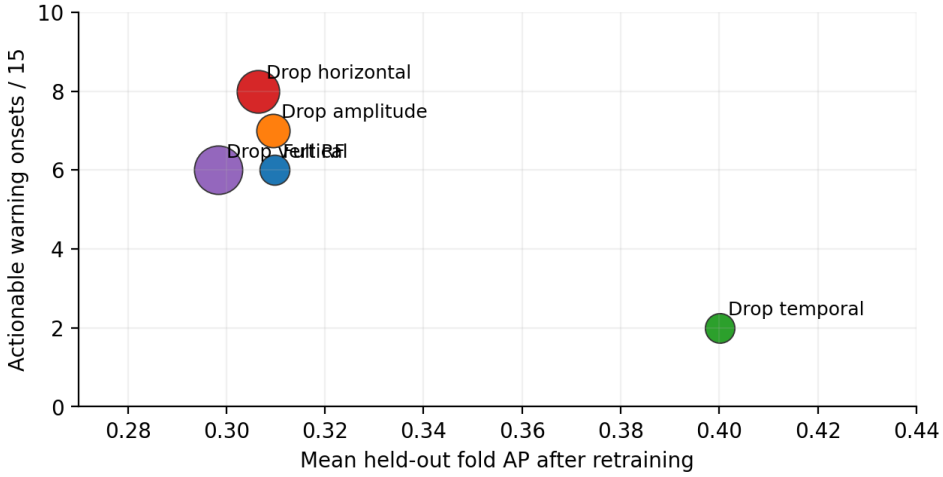


Figure 2. Post-hoc ablation trade-off: improved mean AP does not ensure improved actionable coverage; marker area indicates premature episodes.

Removal of the temporal features increased average precision from about 0.310 to 0.400 but decreased actionable warning coverage from six to two units. Conversely, removing horizontal features yielded eight covered units but substantially more premature episodes. These patterns are consistent with a difference between classification ranking and episode-onset behavior. Since the ablations were examined after primary test results and share the same units, they cannot establish which feature group should be retained for independent future deployment.

4.6 Exploratory operating-condition transfer

Leaving an entire operating condition unseen during training made actionably timed warnings substantially less frequent. Under the same H10 validation-F1/P1 convention, RMS generated zero new actionable onsets across 15 held-out units, while logistic regression and random forest each covered three. RMS retained a mean held-out-condition AP of 0.4650, underscoring that good ranking can coexist with no actionably timed alarm episodes. This stress test was planned after earlier H10 results were reviewed,

reuses the same 15 bearings, and is NOT independent confirmation.

Table 4. Operating-condition holdout stress test (exploratory; validation-F1 threshold, P1).

Model	Mean held-out condition AP	Actionable /15	Premature episodes	Premature/h
RMS	0.4650	0	2	0.0130
Logistic regression	0.4059	3	21	0.1369
Random forest	0.3708	3	21	0.1369

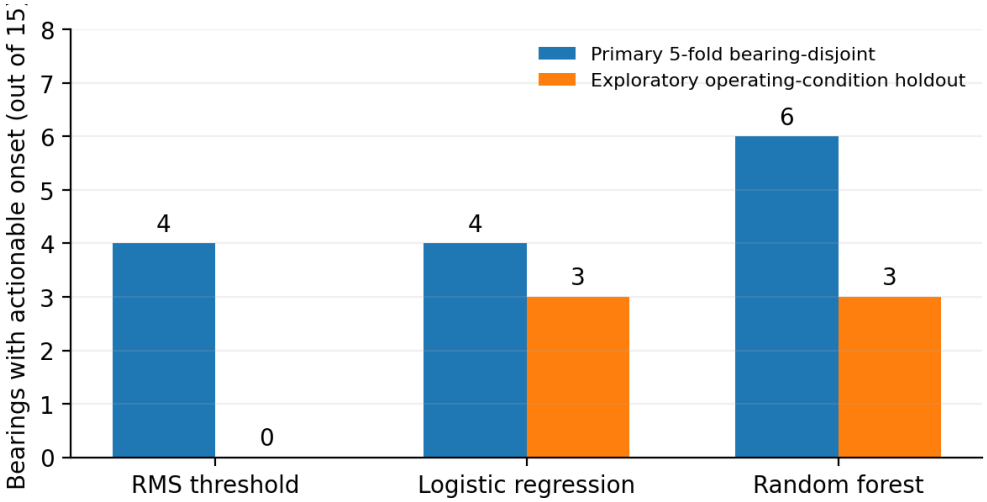


Figure 3. Actionable alert coverage under original bearing-disjoint folds versus exploratory operating-condition holdout; both use the same original bearings.

