Journal of Advanced Marine Engineering and Technology, Vol. 50, No. 5, pp. 429~448, 2026
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2026.50.5.429

ISSN 2234-7925 (Print)
ISSN 2765-4796 (Online)
Original Paper

Severity diagnosis of propulsion-shaft misalignment on a 50 kW-class electric-propulsion land-based test system: a detection–severity two-stage pipeline with multi-domain features and gradient boosting

Jin-Man Kim[1] · Heon-Hui Kim[2] · Hyoung-Ky Oh[3] · Taek-Kun Nam[†]

(Received July 1, 2026； Revised July 28, 2026； Accepted October 1, 2026)

Abstract: In electric-propulsion ships, propulsion-shaft misalignment causes secondary damage such as bearing wear, shaft bending, and coupling loss, making early detection and severity grading essential. Using six-channel acceleration data (32,768 Hz) acquired on a 50 kW-class electric-propulsion land-based test system (LBTS) that emulates such a shafting under constant-speed, no-load conditions, this study performs three-class severity diagnosis—Normal, 2 mm, and 4 mm misalignment—at 1,500 and 2,000 rpm, with five repeated measurement sessions per class and speed, by combining a 372-dimensional multi-domain feature set with LightGBM and XGBoost. Two evaluation protocols are used with the same feature set: the leakage-free chronological split applied within every session, and a session-disjoint five-fold cross-validation in which no window of a test session appears in training. The two protocols agree closely — at 2,000 rpm the binary-detection area under the receiver operating characteristic curve (ROC-AUC) is 0.885 and 0.873 ± 0.033 respectively, and at 1,500 rpm 0.887 and 0.867 ± 0.041 — so the diagnosis does not rest on session identity. At the 15.6 ms window level the three-class accuracy at 2,000 rpm is 0.603 under the chronological split and 0.578 ± 0.026 under session-disjoint validation, and the errors divide almost evenly between the detection axis and the adjacent-severity axis. When window probabilities are aggregated over one second, session-disjoint three-class accuracy rises to 0.903 and detection-side errors fall to 4.8 % of all errors, so that severity grading becomes the sole bottleneck; the decision-window length, not the classifier, sets the achievable accuracy. Feature-importance analysis of the boosted models places the 128–256 Hz band first and a 4.9 ms autocorrelation lag among the three leading families, both of which coincide with the six-times shaft-rotation component identified in the spectra; the amplitude of that component varies monotonically with the imposed offset, but the sign of the variation depends on the drive-train configuration, which is why a learned model rather than a fixed spectral threshold is required. Because detection and grading carry different decision costs, separating them at inference is a viable operational design option rather than an accuracy-optimal method. All results reported here were obtained on the LBTS; external propulsion load, propeller-load variation, hull vibration and wave-induced effects are outside the scope of this study.

Keywords: Propulsion shaft misalignment diagnosis, Binary detection, Severity grading, Error decomposition, Multi-domain feature engineering, LightGBM, XGBoost, Session-disjoint cross-validation, Decision-window aggregation

1. Introduction

In line with the International Maritime Organization (IMO) agenda for greenhouse-gas reduction, the shipping industry is accelerating its transition from combustion-engine-based propulsion to electric-propulsion systems. In electric-propulsion ships, ensuring the reliability and stability of the propulsion shafting that connects the propulsion motor to the propeller (or load) is a key requirement for safe operation; misalignment of the shafting can induce secondary damage such as bearing-coupling damage, shaft bending, and coupling loss, which may in turn lead to operational failures **[1]-[3]**. Continuously monitoring shafting condition and detecting and

† Corresponding Author (ORCID: http://orcid.org/ 0000-0003-0432-3504): Professor, Division of Marine Engineering System, Mokpo National Maritime University, 91, Haeyangdaehak-ro, Mokpo-si, Jeonnam-Gwangju Special Metropolitan City 58628, Korea, E-mail: tknam@mmu.ac.kr, Tel: +82-61-240-7225

1 Head Principal Researcher, Ship Repair Supporting Center, Mokpo National Maritime University, E-mail: papa713@mmu.ac.kr, Tel: +82-61-240-7543

2 Professor, Division of Marine Engineering System, Mokpo National Maritime University, E-mail: heonhuikim@mmu.ac.kr, Tel: +82-61-240-7256

3 Department Manager, Maritime Mobility Department, Korea Maritime Transportation Safety Authority, E-mail: ohk0131@komsa.or.kr, Tel: +82-44-330-2556

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright ⓒ The Korean Society of Marine Engineering

grading anomalies at an early stage has therefore become a central task of Prognostics and Health Management (PHM) **[4]-[7]**.

Misalignment diagnosis generally comprises two stages: detection, which determines whether a fault is present, and severity diagnosis, which quantitatively grades the degree of fault progression. The two tasks have different decision-cost structures, in that false-alarm cost dominates detection whereas mis-grading cost dominates grading. However, few studies have built a controlled test environment, quantitatively compared graded severities—such as normal and abnormal cases—from a classifier's perspective, and reported the causes of accuracy degradation by decomposing them into detection errors and adjacent-severity confusion. Moreover, although the gradient-boosting models LightGBM and XGBoost are robust and efficient on tabular feature vectors **[8][9]**, few studies in this application setting have directly compared the two models on the same test set under the same split while also reporting the error structure.

The primary objective of this study is to determine, on a controlled land-based test system, whether the accuracy limit of three-class propulsion-shaft misalignment diagnosis originates from Normal/Abnormal detection or from adjacent-severity (2 mm ↔ 4 mm) grading, and to quantify this decomposition so that reinforcement effort can be directed to the stage that actually limits performance. The contributions that follow from this objective are fourfold. First, on data measured with six-channel acceleration (32,768 Hz) on a 50 kW-class Land-Based Test System (LBTS), we quantitatively compare the performance of three-class severity diagnosis on a common test set using a 372-dimensional multi-domain feature set combined with LightGBM and XGBoost. Second, we directly compare a detection–severity two-stage cascade—which separates detection and grading in series—against a single-stage three-class classifier under the same split, presenting the trade-off between window-level performance and operational interpretability with quantitative evidence; the cascade is thus evaluated as an operational design option rather than as an accuracy-optimal method. Third, we decompose the three-class accuracy gap into two components, detection-side errors and adjacent-severity confusion, to clarify from a system-operation perspective where reinforcement efforts should be concentrated. Fourth, the decomposition is traced as a function of the decision-window length: the predicted probabilities are averaged over consecutive windows of the same session and the detection-side and severity-side error shares are recomputed at each length. The analysis is carried out under a session-disjoint protocol in which no window of a test session appears in training; that protocol is adopted here as the evaluation condition for the decomposition rather than proposed as a result in its own right. This study is confined to diagnosis under a single, fixed operating condition. A comparison with non-boosting model families is outside the scope of this paper, which is confined to the two gradient-boosting models.

Restated in terms of method design, the gap is this: because the Normal/Abnormal boundary and the severity boundary are learned within a single loss function, the two decision costs — false alarm and missed detection on one side, adjacent-level mis-grading on the other — cannot be tuned independently, and an aggregate accuracy figure conceals which of the two boundaries actually limits performance. The scope of this study is correspondingly narrow, and its limits are stated here at the outset: the experiments are confined to a single 50 kW-class LBTS operated at constant speed without external propulsion load, to three alignment offsets (0, 2 and 4 mm) and to two rotational speeds; every condition was measured five times, so that session-disjoint validation is possible, but all sessions of one condition were acquired consecutively, so that condition and elapsed time are confounded by the measurement order; and the preprocessing uses a simple block-averaging reduction rather than a filtered decimation. Load variation, transfer to a full-scale shafting and non-boosting baselines are therefore outside the present scope and are discussed in Section 5.9.

The remainder of this paper is organized as follows. Section 2 reviews the theory of vibration-based diagnosis, multi-domain features, gradient-boosting classifiers, the general formulation of the detection–severity cascade, and the error-decomposition framework. Section 3 describes the LBTS test system and the data-session configuration, and Section 4 presents the evaluation protocol and the procedure for determining the cascade threshold. Section 5 analyzes the results for binary detection, three-class classification, the cascade, and the error decomposition, and discusses operational implications and limitations. Section 6 concludes the paper.

2. Theory and Methods

Figure 1 summarizes the overall research process of this study, from the LBTS experiment through preprocessing, feature extraction and model training to the evaluation and error-decomposition steps. The remainder of this section describes the individual blocks of that process.

2.1 Vibration Characteristics of Propulsion-Shaft Misalignment and Multi-Domain Feature Engineering

When misalignment occurs in the shafting, non-uniform loading and deformation recur once per revolution, so the vibration signal exhibits the fundamental rotational frequency and its harmonics (particularly the 2X component), together with

