Journal of Industrial Intelligence
https://www.acadlore.com/journals/JII

# Geometry-Anchored Features and Machine-Level Conformal Calibration for Cross-Machine Bearing Fault Diagnosis

Aswin Karkadakattil*, Arjun Kalath

Independent Researcher, 671313 Kerala, India

* Correspondence: Aswin Karkadakattil (ashwinharik20000@gmail.com)

**Received:** 04-23-2026 **Revised:** 06-08-2026 **Accepted:** 06-23-2026

**Citation:** A. Karkadakattil and A. Kalath, "Geometry-anchored features and machine-level conformal calibration for cross-machine bearing fault diagnosis," *J. Ind Intell.*, vol. 4, no. 2, pp. 173–212, 2026. https://doi.org/10.56578/jii040205.

© 2026 by the author(s). Published by Acadlore Publishing Services Limited, Hong Kong. This article is available for free download and can be reused and cited, provided that the original published version is credited, under the CC BY 4.0 license.

**Abstract:** Bearing fault diagnosis remains challenging when models trained on one set of machines are applied to equipment with different bearing geometries, operating speeds, and vibration characteristics. Conventional evaluation protocols often overestimate cross-machine performance, while prediction uncertainty becomes difficult to calibrate under machine-level distribution shift. This study investigates whether geometry-anchored fault features and machine-level conformal calibration can jointly improve diagnostic accuracy and prediction-set reliability on previously unseen machines. A structured review was conducted to identify methodological gaps, followed by controlled simulations involving eight machines with different bearing geometries, shaft speeds, sampling rates, structural resonances, and noise conditions. Three feature representations and four calibration schemes were evaluated using a leave-one-machine-out protocol. The geometry-anchored representation achieved a mean classification accuracy of 0.908 $\pm$ 0.014, compared with 0.585 $\pm$ 0.016 and 0.575 $\pm$ 0.023 for frequency-domain and order-domain features, respectively. Random window splitting overestimated the accuracy of these baselines by 0.25 and 0.29. At a nominal coverage level of 90%, machine-level conformal calibration achieved mean coverage between 0.903 and 0.931, compared with 0.642–0.879 for conventional split calibration. With geometry-anchored features, the calibrated prediction sets contained 0.99 classes on average, whereas the other representations required 2.77–2.96 classes. Conservative calibration increased worst-machine coverage from 0.751 to 0.866, although individual-machine coverage remained variable. The results indicate that physically indexed fault features and machine-level calibration address complementary aspects of cross-machine diagnosis: the former improves fault discrimination, while the latter supports more reliable marginal coverage under machine shift. The study establishes a controlled evaluation framework for examining diagnostic generalization and uncertainty calibration at the machine level, providing a methodological basis for subsequent validation using experimental and industrial bearing data.

**Keywords:** Bearing fault diagnosis; Cross-machine generalization; Geometry-anchored features; Conformal prediction; Machine-level calibration; Distribution shift; Industrial condition monitoring

## 1 Introduction

Rolling-element bearings are essential components of rotating machinery, and their operating condition has a direct influence on equipment reliability, production continuity, and maintenance requirements. Vibration-based fault diagnosis has become an established application of machine learning in industrial condition monitoring, with numerous studies reporting classification accuracies approaching 100% on publicly available bearing datasets [1, 2]. Despite these results, diagnostic performance measured under benchmark conditions does not necessarily reflect the reliability of a model deployed on unfamiliar equipment. Two closely related issues are particularly important: whether the evaluation procedure adequately represents the differences between machines and whether the uncertainty associated with diagnostic predictions remains reliable when those differences arise.

The first issue concerns the unit of evaluation. Bearing vibration datasets commonly consist of long recordings divided into shorter, sometimes overlapping, signal windows. When windows from the same recording are assigned to both training and testing, a classifier may exploit recording-specific characteristics rather than learning fault patterns that generalize across equipment. Smith and Randall [3] examined the diagnostic properties of the Case Western Reserve University (CWRU) bearing dataset, highlighting the importance of its experimental characteristics

when interpreting classification results. Hendriks et al. [4] further identified inconsistencies in dataset selection, preprocessing, and evaluation procedures that complicate comparisons between diagnostic studies. Separating recordings or operating conditions provides a more demanding evaluation than random window splitting, but it does not fully reproduce deployment on another machine. In industrial applications, the target equipment may have a different bearing geometry, rotational speed, sensor arrangement, and vibration transmission path. Evaluating generalization therefore requires the machine itself to be treated as an independent testing unit.

The second issue concerns the reliability of diagnostic uncertainty. A single predicted fault label does not indicate whether alternative diagnoses remain plausible, particularly when the target machine differs from the equipment represented during training. Conformal prediction provides a framework for constructing prediction sets with finite-sample marginal coverage under exchangeability [5, 6]. However, this assumption is generally not assured when calibration data originate from known machines and predictions are required for an unseen machine with different signal characteristics. Methods developed to address distribution shift, including weighted conformal prediction, offer possible corrections under specified assumptions [7, 8], but their application becomes difficult when source and target feature distributions have limited overlap. These difficulties connect diagnostic generalization with uncertainty calibration. Machine-dependent features may produce both inaccurate classifications and poorly calibrated prediction scores, while increasing prediction-set size to recover empirical coverage can leave several competing fault classes unresolved. Existing reviews have examined diagnostic architectures, transfer-learning methods, and their performance across operating conditions [1, 2, 9], whereas the relationship between the evaluation unit, feature representation, and prediction-set reliability has received less attention. Research combining conformal prediction with cross-domain fault diagnosis has begun to address uncertainty under distribution shift [10], but machine-specific coverage and the diagnostic informativeness of prediction sets remain important questions.

Against this background, the present study addresses four research questions (RQs):

- RQ1. To what extent do conventional data-splitting protocols overestimate diagnostic accuracy on previously unseen machines, and how does feature representation affect cross-machine generalization?
- RQ2. Can machine-level calibration improve the empirical coverage of prediction sets without labelled observations from the target machine, and how does this affect prediction-set size?
- RQ3. How does prediction-set coverage vary across machines and fault classes, and what are the implications for unknown-fault detection and cost-sensitive diagnostic decisions?
- RQ4. Which evaluation and reporting practices are needed to assess diagnostic generalization and prediction reliability under machine-level distribution shift?

To investigate these questions, the study combines a structured narrative review with controlled simulation and analytical evaluation, as illustrated in Figure 1. The literature review in Section 2 examines the methodological foundations of cross-machine diagnosis and identifies five research gaps from which seven hypotheses are formulated. Sections 3 and 4 describe and evaluate a simulation involving eight machines with different bearing geometries, rotational speeds, sampling rates, structural resonances, and noise conditions. The experiments compare three feature representations and four conformal calibration schemes under leave-one-machine-out evaluation, allowing diagnostic performance and uncertainty calibration to be assessed on machines excluded from model development. Two complementary analytical investigations examine the ambiguity of fault frequencies across bearing configurations and the relationship between the number of source machines and machine-level coverage. Section 5 brings together the experimental findings, hypothesis assessments, and reporting recommendations.

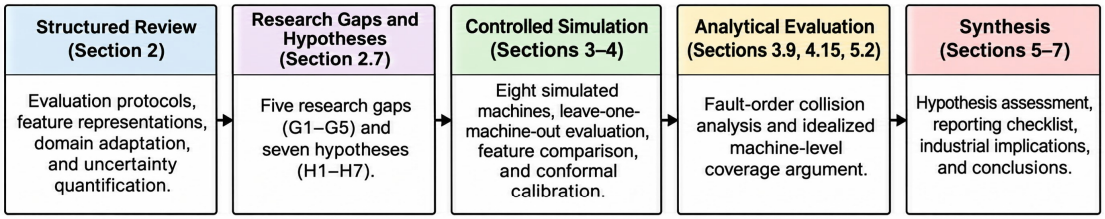



**Figure 1.** Research framework integrating literature review, controlled simulation, analytical evaluation, and synthesis

The study makes three principal contributions.

- Machine-level evaluation and conformal calibration. A leave-one-machine-out evaluation protocol is combined with a leakage audit and a machine-level calibration procedure, termed leave-machine-out conformal prediction (LMO-CP). Calibration scores are obtained through nested machine holdouts, with each calibration machine evaluated using a model trained without its observations. The procedure is examined alongside conventional calibration schemes and conservative, class-conditional, and adaptive-score variants, allowing average coverage, machine-specific coverage, and prediction-set size to be compared within a common experimental framework.

- Geometry-anchored representation for cross-machine diagnosis. A kinematic feature representation is constructed by extracting envelope-spectrum amplitudes at characteristic fault orders determined from each machine's bearing geometry and rotational speed. Unlike features defined at fixed frequencies or fixed spectral orders, the resulting variables retain their association with specific bearing fault mechanisms across machine configurations. The representation is evaluated through feature ablation, geometric and speed perturbations, model-mismatch experiments, classifier substitution, and a structural analysis of fault-order collisions.
- Joint assessment of diagnostic performance and prediction reliability. The study examines classification accuracy together with prediction-set coverage and informativeness, including differences between individual machines and fault classes. Density-ratio weighting, unknown-fault detection, and cost-sensitive abstention are also investigated to establish how uncertainty-aware diagnostic procedures behave under the simulated machine shifts. The findings are consolidated into a twelve-item reporting checklist covering evaluation protocols, calibration, robustness, and operational decision requirements.

The methodological basis of the study draws on established developments in conformal prediction, signal processing, and transfer learning. Split conformal prediction provides the foundation for prediction-set construction [5, 6], while cross-conformal and jackknife+ methods illustrate how held-out observations can contribute to predictive inference [11, 12]. Weighted conformal prediction addresses covariate shift under assumptions concerning the relationship between source and target distributions [7]. In bearing diagnosis, envelope analysis and characteristic fault-frequency calculations provide the physical basis for identifying fault-related vibration components [13, 14], whereas domain-adaptation methods address differences between training and target distributions through learned representations or distribution alignment [15–17]. The present study brings these established elements together by aligning the feature representation and calibration procedure with the machine-level evaluation setting. Rather than treating windows as independent units of transfer, it examines the consequences of excluding entire machines from model training and calibration, while reporting prediction-set size and machine-specific coverage alongside diagnostic accuracy.

Several related studies provide useful points of comparison. Cross-domain bearing diagnosis with conformal prediction has been investigated using labelled target-domain calibration data [10], whereas the present evaluation excludes target-machine labels from calibration. Physics-informed diagnostic approaches have incorporated bearing geometry, rotational speed, and mechanical relationships into signal representations, training data, or learning constraints [18–20]. These studies establish the relevance of physical information to fault classification, but the evaluation considered here focuses specifically on the interaction between geometry-indexed features and prediction-set calibration when complete machines are held out. Recent domain-adaptation approaches have examined feature alignment across devices and operating conditions [21, 22], generally using target-domain observations during adaptation and emphasizing classification performance. Group-conditional conformal prediction provides another relevant statistical perspective by establishing coverage under specified group structures or classes of distribution shifts [23]. Its assumptions differ from those encountered when only a small number of source machines are available and the target machine contributes no labelled calibration observations. Within the literature examined in Section 2, the combined evaluation of geometry-anchored features, leave-machine-out calibration, machine-specific coverage, and prediction-set size has not been addressed as a unified diagnostic problem.

The present investigation is based on controlled simulation, which permits machine characteristics and fault mechanisms to be varied while retaining known ground truth. This design supports a systematic examination of the relationship between feature representation, machine shift, and conformal calibration. The resulting findings establish an experimental and analytical basis for assessing cross-machine diagnostic reliability and for designing subsequent evaluations using independently collected bearing vibration measurements.

## 2 Structured Review of the Literature

### 2.1 Scope and Approach

The literature review was organized around the methodological requirements of bearing fault diagnosis on previously unseen machines. Rather than surveying diagnostic algorithms solely according to model architecture or reported classification accuracy, the review examined how existing studies defined the evaluation unit, represented fault-related vibration information, addressed differences between machines, and assessed prediction uncertainty. These considerations are closely connected in industrial applications, where the bearing configuration and operating environment of a target machine may differ from those encountered during model development.