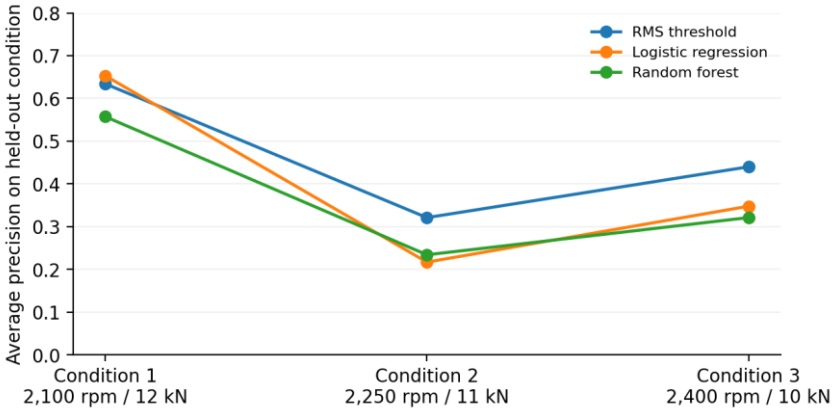


Figure 4. Average precision varies substantially by held-out laboratory operating condition. Three condition means, not three independent datasets.

Notably, all three baselines produced zero actionable onsets when condition 1 (2,100 rpm/12 kN) was held out. Logistic regression covered one condition-2 unit and two condition-3 units; random forest covered two condition-2 units and one condition-3 unit. This heterogeneity supports a limitation on condition transfer rather than a generalized warning claim.

4.7 Why prediction ranking and warning decisions diverge

A risk score is evaluated at the snapshot level by average precision or by F1 following thresholding. A maintenance-facing alarm policy instead operates on runs of scores and creates activation episodes. Even a predictor with good ranking can generate its first threshold crossing too early and remain active throughout the intended

intervention window. Conversely, a weaker global ranker may activate at a favorable time for some units. The H10 baseline and ablation comparisons demonstrate this metric mismatch descriptively. These observations justify including onset timing, per-bearing coverage and premature episode burden alongside conventional performance metrics. They do not prove that any one model is objectively best.

4.8 Interpreting laboratory lead-time windows

The five-minute minimum lead-time requirement is an analytical choice for this laboratory protocol, not an observed maintenance work-order response time in a factory. Since XJTU-SY has discrete one-minute acquisitions and accelerated trajectories, a 10-minute retrospective horizon should not be interpreted as a ten-minute industrial intervention budget. Equally, a premature episode in a run-to-termination record is not a healthy-population false positive: every bearing in this dataset eventually approaches its terminal test rule. Transfer to real maintenance decisions would require incident annotations, never-failing operational exposure, intervention response times and production impact data.

4.9 Comparison with research published through 2023

Table 5. Closest prior work (all published no later than 2023) and comparability with the present analysis.

Prior work	Task and dataset	Relevant precedent	Reason direct score comparison is invalid
Zhu et al. 2023 [3]	XJTU-SY; selected early bearing faults	Dynamic asynchronous peak-factor-ratio warning; envelope-spectrum corroboration	Different sampled bearings, standards, early fault definitions and lead-time endpoints
Hermansa et al. 2022 [5]	Industrial sensor-driven maintenance	Reduction of false alarms and explainable diagnostics	Real industrial false alarms differ from premature lab end-of-test episodes
Baptista et al. 2022 [7]	Prognostic model interpretability	SHAP relation to prognostic evaluation metrics	Different task, dependent variables and explained model
Brusa et al. 2023 [8]	Rotating machinery diagnosis	SHAP-based explanation of feature contribution	Diagnosis classes rather than onset-timed termination alert
Kundu & Hoque 2023 [9]	Prognostic XAI trust	Explanation stability/trust assessment	Different trust construct and dataset

The XJTU-SY source paper and bearing prognostic research [1,2] predate this analysis, as does attention-based RUL work on rotating machinery [10]. Zhu et al. [3] already evaluated XJTU-SY early fault condition warning, selecting data from specific outer-race-fault cases and using envelope-spectrum corroboration. Their early-warning reference and our recorded-test-termination H10 classification target are not identical; direct numerical comparison of their warning times to our event coverage would be methodologically invalid. Florian et al. [4] and Hermansa et al. [5] preclude treating cost-oriented maintenance or false-alarm reduction with XAI as original in themselves. Baptista et al. [7], Brusa et al. [8] and Kundu and Hoque [9] similarly provide existing work on explanatory reliability and SHAP features. This draft therefore frames its evidence as a transparent joint evaluation of group-disjoint alarm onset and explanation heterogeneity, not a claim

that any single component is first-of-its-kind.