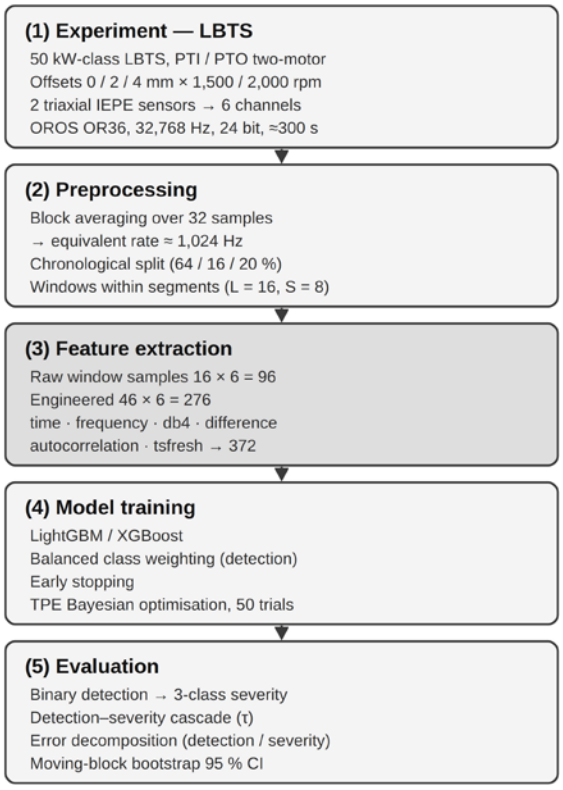

Figure 1: Overall research process, from LBTS measurement to error decomposition

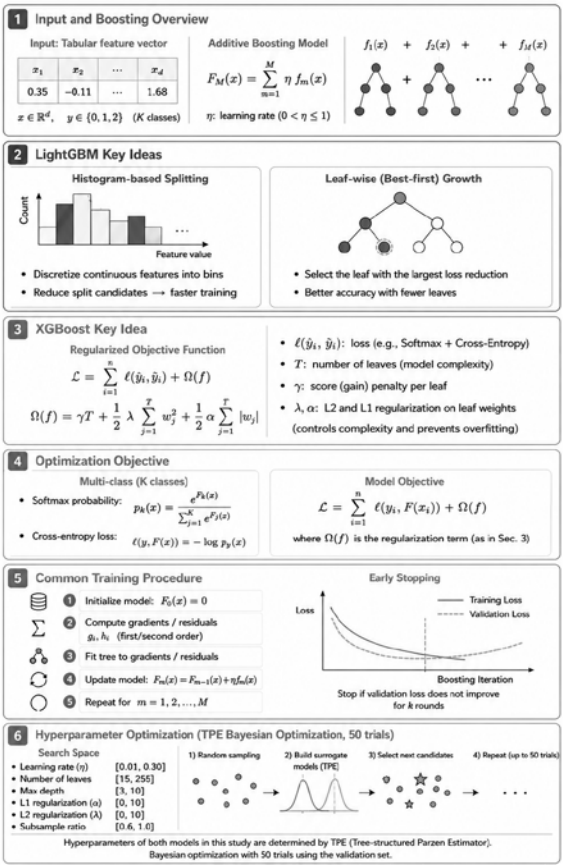

Figure 2: LightGBM and XGBoost for misalignment diagnosis

sidebands around the fundamental **[10][11]**. Although the 2X rotational harmonic is the classical misalignment signature, on the converter-driven test rig used in this study the misalignment energy also couples into electrical and structural bands: as shown in Section 3.3, the discrete components that stand out above the broadband floor are the 60 Hz supply component together with its harmonics and a component at six times the shaft rotational frequency (200 Hz at 2,000 rpm and 150 Hz at 1,500 rpm), whereas the 1X and 2X rotational harmonics do not rise measurably above that floor; at 2,000 rpm the amplitude of the six-times component varies monotonically with the misalignment stage, although the sign of that variation is a property of the drive-train configuration rather than of misalignment as such (Section 3.3). However, because the absolute vibration amplitude alone has been reported to be insufficient for distinguishing normal and abnormal stages with adequate reliability, an approach that comprehensively exploits multi-domain features is required.

The diagnostic information in a vibration signal is distributed across distinct domains: the time domain (amplitude distribution, impulsiveness, and statistical moments), the frequency domain (spectral band energy, centroid, and roll-off), the time–frequency domain (statistics of wavelet coefficients), rate-of-change (first- and second-difference statistics), and autocorrelation structure.

Features from any single domain are insufficient to adequately separate graded severities such as normal, 2 mm, and 4 mm. For each channel, this study combines time-domain statistics, FFT-based band energy and spectral-shape indices, discrete-wavelet (db4) coefficient statistics, difference statistics, autocorrelation statistics, and tsfresh-family features **[12]** to construct 46 engineered features per channel; together with the raw window samples these form the multi-domain feature vector of 372 dimensions used in this paper, whose composition is given in Section 3.3.

Finally, the three conditions examined in this study — a 0 mm reference and shaft-center offsets of 2 mm and 4 mm — are treated as controlled severity stages of a seeded parallel misalignment rather than as scaled representations of the alignment tolerance of a full-scale shafting system. The reason for this distinction, the reference against which the two abnormal levels are normalized, and the way the offsets were physically imposed and verified are given in Section 3.1.

2.2 Gradient-Boosting Classifiers

LightGBM **[8]** and XGBoost **[9]** are decision-tree-based gradient-boosting algorithms that provide fast and robust classification on tabular feature vectors of several hundred

dimensions. LightGBM shortens training time through histogram-based splitting and a leaf-wise growth strategy, whereas XGBoost suppresses overfitting by explicitly including regularization terms in its loss. Through balanced class weighting, early stopping, and Bayesian optimization of tree depth, number of leaves, learning rate, and regularization coefficients, both models adapt effectively to the imbalanced, multi-domain inputs encountered in tasks such as misalignment diagnosis. As shown in **Figure 2**, the hyperparameter search specified for both gradient-boosting models consists of 50 trials of validation-set-based TPE (Tree-structured Parzen Estimator) Bayesian optimization, with the search space set to a learning rate of [0.01, 0.3], number of leaves of [15, 255], tree depth of [3, 10], L1 and L2 regularization coefficients of [0, 10], and a subsampling ratio of [0.6, 1.0]. On the present dataset this search was not repeated separately for each model, so the two models were not tuned under identical budgets (Section 5.9).

2.3 Error Decomposition under the Detection–Severity Cascade

In operational settings, detection—which decides whether a fault is present—and grading—which decides how severe it is—entail different decision-cost structures. For the former, false-alarm and missed-detection costs dominate; for the latter, the cost of mis-grading adjacent levels dominates. A single multi-class classifier learns the Normal/Abnormal boundary and the 2 mm/4 mm boundary simultaneously within a single loss function, making it difficult to tune the two decision costs independently. The detection–severity cascade is a serial structure that classifies Normal/Abnormal in the first stage and classifies 2 mm/4 mm only for windows judged abnormal in the second stage; its operational advantages are stage-wise threshold tuning and stage-wise interpretation. However, because abnormal windows missed in the first stage cannot be recovered in the second stage—producing cumulative error—the accuracy gap relative to a single three-class classifier must be evaluated quantitatively in its own right.

The errors of a three-class classifier can be decomposed into (i) detection-side errors arising at the Normal/Abnormal boundary and (ii) adjacent-severity confusion errors arising among the abnormal classes (cf. **[13]**). The decomposition procedure is as follows. First, in the three-class predicted labels, 2 mm and 4 mm are merged into a single “Abnormal” class, yielding binary Normal/Abnormal labels. The accuracy computed on these converted labels then corresponds to the detection-separation performance that the three-class classifier exhibits at the Normal/Abnormal boundary. The difference between this converted accuracy and the original three-class accuracy can be interpreted as an additional error—the adjacent-severity grading loss incurred when detection succeeded but 2 mm and 4 mm were confused. By making visible the relative shares of detection-side errors and adjacent-severity confusion errors, this decomposition provides a quantitative basis for deciding which stage should receive priority in reinforcement efforts such as threshold tuning, window aggregation, and retraining.

3. Experimental Setup and Data

3.1 LBTS Test System and Misalignment Setup

The experiments were conducted on a 50 kW-class electric-propulsion Land-Based Test System (LBTS) built at the Ship Repair Support Center of Mokpo National Maritime University (**Figure 3**). The test bench has two synchronous reluctance permanent-magnet (SRPM) motors directly coupled by a rigid shaft; one is operated as the propulsion-side motor (PTI) and the other as the generation-side motor (PTO). Because the bench ratings, shafting configuration, instrumentation, operating procedures, and basic validation results are presented in detail in Kim *et al.* **[14]**, this paper summarizes only what is needed for the misalignment-diagnosis experiments and the interpretation of the results. These misalignment experiments were performed at fixed rotational speeds, without external propulsion load or energy-storage-system (ESS) charging/discharging and DC/DC conversion.

Misalignment was applied not by adjusting the coupling gap but by shifting the position of the sliding base on which the generation-side motor (PTO) is mounted. The sliding base was designed so that its position can be adjusted within ±20 mm in the up, down, left, and right directions and then locked. Using a gauge meter and a digital vernier caliper, the shaft-center deviation was verified, and offset conditions of 0 mm (normal), 2 mm, and 4 mm were realized. At each offset condition, the base was locked and the coupling was then fastened so that no further positional drift occurred during testing.

The connecting shaft is shrink-fitted onto each motor shaft, so that a motor and its shaft form a rigid assembly. During the test campaign a bearing on the generation side had to be replaced, and a flexible coupling (UCM87, nominal bore 30 × 30 mm) was

Table 1: Dataset inventory at 1,500 and 2,000 rpm (five repeated sessions per class) (LBTS, IEPE 6-channel, OROS OR36 @ 32,768 Hz · 24 bit; alignment offsets 0 / 2 / 4 mm; sliding-window length 16 with stride 8, every second window retained for evaluation)

rpm	Class	Sessions	Test windows (N_test)	Class ratio
1,500	Normal	5	19200	0.333
1,500	Abnormal-2mm	5	19200	0.333
1,500	Abnormal-4mm	5	19200	0.333
1,500	Subtotal	15	57600	1.0
2,000	Normal	5	19200	0.333
2,000	Abnormal-2mm	5	19200	0.333
2,000	Abnormal-4mm	5	19200	0.333
2,000	Subtotal	15	57600	1.0
Total	All	30	115200	1.0

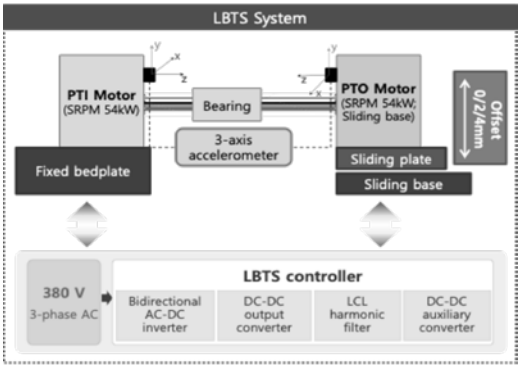

Figure 3: LBTS PTI/PTO two-motor configuration and instrumentation

inserted at the mid-shaft joint so that the bearing could be withdrawn. All measurements reported in this paper were acquired after that installation, so the drive train consists of two shrink-fitted motor-shaft assemblies joined by a single flexible coupling whose catalog allowances are an angular misalignment of 0.6° and an end play of ±0.5 mm. The prescribed offset is therefore taken up partly by the coupling and partly by the shafting itself. The two motor shafts are joined over a span of 1.6 m; the connecting shaft has a diameter of 30 mm, matching the nominal bore of the coupling; and a single support bearing is installed 0.6 m from the propulsion-side motor. At each condition the sliding base was displaced by the nominal amount and locked before the shaft connection was tightened, so that the offset persists as an imposed static condition throughout the measurement. The consequence of this configuration for the interpretation of the spectra is examined in Section 3.3.

Two points should be made explicit about the choice of 0, 2 and 4 mm. First, alignment acceptance for a full-scale marine shafting is not expressed as a lateral offset in millimeters. Classification-society guidance defines acceptability through the bearing reaction distribution — reactions positive under all operating and loading conditions, with a desired margin of at least about 10 % of the allowable bearing load — and through the relative misalignment slope between the shaft and the tail-shaft bearing, for which 0.3×10^{-3} rad is the value above which corrective measures such as slope boring are to be considered **[15]**. Designed bearing offsets reported for a small electric-propulsion ship are correspondingly of the order of a few tenths of a millimeter **[16]**. The 2 mm and 4 mm conditions used here are therefore not scaled representations of an alignment tolerance; they are deliberately seeded fault levels, set well beyond any alignment tolerance so that repeatable and clearly separated stages of fault progression could be produced on a rigid shafting. Referred to the 1.6 m span between the two motors, the imposed offsets correspond to first-order angular deviations of 1.25×10^{-3} rad at 2 mm and 2.50×10^{-3} rad at 4 mm, about four and eight times the 0.3×10^{-3} rad value quoted above. The two quantities are not interchangeable — the classification-society figure is a relative slope between the shaft and the tail-shaft bearing, whereas the figure given here is a rigid-body angle over the whole span — but the comparison places both seeded levels well outside any alignment tolerance, which is what they were intended to be.