A structured narrative synthesis was adopted to bring together three categories of research: foundational studies on bearing vibration analysis, conformal prediction, and domain adaptation; established reviews and methodological assessments of bearing diagnosis benchmarks; and recent investigations of cross-domain fault diagnosis and uncertainty quantification. The review was not conducted as a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-style systematic review and did not involve a formal database screening procedure. Among the selected sources, a recent preprint examining conformal prediction under cross-dataset bearing fault

diagnosis was included because of its direct relevance to the calibration problem considered here [10]. Each research thread was examined in terms of its methodological assumptions, evaluation protocol, treatment of distribution shift, and implications for diagnosis on an unseen machine. The resulting synthesis was used to identify the gaps summarized in Section 2.7. Particular attention was also given to recent studies and preprints from 2024 to 2026 concerning cross-domain transfer, physics-informed diagnostic representations, and group-aware uncertainty calibration [20–24].

### 2.2 Evaluation Practice and the Leakage Problem

The development of learning-based bearing diagnosis has been accompanied by increasingly high classification accuracies on established experimental datasets. Reviews of machine fault diagnosis have documented the extensive use of deep learning and the near-perfect performance frequently reported on benchmarks such as the CWRU dataset [1, 2]. These results must, however, be interpreted in relation to the way the experimental data are selected and divided. Smith and Randall demonstrated that individual CWRU recordings differ substantially in their diagnostic characteristics, making the choice of recordings an important factor in benchmark evaluation [3]. Hendriks et al. [4] subsequently identified considerable variation in data selection, preprocessing, and train–test splitting across studies using the same dataset, and proposed guidelines for more consistent benchmarking.

A particularly important distinction concerns whether the observations assigned to training and testing originate from independent recordings or machines. Vibration recordings are commonly divided into shorter windows, and adjacent windows may share fault signatures, background noise, operating conditions, and recording-specific characteristics. When windows from the same recording occur in both datasets, the evaluation partly measures the recognition of previously encountered signal characteristics. Separating complete recordings reduces this dependence, but the resulting test still concerns generalization within an already represented machine. A further distinction arises when operating conditions, rather than recordings, are held out. Public bearing datasets, including Paderborn [25], XJTU-SY [26], and CWRU [3, 4], have been widely used in fault diagnosis research. Transfer-learning studies have also examined diagnostic performance across changes in operating conditions and data distributions [9, 27]. Nevertheless, transferring between conditions of the same test rig differs from diagnosing faults on equipment with a different bearing geometry, sensor arrangement, and structural transmission path.

These distinctions establish a hierarchy of evaluation protocols, summarized in Table 1. Random window splitting primarily examines recognition within the available recordings, whereas recording-level separation assesses generalization to new observations from an existing machine. Condition-level holdout introduces changes in operating conditions, and machine-level holdout evaluates the ability to diagnose faults on equipment excluded from model development. Cross-dataset evaluation introduces additional differences between laboratories, acquisition systems, and experimental configurations. A recent cross-dataset investigation using convolutional neural networks reported classification accuracies of approximately 24%–36% when transferring between CWRU and Southeast University (SEU) bearing datasets [10]. Such findings illustrate the substantial dependence of reported diagnostic performance on the evaluation setting. They also indicate why machine-level separation is necessary when the intended application concerns previously unseen equipment.

**Table 1.** Evaluation protocols for bearing fault diagnosis

| Protocol | Held-Out Unit | Evaluation Objective | Typical Application |
|---|---|---|---|
| P1 Random window split | No recording-level separation | Recognition within available recordings | Common benchmark evaluation [2, 3] |
| P2 Recording held out | Complete recordings from the same machine | Generalization across recordings | Within-machine diagnosis |
| P3 Condition held out | Speed or load condition of an existing test rig | Transfer across operating conditions | Transfer-learning studies [9, 27] |
| P4 Machine held out | Complete machine, including its bearing and vibration path | Generalization to unseen equipment | Cross-machine diagnosis examined here |
| P5 Dataset held out | Independent dataset or laboratory | Transfer across acquisition and experimental settings | Cross-dataset evaluation [10] |

### 2.3 Feature Representations for Cross-Machine Diagnosis

The transferability of a diagnostic model depends not only on the learning algorithm but also on whether its input features retain consistent physical meaning across machines. Classical bearing diagnosis identifies fault-related components in the envelope spectrum, often after selecting a suitable resonance band using spectral kurtosis.

Characteristic fault frequencies are calculated from the bearing geometry and rotational speed, allowing the measured spectral response to be interpreted in relation to specific mechanical fault mechanisms [13, 14]. This approach establishes a direct connection between the feature location and the physical source of the vibration.

Many learning-based diagnostic methods instead use raw vibration signals, conventional spectra, or time-frequency representations as model inputs [1, 2, 28]. Such representations can retain substantial information about fault-related patterns, but the classifier must learn how those patterns correspond to fault categories under different operating and machine configurations. Other studies have retained more of the physical structure during feature construction or model development. For example, cyclic spectral coherence has been combined with convolutional neural networks for bearing fault classification [18], while dynamic bearing models have been used to generate training data for operating conditions in which measured fault samples are unavailable [19]. These approaches demonstrate alternative ways of incorporating mechanical information into data-driven diagnosis. Their evaluation settings, however, differ from the machine-level holdout and uncertainty-calibration framework considered in this study.

Order-domain representations provide an intermediate approach by expressing spectral frequency relative to shaft rotational frequency. This transformation reduces the dependence of fault signatures on rotational speed, but characteristic bearing fault orders remain functions of bearing geometry. Differences in rolling-element number, diameter ratio, and contact angle can therefore cause the same order to correspond to different faults on different machines. Conversely, the same fault mechanism may appear at different orders when bearing geometry changes. These relationships are examined through the structural fault-order analysis in Section 4.15.

Recent physics-informed learning approaches have sought to incorporate these mechanical relationships more directly. A multimodal convolutional framework, for example, introduced constraints associated with ball-pass frequencies calculated from bearing geometry and rotational speed, and examined diagnostic transfer across operating conditions and datasets [20]. This approach uses fault-frequency relationships to constrain a learned model, whereas the geometry-anchored representation investigated here extracts spectral information directly at the calculated fault orders. The distinction concerns how physical knowledge enters the diagnostic process. The related studies discussed above [18, 19] provide a basis for physically informed diagnosis, while the present work examines whether explicitly indexed fault features retain their diagnostic meaning when complete machines are excluded from training.

### 2.4 Domain Adaptation and Cross-Machine Generalization

Differences between machines are commonly treated as a domain-shift problem. Domain-adaptation methods attempt to reduce discrepancies between source and target distributions through approaches such as adversarial training [15], correlation alignment [16], and adaptive batch normalization [17]. Their application to machinery fault diagnosis has been extensively reviewed and evaluated [9, 27]. Transfer learning from pretrained image models, using time-frequency representations of vibration signals, provides another means of adapting learned features to diagnostic tasks [28].

The effectiveness of these approaches depends on the transfer setting. In many bearing diagnosis studies, the source and target domains correspond to different operating conditions of the same experimental rig. Such evaluations are useful for examining changes in speed or load, but they do not necessarily capture differences in bearing geometry, structural response, and measurement configuration between independent machines. Moreover, classification accuracy is commonly the principal evaluation metric, leaving the reliability of predictive uncertainty under transfer less clearly characterized [9, 27].

The theoretical difficulty of cross-domain learning is reflected in the generalization analysis of Ben-David et al. [29], which relates target-domain error to source-domain error, distributional discrepancy, and the error of a hypothesis suitable for both domains. When source and target machines produce substantially different feature distributions, the discrepancy term becomes important, even if their fault categories are nominally identical. Domain generalization addresses the related problem of learning representations without using target-domain observations. Existing surveys have organized the available approaches [30], while comparative work under controlled model-selection procedures has shown that conventional empirical risk minimization can remain competitive with specialized domain-generalization algorithms [31]. Although these findings originate from broader machine-learning settings, they underline the importance of separating the contribution of a learning algorithm from the influence of the evaluation protocol.

For bearing diagnosis, an additional difficulty arises when the available source machines provide only limited variation in bearing geometry. Under such conditions, a model trained on fixed-frequency or fixed-order features must infer the relationship between geometry and fault signatures from a small number of machine configurations. The experiments in Section 4.2 examine this issue by comparing geometry-anchored features with order-domain features supplemented by geometric parameters. Recent studies have continued to investigate distribution alignment for cross-domain bearing diagnosis, including multi-adversarial adaptation across motor-bearing devices [21] and methods based on improved feature-distribution alignment [22]. These approaches generally rely on target-domain

observations during adaptation and assess their performance primarily through diagnostic classification. The setting considered here differs in one important respect: the target machine contributes no labelled data to model training or calibration, and the central question is whether reliable diagnostic predictions can be obtained using information from the available source machines.

### 2.5 Prediction Uncertainty and Conformal Calibration

Accurate fault classification does not necessarily imply reliable predictive confidence. Neural-network classifiers can produce poorly calibrated probabilities [32], and empirical investigations have shown that predictive uncertainty may deteriorate as the difference between training and testing distributions increases [33]. These observations are particularly relevant to cross-machine diagnosis, where changes in geometry and vibration transmission can alter both the feature distribution and the relationship between predicted probabilities and diagnostic correctness.

Conformal prediction provides a model-independent approach to constructing prediction sets with marginal coverage under exchangeability [5, 6]. The informativeness of these sets depends on the choice of nonconformity score and the underlying classifier. Methods for constructing less ambiguous prediction sets [34] and adaptive prediction sets [35] have been developed, while Mondrian conformal procedures use separate calibration thresholds for different classes [36]. These methods address aspects of prediction-set construction, but their conventional coverage guarantees depend on the statistical relationship between calibration and test observations.

Several extensions of conformal prediction are relevant to distribution shift. Weighted conformal prediction accounts for covariate shift through density-ratio information [7], while subsequent work has examined coverage beyond exchangeability [8], label-distribution shift [37], and adaptive inference under changing data distributions [38]. A related line of research uses held-out observations to construct predictive inference procedures. Cross-conformal prediction [11] and jackknife+ methods [12] provide relevant methodological foundations, although their theoretical properties and prediction-set constructions differ. In particular, the general jackknife+ result provides a weaker coverage bound than conventional split conformal prediction, with a bound involving twice the nominal error level [12]. These distinctions are important because excluding complete machines from calibration-model training does not automatically establish the finite-sample guarantees associated with established conformal procedures.

The interpretation of coverage also requires care. The distinction between marginal and conditional coverage is fundamental to distribution-free predictive inference [39]. In cross-machine fault diagnosis, this distinction is particularly relevant because average coverage across machines does not ensure comparable coverage on each individual machine. In cross-machine diagnosis, this distinction separates coverage averaged across machines from coverage obtained on a particular machine. A procedure may achieve the nominal level when averaged over the evaluated machine population while producing substantially lower coverage for equipment with unusual noise, geometry, or vibration characteristics. Conformal risk control extends related ideas toward controlling expected monotone losses [40], providing a statistical perspective for diagnostic decisions in which different classification errors carry different operational costs.

Applications of conformal prediction to machinery diagnosis remain relatively limited. In cross-dataset bearing classification, one study reported empirical coverage near the nominal level even when the underlying classifier performed poorly, but the resulting prediction sets contained more than three of the four available classes at a nominal error level of 0.1 [10]. This illustrates the distinction between statistical coverage and diagnostic informativeness. The reported procedure used labelled target-domain observations for calibration, whereas a newly encountered machine may have no such data available. Conformal anomaly detection has also been investigated for industrial fleet monitoring, including district-heating systems [41], demonstrating the relevance of uncertainty-aware monitoring at the equipment level.

Further developments have examined conditional coverage for specified groups or distribution shifts. Group-aware conformal methods can establish guarantees over predefined collections of groups or structured classes of covariate shifts [23], provided the corresponding assumptions and calibration requirements are satisfied. Other work has proposed diagnostics for identifying vulnerability to distribution shift before model deployment, including investigations involving gradient-boosted classifiers [24]. These studies address important aspects of uncertainty under changing conditions, but their assumptions differ from a setting involving only a small number of source machines and an entirely unseen target machine. In the present study, the calibration units are complete machines, and the available source population contains at most seven machines in each leave-one-machine-out evaluation. This motivates examining machine-level calibration empirically, with particular attention to average coverage, machine-specific coverage, and prediction-set size.