Earlier maintenance-decision research [11]–[15] and [17] and interpretability methods [19]–[21] make it essential to distinguish the present evaluation contribution from already established classifiers, warning strategies, and explainability algorithms. In particular, the direct XJTU-SY early-warning precedent [3] employs different selected bearings and event criteria; numerical claims of superiority would therefore be unsupported. Grouped-validation and class-imbalance literature [16], [18], [23], [26], [27] likewise motivate caution in comparing overall snapshot-level metrics with bearing-level events.

4.10 Explainability heterogeneity and ablation trade-offs

The moderate median pairwise SHAP rank correlation shows that the 16-feature importance ordering is not uniform across experimental units. This matters because a single global SHAP chart may obscure unit-level variation. Nevertheless, the exact mechanism of this variability is undetermined: vibration distributions differ across bearing conditions, and correlated features can share attribution in model-specific ways. Group ablation supplies a complementary perturbational test of predictive relevance, but refitting after a group deletion changes the learned decision function and can alter the selected threshold. Hence the observed AP and warning changes cannot be read as direct causal effects of removing vibration sources or a proof of explanation faithfulness.

4.11 Study limitations and reproducibility

The independent sample size is 15 accelerated laboratory bearings. Repeated snapshots within each bearing are correlated and unevenly distributed across operating conditions. The dataset contains no never-failing units, no independent first-damage-onset annotations and no observed industrial response, cost, downtime, throughput or energy metrics. Model calibration was not separately validated. A proposed post-hoc comparison with a uniform 5-10-minute action window was not executed; no result from that comparison is included. The H5/H20 sensitivity experiments used different positive prevalences and actionable window widths, preventing naive ranking across horizons. The added condition-holdout stress test also found marked weakness: even with favorable AP in some conditions, RMS had no actionable alarm onsets and the learned models reached only three units each out of 15. Those results should be treated as exploratory evidence of distribution shift, not as an independent external estimate.

SHAP model selection and feature-group ablations were undertaken after H10 held-out outcomes had been observed, so their performance interpretations may be optimistic and require independent replication. Even the fivefold grouped evaluation reuses the same 15 experimental units over folds, and uncertainty is primarily at the bearing-unit level. Additional datasets, held-out machines/conditions, independent physical degradation labels and a prospective alarm-response study would be necessary for confirmation and applicability to industry.

The reproducibility record contains the original unmodified waveform archives, full numeric quality control, SHA-256 checksums, deterministic causal feature code, unit-disjoint split manifest, saved validation thresholds, per-fold and per-bearing model scores, SHAP contributions, group-drop refit results, independent re-computation reports, and the compact consolidated Step 6 tables. Primary H10 and exploratory analyses remain versioned and separate. The historical literature positioning cites research published no later than 2023; the computational experiments reported in this manuscript were performed in 2026.

An additional methodological limitation is that the primary title refers to industrial equipment failures broadly, whereas direct evidence here comes from a single bearing test rig. A new model that performs well at predicting a recorded stop rule does not necessarily predict the physical onset of component damage. Nor do alarm rates from units that all eventually terminate estimate false alerts in a healthy operating fleet. These distinctions are central to interpreting the framework as a laboratory bearing-based case study.

5. CONCLUSION

On a laboratory bearing end-of-test forecasting task, prediction ranking and the onset timing of warning episodes produced different model preferences. A simple RMS score achieved the strongest average precision under the original grouped evaluation, whereas random forest covered more bearings with appropriately timed onsets but imposed greater premature-alarm burden. Exploratory feature-group ablations confirmed that gains in ranking can coincide with fewer timely warnings. A post-hoc whole-operating-condition holdout experiment exposed substantial additional fragility, with no actionable RMS alarms and only three out of 15 timely units for each learned model. SHAP rankings also varied by bearing. These findings support event-level reporting and transparent shift testing as complements to conventional metrics, but do not establish a novel predictive algorithm, independent physical fault onset, superior deployment performance or actual resilience improvements.