Second, and consequently, the severity scale of this study is defined relative to the as-installed condition of the bench rather than against an absolute alignment criterion. The 0 mm condition is the aligned reference state of the shafting as assembled and verified before the experiments; the 2 mm and 4 mm conditions are obtained from that same reference by displacing the sliding base by a known amount, with the shafting, the instrumentation and the operating point otherwise unchanged. What the diagnosis learns is therefore the change in the vibration response produced

by a known increment of imposed offset on one and the same shafting, and this comparison does not require the absolute geometry of the shaft system to be known. The corollary is that the resulting severity scale is specific to this bench: mapping an offset in millimeters on the LBTS onto a severity grade for a full-scale shafting system, whose alignment acceptance is defined through bearing reactions and relative misalignment slope, requires a separate transfer study that is outside the scope of this paper (Section 5.9). The bench ratings, the shafting configuration and the instrumentation are described in detail in Kim *et al.* **[14]**.

3.2 Acceleration Measurement and Data-Session Configuration

Vibration signals were measured using two IEPE (integrated-electronics piezoelectric) triaxial accelerometers installed near the bearing housings of the SRPM motors. As shown in **Figure 3**, the two sensors were mounted on the front-end bearing-housing faces of the propulsion-side motor (PTI) and the generation-side motor (PTO), respectively. On the propulsion-side bearing housing, which exhibits a relatively large impulsive response, a wide-dynamic-range sensor rated at ±600 g and 10 mV/g was used; on the generation-side bearing housing, a sensor rated at ±60 g and 100 mV/g was used to measure the fine vibration response at higher sensitivity. The measured signals were acquired through an OROS OR36 vibration analyzer at 32,768 Hz per channel with 24-bit resolution.

In this study, a total of six acceleration channels (Motor_X/Y/Z and Propeller_X/Y/Z) were used for analysis. Here, the X, Y, and Z axes denote the axial, horizontal, and vertical directions, respectively, relative to the bearing housing. The triaxial signals from the propulsion-side sensor are denoted Motor_X/Y/Z, and those from the generation-side sensor are denoted Propeller_X/Y/Z for consistency with the data channel names. Whether these six recorded channels in fact carry six mutually independent directional responses is examined quantitatively in Section 3.3.

The dataset comprises three classes—Normal, 2 mm misalignment (Abnormal_2), and 4 mm misalignment (Abnormal_4)—at rotational speeds of 1,500 rpm and 2,000 rpm. Each measurement session is about 300 s long and contains approximately 9.83×10^6 samples per file. Each of the three classes was measured five times at each rotational speed, giving 15 sessions per speed and 30 sessions in total for the two speeds analyzed here (**Table 1**, **Figure 4**). The five repeats of a

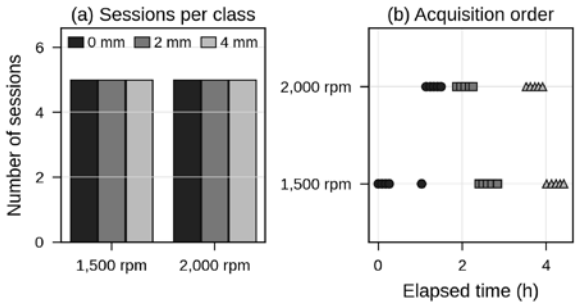

Figure 4: Session inventory: (a) sessions per alignment condition and rotational speed; (b) acquisition time of each session, showing block-wise measurement

condition were acquired consecutively, so that the within-condition variability of every feature can be quantified directly.

Because the number of sessions is balanced across classes and speeds, no class carries more measurement sessions than another, and each rotational-speed condition is treated as an independent test condition in which models and evaluation protocols are compared within the same speed. One property of the measurement order does require attention: the five 0 mm sessions were acquired before the five 2 mm sessions, which in turn preceded the five 4 mm sessions, so that alignment condition and elapsed time are confounded by design. Section 3.3 reports two checks that separate the two effects.

3.3 Vibration Signatures and Construction of the Feature Vector

This section reports the measured spectral characteristics of the recorded signals and then specifies exactly how the 372-dimensional feature vector is constructed. The two are described together because the spectral properties of the measurement determine which feature families can carry diagnostic information.

For the spectral analysis the original 32,768 Hz recordings were used rather than the reduced signal on which the classifiers operate. For each measurement session the first 30 s of each channel were divided into 32,768-sample Hann-windowed segments with 50 % overlap, the single-sided amplitude spectra of the segments were averaged, giving a frequency resolution of 1 Hz, and the five sessions belonging to the same class and rotational speed were then averaged. Session-by-session values are reported below because the within-class spread is itself part of the evidence. This analysis serves physical interpretation only; the classification pipeline described in Section 4 operates on the

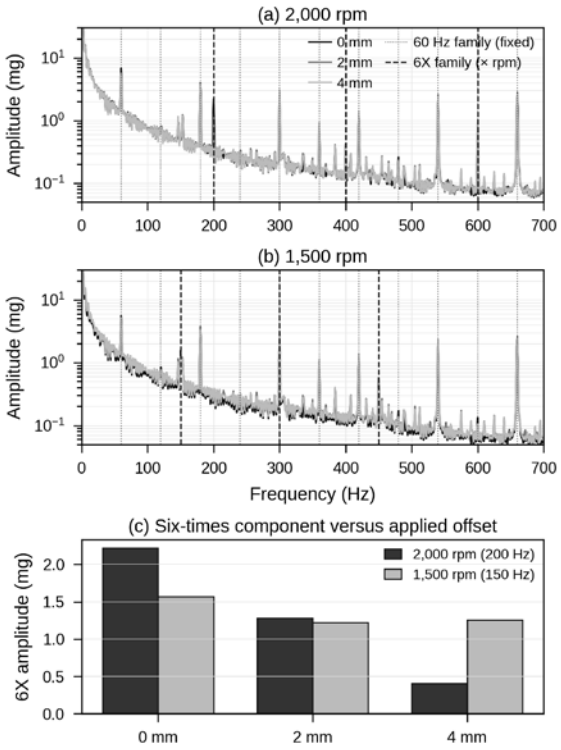

Figure 5: Amplitude spectra averaged over the five sessions of each class, from the original 32,768 Hz recordings (channel 1, Motor_X; Hann window of 32,768 samples, 50 % overlap, 30 s per session): (a) 2,000 rpm and (b) 1,500 rpm, with the speed-independent 60 Hz family and the speed-proportional 6X family marked; (c) amplitude of the six-times component versus applied offset

block-averaged signal at an equivalent sampling rate of about 1,024 Hz.

Figure 5 shows the resulting spectra, and three observations are relevant to the diagnosis task. First, below approximately 20 Hz the response is dominated by a broadband component that decays monotonically with frequency and does not differ appreciably among the three conditions. Second, the discrete components that rise clearly above this floor separate into two families, which can be told apart by comparing the two rotational speeds. One family stays at the same frequency at both speeds and consists of the 60 Hz supply component and its harmonics (180, 300, 420, 540, 660, 780 and 900 Hz); the other scales with rotational speed and consists of a component at six times the shaft rotational frequency together with its multiples — 200, 400 and 600 Hz at 2,000 rpm and 150 and 450 Hz at 1,500 rpm. Re-examined at a frequency resolution of 0.25 Hz, the six-times component lies at exactly 200.00 Hz and 150.00 Hz, so that the ratio of the two frequencies reproduces the ratio of the two rotational speeds. Third, and contrary to the classical picture of misalignment, the 1X and 2X rotational harmonics (33.3 Hz and 66.7 Hz at 2,000 rpm; 25.0 Hz and 50.0 Hz at 1,500 rpm) do not appear as discrete lines at all: in their neighborhood the spectrum shows only broadband ripple whose largest local maxima reach about twice the local median level and do not change systematically with the applied offset.

The component that varies systematically with the offset is the six-times component. At 2,000 rpm its amplitude changes monotonically across the three stages, the five-session means being 2.221 mg at 0 mm, 1.279 mg at 2 mm and 0.406 mg at 4 mm; no session departs from its class mean by more than 5.2 %, so the fifteen sessions separate without overlap (Figure 6(a)). Expressed on the block-averaged signal on which the classifiers operate, the corresponding means are 2.040, 1.179 and 0.276 mg with a pooled within-class standard deviation of 0.018 mg, so the spacing between adjacent stages is about 47 and 49 standard deviations. At 1,500 rpm the five-session means are 1.568, 1.223 and 1.258 mg: the component separates the normal from the abnormal condition but the ordering of the two abnormal stages is reversed, so it does not grade severity at that speed. The 60 Hz component decreases over the same sequence at 2,000 rpm, from 6.77 mg to 5.76 mg and 5.39 mg.

Two cautions attach to this observation. The first is that the direction of the change is not a property of misalignment as such. In a preliminary campaign acquired on the same bench before the flexible coupling described in Section 3.1 was installed, the six-times amplitude at 2,000 rpm rose over the same sequence, from 2.73 mg at 0 mm to 3.86 mg at 2 mm and 5.21 mg at 4 mm. Re-processing both campaigns with the identical procedure reproduces both results, so the reversal reflects the change of drive-train configuration and not an artifact of the analysis. A diagnostic rule that fixed the sign of the six-times deviation would therefore not transfer between configurations, whereas a model that learns the mapping from data can be retrained for each configuration. This is one of the reasons the present study uses learned classifiers rather than a spectral threshold.

The second caution is the confounding of condition with elapsed time noted in Section 3.2. Two checks separate the two effects. Within a condition block the five repeats span about 20 minutes; extrapolating the linear drift rate observed inside a block to the interval between blocks accounts for only 1/37 of the observed change in the six-times amplitude at 2,000 rpm and 1/38 at 1,500 rpm, whereas for the broadband r.m.s., the 1X and 2X harmonics and the 300 Hz line the same ratio lies between

0.14 and 2.1, and for no other feature examined does it exceed 6.9. In addition, the interval between the 0 mm and the 2 mm block differs by rotational speed — 0.74 h at 2,000 rpm, 2.30 h at 1,500 rpm and 3.39 h at 1,000 rpm, the last speed being measured on the same day and used here only for this control. The relative change of the r.m.s., of the 1X harmonic and of the 300 Hz line grows with that interval (−0.3 %, +4.5 % and +29.8 % for the r.m.s.), as a time drift would, whereas the six-times component changes most where the interval is shortest (−42.2 %, −23.2 % and −9.0 %); see **Figure 6(b)**. The six-times component is therefore attributed to the imposed offset, while level and broadband features in these recordings cannot be separated from the drift of the measurement state. This distinction is taken up again in Section 5.8.