### 2.6 Unknown-Fault Detection and Cost-Sensitive Decisions

Cross-machine diagnostic reliability also depends on how a classifier responds to faults outside the categories represented during training. Maximum softmax probability provides a conventional baseline for identifying misclassified or out-of-distribution observations [42], while class-conditional Mahalanobis distance offers an

alternative based on the distribution of learned features [43]. Selective classification addresses a related decision problem by allowing a model to abstain from prediction and evaluating the relationship between classification accuracy and the proportion of observations receiving a decision [44].

These approaches become more difficult to interpret when an unfamiliar machine introduces distribution shift independently of fault novelty. A low-confidence prediction may arise because the machine differs from those used in training, because the observed fault is unknown, or because the measured signal has poor diagnostic quality. Conversely, an unfamiliar fault may still receive a high-confidence label from one of the known categories. Distinguishing these situations is important in industrial maintenance, where missed faults, unnecessary inspections, and incorrect fault identification have different consequences.

A diagnostic system should therefore be evaluated not only by its classification and detection metrics but also by the operational costs associated with its decisions. Abstention may be preferable when the expected cost of an incorrect diagnosis exceeds the cost of inspection, although this relationship depends on fault prevalence, probability calibration, and the assumed cost matrix. The combined assessment of unknown-fault recognition, selective prediction, and cost-sensitive decisions under machine-level distribution shift provides an additional perspective on the practical reliability of cross-machine diagnostic models.

### 2.7 Research Gaps and Hypotheses

The preceding review identifies a common methodological issue across the literature: diagnostic accuracy, feature transferability, and prediction uncertainty are often evaluated separately, even though they interact when a model is deployed on previously unseen equipment. The evaluation protocol determines which forms of machine variation are represented in testing; the feature representation determines whether fault information retains consistent meaning across machines; and the calibration procedure determines how prediction uncertainty is translated into diagnostic decisions.

**Table 2.** Literature synthesis and identified research gaps

| Research Direction | Existing Methodological Focus | Uncertainty Assessment | Research Gap |
|---|---|---|---|
| Evaluation protocols [3, 4] | Classification performance under benchmark data partitions | Usually absent or implicit | G1: Evaluation on known recordings does not establish generalization to unseen machines |
| Feature representations [13, 14, 18–20] | Spectral, learned, and physically informed fault representations | Generally not central to evaluation | G2: Physical fault-frequency relationships are not consistently incorporated into cross-machine representations |
| Domain adaptation and generalization [9, 15–17, 21, 22, 27, 31] | Transfer across operating conditions or domains | Commonly focused on classification performance | G2, G3: Target-domain information may be required, and machine-level uncertainty remains insufficiently characterized |
| Conformal calibration [5, 7, 10–12, 23, 24, 39] | Prediction-set coverage under specified statistical assumptions | Primarily marginal or structured conditional coverage | G3, G4: Calibration on known machines does not directly characterize coverage and set size on unseen machines |
| Unknown-fault detection and decisions [40, 42–44] | Novelty detection, selective prediction, and risk–coverage relationships | Detection scores and risk–coverage relationships | G5: Unknown faults and operational costs require joint evaluation under machine shift |

Five research gaps follow from this synthesis. G1 concerns the distinction between evaluation on previously encountered machines and evaluation on genuinely unseen equipment. G2 concerns the loss of machine-independent physical meaning when fault signatures are represented on fixed frequency or order axes. G3 concerns the mismatch between calibration on known machines and prediction under machine-level distribution shift. G4 concerns the need

to examine marginal coverage, machine-specific coverage, and prediction-set size together. G5 concerns the limited integration of unknown-fault recognition and operational decision costs into cross-machine diagnostic evaluation. Table 2 summarizes the relationship between the reviewed research directions and these gaps.

These gaps provide the basis for seven research hypotheses, summarized in Table 3. The hypotheses were formulated as part of the structured interpretation of the literature and the completed simulation experiments, rather than being registered before the experiments. They are therefore used to organize the evidence and examine the relationships between evaluation protocol, feature representation, calibration, and diagnostic decisions. Their assessment in Section 5.1 follows this exploratory interpretation.

**Table 3.** Research hypotheses and corresponding experimental analyses

| ID | Research Hypothesis | Gap and Sources | Evaluated in |
|---|---|---|---|
| H1 | Random-window and recording-level splits overestimate diagnostic performance on unseen machines when the feature representation retains machine-dependent characteristics. | G1; [3, 4, 10] | Section 4.1 |
| H2 | Geometry-anchored features retain fault-specific physical meaning across machines and provide better diagnostic generalization than fixed-frequency or fixed-order representations. With limited source machines, appending geometric parameters to conventional features does not provide an equivalent representation. | G2; [13, 19, 29] | Sections 4.2, 4.9, 4.10, 4.14, 4.15 |
| H3 | Calibration using observations from previously encountered machines produces insufficient coverage under machine shift, whereas calibration based on nested machine holdouts improves average coverage on unseen machines. | G3; [11, 12, 33] | Sections 4.3, 4.5, 4.9 |
| H4 | Prediction-set coverage alone does not establish diagnostic informativeness. Small and informative prediction sets depend on feature representations that preserve fault discrimination across machines. | G4; [10, 34, 35] | Sections 4.3, 4.7, 4.12 |
| H5 | Marginal coverage averaged across machines does not ensure nominal coverage for each machine. Conservative machine-level thresholds improve coverage on poorly represented machines at the cost of larger prediction sets. | G4; [36, 39] | Sections 4.4, 4.12, 5.2 |
| H6 | Under machine-level distribution shift and with a limited number of source machines, density-ratio weighting, conventional open-set detection, and cost-sensitive abstention do not consistently improve diagnostic reliability. | G5; [7, 40, 42–44] | Sections 4.3, 4.8, 4.11 |
| H7 | The relative performance of geometry-anchored features and machine-level calibration remains consistent across the examined model variations and moderate departures from nominal kinematic assumptions, although individual machines may experience reduced reliability. | G2–G4; [12, 39] | Sections 4.6, 4.13, 4.14 |

Note: ID = identifier; H1–H7 denote the seven research hypotheses.

Taken together, these hypotheses establish a common basis for examining the interaction between evaluation protocol, feature representation, and uncertainty calibration. H1 and H2 address the sources of apparent diagnostic generalization and the role of physically indexed fault information. H3–H5 concern the relationship between machine-level calibration, prediction-set informativeness, and coverage variation across machines. H6 extends the evaluation to unknown faults and operational decisions, while H7 examines the stability of the principal findings under changes in signal characteristics and classification models. The controlled simulation described in Section 3 was designed to provide comparable evidence across these aspects of cross-machine diagnosis, with the corresponding experimental

results presented in Section 4.

## 3 Hybrid Study Design and Methods

The study employed a controlled simulation framework to investigate the relationship between feature representation, machine-level distribution shift, and uncertainty calibration in cross-machine bearing fault diagnosis. Public bearing datasets provide valuable experimental evidence, but relatively few contain multiple independent machines with fully documented bearing geometries. Moreover, differences in fault characteristics, sensor configurations, and structural transmission paths often occur simultaneously, making their individual effects difficult to distinguish. Controlled simulation allowed these factors to be examined under known ground-truth conditions while maintaining consistent evaluation procedures across diagnostic methods.

The methodological framework comprised leave-one-machine-out evaluation, signal processing and feature construction, conformal calibration, robustness and decision analyses, and analytical examination of fault-frequency ambiguity. Sections 3.1–3.8 describe the simulation, feature extraction, calibration, and evaluation procedures, while Section 3.9 presents the deterministic analysis of fault-frequency ambiguity across the simulated machines. The findings are synthesized in Section 5, followed by a discussion of their implications for industrial condition monitoring in Section 6.

### 3.1 Problem Setting

The diagnostic problem involved $K$ labelled source machines and one target machine excluded from model training and calibration. Each source machine provided vibration windows associated with four known conditions: normal operation, inner-race fault, outer-race fault, and ball fault. Mechanical looseness was introduced as a fifth condition that was absent from training and served as an unknown fault during evaluation.

For each vibration window $x$, the diagnostic model produced a prediction set $C(x)$ containing the fault classes considered plausible at a prescribed miscoverage level $\alpha$. The nominal objective was to obtain a coverage probability of at least $1-\alpha$ for observations from a newly encountered machine. Such a coverage statement concerns the joint distribution of machines and their observations and requires suitable exchangeability assumptions at the machine level. It does not imply that every individual machine will achieve the same coverage.

The evaluation therefore distinguished between average coverage and machine-specific coverage. Average coverage was calculated across the held-out machines and random seeds, whereas machine-specific coverage represented the proportion of correctly covered observations on a particular machine. The machine-level calibration procedure investigated in this study was assessed empirically, and its performance was interpreted in relation to these two measures.

### 3.2 Controlled Bearing Signal Simulation

A controlled vibration simulator was developed to generate bearing signals with known fault mechanisms and machine-dependent operating characteristics. Each signal window contained 16,384 samples. Fault-related vibration was represented by a sequence of excitation impulses convolved with a two-mode structural impulse response. White noise and randomly occurring non-periodic interference impulses were added to reproduce variation in the measured vibration response.

The characteristic fault orders were calculated from conventional bearing kinematics. Let $n$ denote the number of rolling elements, $d$ the rolling-element diameter, $D$ the bearing pitch diameter, and $\phi$ the contact angle. Defining $\xi = (d/D)\cos\varphi$, the characteristic bearing fault frequencies include the ball pass frequency of the outer race (BPFO), the ball pass frequency of the inner race (BPFI), the ball spin frequency (BSF), and the fundamental train frequency (FTF). Their corresponding characteristic orders were calculated from conventional bearing kinematics as follows:

Outer-race fault:

$$\mathrm{BPFO} = \frac{n}{2}(1-\xi)$$

Inner-race fault:

$$\mathrm{BPFI} = \frac{n}{2}(1+\xi)$$

Ball fault:

$$2\mathrm{BSF} = \frac{D}{d}\left(1-\xi^2\right)$$

Fundamental train order:

$$\mathrm{FTF} = \frac{1}{2}(1 - \xi)$$

All characteristic orders were expressed relative to shaft rotational frequency. Inner-race fault signals were amplitude-modulated at shaft frequency, whereas ball-fault signals were modulated at the cage frequency. Mechanical looseness, which served as the unknown fault condition, was represented by excitation components at 0.5 and 1 shaft orders.

Timing jitter of 1% was introduced to approximate rolling-element slip. Fault severity was sampled from a log-uniform distribution at the recording level, while rotational speed, noise amplitude, and interference characteristics varied between recordings. Each machine had its own bearing geometry, shaft speed, sampling rate, structural resonance, noise level, signal gain, and interference rate. Shaft frequencies ranged from 20 to 45 Hz, sampling frequencies from 12 to 25.6 kHz, and structural resonances from 2.5 to 7 kHz. The individual machine parameters are provided in Appendix A.

Eight machines, designated M1–M8, were simulated under five conditions: normal operation, inner-race fault, outer-race fault, ball fault, and mechanical looseness. Four recordings were generated for each machine–condition combination, with twelve non-overlapping windows per recording. This produced 48 windows per machine and condition, 240 windows per machine, and 1,920 windows per random seed. Windows belonging to the same recording shared their speed, noise characteristics, interference amplitude, and fault severity.

The simulated configurations also included overlapping characteristic fault orders between different machines. For example, the inner-race fault order of M5 was close to the outer-race fault order of M6. Such overlap introduces ambiguity when fault signatures are represented on a common order axis without accounting for bearing geometry. Figure 2 illustrates the characteristic fault orders of the eight machines, while Table 4 summarizes the simulation design.

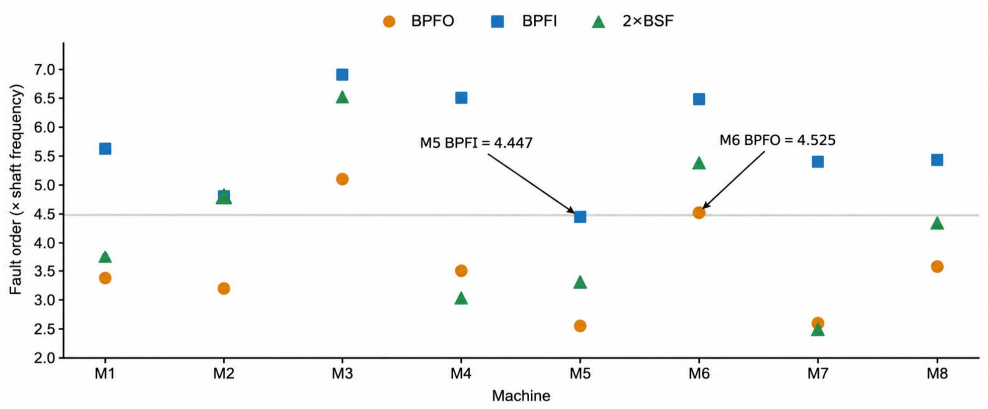



**Figure 2.** Characteristic fault orders of the eight simulated machines, showing cross-machine fault-order overlap

Note: BPFO = ball pass frequency of outer race; BPFI = ball pass frequency of inner race; BSF = ball spin frequency.