DATA AVAILABILITY AND REPRODUCIBILITY

The public XJTU-SY data were originally provided by Lei et al. [1]. Original analyses, preprocessing routines, frozen split and threshold definitions, per-bearing evaluation tables, independent audit logs, and SHAP/ablation outputs are retained in the research project archive. A public repository identifier and permanent reproducibility link have not yet been assigned. Corresponding authors should provide these materials in accordance with ESISCS submission and data-sharing procedures.

REFERENCES

[1] Y. Lei, T. Han, B. Wang, N. Li, T. Yan, and J. Yang, "XJTU-SY Rolling Element Bearing Accelerated Life Test Datasets: A Tutorial," *J. Mech. Eng.*, vol. 55, no. 16, pp. 1–6, 2019, doi: 10.3901/JME.2019.16.001.

[2] B. Wang, Y. Lei, N. Li, and N. Li, "A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings," *IEEE Trans. Reliab.*, vol. 69, no. 1, pp. 401–412, 2020, doi: 10.1109/TR.2018.2882682.

[3] G. Zhu, Q. Huang, and Z. Zhang, "A Condition-Monitoring Method for Rolling Bearings Based on Dynamic Asynchronous Peak-Factor Ratios," *Sensors*, vol. 23, no. 21, 2023, doi: 10.3390/s23218939.

[4] E. Florian, F. Sgarbossa, and I. Zennaro, "Machine learning-based predictive maintenance: A cost-oriented model for implementation," *Int. J. Prod. Econ.*, vol. 236, 2021, doi: 10.1016/j.ijpe.2021.108114.

[5] M. Hermansa, M. Kozielski, M. Michalak, K. Szczyrba, Ł. Wróbel, and M. Sikora, "Sensor-Based Predictive Maintenance with Reduction of False Alarms—A Case Study in Heavy Industry," *Sensors*, vol. 22, no. 1, 2022, doi: 10.3390/s22010226.

[6] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[7] M. L. Baptista, K. Goebel, and E. M. P. Henriques, "Relation between prognostics predictor evaluation metrics and local interpretability SHAP values," *Artif. Intell.*, vol. 306, 2022, doi: 10.1016/j.artint.2022.103667.

[8] E. Brusa, L. Cibrario, C. Delprete, and L. G. Di Maggio, "Explainable AI for Machine Fault Diagnosis: Understanding Features' Contribution in Machine Learning Models for Industrial Condition Monitoring," *Appl. Sci.*, vol. 13, no. 4, 2023, doi: 10.3390/app13042038.

[9] R. K. Kundu and K. A. Hoque, "Explainable Predictive Maintenance is Not Enough: Quantifying Trust in Remaining Useful Life Estimation," in *Annual Conference of the PHM Society*, 2023. doi: 10.36001/phmconf.2023.v15i1.3472.

[10] Y. Deng, C. Guo, Z. Zhang, L. Zou, X. Liu, and S. Lin, "An Attention-Based Method for Remaining Useful Life Prediction of Rotating Machinery," *Appl. Sci.*, vol. 13, no. 4, 2023, doi: 10.3390/app13042622.

[11] A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing

condition-based maintenance," *Mech. Syst. Signal Process.*, vol. 20, no. 7, pp. 1483–1510, 2006, doi: 10.1016/j.ymssp.2005.09.012.

[12] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mech. Syst. Signal Process.*, vol. 104, pp. 799–834, 2018, doi: 10.1016/j.ymssp.2017.11.016.

[13] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, R. da P. Francisco, J. P. T. V. Basto, and S. G. S. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," *Comput. Ind. Eng.*, vol. 137, 2019, doi: 10.1016/j.cie.2019.106024.

[14] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine Learning for Predictive Maintenance: A Multiple Classifier Approach," *IEEE Trans. Ind. Informatics*, vol. 11, no. 3, pp. 812–820, 2015, doi: 10.1109/TII.2014.2349359.

[15] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, 2019, doi: 10.1016/j.ymssp.2018.05.050.

[16] C. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Comput. Stat. Data Anal.*, vol. 120, pp. 70–83, 2018, doi: 10.1016/j.csda.2017.11.003.

[17] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[18] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, 2015, doi: 10.1371/journal.pone.0118432.

[19] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[20] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[22] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[23] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *Neuroimage*, vol. 180, pp. 68–77, 2018, doi: 10.1016/j.neuroimage.2017.06.061.

[24] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, 2020, doi: 10.1186/s12864-019-6413-7.

[25] X. Li, Q. Ding, and J.-Q. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," *Reliab. Eng. Syst. Saf.*, vol. 172, pp. 1–11, 2018, doi: 10.1016/j.ress.2017.11.021.

[26] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[27] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.