The degree of independence among the six recorded channels was also examined. Over every session used in this study, all pairwise Pearson correlations between the six channels exceed 0.999 — in the additional 1,000 rpm sessions used only for the control described above the lowest value is 0.982 — and channels 4–6 differ from channels 1–3 by a factor that is almost constant within each axis but differs between axes — 10.21 to 10.22 for the axial pair, 10.26 to 10.27 for the horizontal pair and 9.45 for the vertical pair — all close to the ratio of the nominal sensitivities of the two accelerometers, which is ten. The RangePeak entries recorded by the analyzer differ by the same axis-wise factors. The six recorded channels therefore do not provide six independent directional responses, and the per-channel feature blocks described below are correspondingly redundant. This is a property of the measurement chain rather than of the analysis, and its consequences for the interpretation of the results are stated in Section 5.9.

The feature vector is formed per window and comprises two parts. The first part is the raw window itself: with a window length of 16 samples over 6 channels, 96 values are retained without transformation. The second part comprises 46 engineered features computed independently for each channel, giving 276 values. The two parts together give the 372 dimensions used throughout this paper, and **Table 2** lists their composition by domain. The wavelet block uses the db4 basis; for a 16-sample window the admissible decomposition depth is one level, so this block consists of the mean absolute value, the standard deviation and the maximum absolute value of the approximation and detail coefficients of that single level. The autocorrelation block uses lags of 1, 5 and 10 samples.

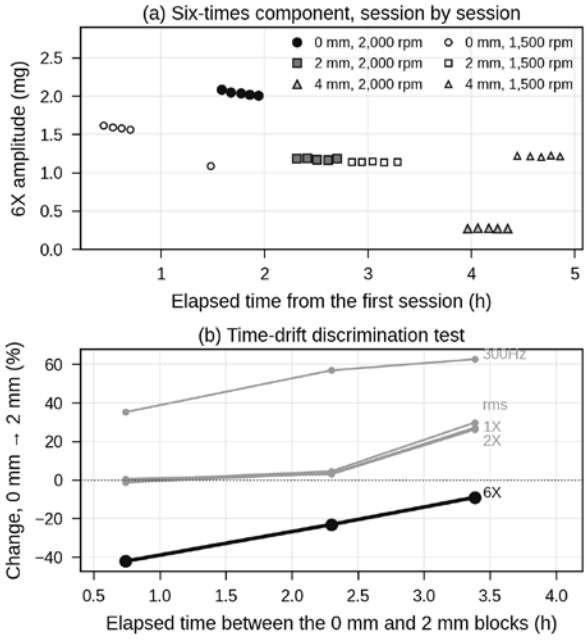

Figure 6: (a) Amplitude of the six-times component of every measurement session against the time at which the session was acquired, at 2,000 rpm (filled) and 1,500 rpm (open); (b) relative change of selected features between the 0 mm and 2 mm blocks against the elapsed time between those blocks, at the three rotational speeds

Table 2: Composition of the 372-dimensional multi-domain feature vector (window length 16, 6 recorded channels)

Domain	Per ch.	Total	Representative features
Raw window samples	16	96	window samples, no transformation
Time-domain statistics	13	78	mean, std, max, min, range, median, kurtosis, skewness, p25, p75, RMS, MAV, variance
Waveform shape	4	24	peak count, zero crossings, energy, RMS of first difference
Frequency domain (FFT)	8	48	spectral max / mean / sum, low / mid / high band energy, centroid, roll-off
Rate of change	5	30	mean and max $\|\Delta\|$, std Δ, mean $\|\Delta^2\|$, std Δ^2
tsfresh-family	7	42	absolute energy, mean absolute change, mean change, count above / below mean, first location of max / min
Wavelet (db4, level 1)	6	36	mean $\|c\|$, std c, max $\|c\|$ for approximation and detail coefficients
Autocorrelation	3	18	coefficients at lags 1, 5 and 10
Total	**62**	**372**	—

Two properties of this construction should be stated explicitly, because they bound what the frequency-domain features can represent. A 16-sample window at an equivalent sampling rate of about 1,024 Hz spans approximately 15.6 ms, which corresponds to about 0.52 of a shaft revolution at 2,000 rpm and about 0.39 of a revolution at 1,500 rpm, so no single window contains a complete revolution. Consequently, the FFT-based features are coarse band-energy and spectral-shape descriptors — eight spectral lines spaced about 64 Hz apart — rather than resolved harmonic components, and the discrimination they provide rests on short-time amplitude and waveform statistics rather than on an explicit separation of rotational orders. Information at the rotational time scale is recovered instead by the decision-aggregation step of Section 5.7: at 2,000 rpm the session-disjoint three-class accuracy rises from 0.578 for a single 15.6 ms window to 0.903 when the window probabilities are averaged over one second.

4. Evaluation Protocol

4.1 Leakage-Free Chronological Splitting

In training based on overlapping sliding windows, data leakage can occur because of sample overlap between adjacent windows. To prevent this, we first split the raw time series of each measurement session chronologically into training, validation, and test segments, and then generate sliding windows only within each segment. That is, letting the raw time-series session be $S = \{x_t\}_{t=1}^{T}$, the training, validation, and test segments are defined as in **Equation (1)**.

$$\begin{cases} S_{\text{train}} = \{x_t\}_{t=1}^{\lfloor 0.64T \rfloor} \\ S_{\text{val}} = \{x_t\}_{t=\lfloor 0.64T \rfloor+1}^{\lfloor 0.80T \rfloor} \\ S_{\text{test}} = \{x_t\}_{t=\lfloor 0.80T \rfloor+1}^{T} \end{cases} \tag{1}$$

Sliding windows of length L and stride s are then generated only within each segment. The process of generating window $\mathbf{w}_i$ can be expressed as in **Equation (2)**.

$$\mathbf{w}_i = [x_i, x_{i+1}, \dots, x_{i+L-1}], \quad i = 1, 1+s, 1+2s, \dots \tag{2}$$

In this study, the raw signal was reduced by non-overlapping averaging in blocks of 32 samples, lowering the equivalent sampling frequency to about 1,024 Hz, and sliding windows with $L = 16$ and $s = 8$ were then generated on the reduced signal. The key point is that the windows are not split randomly after being generated; rather, the time segments are separated first, and windows are generated only within each segment. This fundamentally prevents overlapping adjacent windows from being included in both the training and test sets, thereby avoiding the resulting optimistic performance bias.

Two evaluation protocols are used. Protocol P is the chronological split described above, applied within every measurement session; it is the protocol under which the per-window results of Sections 5.1 to 5.5 are reported. Protocol L is a session-disjoint five-fold cross-validation that tests whether the diagnosis depends on session identity rather than on alignment: in fold k the k-th session of each class is held out for testing, the (k+1)-th session of each class is used for validation, and the remaining nine sessions are used for training, so that no window of a test session ever appears in training. Windows are generated with L = 16 and s = 8 as above; every second window is retained under protocol P, which makes the evaluated windows non-overlapping, and every fourth window under protocol L. Results under protocol L are reported as the mean and standard deviation over the five folds in Section 5.6. As a further control, a classifier was trained to identify which of the five sessions of a single class a window came from; if sessions carried a strong identity signature this five-way problem would be easy, and its accuracy is reported alongside the protocol-L results.

4.2 Metric Definitions

For binary detection, we use accuracy, the area under the receiver operating characteristic curve (ROC-AUC), the area under the precision–recall curve (PR-AUC), and the precision and recall with respect to the abnormal class to evaluate the performance of distinguishing normal from abnormal. Letting the predicted label be $\hat{y}_i$, the true label be y_i, and the total number of test windows be N, accuracy is defined as in **Equation (3)**.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\,(\hat{y}_i = y_i) \tag{3}$$

Taking the abnormal class as the positive class, precision and recall are given by **Equation (4)**.

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN} \tag{4}$$

For three-class severity diagnosis over the three classes—Normal, 2 mm misalignment, and 4 mm misalignment—we report accuracy, macro-F1, the confusion matrix, balanced

accuracy, and Cohen's κ. The F1-score for class c is defined as in **Equation (5)**.

$$\mathrm{F1}_c = \frac{2 \cdot \mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c} \tag{5}$$

Macro-F1 is the equally weighted average of the per-class F1-scores and is computed by **Equation (6)**.

$$\mathrm{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{F1}_c \tag{6}$$

Here, C is the number of classes. Balanced accuracy is defined as the arithmetic mean of the per-class recalls and is used as a metric that complements overall accuracy when class imbalance is present. In this paper, balanced accuracy is computed according to **Equation (7)**.

$$\mathrm{Balanced\ Accuracy} = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c} \tag{7}$$

Cohen's κ is a metric that corrects the observed agreement p_o for the chance agreement p_e. In this paper, we adopt the unweighted Cohen's κ, which assigns no separate weights to the costs of misclassification between classes, and define it as in **Equation (8)**.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{8}$$

In the error-decomposition analysis, we additionally report the converted accuracy obtained after converting the three-class predicted labels into binary Normal/Abnormal labels. Letting the three-class label set be $\mathcal{Y}_3 = \{\mathrm{Normal}, \mathrm{Abnormal_2}, \mathrm{Abnormal_4}\}$, the binary conversion function $g(\cdot)$ is defined as in **Equation (9)**.

$$g(y) = \begin{cases} 0, & y = \mathrm{Normal}, \\ 1, & y \in \{\mathrm{Abnormal_2}, \mathrm{Abnormal_4}\} \end{cases} \tag{9}$$

Accordingly, the converted accuracy is computed as in **Equation (10)**.

$$\mathrm{Accuracy}_{\mathrm{conv}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(g(\hat{y}_i) = g(y_i)\right) \tag{10}$$

Using **Equation (10)**, the three-class error can be decomposed into detection-side error and severity-grading error. The detection-side error rate arising at the Normal/Abnormal boundary and the additional loss due to adjacent-severity grading are defined as in **Equation (11)**.

$$\begin{aligned} E_{\mathrm{det}} &= 1 - \mathrm{Accuracy}_{\mathrm{conv}}, \\ E_{\mathrm{sev}} &= \mathrm{Accuracy}_{\mathrm{conv}} - \mathrm{Accuracy}_{3\text{-}class} \end{aligned} \tag{11}$$

Thus, in **Equation (11)**, E_{det} denotes the detection-side error rate arising at the Normal/Abnormal boundary, and E_{sev} can be interpreted as the additional loss incurred in severity grading between 2 mm and 4 mm when Normal/Abnormal detection succeeded.