**Table 4.** Configuration of the simulated bearing dataset

| Item | Value |
|---|---|
| Machines | 8 (M1–M8; parameters in Table A1) |
| Conditions per machine | 5: normal, inner-race fault, outer-race fault, ball fault, and looseness |
| Recordings per machine and condition | 4 |
| Windows per recording | 12 |
| Windows per machine and condition | 48 |
| Windows per machine | 240 (192 known-class and 48 unknown-class windows) |
| Windows per random seed | 1,920 |
| Window length | 16,384 samples |
| Window overlap | None; windows were non-overlapping and shared recording-level parameters |
| Random seeds | 5 for the reference experiment; 2 for the source-machine-number sweep and additional noise levels; 3 for geometry and speed perturbations, stress testing, and classifier comparisons |

### 3.3 Signal Processing and Feature Representations

All feature representations were extracted from a common signal-processing pipeline to ensure that differences in diagnostic performance could be attributed primarily to feature construction rather than to separate preprocessing procedures.

For each vibration window, the resonance band was selected from fourteen candidate bands according to maximum envelope kurtosis. The Hilbert transform was then used to obtain the signal envelope, from which the envelope spectrum was calculated. Spectral amplitudes were normalized by the median spectral level between 3 and 800 Hz, yielding features expressed relative to the background spectral floor.

Three representations were constructed.

Frequency-domain representation. The envelope spectrum between 0 and 500 Hz was divided into fifty frequency intervals of 10 Hz. The maximum amplitude within each interval was retained and logarithmically transformed. This representation preserved spectral information at fixed physical frequencies without explicitly incorporating shaft speed or bearing geometry.

Order-domain representation. Spectral frequencies were normalized by the estimated shaft rotational frequency. Features were extracted between 0.3 and 10 shaft orders at intervals of 0.1, producing 98 features. The shaft-frequency estimate included a random error of 0.3%. This representation accounted for differences in rotational speed while retaining a common order axis across machines.

Geometry-anchored kinematic representation. Twenty features were calculated using characteristic fault orders derived from each machine's bearing geometry and rotational speed. Spectral amplitudes were extracted within a $\pm 3\%$ tolerance band around the calculated fault orders. Twelve features represented the first three harmonics of FTF, twice the ball-spin frequency, BPFO, and BPFI. Four additional features described shaft-order and cage-related sidebands around BPFI and twice the ball-spin frequency. Three features represented amplitudes at 0.5, 1, and 2 shaft orders, and the remaining feature was envelope kurtosis.

The central distinction between these representations concerns the physical meaning of their feature coordinates. Frequency-domain features were indexed by fixed frequencies, and order-domain features by normalized frequencies. In contrast, geometry-anchored features were indexed by characteristic fault mechanisms, allowing the same feature position to refer to the same physical quantity across different bearing configurations.

The kinematic representation required bearing geometry and shaft speed to be available for each machine. These requirements were retained throughout the evaluation, and their influence was examined through controlled geometry and speed perturbations in Section 4.6. Their practical implications for industrial condition monitoring are discussed in Section 6.3.

### 3.4 Classification Models and Baselines

The principal diagnostic classifier was a histogram-based gradient-boosting model implemented using scikit-learn [45]. The model used 60 boosting iterations, a maximum tree depth of 3, a learning rate of 0.1, and an L2 regularization parameter of 1. The same configuration was applied to the three feature representations without representation-specific hyperparameter optimization.

Several additional configurations were evaluated to distinguish the influence of feature representation from that of the classifier or adaptation procedure.

First, per-machine feature standardization was applied using unlabelled target-machine statistics, following an adaptive-normalization strategy related to Adaptive Batch Normalization (AdaBN) [17]. This configuration examined whether adjusting feature distributions using target-machine observations could improve cross-machine diagnosis.

Second, a multilayer perceptron with one hidden layer containing 64 units was evaluated on the same feature representations. This provided a neural-network comparison with the principal gradient-boosting classifier.

Third, the order-domain representation was supplemented with a vector of bearing geometry parameters. This configuration, denoted order+geo, examined whether a classifier trained on a limited number of machines could learn the relationship between geometric parameters and characteristic fault orders without explicitly constructing geometry-anchored features.

The comparisons were designed to isolate the effects of feature construction, target-machine normalization, and classifier selection while maintaining consistent machine-level evaluation conditions.

### 3.5 Conformal Prediction and Machine-Level Calibration

Conformal prediction was used to construct sets of plausible fault classes from the probability estimates of the diagnostic classifier. For a vibration window $x$ and candidate fault class $y$, the nonconformity score was defined as

$$s(x, y) = 1 - p(y \mid x)$$

where, $p(y \mid x)$ is the probability assigned to class $y$.

For $n$ calibration scores $s_1, \ldots, s_n$, the conformal threshold was calculated as

$$\hat{q}_\alpha = s_{([(n+1)(1-\alpha)])}$$

where, $s_{(k)}$ denotes the $k$-th smallest calibration score. If $k > n$, the threshold is defined as $+\infty$.

The prediction set contained every candidate class whose nonconformity score did not exceed the resulting calibration threshold. Four calibration procedures were compared.

Naive split calibration. The pooled windows from the source machines were randomly divided into training and calibration subsets using a 60:40 ratio. Because the split was performed at the window level, observations from the same recording could occur in both subsets.

Recording-held-out calibration. Complete recordings were assigned to either training or calibration, ensuring that windows from the same recording were not shared between the two subsets. However, both subsets still contained observations from the same source machines.

Weighted calibration. Recording-held-out calibration was combined with density-ratio weighting based on a domain classifier [7]. Logistic regression was used to distinguish source calibration windows from unlabelled target-machine windows. The resulting domain probabilities were clipped to the interval 0.01–0.99 before being converted into calibration weights.

LMO-CP. Each source machine was excluded in turn from classifier training. For source machine $j$, a classifier was fitted using the remaining source machines, and the resulting model was used to calculate nonconformity scores for the windows of machine $j$. Scores from all source-machine holdouts were pooled to determine the calibration threshold. The final classifier was then trained using all available source machines and applied to the unseen target machine.

The distinction between LMO-CP and conventional window-level calibration is the unit excluded during score generation. By removing an entire machine from each nested training subset, LMO-CP produces calibration scores under a machine-shift setting closer to that encountered during deployment. The procedure is related in motivation to cross-conformal prediction [11] and jackknife+ inference [12], although its construction and theoretical properties are not identical to those established methods.

Because the calibration scores were obtained from nested models trained on $K - 1$ source machines, while the deployed classifier was trained on all $K$ source machines, a separate finite-sample coverage theorem was not established for LMO-CP. Its performance was therefore assessed empirically through average coverage, machine-specific coverage, and prediction-set size. The experiments in Sections 4.3 and 4.4 examined whether the machine-level calibration scores provided an appropriate approximation to prediction errors on unseen machines.

### 3.6 Unknown-Fault Detection and Decision Analysis

The evaluation was extended beyond closed-set fault classification to examine unknown-fault recognition and the consequences of diagnostic decisions. Mechanical looseness served as the unknown condition because it was excluded from the four-class training label space.

For conformal prediction, a window was flagged as potentially unknown when its prediction set was empty. This rule was evaluated together with two conventional novelty scores: maximum predicted class probability and minimum class-conditional Mahalanobis distance. The Mahalanobis calculation used a shared covariance matrix estimated using the Ledoit–Wolf procedure, with and without per-machine feature standardization. Detection performance was assessed using the area under the receiver operating characteristic curve (AUROC).

A cost-sensitive decision analysis was also conducted to examine the relationship between diagnostic predictions and maintenance actions. The available actions consisted of assigning one of the four known fault labels or abstaining from classification and referring the observation for inspection.

The decision costs were specified as follows: correct classification, 0; classifying a fault as normal, 10; falsely identifying a fault in a normal window, 1; assigning an incorrect fault category, 2; and abstaining for inspection, 0.3.

Four decision policies were compared: direct classification using the highest predicted probability, a Bayes decision rule minimizing expected cost under the estimated class probabilities, and two singleton-set decision rules using LMO-CP and naive conformal thresholds, respectively. Under the singleton-set policies, a class label was assigned only when the prediction set contained exactly one class; otherwise, the system abstained.

A policy that always referred observations for inspection was included as a reference, with a constant cost of 0.3 per window. This comparison allowed the decision analysis to distinguish improvements in classification performance from reductions in expected operational cost under the specified cost assumptions.

### 3.7 Machine-Level Evaluation Protocol

Generalization was evaluated using leave-one-machine-out testing. Each of the eight simulated machines was excluded from model development in turn and used as the target machine, while the remaining seven machines provided the source observations.

Classification accuracy, macro-F1, expected calibration error, and conformal coverage were calculated using the known-class windows of the held-out machine. Unknown-fault windows were included in the open-set and relevant decision analyses. The principal results were summarized as the mean and standard deviation, across five random seeds, of the average performance over the eight held-out machines.

To characterize differences between machines, the worst-machine coverage was calculated as the lowest coverage among the eight machines for each seed and then averaged across seeds. The proportion of machine–seed cases achieving coverage within 0.05 below the nominal target was also reported. For a nominal miscoverage level $\alpha$, the corresponding threshold was the nominal coverage minus 0.05.

With eight machines and five seeds, the principal experiment contained forty machine–seed evaluation cases.

A separate leakage audit compared three evaluation protocols: random 80:20 window splitting, 80:20 recording-level splitting, and leave-one-machine-out testing. This comparison examined how the unit of data separation influenced apparent diagnostic performance.

Table 5 summarizes the assignment of observations to training, calibration, and testing under the four conformal procedures. Under naive calibration, windows from the same recording could be shared between training and calibration. Recording-level separation removed this overlap, while LMO-CP excluded the complete calibration machine from the corresponding nested model. The target machine was excluded from supervised training and calibration in all procedures. Unlabelled target observations were used only in the weighted calibration and per-machine standardization configurations.

**Table 5.** Training, calibration, and testing arrangements

| Calibration Scheme | Training Observations | Calibration Observations | Target Evaluation | Recording Overlap |
|---|---|---|---|---|
| Naive split | Random 60% of pooled source-machine windows | Remaining 40% of pooled source-machine windows | Known-class windows of the held-out machine | Possible |
| Recording held out | Windows from selected source-machine recordings | Windows from separate recordings of the same source machines | Same held-out target machine | No |
| Weighted | Same as recording-held-out calibration | Recording-held-out scores weighted using a domain classifier fitted with unlabelled target observations | Same held-out target machine | No |
| LMO-CP | Nested models: $K-1$ source machines; final model: all $K$ source machines | Scores obtained from each excluded source machine and subsequently pooled | Same held-out target machine | No; calibration machine excluded from nested training |

Note: LMO-CP = leave-machine-out conformal prediction.

### 3.8 Statistical, Ablation, and Robustness Analyses

Additional analyses were conducted to examine the consistency of the principal findings across machines, identify the contribution of individual kinematic feature groups, and assess the sensitivity of diagnostic performance and conformal calibration to changes in signal characteristics and classifier selection. Six complementary analyses were performed within the experimental framework described above.

Statistical comparisons. Paired differences between methods were calculated across 40 machine–seed evaluation cases obtained from eight held-out machines and five random seeds. Bootstrap confidence intervals were estimated using 10,000 resamples of these cases and describe variability within the simulated evaluation design. Because repeated evaluations of the same machine were not independent machine-level observations, statistical significance was assessed separately using an exact two-sided Wilcoxon signed-rank test on the eight machine-level mean differences. With eight nonzero paired differences and no tied absolute differences, the smallest attainable two-sided $p$-value is 0.0078. These comparisons examined the consistency of differences between feature representations and calibration schemes across the simulated machines.