To quantify the uncertainty of performance differences between models or between approaches, we computed 95 % confidence intervals (CIs) by applying a moving-block bootstrap over the same test windows. Because the test-set windows are generated within the same measurement session, autocorrelation between adjacent windows may remain. Accounting for this, instead of the ordinary bootstrap—which assumes independent and identically distributed samples—we used a moving-block bootstrap with a block length of $B = 100$. The bootstrap block is defined as in **Equation (12)**.

$$\mathcal{B}_j = \{\mathbf{w}_j, \mathbf{w}_{j+1}, \ldots, \mathbf{w}_{j+B-1}\}, j = 1,2, \ldots, N - B + 1 \tag{12}$$

Bootstrap test sets were constructed by resampling the blocks in **Equation (12)** with replacement, and the number of iterations was set to 2,000. The resulting confidence intervals are presented together with the performance comparisons in Sections 5.2. The session-level cluster bootstrap is left to a subsequent report; cross-session generalization is not deferred but is evaluated directly by the session-disjoint protocol defined in Section 4.1.

4.3 Cascade Threshold Determination

In the detection–severity cascade, the first-stage detector outputs, for an input window $\mathbf{w}_i$, an abnormality probability $p_i = P(y_i = \mathrm{Abnormal} \mid \mathbf{w}_i)$. The binary decision of the first stage is then determined by a threshold τ as in **Equation (13)**.

$$\hat{z}_i = \begin{cases} 1, & p_i \geq \tau, \\ 0, & p_i < \tau. \end{cases} \tag{13}$$

Here, $\hat{z}_i = 1$ denotes an abnormal decision and $\hat{z}_i = 0$ a normal decision. The second-stage severity classifier classifies 2 mm versus 4 mm misalignment only for the windows judged abnormal by **Equation (13)**. The first-stage threshold τ is therefore a key operational variable that simultaneously determines the false alarms, missed detections, and the number of windows passed to severity classification for the entire cascade.

The threshold τ should be fixed on data that are not used for the final performance claim. In this paper the baseline results of Sections 5.1 to 5.4 are reported at a single fixed value of this threshold, and the candidate operating points below are computed on the test segments of protocol P and are reported as a sensitivity analysis of the cost trade-off rather than as tuned values. They are obtained by three criteria. First, we use the operating point on the ROC curve that maximizes Youden's J statistic. Youden's J statistic, the difference between the true-positive rate (TPR) and the false-positive rate (FPR) at a given threshold, is defined as in **Equation (14)**.

$$J(\tau) = \mathrm{TPR}(\tau) - \mathrm{FPR}(\tau) \tag{14}$$

The corresponding threshold is determined as in **Equation (15)**.

$$\tau_J = \arg\max_{\tau} J(\tau) \tag{15}$$

Second, we use the operating point that maximizes the F1-score on the same test segments; the corresponding threshold is defined as in **Equation (16)**.

$$\tau_{F1} = \arg\max_{\tau} \mathrm{F1}(\tau) \tag{16}$$

Third, to reflect the cost structure required under actual operating conditions, we separately derive cost-based operating points that satisfy an abnormal recall $\mathrm{Recall}_{abn} \geq 0.95$ or an abnormal precision $\mathrm{Precision}_{abn} \geq 0.95$.

The central operating point is defined as the one that simultaneously reflects τ_J from **Equation (15)** and τ_{F1} from **Equation (16)**. Specifically, the central threshold is computed as in **Equation (17)**.

$$\tau_{\mathrm{base}} = \frac{\tau_J + \tau_{F1}}{2} \tag{17}$$

Equation (17) is a compromise choice intended to prevent the selection of a threshold that is excessively biased toward either sensitivity or specificity by a single metric. In particular, when τ_J and τ_{F1} are close to each other, the central operating point in **Equation (17)** functions as a balanced threshold that reflects both criteria while being less sensitive to local fluctuations in either one. On the present test set, τ_{base} lies at about 0.49 for LightGBM at 1,500 rpm and at about 0.42 at 2,000 rpm; for XGBoost the corresponding values are 0.47 and 0.42. All baseline results in Sections 5.1 to 5.4 were therefore obtained at a fixed reference threshold of $\tau = 0.5$ rather than at the τ_{base} value given by **Equation (17)**; the operating points of **Equations (14)** to **(17)** are reported in Section 5.5 as a sensitivity analysis of the cost trade-off. In actual system deployment, however, the threshold can be readjusted to the cost-based operating points presented in Section 5.5 or to a user-defined operating point, depending on the relative costs of false alarms and missed detections, the maintenance-intervention criteria, and the safety margin.

5. Results and Discussion

5.1 Normal/Abnormal Binary Detection

At 2,000 rpm, Normal/Abnormal binary detection separates the two classes well at the level of a single 15.6 ms window (**Table 3**, **Figure 7**). LightGBM achieved an accuracy of 0.795, a ROC-AUC of 0.885 and a PR-AUC of 0.941, with an abnormal-class precision of 0.892 and recall of 0.787. XGBoost achieved an accuracy of 0.770, a ROC-AUC of 0.861 and a PR-AUC of 0.928, with an abnormal-class precision of 0.879 and recall of 0.760. These are per-window figures; Section 5.7 shows that the same models reach an accuracy above 0.98 once the decision is taken over one second of signal.

At 1,500 rpm the corresponding accuracies were 0.814 for LightGBM and 0.793 for XGBoost, with ROC-AUCs of 0.887 and 0.866 — slightly higher than at 2,000 rpm rather than lower. The difference between the two speeds is therefore not a matter of data availability, since both speeds carry five sessions per class, but of which spectral components carry the discriminative information: at 2,000 rpm the six-times component separates the three stages without overlap, whereas at 1,500 rpm it does not order them (Section 3.3), and the models fall back on broadband

Table 3: Binary detection performance per rpm and model on test segments (τ = 0.5; 'Abn' denotes Abnormal-2mm and Abnormal-4mm merged into a single Abnormal class)

rpm	Model	Accuracy	Balanced Acc	Cohen κ	macro-F1	ROC-AUC	PR-AUC	Precision (Abn)	Recall (Abn)
1,500	LightGBM	0.814	0.798	0.587	0.794	0.887	0.940	0.872	0.845
1,500	XGBoost	0.793	0.778	0.543	0.771	0.866	0.928	0.860	0.823
2,000	LightGBM	0.795	0.798	0.564	0.780	0.885	0.941	0.892	0.787
2,000	XGBoost	0.770	0.775	0.516	0.756	0.861	0.928	0.879	0.760

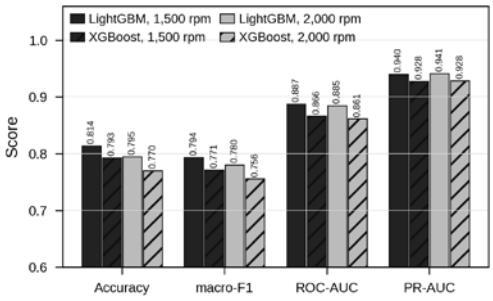

Figure 7: Binary detection performance on the held-out windows of protocol P ($\tau = 0.5$)

and level features (Section 5.8). Each speed is accordingly treated as an independent test condition, and comparisons are drawn between models and between protocols within the same speed.

5.2 Three-Class Severity Diagnosis

In three-class severity diagnosis distinguishing Normal, 2 mm and 4 mm misalignment, the 2,000 rpm per-window accuracy was 0.603 for LightGBM and 0.561 for XGBoost (**Table 4**, **Figure 8**). Balanced accuracy was 0.603 and 0.561, and Cohen's κ was 0.404 and 0.342. LightGBM is ahead of XGBoost on every metric here; the two models were not tuned identically in the present comparison, so this ordering should be read as provisional (Section 5.9).

At 1,500 rpm the three-class accuracy was 0.684 for LightGBM and 0.639 for XGBoost, with balanced accuracies of 0.684 and 0.639 and Cohen's κ of 0.525 and 0.458. Per-window three-class accuracy is thus higher at 1,500 rpm than at 2,000 rpm, even though the six-times component orders the three stages only at 2,000 rpm. The two observations are consistent: at 1,500 rpm the separation rests on broadband and level features that are numerous in the 372-dimensional vector, whereas at 2,000 rpm it rests on a single narrow band that a 16-sample window resolves poorly (Section 3.3), and it is exactly that component which the decision aggregation of Section 5.7 recovers.

Because the test windows are drawn from the same measurement sessions as the training windows under protocol P, a significance assessment based on an independent-sample assumption is not appropriate. Two corrections are applied. Within protocol P, 95 % confidence intervals were obtained from a moving-block bootstrap of block length 100 over 2,000 iterations, as described in Section 4.2. At 2,000 rpm the interval for balanced accuracy is [0.595, 0.610] for LightGBM and [0.554, 0.569] for XGBoost, and that for Cohen's κ is [0.387, 0.420] and [0.324, 0.359]; at 1,500 rpm the corresponding intervals are [0.676, 0.691] and [0.632, 0.646] for balanced accuracy and [0.513, 0.537] and [0.446, 0.470] for Cohen's κ.

Table 4: Three-class severity classification metrics (Normal / Abnormal-2mm / Abnormal-4mm) on test segments, including per-class precision (P) / recall (R) / F1 and converted detection accuracy (Normal vs. Abnormal view of the 3-class predictions)

rpm	Model	Accuracy	Balanced Acc	Cohen κ	Det. Acc (N/Abn view)	Normal P	Normal R	Normal F1	Ab-2 P	Ab-2 R	Ab-2 F1	Ab-4 P	Ab-4 R	Ab-4 F1
1,500	LightGBM	0.684	0.684	0.525	0.815	0.732	0.701	0.716	0.735	0.707	0.721	0.596	0.643	0.618
1,500	XGBoost	0.639	0.639	0.458	0.787	0.685	0.666	0.676	0.686	0.657	0.671	0.554	0.593	0.573
2,000	LightGBM	0.603	0.603	0.404	0.797	0.670	0.769	0.716	0.546	0.449	0.493	0.573	0.589	0.581
2,000	XGBoost	0.561	0.561	0.342	0.764	0.624	0.731	0.674	0.510	0.392	0.443	0.529	0.561	0.544

Table 5: Two-stage cascade (Stage1 binary detector → Stage2 severity classifier) vs. direct 3-class baseline on test segments

rpm	Stage1 binary Acc	Stage2 severity Acc (true Abn)	Cascade 3-class Acc	Cascade macro-F1	Direct 3-class Acc (LGBM)	Direct 3-class macro-F1	XGB direct Acc	XGB total errors	XGB detection-axis errors	XGB severity-axis errors	XGB severity-share	XGB detection-share
1,500	0.814	0.753	0.676	0.675	0.684	0.685	0.639	20800	12283	8517	0.409	0.591
2,000	0.795	0.651	0.612	0.605	0.603	0.597	0.561	25265	13606	11659	0.461	0.539

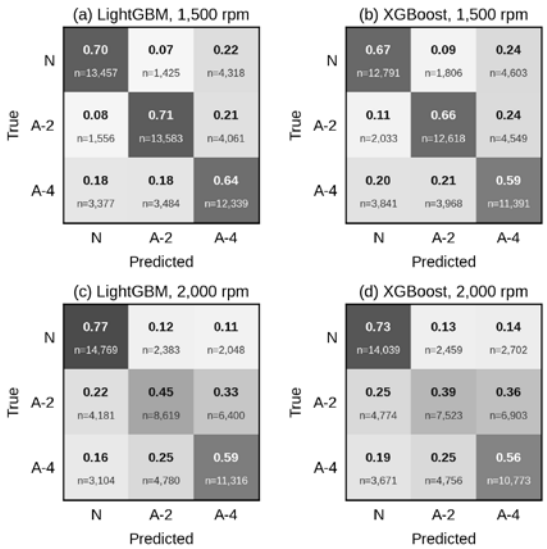

Figure 8: Three-class confusion matrices on the held-out windows of protocol P (row-normalized value on top, raw count below)

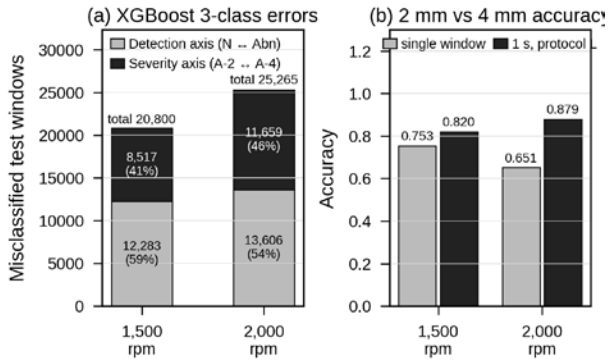

Figure 9: Error decomposition: (a) three-class errors of XGBoost split into the detection axis and the adjacent-severity axis; (b) accuracy of the 2 mm versus 4 mm classifier for a single window and for a one-second decision under the session-disjoint protocol

The intervals of the two models do not overlap at either speed, but the separation is small in absolute terms. Independently of that, protocol L removes the session overlap altogether and reports the fold-to-fold spread directly; those results are given in Section 5.6 and are the ones on which the conclusions of this paper rest.

The between-model differences at both speeds are of the order of 0.02 to 0.04 in accuracy and lie within, or close to, the fold-to-fold spread reported under protocol L. They are therefore reported as observations on this test set rather than as a general ranking of the two gradient-boosting implementations.

5.3 Error Decomposition: The Difficulty Lies Not in Detection but in Severity

As the core analysis of this paper, the three-class predicted labels were converted to a Normal/Abnormal view and the errors decomposed into detection-side errors and adjacent-severity confusion errors (**Figure 9**). At 2,000 rpm the converted accuracy was 0.797 for LightGBM and 0.764 for XGBoost, against three-class accuracies of 0.603 and 0.561. The three-class classifier therefore separates the Normal/Abnormal boundary far better than it grades severity, even at the single-window level.

For XGBoost at 2,000 rpm, 46 % of all three-class errors originated from adjacent-severity confusion and 54 % from detection-side errors at the Normal/Abnormal boundary; at 1,500 rpm the corresponding shares were 41 % and 59 %. At the single-window level the two error sources are thus of comparable size.

The balance changes decisively with the length of the decision window. When the window probabilities of the LightGBM three-class model are averaged over one second under the session-disjoint protocol at 2,000 rpm, the three-class accuracy rises to 0.903 while the converted detection accuracy rises to 0.995, so that detection contributes only 4.8 % of the remaining error and adjacent-severity confusion 95.2 %. The statement that severity grading rather than detection limits the diagnosis is therefore a statement about the operating regime in which a practical monitoring system would run.

The severity-conditional accuracy restricted to abnormal windows, that is the stand-alone accuracy of a 2 mm versus 4 mm classifier, was 0.753 at 1,500 rpm and 0.651 at 2,000 rpm (**Figure 9(b)**). Under one-second aggregation and the session-disjoint protocol the same quantity rises to 0.820 and 0.879 respectively.

To examine the distribution of the errors, the mean prediction confidence of the three-class classifier was compared between correctly and incorrectly classified windows. For the LightGBM three-class model at 2,000 rpm the mean confidence was 0.587 for correct windows and 0.503 for incorrect windows, so a substantial fraction of the misclassifications occur at windows close to the decision boundary. This is the mechanism by which averaging the predicted probabilities over consecutive windows removes most of the adjacent-severity confusion, as quantified in Section 5.7.

This error decomposition offers two implications for system operation. First, a drop in three-class accuracy does not imply a failure of Normal/Abnormal detection; on this test set the difficulty is distinguishing 2 mm from 4 mm within windows already identified as abnormal, and that difficulty is what survives when the decision window is lengthened. Second,

Table 6: Detection-threshold operating point τ by rpm × model, computed on test segments as a sensitivity analysis of the cost trade-off

rpm	Model	τ (ROC-optimal)	τ (F1-optimal)	τ (recall≥0.95)	τ (precision≥0.95)	τ (fixed)	FP rate @ fixed τ	FN rate @ fixed τ	Recall_abn @ ROC-opt	Precision_abn @ ROC-opt	F1_abn @ ROC-opt	Recall_abn @ cost-recall	Precision_abn @ cost-recall	Recall_abn @ cost-prec	Precision_abn @ cost-prec
1,500	LightGBM	0.578	0.403	0.298	0.758	0.500	0.248	0.155	0.787	0.899	0.839	0.950	0.788	0.599	0.950
1,500	XGBoost	0.561	0.373	0.306	0.754	0.500	0.268	0.177	0.767	0.884	0.821	0.950	0.770	0.532	0.950
2,000	LightGBM	0.523	0.325	0.261	0.669	0.500	0.190	0.213	0.767	0.901	0.829	0.950	0.782	0.616	0.950
2,000	XGBoost	0.511	0.335	0.267	0.681	0.500	0.209	0.240	0.747	0.885	0.810	0.950	0.761	0.539	0.950

reinforcement effort should therefore be concentrated on the severity-grading stage, and the cheapest available reinforcement is not a different classifier but a longer decision window.

5.4 Detection–Severity Cascade vs. Single Three-Class Classifier

The cascade and the single three-class classifier are compared here from the standpoint of propagated error and cost separability rather than of window-level accuracy alone: what the cascade costs when the first stage misses an abnormal window, and what it buys by exposing the operating point as an independently tunable variable. On the same test set under the same chronological split, the two structures were evaluated at the reference threshold $\tau = 0.5$ with a gradient-boosting second stage (**Table 5**, **Figure 10**). At 2,000 rpm the first-stage binary detector (LightGBM) had an accuracy of 0.795, and the stand-alone accuracy of the second-stage 2 mm versus 4 mm classifier was 0.651. The cascade formed by connecting the two stages in series achieved a three-class accuracy of 0.612 and a macro-F1 of 0.605, against 0.603 and 0.597 for the single three-class LightGBM classifier under the same condition. At 1,500 rpm the cascade reached 0.676 and 0.675 against 0.684 and 0.685 for the single classifier.

Neither structure dominates. The cascade loses abnormal windows that the first stage misses, because they cannot be recovered in the second stage, and this error propagation is what makes it fall behind the single classifier at 1,500 rpm. At 2,000 rpm the loss is offset by the fact that the second stage is trained only on abnormal windows and therefore devotes its whole capacity to the 2 mm/4 mm boundary, which is the boundary that limits performance (Section 5.3); there the cascade is marginally ahead. The difference in either direction is smaller than the fold-to-fold spread of protocol L, so the choice between the two structures should be made on operational grounds rather than on window-level accuracy.

There are two implications for system operation. First, the two structures are close enough in window-level accuracy that neither can be recommended on accuracy alone. Second, in monitoring settings where false-alarm and missed-detection costs must be tuned independently, the cascade retains a concrete advantage: the first-stage threshold balances missed detections against false alarms, and the second-stage grading can be interpreted and retrained separately from it. The stand-alone second-stage accuracy shows that this separation costs little, and Section 5.7

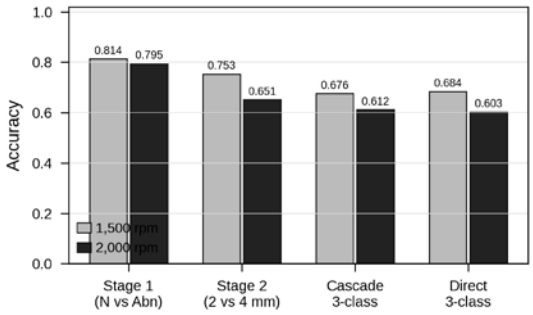

Figure 10: Two-stage cascade versus direct three-class classification (LightGBM, protocol P)

shows that both structures reach a practically usable level once the decision is taken over a second of signal.

5.5 Threshold Operating Points

In a maintenance decision-making setting, the first-stage detection threshold τ can be adjusted according to the relative costs of missed detections and false alarms (Table 6). The baseline results in this paper were obtained at $\tau = 0.5$. When the cost of a missed detection is high the threshold can be lowered: for LightGBM at 2,000 rpm the operating point targeting an abnormal recall of 0.95 is $\tau = 0.261$, at which precision falls to 0.782 while F1 rises to 0.858 (from 0.836 at $\tau = 0.5$). Conversely, where false-alarm cost dominates, the operating point satisfying an abnormal precision of 0.95 is $\tau = 0.669$, at which recall falls to 0.616. XGBoost behaves similarly, with $\tau = 0.267$ and $\tau = 0.681$ for the two criteria at the same speed.

The spread between these operating points is wide, which reflects the fact that a single 15.6 ms window carries little evidence: the predicted probabilities are concentrated near the decision boundary, so a small change of threshold moves many windows. This is a further reason to treat the threshold as an operational variable of the aggregated decision rather than of the individual window, and it motivates the analysis of Section 5.7.

5.6 Session-Disjoint Validation

Protocol L is reported here as the evaluation condition under which the error decomposition of Section 5.3 and the decision-window analysis of Section 5.7 are made, rather than as a contribution in its own right. Protocol P splits each session chronologically, so training and test windows come from the same recordings. If the sessions of one class shared an incidental signature — a slightly different mounting, a different thermal state — a classifier could exploit it instead of the alignment.