Feature-group ablation. The 20 geometry-anchored kinematic features were divided into seven groups according to their physical interpretation: harmonics of FTF, twice the ball-spin frequency (2×BSF), BPFO, and BPFI; modulation sidebands; shaft-order components; and envelope kurtosis. Each group was removed separately, and

the classifier was retrained using the remaining features. The resulting changes in accuracy were used to assess the contribution of each feature group after the model had adapted to the reduced representation.

A complementary permutation analysis was performed using the original trained classifier. For each held-out machine, one feature group was permuted while the remaining features were kept unchanged. Three permutations were conducted per evaluation fold, and the corresponding reductions in diagnostic accuracy were recorded. Unlike feature removal followed by retraining, this procedure measured the dependence of the fitted classifier on the information contained in each group.

Two reduced representations were also examined. The first retained only the four fundamental fault-order amplitudes, while the second included all twelve harmonic features but excluded modulation sidebands, shaft-order components, and envelope kurtosis. These comparisons were used to determine how much of the diagnostic performance could be retained by the characteristic fault-frequency information alone.

Confusion and unknown-fault analysis. Class-level diagnostic performance was examined using confusion matrices obtained from the held-out machines. The analysis considered the four known conditions and the distribution of predicted labels assigned to the unknown mechanical-looseness condition. Particular attention was given to the proportion of unknown-fault windows classified as normal, since these errors correspond to missed fault indications in the simulated diagnostic setting.

Alternative conformal calibration procedures. Three variants of LMO-CP were examined at a nominal miscoverage level of $\alpha = 0.1$. The first was a class-conditional, or Mondrian, calibration procedure [36], in which separate thresholds were calculated for each known fault class using the corresponding nested machine-level nonconformity scores. This variant examined whether class-specific calibration reduced differences in coverage between fault categories.

The second variant, denoted LMO-max, used a conservative calibration threshold equal to the largest of the machine-specific quantiles. Each quantile was obtained from the nonconformity scores of one source machine evaluated by a classifier trained without that machine. The conservative threshold was used to examine whether accounting for the least favorable source-machine calibration distribution improved coverage on atypical target machines, and how this affected prediction-set size.

The third variant replaced the conventional probability-based nonconformity score with a non-randomized adaptive score based on cumulative class probability [35]. This comparison examined whether accounting for the relative probability mass of candidate classes altered prediction-set coverage and informativeness under machine-level distribution shift.

For these variants, worst-class coverage was defined as the lowest class-specific coverage within each held-out machine, averaged across machine–seed evaluation cases. Each fault class contributed 48 windows per machine. Worst-machine coverage was defined as the minimum coverage across the eight held-out machines within each seed, averaged over seeds. Both measures were reported alongside overall coverage and mean prediction-set size to distinguish average calibration performance from variation across machines and fault categories.

Kinematic model-mismatch stress test. A separate stress test was conducted to examine the sensitivity of geometry-anchored features to departures from the nominal fault-frequency relationships used in the reference simulation. Three random seeds were regenerated with timing jitter increased from 1% to 3%. A multiplicative perturbation was also applied to the true fault orders at the recording level, with a standard deviation of 2% and values clipped to $\pm 5\%$. In addition, shaft speed was allowed to drift by 1%–3% within each signal window.

During feature extraction, the nominal kinematic fault orders and their original 3% tolerance bands were retained. Consequently, the extracted features were based on imperfect estimates of the fault-frequency locations present in the generated signals. The stress test examined how these deviations affected classification accuracy, conformal coverage, prediction-set size, and coverage on individual machines.

Classifier substitution. The diagnostic pipeline was repeated using standardized logistic regression and a random forest to examine whether the principal findings depended on the histogram-based gradient-boosting classifier. The random forest contained 100 trees, with a minimum leaf size of three observations. Both classifiers were evaluated using order-domain and geometry-anchored features under the same leave-one-machine-out protocol. Three random seeds and the corresponding data partitions were used for these comparisons.

Together, these analyses examined the relative contributions of physically indexed features and machine-level calibration, the variation in prediction reliability across machines and fault classes, and the stability of the findings under changes in signal characteristics and classifier selection. A Colab notebook containing a reduced-size version of the reference experiment was supplied as supporting material.

### 3.9 Structural Audit of Fault-Line Ambiguity

The final analysis examined the extent to which characteristic bearing fault frequencies could be distinguished across machines using fixed frequency and order coordinates. Unlike the preceding simulation experiments, this

assessment was deterministic and used only the machine parameters listed in Table A1. No random signal generation or additional classifier training was required to calculate the fault-line collisions.

Three characteristic fault orders were considered for each machine: BPFO for outer-race faults, BPFI for inner-race faults, and twice the ball-spin frequency ($2 \times$ BSF) for ball faults. The orders were calculated using the bearing kinematic relationships described in Section 3.2, producing 24 fault-defining lines across the eight simulated machines.

A cross-machine collision was identified when a characteristic fault order lay within 3% of the order associated with a different fault on another machine. The relative tolerance corresponded to the search band used by the geometry-anchored feature extractor. Collisions between different fault categories within the same machine were treated separately and were not included in the cross-machine collision count.

For the frequency-domain representation, a collision was identified when characteristic fault frequencies associated with different faults on different machines occupied the same 10 Hz interval used by the fixed-frequency features. These two criteria allowed structural ambiguity to be examined independently on the normalized order axis and the physical frequency axis.

To assess the diagnostic consequences of this ambiguity, an idealized leave-one-machine-out nearest-line classifier was evaluated. Each machine was excluded in turn, and its characteristic fault lines were compared with those of the remaining seven machines. For each fault line of the held-out machine, the predicted category was assigned according to the nearest characteristic line among the source machines, with distance measured on a logarithmic axis.

The classifier operated on three fault categories, giving a chance-level accuracy of 1/3. It was evaluated without noise, interference, or amplitude variation, so its results represented the structural discriminative information available from fault-line positions rather than the performance of a complete vibration-based diagnostic model.

The audit also distinguished between two properties of fixed-axis representations. Cross-fault collision described cases in which similar frequency or order coordinates corresponded to different fault mechanisms on different machines. Same-fault agreement described cases in which the characteristic lines of the same fault occurred at similar coordinates across different machines. The former measures ambiguity in fault identification, whereas the latter indicates the consistency of fault-related information available to a classifier operating on a fixed axis.

The results of this analysis are presented in Section 4.15, with the individual fault-line collisions reported in Appendix C. Together with the classification experiments, the audit provides a physical basis for interpreting the differences between fixed-axis and geometry-anchored feature representations under machine-level distribution shift.

## 4 Results

### 4.1 Influence of Evaluation Protocol on Diagnostic Accuracy

The effect of data partitioning was first examined by comparing random window splitting, recording-level separation, and leave-one-machine-out evaluation using the same classifier and feature representations. As shown in Table 6 and Figure 3, the apparent performance of the frequency-domain and order-domain representations depended strongly on the evaluation protocol. Under random window splitting, the Hz-domain and order-domain features achieved accuracies of $0.834 \pm 0.025$ and $0.866 \pm 0.032$, respectively. When entire machines were excluded from training, these values fell to $0.585 \pm 0.016$ and $0.575 \pm 0.023$, corresponding to overestimations of approximately 0.25 and 0.29.

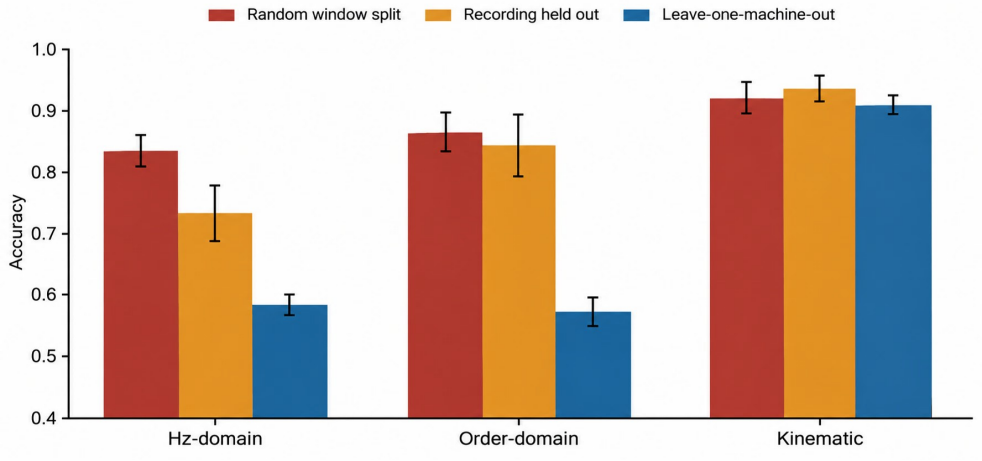



**Figure 3.** Diagnostic accuracy under three evaluation protocols (mean $\pm$ SD over five seeds)

Note: SD = standard deviation.

**Table 6.** Diagnostic accuracy under three evaluation protocols (mean ± SD over five seeds)

| Features | Random Window Split | Recording Held Out | Leave-One-Machine-Out |
|---|---|---|---|
| Hz-domain | $0.834 \pm 0.025$ | $0.737 \pm 0.050$ | $0.585 \pm 0.016$ |
| Order-domain | $0.866 \pm 0.032$ | $0.847 \pm 0.047$ | $0.575 \pm 0.023$ |
| Kinematic (geometry-anchored) | $0.921 \pm 0.025$ | $0.935 \pm 0.020$ | $0.908 \pm 0.014$ |

Note: SD = standard deviation.

Separating complete recordings reduced the apparent accuracy of Hz-domain features to $0.737 \pm 0.050$, but the order-domain representation still achieved $0.847 \pm 0.047$. Thus, recording-level separation reduced direct overlap between training and test recordings without fully accounting for differences between machines. The geometry-anchored representation showed considerably smaller variation across the three protocols, with accuracies of $0.921 \pm 0.025$, $0.935 \pm 0.020$, and $0.908 \pm 0.014$, respectively. Its performance was therefore less dependent on whether the evaluation used previously encountered or entirely held-out machines.

### 4.2 Diagnostic Performance on Unseen Machines

The leave-one-machine-out results were examined in greater detail to determine whether the differences observed in Section 4.1 persisted across classification metrics and alternative model configurations. The geometry-anchored representation achieved an accuracy of $0.908 \pm 0.014$ and a macro-F1 score of $0.909 \pm 0.013$, with an expected calibration error (ECE) of $0.052 \pm 0.015$. In comparison, the Hz-domain and order-domain representations achieved accuracies of 0.585 and 0.575, respectively (Table 7).

**Table 7.** Leave-one-machine-out diagnostic performance (mean ± SD over five seeds)

| Features/Model | Accuracy | Macro-F1 | ECE | Worst-Machine Accuracy |
|---|---|---|---|---|
| Hz-domain | $0.585 \pm 0.016$ | $0.560 \pm 0.017$ | $0.151 \pm 0.017$ | $0.418 \pm 0.042$ |
| Hz-domain + per-machine norm. | $0.513 \pm 0.014$ | $0.501 \pm 0.015$ | $0.131 \pm 0.015$ | $0.397 \pm 0.038$ |
| Hz-domain, perceptron | $0.501 \pm 0.023$ | $0.476 \pm 0.024$ | $0.375 \pm 0.026$ | $0.382 \pm 0.030$ |
| Order-domain | $0.575 \pm 0.023$ | $0.527 \pm 0.025$ | $0.171 \pm 0.008$ | $0.454 \pm 0.047$ |
| Order-domain + per-machine norm. | $0.596 \pm 0.030$ | $0.586 \pm 0.030$ | $0.124 \pm 0.008$ | $0.477 \pm 0.036$ |
| Order-domain + geometry vector | $0.573 \pm 0.027$ | $0.526 \pm 0.030$ | $0.182 \pm 0.015$ | $0.461 \pm 0.045$ |
| Order-domain, perceptron | $0.426 \pm 0.043$ | $0.393 \pm 0.039$ | $0.437 \pm 0.037$ | $0.277 \pm 0.026$ |
| Kinematic (geometry-anchored) | $0.908 \pm 0.014$ | $0.909 \pm 0.013$ | $0.052 \pm 0.015$ | $0.795 \pm 0.059$ |
| Kinematic + per-machine norm. | $0.915 \pm 0.019$ | $0.915 \pm 0.019$ | $0.045 \pm 0.007$ | $0.813 \pm 0.064$ |

Note: SD = standard deviation; ECE = expected calibration error; norm. = normalization.