Table 7: Chronological split (protocol P) versus session-disjoint five-fold cross-validation (protocol L) on the same feature set; protocol L values are the mean and standard deviation over the five folds

rpm	Task	Model	Protocol P (accuracy)	Protocol L (accuracy, mean ± s.d.)	Protocol P (ROC-AUC)	Protocol L (ROC-AUC, mean ± s.d.)
1,500	Detection	LightGBM	0.814	0.792 ± 0.033	0.887	0.867 ± 0.041
1,500	Detection	XGBoost	0.793	0.777 ± 0.030	0.866	0.852 ± 0.039
1,500	Three-class	LightGBM	0.684	0.632 ± 0.049	—	—
1,500	Three-class	XGBoost	0.639	0.597 ± 0.047	—	—
1,500	2 vs 4 mm	LightGBM	0.753	0.637 ± 0.106	0.834	0.668 ± 0.156
1,500	2 vs 4 mm	XGBoost	0.740	0.619 ± 0.109	0.814	0.661 ± 0.139
2,000	Detection	LightGBM	0.795	0.783 ± 0.023	0.885	0.873 ± 0.033
2,000	Detection	XGBoost	0.770	0.764 ± 0.020	0.861	0.856 ± 0.036
2,000	Three-class	LightGBM	0.603	0.578 ± 0.026	—	—
2,000	Three-class	XGBoost	0.561	0.552 ± 0.030	—	—
2,000	2 vs 4 mm	LightGBM	0.651	0.619 ± 0.026	0.715	0.669 ± 0.037
2,000	2 vs 4 mm	XGBoost	0.620	0.603 ± 0.020	0.674	0.647 ± 0.030

Table 8: Accuracy as a function of the decision-window length, obtained by averaging the predicted class probabilities over consecutive windows of the same session (LightGBM); protocol L values are the mean over the five folds

rpm	Task	Protocol	0.016 s	0.031 s	0.062 s	0.125 s	0.25 s	0.5 s	1 s	2 s	4 s	8 s	16 s	32 s
1,500	Detection	P	0.814	—	0.926	—	0.970	—	0.974	—	0.982	—	0.978	—
1,500	Detection	L	—	0.792	—	0.915	—	0.973	—	0.979	—	0.977	—	0.970
1,500	Three-class	P	0.684	—	0.868	—	0.976	—	0.992	—	1.000	—	1.000	—
1,500	Three-class	L	—	0.632	—	0.802	—	0.934	—	0.979	—	0.995	—	1.000
2,000	Detection	P	0.795	—	0.948	—	0.996	—	1.000	—	1.000	—	1.000	—
2,000	Detection	L	—	0.783	—	0.933	—	0.985	—	0.993	—	1.000	—	1.000
2,000	Three-class	P	0.603	—	0.775	—	0.915	—	0.963	—	0.969	—	0.978	—
2,000	Three-class	L	—	0.578	—	0.740	—	0.872	—	0.924	—	0.939	—	0.941

Protocol L removes that possibility by holding out whole sessions. **Table 7** compares the two protocols. Binary detection with LightGBM gives 0.795 under protocol P and 0.783 ± 0.023 under protocol L at 2,000 rpm, with ROC-AUCs of 0.885 and 0.873 ± 0.033; at 1,500 rpm the corresponding figures are 0.814 and 0.792 ± 0.033, with ROC-AUCs of 0.887 and 0.867 ± 0.041. Three-class severity behaves the same way, 0.603 against 0.578 ± 0.026 at 2,000 rpm and 0.684 against 0.632 ± 0.049 at 1,500 rpm.

For detection and for three-class severity the two protocols differ by less than 0.06, which is of the order of the fold-to-fold spread itself. The diagnosis therefore transfers to sessions the model has never seen, and the performance reported under protocol P is not an artifact of within-session splitting. The auxiliary control points the same way: a classifier trained to identify which of the five sessions of a single class a window came from reaches an accuracy of 0.29 to 0.35 against a chance level of 0.20, so a session signature exists but is far too weak to carry the class decision.

One exception should be recorded. Discriminating 2 mm from 4 mm at 1,500 rpm gives 0.637 ± 0.106 under protocol L, and one of the five folds falls to a ROC-AUC of 0.546, which is close to chance. At 2,000 rpm the same task is stable at 0.619 ± 0.026. This is the same asymmetry seen in the spectra: the six-times component orders the three stages at 2,000 rpm but not at 1,500 rpm (Section 3.3). Severity grading at 1,500 rpm is therefore reported here but is not claimed to generalize.

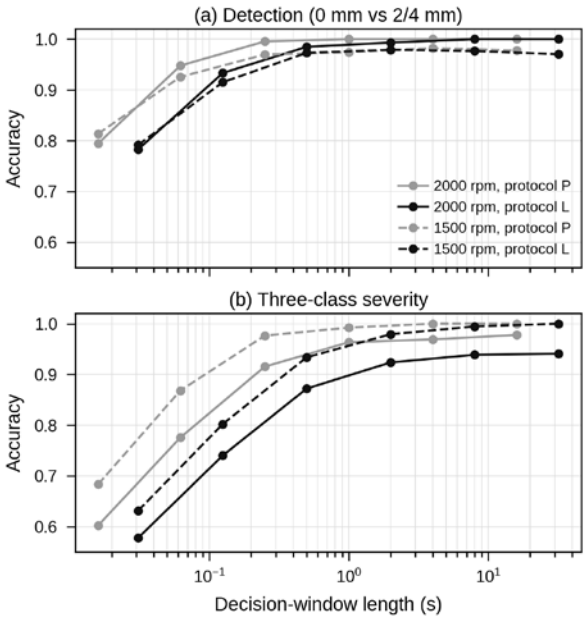

Figure 11: Accuracy against decision-window length for (a) binary detection and (b) three-class severity, under the chronological split (protocol P, gray) and session-disjoint cross-validation (protocol L, black)

Table 9: Split-gain contribution of the leading feature families in the 2,000 rpm binary detector (LightGBM), aggregated over the six recorded channels

Feature family	Gain share	Definition	Corresponding frequency
fft_mid	21.1 %	Sum of bins 2–4 of the 16-point FFT	128–256 Hz (bin 3 = 192 Hz)
acorr_lag1	10.7 %	Autocorrelation at a lag of 1 sample	period 0.977 ms
acorr_lag5	9.7 %	Autocorrelation at a lag of 5 samples	period 4.883 ms = 204.8 Hz
range	8.6 %	Window maximum minus minimum	broadband
std	5.2 %	Window standard deviation	broadband
acorr_lag10	5.0 %	Autocorrelation at a lag of 10 samples	period 9.766 ms = 102.4 Hz

5.7 Decision-Window Aggregation

The decision unit of the pipeline is a 16-sample window, which at an equivalent sampling rate of about 1,024 Hz spans 15.6 ms — about half a shaft revolution — and gives a frequency resolution of 64 Hz. Nothing in a monitoring application requires an independent decision every 15.6 ms. Averaging the predicted class probabilities over K consecutive windows of the same session and taking the arg-max of the average yields a decision over K × 15.6 ms of signal at protocol-P sampling, and the accuracy as a function of that length is given in **Table 8** and **Figure 11**. Under protocol L, where every fourth window is retained (Section 4.1), the windows are 31.25 ms apart, so K windows cover K × 31.25 ms. The one-second protocol-L values quoted in this paper — the three-class accuracy of 0.903 at 2,000 rpm with the detection figures derived from the same decisions (0.995, 4.8 % and 95.2 %), and the 2 mm versus 4 mm accuracies of 0.820 and 0.879 — were obtained in this way with K = 32. They fall between the 0.5 s and 2 s columns of Table 8 and are therefore not listed there; 0.903 lies between the corresponding entries of 0.872 and 0.924.

The gain is large and it is concentrated below one second. At 2,000 rpm the binary-detection accuracy rises from 0.795 for a single window to 1.000 at a one-second decision under protocol P, and from 0.783 to 0.993 at a two-second decision under protocol L. Three-class severity rises from 0.603 to 0.963 and from 0.578 to 0.924 respectively. At 1,500 rpm the same pattern holds, three-class severity reaching 0.979 at a two-second decision under protocol L. The curves flatten beyond a few seconds, so the useful operating range is of the order of one to four seconds.

Two qualifications apply. The number of aggregated decisions falls in proportion to the window length, so the longest windows in Table 8 rest on a small number of decisions and their accuracies carry a correspondingly wide interval; the values at and beyond eight seconds are reported for completeness rather than as precise estimates. And aggregation assumes the alignment state is constant over the averaging interval, which holds for a progressive fault but not for a transient event. Within those limits the result is unambiguous: the per-window accuracies of Sections 5.1 to 5.4 understate what the same models deliver in an operational setting, and the binding constraint on the reported performance was the length of the decision window rather than the choice of classifier.

5.8 What the Boosted Models Use

Gradient-boosting models expose the contribution of each input directly, and the split-gain ranking provides a check on whether the learned decision rests on the components identified in the spectra. Table 9 aggregates the gain of the 372 inputs by feature family for the 2,000 rpm binary detector.

Two of the three leading families point to the same place in the spectrum. The mid-band FFT feature, which ranks first at 21.1 %, sums bins 2 to 4 of a 16-point transform, which at an equivalent sampling rate of 1,024 Hz covers 128 to 256 Hz with its central bin at 192 Hz; the autocorrelation at a lag of five samples, third at 9.7 %, measures repetition at a period of 4.883 ms, that is at 204.8 Hz. Both coincide with the six-times component at 200 Hz that Section 3.3 identified from the spectra by an entirely separate route. Between them stands the lag-1 autocorrelation at 10.7 %, which measures sample-to-sample smoothness and corresponds to no single frequency. In the three-class model the same two spectral families rank first and second, with gains of 14.6 % and 7.9 %. The learned decision is thus not opaque: it rests on the component whose amplitude separates the three alignment stages without overlap.

At 1,500 rpm the ranking is different. The leading families are the high-band FFT sum (320–448 Hz, 11.1 %), the autocorrelation at lag 10 (102.4 Hz, 10.9 %), the spectral maximum (10.1 %) and the window standard deviation (7.4 %)

— none of which corresponds to the six-times component at 150 Hz. These are broadband and level features, and Section 3.3 showed that features of exactly this kind cannot be separated from the drift of the measurement state in this experiment. The 1,500 rpm results are therefore reported with the caveat that part of the separation may reflect measurement state rather than alignment; the 2,000 rpm results, which rest on the six-times component, are not subject to that reservation.

5.9 Limitations and Scope of Further Validation

The limitations of this paper are as follows. First, the experiments were conducted on a single 50 kW-class LBTS and are confined to constant-speed operation with limited external load variation and to misalignment conditions of 0/2/4 mm. Generalizing these results to full-scale ship environments with large load variations and diverse operating profiles therefore requires additional validation. Second, and most consequentially, all sessions of one alignment condition were acquired consecutively, so alignment condition and elapsed time are confounded by the measurement order. The two checks reported in Section 3.3 exclude that confounding for the six-times component but not for broadband and level features, and the 1,500 rpm results rest largely on the latter (Section 5.8). A design that interleaves the conditions in random order, with at least one return to a previously measured condition, is required to remove this limitation, and is left to a further measurement campaign.