Adding geometric parameters directly to the order-domain feature vector produced an accuracy of $0.573 \pm 0.027$, almost identical to that of the original order-domain representation. Under the eight-machine simulation design, supplying geometric parameters as additional inputs did not reproduce the performance obtained when those parameters were used to define the locations of fault-related spectral features. This distinction indicates that explicit construction of physically indexed features was more effective than leaving the classifier to learn the relationship between geometry and fault frequencies from the available source configurations.

Per-machine normalization produced relatively small changes for the order-domain and kinematic representations, increasing their accuracies to 0.596 and 0.915, respectively. In contrast, Hz-domain accuracy decreased to 0.513. The perceptron baselines also performed below the corresponding gradient-boosting models, achieving accuracies of 0.501 for Hz-domain features and 0.426 for order-domain features, together with ECE values of 0.375 and 0.437. These results indicate that neither target-machine feature normalization nor substitution with the tested perceptron compensated for the machine dependence of fixed-axis representations.

The advantage of the kinematic representation was also reflected in worst-machine accuracy, which reached $0.795 \pm 0.059$ compared with $0.418 \pm 0.042$ and $0.454 \pm 0.047$ for the Hz-domain and order-domain baselines. The comparison concerns the classifiers and configurations evaluated here; deep convolutional and adversarial domain-adaptation models were not included in this experiment.

### 4.3 Prediction-Set Coverage and Informativeness on Unseen Machines

The calibration experiments examined whether the diagnostic representations could produce prediction sets with empirical coverage close to the nominal level while retaining useful fault discrimination. At a nominal coverage of

0.90, naive split calibration achieved $0.705 \pm 0.042$ coverage with Hz-domain features, $0.642 \pm 0.020$ with order-domain features, and $0.879 \pm 0.008$ with kinematic features. Recording-held-out calibration produced corresponding values of 0.797, 0.682, and 0.877. In contrast, LMO-CP achieved $0.931 \pm 0.006$, $0.928 \pm 0.009$, and $0.903 \pm 0.007$, respectively (Table 8 and Figure 4).

**Table 8.** Conformal prediction performance at 90% nominal coverage

| Features/Calibration | Coverage (Mean ± SD) | Worst-Machine Coverage | Mean Set Size | Machine–Seed Cases With Coverage ≥ 0.85 |
|---|---|---|---|---|
| Hz-domain/Naive split | $0.705 \pm 0.042$ | 0.508 | 1.51 | 15% |
| Hz-domain/Recording split | $0.797 \pm 0.046$ | 0.644 | 2.29 | 30% |
| Hz-domain/Weighted | $0.811 \pm 0.038$ | 0.561 | 2.49 | 45% |
| Hz-domain/LMO-CP | $0.931 \pm 0.006$ | 0.845 | 2.96 | 90% |
| Order-domain/Naive split | $0.642 \pm 0.020$ | 0.478 | 1.23 | 5% |
| Order-domain/Recording split | $0.682 \pm 0.046$ | 0.545 | 1.38 | 5% |
| Order-domain/Weighted | $0.826 \pm 0.037$ | 0.621 | 2.55 | 50% |
| Order-domain/LMO-CP | $0.928 \pm 0.009$ | 0.863 | 2.77 | 93% |
| Kinematic/Naive split | $0.879 \pm 0.008$ | 0.729 | 0.95 | 68% |
| Kinematic/Recording split | $0.877 \pm 0.013$ | 0.710 | 0.98 | 68% |
| Kinematic/Weighted | $0.903 \pm 0.010$ | 0.769 | 1.32 | 83% |
| Kinematic/LMO-CP | $0.903 \pm 0.007$ | 0.751 | 0.99 | 80% |

Note: SD = standard deviation; LMO-CP = leave-machine-out conformal prediction.

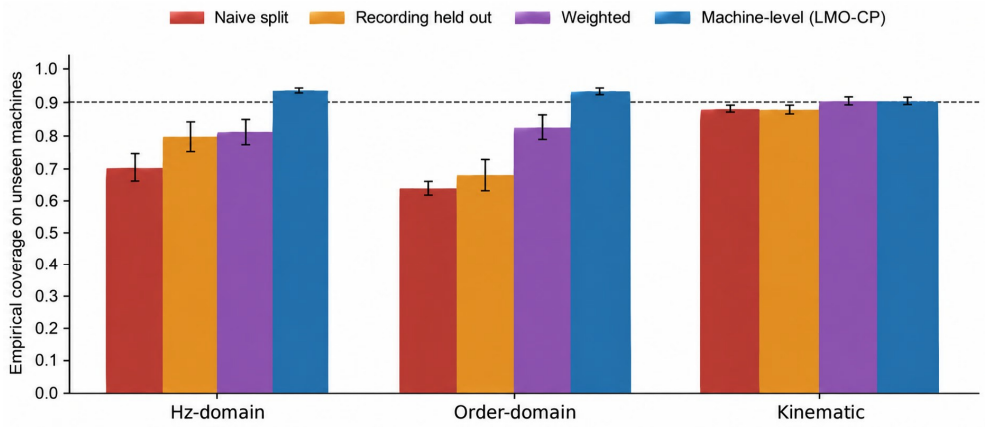



**Figure 4.** Empirical coverage by feature representation and calibration method at 90% nominal coverage

Note: LMO-CP = leave-machine-out conformal prediction.

The effect of machine-level calibration differed substantially between feature representations. For the two fixed-axis representations, the increase in coverage was accompanied by considerably larger prediction sets. LMO-CP returned an average of 2.96 classes for Hz-domain features and 2.77 classes for order-domain features, out of four possible classes. Although these sets approached or exceeded the nominal coverage target, they retained considerable ambiguity about the underlying fault. With geometry-anchored features, LMO-CP achieved 0.903 coverage with an average set size of only 0.99 classes. At nominal coverage levels of 0.95 and 0.80, the same representation and calibration procedure achieved empirical coverage of 0.951 and 0.794, respectively.

The average set size of 0.99 included empty, singleton, and multi-class prediction sets and should not be interpreted as the proportion of observations receiving a unique diagnosis. In the decision analysis reported in Section 4.8, the kinematic LMO-CP singleton policy abstained on $0.06 \pm 0.03$ of the evaluated windows, indicating that approximately 94% received a single-class prediction set. The remaining observations produced either empty or multi-class sets. These outcomes have different diagnostic interpretations: singleton sets identify one candidate fault, multi-class sets retain several plausible conditions, and empty sets indicate that none of the known classes satisfied the selected conformity threshold. However, the empty-set rule detected only 8.7% of the unknown-fault windows in the corresponding experiment, so a non-empty prediction set did not establish that the observed condition

belonged to the training classes.

Weighted calibration provided smaller coverage improvements for the fixed-axis representations, reaching 0.811 for Hz-domain features and 0.826 for order-domain features. The domain classifier distinguished source and target observations with high confidence, leaving normalized effective sample sizes of approximately 7%, 4%, and 14% for the Hz-domain, order-domain, and kinematic representations, respectively. With kinematic features, weighted calibration achieved $0.903 \pm 0.010$ coverage but required an average of 1.32 classes per prediction set, compared with 0.99 for LMO-CP. Thus, the machine-level calibration procedure achieved the most favorable combination of average coverage and set size among the evaluated schemes when paired with geometry-anchored features.

### 4.4 Machine-Specific Variation in Prediction-Set Coverage

The results in Section 4.3 establish the average empirical performance of the calibration procedures, but they do not describe the reliability experienced by every individual machine. Marginal coverage concerns observations averaged over the evaluated machine population, whereas machine-specific coverage is calculated for a particular machine. These quantities are distinct. The general limitations of distribution-free conditional coverage [39] also motivate distinguishing average coverage from coverage conditional on a particular machine.

With kinematic features and LMO-CP, mean coverage across individual machines ranged from 0.777 for M4 to 0.994 for M3 (Table 9 and Figure 5). The average of the lowest machine coverage within each seed was 0.751, showing that the minimum within a seed could fall below the minimum of the machine-level means reported in Table 9. Nevertheless, 80% of the forty machine–seed cases achieved coverage of at least 0.85, compared with 68% under naive calibration.

**Table 9.** Machine-specific empirical coverage at 90% nominal coverage (five seeds)

| Held-out Machine | Order/Naive | Order/LMO-CP | Kinematic/Naive | Kinematic/LMO-CP |
|---|---|---|---|---|
| M1 | 0.578 | 0.895 | 0.873 | 0.909 |
| M2 | 0.649 | 0.925 | 0.820 | 0.862 |
| M3 | 0.685 | 0.977 | 0.988 | 0.994 |
| M4 | 0.492 | 0.928 | 0.758 | 0.777 |
| M5 | 0.805 | 0.989 | 0.935 | 0.951 |
| M6 | 0.659 | 0.888 | 0.936 | 0.954 |
| M7 | 0.602 | 0.898 | 0.833 | 0.877 |
| M8 | 0.669 | 0.928 | 0.884 | 0.897 |

Note: LMO-CP = leave-machine-out conformal prediction.

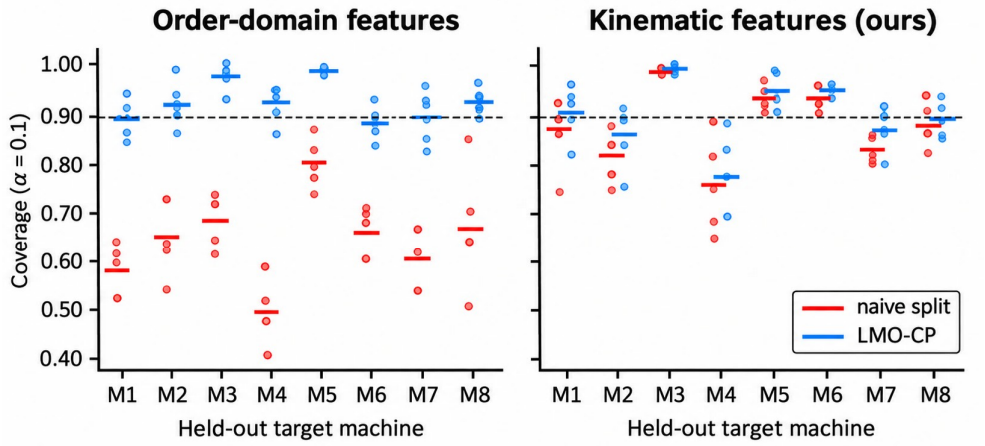



**Figure 5.** Machine-specific coverage under naive and LMO-CP calibration at 90% nominal coverage

Note: LMO-CP = leave-machine-out conformal prediction. Dots indicate individual seeds and bars indicate machine means.

The largest departure from nominal coverage occurred on M4, which combined the lowest signal gain with the highest noise level among the simulated machines. Its kinematic LMO-CP coverage was 0.777, compared with 0.758 under naive calibration. M2 had the second-lowest kinematic LMO-CP coverage at 0.862. Its inner-race and ball-fault characteristic orders coincided at 4.8 shaft orders, creating an additional source of spectral ambiguity examined in Section 4.15. In contrast, M3 achieved coverage of 0.994, while M5 and M6 reached 0.951 and 0.954, respectively.

Seven of the eight machines achieved mean kinematic LMO-CP coverage of at least 0.85. The difference between M4 and M3 nevertheless demonstrates that average coverage close to the nominal level can coexist with substantial variation across machine configurations. The results therefore support evaluating machine-specific coverage alongside the overall average, particularly when the target machine differs markedly from the source population.

### 4.5 Effect of the Number of Source Machines

The dependence of conformal coverage on the number of available source machines was examined by varying $K$ from two to seven, using two random seeds for each configuration. Figure 6 shows that the calibration schemes responded differently as additional machines became available.

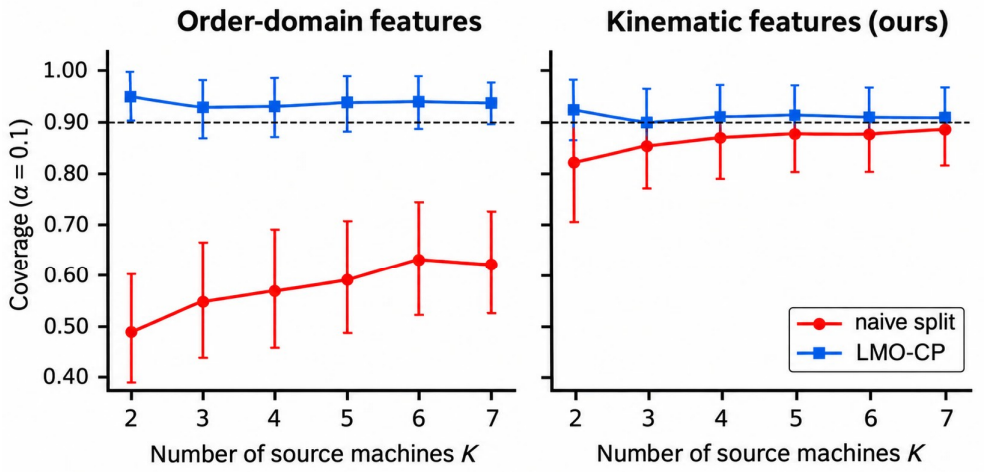



**Figure 6.** Effect of the number of source machines on empirical coverage (two seeds; nominal coverage 90%)
Note: LMO-CP = leave-machine-out conformal prediction.

For geometry-anchored features, naive calibration coverage increased gradually from 0.82 with two source machines to 0.88 with seven. LMO-CP remained close to nominal coverage across the evaluated configurations, reaching approximately 0.90 at $K = 3$ and 0.91 at $K = 7$. With only two source machines, LMO-CP achieved coverage of 0.92, accompanied by a larger mean prediction-set size of 1.43 classes. The reduction in set size as more machines became available indicates that the additional source configurations supplied more informative calibration observations under the simulated conditions.

The order-domain representation showed a different pattern. Naive calibration remained substantially below nominal coverage throughout the sweep, ranging from 0.49 to 0.63. LMO-CP maintained coverage above the nominal level, reaching 0.94 at $K = 7$, but the corresponding prediction sets contained approximately 2.91 classes on average. Increasing the number of source machines therefore improved the calibration setting without resolving the limited fault discrimination of the fixed-order representation.

These results indicate that source-machine diversity and feature transferability play different roles. Additional machines can support more stable machine-level calibration, but the informativeness of the resulting prediction sets remains closely related to the underlying feature representation.

### 4.6 Sensitivity to Bearing Geometry and Shaft-Speed Errors

The geometry-anchored representation relies on accurate information about bearing geometry and shaft rotational speed. To examine its sensitivity to errors in these inputs, the parameters supplied to the feature extractor were perturbed while the underlying simulation framework was retained. The geometry perturbations involved relative errors in the rolling-element-to-pitch-diameter ratio, $d/D$, with randomly assigned signs for individual machines. The largest perturbation also included a $5^\circ$ contact-angle error. Additional experiments introduced errors in the estimated shaft speed. Each configuration was evaluated using three random seeds.

As shown in Table 10, the geometry-anchored representation achieved an accuracy of $0.870 \pm 0.012$ with exact geometric information. Accuracy decreased to $0.823 \pm 0.033$ and $0.807 \pm 0.028$ when the diameter-ratio error increased to 3% and 6%, respectively. The largest perturbation, combining a 10% diameter-ratio error with a $5^\circ$ contact-angle error, reduced accuracy to $0.762 \pm 0.019$. Shaft-speed errors produced a similar pattern, although the effect was smaller at low perturbation levels. Accuracy remained at $0.867 \pm 0.008$ with a 1% speed error and decreased to $0.834 \pm 0.022$ when the error reached 2%.

**Table 10.** Sensitivity to bearing geometry and shaft-speed errors (three seeds)

| Perturbation | Accuracy | LMO-CP Coverage | Mean Set Size |
|---|---|---|---|
| None (exact) | $0.870 \pm 0.012$ | $0.901 \pm 0.002$ | $1.10 \pm 0.06$ |
| $d/D$ error 3% | $0.823 \pm 0.033$ | $0.903 \pm 0.003$ | $1.31 \pm 0.16$ |
| $d/D$ error 6% | $0.807 \pm 0.028$ | $0.905 \pm 0.005$ | $1.38 \pm 0.16$ |
| $d/D$ error 10% + 5° | $0.762 \pm 0.019$ | $0.900 \pm 0.002$ | $1.61 \pm 0.03$ |
| Speed error 1% | $0.867 \pm 0.008$ | $0.901 \pm 0.002$ | $1.11 \pm 0.05$ |
| Speed error 2% | $0.834 \pm 0.022$ | $0.901 \pm 0.003$ | $1.29 \pm 0.11$ |

Note: LMO-CP = leave-machine-out conformal prediction.

The corresponding conformal results revealed a different response. LMO-CP maintained average coverage close to the nominal level of 0.90 across all perturbation settings, ranging from 0.900 to 0.905. The principal effect of inaccurate input information appeared in prediction-set size, which increased from $1.10 \pm 0.06$ classes with exact geometry to $1.61 \pm 0.03$ under the largest geometric perturbation. A 2% speed error similarly increased mean set size to $1.29 \pm 0.11$ classes. Thus, under the simulated perturbations, machine-level calibration maintained average empirical coverage by producing less selective prediction sets as the diagnostic representation became less accurate.

The exact-geometry accuracy in this experiment was lower than the 0.908 reported in Section 4.2 because the sensitivity analysis used a separate random realization and three seeds rather than the five-seed reference experiment. The relevant comparison is therefore between the perturbation conditions within Table 10. Figure 7 illustrates the corresponding changes in diagnostic accuracy and conformal coverage.

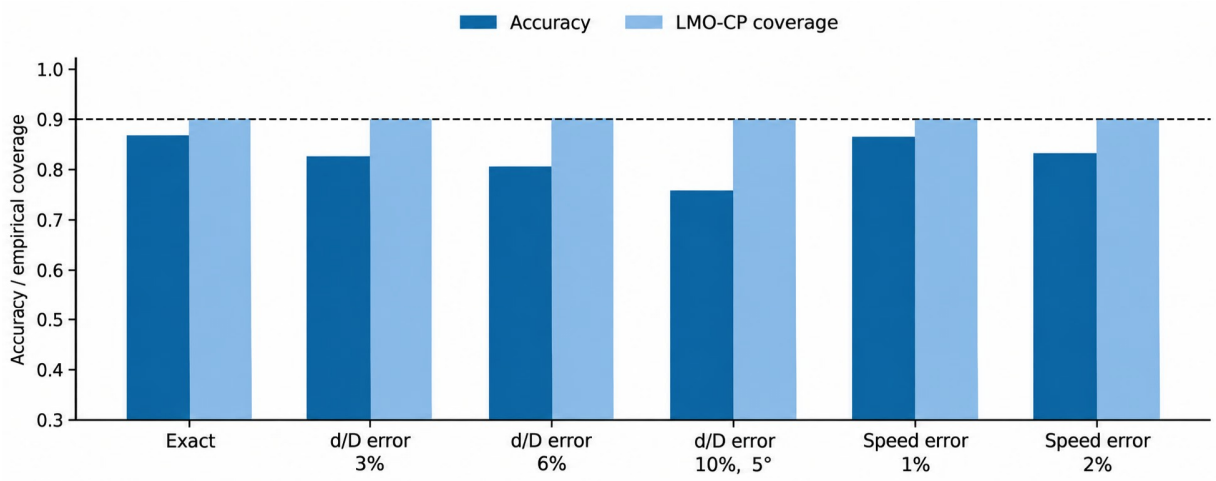



**Figure 7.** Diagnostic accuracy and LMO-CP coverage under bearing geometry and shaft-speed perturbations

Note: LMO-CP = leave-machine-out conformal prediction.

### 4.7 Sensitivity to Signal-to-Noise Ratio

The preceding analysis examined errors in the physical parameters used for feature extraction. A complementary experiment investigated how diagnostic performance changed when fault-related vibration components became more or less distinguishable from background noise. Signal difficulty was adjusted by rescaling fault severity, while the remaining simulation procedures were retained. Two additional random seeds were used for each modified signal-to-noise ratio (SNR) condition, and the five-seed reference results were included for comparison.

The geometry-anchored representation maintained the highest classification accuracy across all three SNR conditions. At high SNR, it achieved an accuracy of 0.95, compared with 0.65 for frequency-domain features and 0.62 for order-domain features. Under the reference condition, the corresponding accuracies were approximately 0.91, 0.58, and 0.57. At low SNR, accuracy decreased to 0.71 for geometry-anchored features, while the frequency-domain and order-domain representations reached 0.42 and 0.41, respectively. As summarized in Table 11, the separation between the representations persisted even when the fault-related signal components became substantially weaker.

The calibration results showed that decreasing SNR affected prediction-set informativeness more strongly than average coverage. With geometry-anchored features, LMO-CP achieved coverage of approximately 0.91, 0.90, and 0.90 under high, reference, and low SNR conditions, respectively. Mean prediction-set size, however, increased from

0.93 classes at high SNR to 0.99 under the reference condition and 1.89 at low SNR. The naive calibration procedure also produced larger sets as signal quality decreased, but its average coverage remained below the nominal level.

**Table 11.** Effects of signal-to-noise ratio on diagnostic performance

| SNR Condition | Accuracy: Hz/Order/Kinematic | Kinematic Coverage: Naive/LMO-CP | Kinematic Set Size: Naive/LMO-CP | Order Coverage: Naive/LMO-CP |
|---|---|---|---|---|
| High SNR (2 seeds) | 0.65/0.62/0.95 | 0.83/0.91 | 0.85/0.93 | 0.58/0.94 |
| Reference (5 seeds) | 0.58/0.57/0.91 | 0.88/0.90 | 0.95/0.99 | 0.64/0.93 |
| Low SNR (2 seeds) | 0.42/0.41/0.71 | 0.88/0.90 | 1.78/1.89 | 0.80/0.91 |

Note: SNR = signal-to-noise ratio; LMO-CP = leave-machine-out conformal prediction.

For order-domain features, LMO-CP maintained coverage between 0.91 and 0.94 across the three SNR conditions, despite substantially lower classification accuracy. These results reinforce the distinction between obtaining nominal empirical coverage and producing diagnostically informative prediction sets. When fault signatures became difficult to distinguish, the calibration procedure maintained coverage primarily by admitting more candidate classes.

### 4.8 Unknown-Fault Detection, Selective Classification, and Cost-Sensitive Decisions

The previous experiments concerned faults represented in the training data. Industrial diagnostic systems must also respond to observations that do not correspond to any known fault category. Mechanical looseness was therefore introduced as an unknown condition, and its detection was examined alongside selective classification and cost-sensitive decision-making.

Selective classification produced the expected relationship between the fraction of windows receiving a prediction and the accuracy of those predictions. As the model abstained from an increasing proportion of low-confidence observations, accuracy on the remaining windows improved. This effect was most apparent for geometry-anchored features, as illustrated in Figure 8. However, the ability to identify misclassified known faults did not translate directly into reliable recognition of the unknown looseness condition.

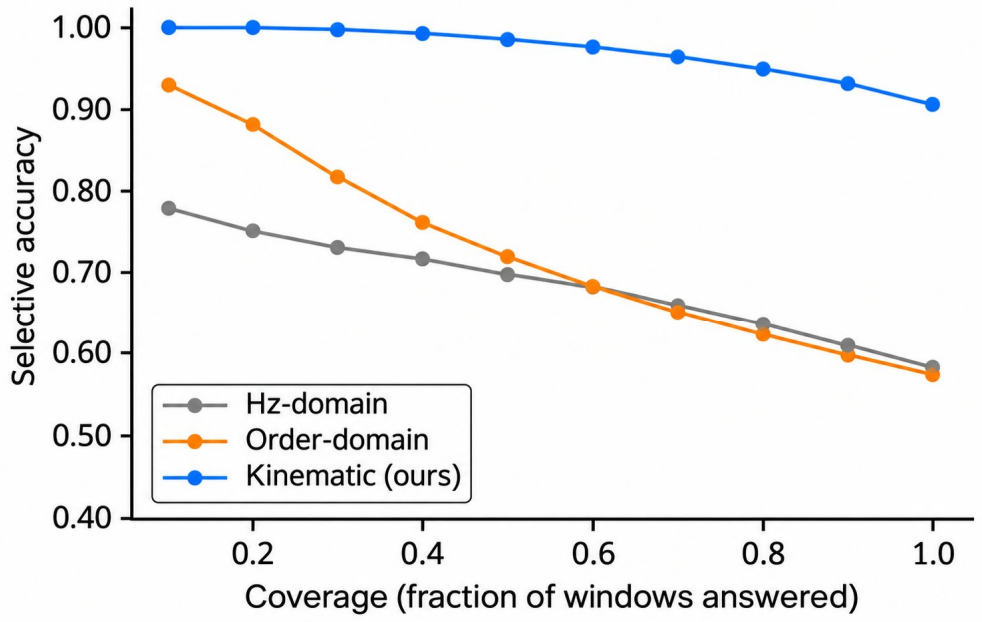



**Figure 8.** Selective accuracy versus the fraction of windows classified, averaged across machines and seeds

Maximum predicted class probability achieved an AUROC of $0.831 \pm 0.023$ for geometry-anchored features, compared with 0.484 and 0.444 for frequency-domain and order-domain features, respectively. Class-conditional Mahalanobis distance produced AUROC values of 0.502, 0.548, and 0.619 for the three representations, increasing

to 0.673 for the geometry-anchored representation after per-machine standardization. These results showed that the usefulness of conventional novelty scores depended strongly on the underlying feature representation.