Third, the preprocessing in this paper is based on a simple reduction procedure using non-overlapping averaging in blocks of 32 samples. For precise frequency-component analysis or spectrum-based comparison, polyphase decimation with an anti-aliasing filter would therefore need to be introduced. The conclusions of this paper should be interpreted within the range in which such preprocessing differences do not alter the main classification tendencies. Fourth, this study compared only the gradient-boosting models LightGBM and XGBoost, which are well suited to 372-dimensional tabular features. A multi-model comparison with non-boosting baselines such as SVM, Random Forest, 1D-CNN, MiniRocket, and InceptionTime is left to future work using separate data and a separate evaluation protocol.

Fifth, the two gradient-boosting models were not tuned under identical budgets in this comparison: the reported ordering, in which LightGBM leads XGBoost at both speeds, was obtained without repeating the Bayesian hyperparameter search separately for each model on the present dataset, and should therefore be read as provisional. The uncertainty of differences between models and between approaches was presented as 95 % confidence intervals via a moving-block bootstrap with a block length of 100 over 2,000 iterations. This is a conservative correction that partially accounts for autocorrelation between windows within the same session, but it does not replace a full session-level statistical test on this session-limited test set. The session-level cluster bootstrap and the paired bootstrap for the difference between the two classifiers need to be further quantified in future work. Cross-session generalization itself is not deferred: it is evaluated directly by the session-disjoint protocol of Section 5.6.

Sixth, the six recorded acceleration channels were found to be strongly co-linear: over every session analyzed in this study all pairwise correlations exceed 0.999, and the two channel groups differ essentially by the ratio of the accelerometer sensitivities (Section 3.3). The measurement therefore does not resolve the axial, horizontal and vertical directions independently, and the results of this paper should be read as having been obtained from a single effective vibration channel replicated across six inputs. Restoring genuinely independent triaxial channels requires a review of the measurement chain and a repeat measurement campaign, and is a prerequisite for any direction-resolved interpretation of misalignment.

Seventh, hull excitation in an electric-propulsion ship differs in kind from that of an internal-combustion-engine ship. In the latter the excitation that dominates the hull originates from the main engine, so the difference between a shop test and an onboard measurement is comparatively small; in the former, generator and auxiliary-machinery excitation and wave-induced impacts can dominate the hull vibration response. The LBTS data of this study were acquired in a controlled land environment in which such external excitation is absent, so its influence on the diagnosis is not represented here. Applying the present approach onboard will therefore require an order-tracking-based cutoff logic that extracts and gates the rotational-order components, together with a robustness-reinforcement framework capable of separating and removing auxiliary-machinery excitation and external impact noise.

Eighth, the severity scale used here is defined on the test shafting. Translating an offset in millimeters on the LBTS into a severity grade for a full-scale shafting system — whose alignment acceptance is defined through bearing reactions and relative misalignment slope rather than through a lateral offset **[15][16]** — requires a separate transfer study that is not attempted in this paper.

In view of these limitations, this paper focuses on the methodology and error-structure analysis of in-distribution three-class diagnosis within a single operating condition. Cross-session generalization is addressed here through the session-disjoint protocol of Section 5.6, whereas randomized measurement order, load variation, multi-model comparison under a common tuning budget, order tracking and latency analysis are left to future work using separate data and a separate evaluation protocol. In this way, the paper clearly presents the detection–severity diagnosis structure and the direction for error reinforcement under in-distribution conditions, while delimiting the research scope so that the broader generalization problem can be addressed systematically in separate studies.

6. Conclusion

Based on six-channel vibration signals (32,768 Hz) measured on a 50 kW-class electric-propulsion Land-Based Test System (LBTS), this study extracted a 372-dimensional multi-domain feature set and applied the gradient-boosting models LightGBM and XGBoost to the detection and 2 mm/4 mm severity grading of propulsion-shaft misalignment. Thirty measurement sessions were used, five for each of the three alignment conditions at each of the two rotational speeds, which made a session-disjoint evaluation possible. At 2,000 rpm binary detection reached an accuracy of 0.795 and a ROC-AUC of 0.885 for a single 15.6 ms window under the chronological split, and 0.783 ± 0.023 with a ROC-AUC of 0.873 ± 0.033 under session-disjoint cross-validation; the close agreement of the two protocols, together with a session-identification control that barely exceeds chance, shows that the diagnosis rests on the alignment condition rather than on session identity.

The core contribution of this study is the decomposition of the three-class error into Normal/Abnormal detection-side errors and 2 mm ↔ 4 mm adjacent-severity confusion, and the demonstration that this decomposition depends strongly on the length of the decision window. For a single 15.6 ms window at 2,000 rpm the two sources are of comparable size, 46 % and 54 % for XGBoost. When the predicted probabilities are averaged over one second under the session-disjoint protocol, the three-class accuracy rises from 0.578 to 0.903 while the converted detection accuracy reaches 0.995, so detection accounts for only 4.8 % of the remaining error. Severity grading within abnormal windows is therefore the bottleneck in the regime in which a monitoring system would actually operate, and the binding constraint on the per-window figures was the observation length rather than the classifier.

The detection–severity two-stage cascade and the single three-class classifier were within 0.9 percentage points of each other at 2,000 rpm and 0.8 points at 1,500 rpm, a difference smaller than the fold-to-fold spread of the session-disjoint protocol, so the choice between them rests on operational grounds. The cascade allows the detection threshold to be tuned independently of severity interpretation, and the stand-alone second-stage accuracy of 0.651 at 2,000 rpm shows that this separation costs little. Finally, the split-gain ranking of the boosted models places the 128–256 Hz band first at 2,000 rpm and the 4.9 ms autocorrelation lag among the three leading families, both of which coincide with the six-times shaft-rotation component identified independently in the spectra, so the learned decision admits a direct physical reading.

The conclusions of this paper are confined to an in-distribution diagnosis problem within a single operating condition, comprising fixed rotational speeds of 1,500 and 2,000 rpm, a no-load condition, and Normal, 2 mm and 4 mm misalignment. Two reservations should be carried forward. The conditions were measured in blocks rather than in randomized order, so that alignment and elapsed time are confounded; the checks of Section 3.3 exclude that confounding for the six-times component but not for broadband and level features, on which the 1,500 rpm results largely rest. And the sign of the change of the six-times component with the imposed offset was found to reverse between two drive-train configurations of the same bench, so it should be used as a deviation from a learned baseline rather than as a signed physical indicator. These out-of-scope directions are summarized in Section 5.9.

Future work should add multi-run measurement data and load-variation conditions to validate the external validity of misalignment severity diagnosis. In addition, cascade-threshold optimization, probability calibration, and the session-level cluster bootstrap should be combined to jointly evaluate missed-detection and false-alarm costs and the reliability of severity grading in real operating environments. Through this, the in-distribution diagnostic performance and error-decomposition results obtained in this study are expected to be extensible to full-scale shipboard operating conditions and long-term condition-monitoring systems. In particular, transferring the present diagnosis to an onboard environment will require an order-tracking-based cutoff logic and a robustness-reinforcement

framework able to separate auxiliary-machinery excitation and wave-induced impact noise from the shafting response, together with a measurement chain that resolves the three directions independently.

Acknowledgement

This research was supported by Korea Institute of Marine Science & Technology Promotion(KIMST) funded by the Ministry of Oceans and Fisheries, Korea(RS-2022-KS221680).

Author Contributions

Conceptualization, J. M. Kim; Methodology, J. M. Kim; Software, J. M. Kim; Formal Analysis, J. M. Kim; Investigation, J. M. Kim; Resources, H. H. Kim and H. K. Oh; Data Curation, J. M. Kim; Writing-Original Draft Preparation, J. M. Kim; Writing-Review & Editing, H. H. Kim, H. K. Oh and T. K. Nam; Visualization, J. M. Kim; Supervision, T. K. Nam; Project Administration, T. K. Nam; Funding Acquisition, H. K. Oh and T. K. Nam.

References

[1] L. Murawski, “The influence of shaft line alignment accuracy on the operational reliability of marine propulsion systems,” Journal of KONES Powertrain and Transport, vol. 23, no. 1, pp. 247-254, 2016.

[2] D. L. Magalhães, D. H. C. de S. S. Martins, B. M. Castro, L. A. Vaz, U. A. Monteiro, and R. H. R. Gutiérrez, “Machine learning performance comparison for main propulsive shafting systems alignment,” Ocean Engineering, vol. 280, 114556, 2023.

[3] G. Vizentin, G. Vukelić, L. Murawski, N. Recho, and J. Orović, “Marine propulsion system failures – A review,” Journal of Marine Science and Engineering, vol. 8, no. 9, 662, 2020.

[4] H. Gharib and G. Kovács, “A review of prognostic and health management (PHM) methods and limitations for marine diesel engines: New research directions,” Machines, vol. 11, no. 7, 695, 2023.

[5] M. G. Pecht and M. Kang, Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things, Hoboken, NJ, USA: Wiley-IEEE Press, 2018.

[6] P. Zhang, Z. Gao, L. Cao, F. Dong, Y. Zou, K. Wang, Y. Zhang, and P. Sun, “Marine systems and equipment prognostics and health management: A systematic review from health condition monitoring to maintenance strategy,” Machines, vol. 10, no. 2, 72, 2022.

[7] T. Berghout, L. H. Mouss, T. Bentrcia, E. Elbouchikhi, and M. Benbouzid, “A deep supervised learning approach for condition-based maintenance of naval propulsion systems,” Ocean Engineering, vol. 221, 108525, 2021.

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” Advances in Neural Information Processing Systems, vol. 30, pp. 3146-3154, 2017.

[9] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[10] H. Yang, J. Mathew, and L. Ma, “Vibration feature extraction techniques for fault diagnosis of rotating machinery: A literature survey,” Proceedings of the Asia-Pacific Vibration Conference, pp. 801-807, 2003.

[11] S. Li, Y. Xin, X. Li, J. Wang, and K. Xu, “A review on the signal processing methods of rotating machinery fault diagnosis,” 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), pp. 1559-1565, 2019.

[12] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, “Time series feature extraction on basis of scalable hypothesis tests (tsfresh – A Python package),” Neurocomputing, vol. 307, pp. 72-77, 2018.

[13] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, “Applications of machine learning to machine fault diagnosis: A review and roadmap,” Mechanical Systems and Signal Processing, vol. 138, 106587, 2020.

[14] J. M. Kim, H. W. Yoon, H. H. Kim, and T. K. Nam, “Development of a land-based test system for simulating shaft-system anomalies in electric-propulsion ships,” Journal of the Korean Society of Marine Environment & Safety, vol. 31, no. 6, pp. 1043-1053, 2025 (in Korean).

[15] American Bureau of Shipping, Guidance Notes on Propulsion Shafting Alignment, Houston, TX, USA: American Bureau of Shipping, 2019.

[16] Y. Kim, “Characteristics of the propulsion shafting alignment for a small class electric propulsion ship equipped

with a reduction gear," Journal of Advanced Marine Engineering and Technology, vol. 48, no. 6, pp. 430-437, 2024.