The conformal empty-set rule was also examined as an indicator of unknown faults. With geometry-anchored features and LMO-CP at $\alpha = 0.1$, the rule flagged $0.087 \pm 0.102$ of unknown-fault windows while producing empty sets for $0.020 \pm 0.024$ of known-fault windows. Increasing the nominal miscoverage level to $\alpha = 0.2$ raised the unknown-fault flagging rate to $0.541 \pm 0.075$, but also increased the empty-set rate for known faults to $0.165 \pm 0.023$. The higher detection rate was therefore accompanied by a substantial increase in false warnings. In this experiment, conformal empty sets provided a selective warning mechanism rather than a reliable standalone detector of previously unseen faults.

The operational consequences of these predictions were examined using the cost matrix defined in Section 3.6. Table 12 reports the mean decision cost, missed-fault rate, and abstention rate for each evaluated decision policy. Under the specified costs, referring every observation for inspection incurred a constant cost of 0.30 per window. The Bayes decision rule with geometry-anchored features achieved the lowest mean cost among the evaluated policies, at $0.289 \pm 0.068$. It abstained on $0.44 \pm 0.04$ of windows and classified $0.01 \pm 0.01$ of fault windows as normal.

**Table 12.** Cost-sensitive diagnostic decisions at 90% nominal coverage

| Features/Decision Policy | Mean Cost | Faults Called Normal | Abstention Rate |
|---|---|---|---|
| Hz-domain/Argmax | $2.760 \pm 0.171$ | $0.27 \pm 0.02$ | $0.00 \pm 0.00$ |
| Hz-domain/Bayes | $0.381 \pm 0.023$ | $0.00 \pm 0.00$ | $0.83 \pm 0.02$ |
| Hz-domain/Singleton, LMO-CP | $0.362 \pm 0.033$ | $0.00 \pm 0.00$ | $0.90 \pm 0.03$ |
| Hz-domain/Singleton, naive | $1.778 \pm 0.220$ | $0.17 \pm 0.02$ | $0.42 \pm 0.13$ |
| Order-domain/Argmax | $3.534 \pm 0.102$ | $0.39 \pm 0.01$ | $0.00 \pm 0.00$ |
| Order-domain/Bayes | $0.398 \pm 0.019$ | $0.00 \pm 0.00$ | $0.74 \pm 0.03$ |
| Order-domain/Singleton, LMO-CP | $0.394 \pm 0.030$ | $0.01 \pm 0.00$ | $0.85 \pm 0.03$ |
| Order-domain/Singleton, naive | $2.620 \pm 0.215$ | $0.28 \pm 0.03$ | $0.22 \pm 0.09$ |
| Kinematic/Argmax | $2.064 \pm 0.085$ | $0.24 \pm 0.01$ | $0.00 \pm 0.00$ |
| Kinematic/Bayes | $0.289 \pm 0.068$ | $0.01 \pm 0.01$ | $0.44 \pm 0.04$ |
| Kinematic/Singleton, LMO-CP | $1.859 \pm 0.170$ | $0.22 \pm 0.02$ | $0.06 \pm 0.03$ |
| Kinematic/Singleton, naive | $1.700 \pm 0.279$ | $0.20 \pm 0.03$ | $0.09 \pm 0.05$ |

Note: LMO-CP = leave-machine-out conformal prediction.

In contrast, always selecting the most probable class produced a mean cost of $2.064 \pm 0.085$ with geometry-anchored features and classified $0.24 \pm 0.01$ of fault windows as normal. The corresponding costs were higher for the frequency-domain and order-domain representations. These results indicate that classification decisions based solely on the highest predicted probability were poorly aligned with the asymmetric penalties assigned to missed faults.

The conformal singleton-set policy also performed less favorably than the Bayes rule. With geometry-anchored features, the LMO-CP singleton policy incurred a mean cost of $1.859 \pm 0.170$ and abstained on only $0.06 \pm 0.03$ of windows. Its relatively low abstention rate meant that many incorrect singleton predictions still resulted in costly decisions. For the weaker representations, the same policy abstained much more frequently, reaching $0.90 \pm 0.03$ for frequency-domain features and $0.85 \pm 0.03$ for order-domain features. Nevertheless, their mean costs remained above the cost of inspecting every observation.

The relative performance of these decision policies depended on the assumed costs. Under the Bayes rule, abstention was selected whenever the lowest expected cost of assigning a class exceeded the inspection cost. Reducing the inspection cost or increasing the penalty for missed faults therefore favored more frequent abstention, bringing the resulting cost closer to that of inspecting every observation. Conversely, higher inspection costs could make automated classification more attractive, provided that the estimated class probabilities were sufficiently reliable.

For geometry-anchored features, the expected calibration error was 0.052 in the reference experiment. Even with this level of probability calibration, the difference between the Bayes decision cost and the constant inspection cost was only 0.011, considerably smaller than the observed variation across machine–seed evaluations. The results therefore showed that improved classification accuracy and conformal coverage did not automatically translate into lower operational cost under the selected decision assumptions.

### 4.9 Statistical Comparison of Diagnostic and Calibration Methods

The preceding results indicated substantial differences between feature representations and calibration schemes. Paired statistical comparisons were performed to determine whether these differences were consistent across the simulated machines. Table 13 summarizes the mean paired differences, bootstrap confidence intervals, and machine-level Wilcoxon signed-rank tests.

**Table 13.** Paired statistical comparisons across 40 machine–seed evaluations

| Comparison (First Minus Second) | Mean Difference | 95% Bootstrap CI | Machine–Seed Cases Favoring First | Machines Favoring First | Exact $p$ |
|---|---|---|---|---|---|
| Accuracy: kinematic − order-domain | +0.333 | [+0.313, +0.353] | 100% | 8/8 | 0.0078 |
| Accuracy: kinematic − Hz-domain | +0.323 | [+0.292, +0.352] | 100% | 8/8 | 0.0078 |
| Accuracy: order-domain − Hz-domain | −0.010 | [−0.044, +0.026] | 43% | 3/8 | 0.46 |
| Coverage: LMO-CP − naive (Hz-domain) | +0.226 | [+0.194, +0.259] | 100% | 8/8 | 0.0078 |
| Coverage: LMO-CP − naive (order-domain) | +0.286 | [+0.259, +0.314] | 100% | 8/8 | 0.0078 |
| Coverage: LMO-CP − naive (kinematic) | +0.024 | [+0.017, +0.032] | 83% | 8/8 | 0.0078 |
| Set size: LMO-CP − naive (kinematic) | +0.040 | [+0.030, +0.050] | 90% | 8/8 | 0.0078 |
| Naive coverage − 0.90 (kinematic) | −0.021 | [−0.047, +0.004] | – | 3/8 at or above | 0.64 |
| LMO-CP coverage − 0.90 (kinematic) | +0.003 | [−0.022, +0.025] | – | 4/8 at or above | 0.84 |
| Naive coverage − 0.90 (order-domain) | −0.258 | [−0.291, −0.224] | – | 0/8 at or above | 0.0078 |
| LMO-CP coverage − 0.90 (order-domain) | +0.028 | [+0.014, +0.043] | – | 5/8 at or above | 0.11 |

Note: CI = confidence interval; LMO-CP = leave-machine-out conformal prediction. The symbol (–) indicates not applicable. Wilcoxon signed-rank tests were performed on eight machine-level mean differences.

Geometry-anchored features achieved higher classification accuracy than both fixed-axis representations across all 40 machine–seed evaluation cases. Relative to order-domain features, the mean accuracy difference was +0.333, with a 95% bootstrap confidence interval of [+0.313, +0.353]. The corresponding difference relative to frequency-domain features was +0.323, with an interval of [+0.292, +0.352]. Both comparisons favored the geometry-anchored representation on all eight machines, with exact machine-level $p$-values of 0.0078.

By comparison, the difference between order-domain and frequency-domain features was −0.010, with a 95% confidence interval of [−0.044, +0.026]. The machine-level test gave $p = 0.46$, indicating that normalizing spectral frequency by shaft speed did not produce a consistent accuracy advantage across the evaluated machines.

Machine-level calibration produced a similarly consistent improvement in empirical coverage. Relative to naive calibration, LMO-CP increased coverage by +0.226 for frequency-domain features and +0.286 for order-domain features. Both improvements were observed on all eight machines. For geometry-anchored features, the increase was smaller, at +0.024, with a 95% confidence interval of [+0.017, +0.032]. This improvement was also consistent at the machine level, although it required an average increase of 0.040 classes per prediction set.

The comparison with nominal coverage provided a more detailed interpretation. For geometry-anchored features, naive calibration produced a mean coverage difference of −0.021 relative to the 0.90 target, with a confidence interval of [−0.047, +0.004] and a machine-level $p$-value of 0.64. LMO-CP produced a difference of +0.003, with an interval of [−0.022, +0.025] and $p = 0.84$. Thus, the main evidence for the effect of machine-level calibration came from the substantial improvement obtained with machine-dependent representations and the consistent paired increase across machines, rather than from a large calibration failure of the geometry-anchored representation.

### 4.10 Contribution of Individual Kinematic Feature Groups

The ablation experiments examined whether the superior cross-machine performance of geometry-anchored features resulted from the physical indexing of fault-related spectral components or simply from the number of features available to the classifier. Table 14 and Figure 9 present the results of feature removal and permutation analyses.

The four fundamental fault-order amplitudes alone achieved an accuracy of $0.877 \pm 0.018$, compared with $0.908 \pm 0.014$ for the complete 20-feature representation. Retaining all twelve harmonic features increased accuracy

to $0.884 \pm 0.017$. These results indicate that much of the diagnostic information was contained in spectral amplitudes extracted at the characteristic fault orders, while modulation sidebands, shaft-order components, and envelope kurtosis provided additional discrimination.

**Table 14.** Ablation and permutation analysis of kinematic feature groups

| Feature Set | Accuracy | Change vs. Full | Accuracy Drop After Permutation |
|---|---|---|---|
| Full representation (20 features) | $0.908 \pm 0.014$ | – | – |
| Without envelope kurtosis | $0.908 \pm 0.016$ | $0.000 \pm 0.004$ | $0.003 \pm 0.005$ |
| Without FTF harmonics | $0.904 \pm 0.016$ | $-0.004 \pm 0.006$ | $0.011 \pm 0.006$ |
| Without shaft-order components | $0.904 \pm 0.011$ | $-0.004 \pm 0.003$ | $0.018 \pm 0.013$ |
| Without sidebands | $0.900 \pm 0.012$ | $-0.007 \pm 0.005$ | $0.033 \pm 0.007$ |
| Without BPFI harmonics | $0.893 \pm 0.021$ | $-0.015 \pm 0.008$ | $0.159 \pm 0.022$ |
| Harmonics only | $0.884 \pm 0.017$ | $-0.024 \pm 0.012$ | – |
| Without 2×BSF harmonics | $0.883 \pm 0.016$ | $-0.024 \pm 0.004$ | $0.172 \pm 0.015$ |
| Fundamental orders only | $0.877 \pm 0.018$ | $-0.030 \pm 0.009$ | – |
| Without BPFO harmonics | $0.748 \pm 0.020$ | $-0.160 \pm 0.009$ | $0.295 \pm 0.017$ |

Note: FTF = fundamental train frequency; BPFI = ball pass frequency of inner race; BSF = ball spin frequency; BPFO = ball pass frequency of outer race. The symbol (–) indicates that no corresponding value was reported.

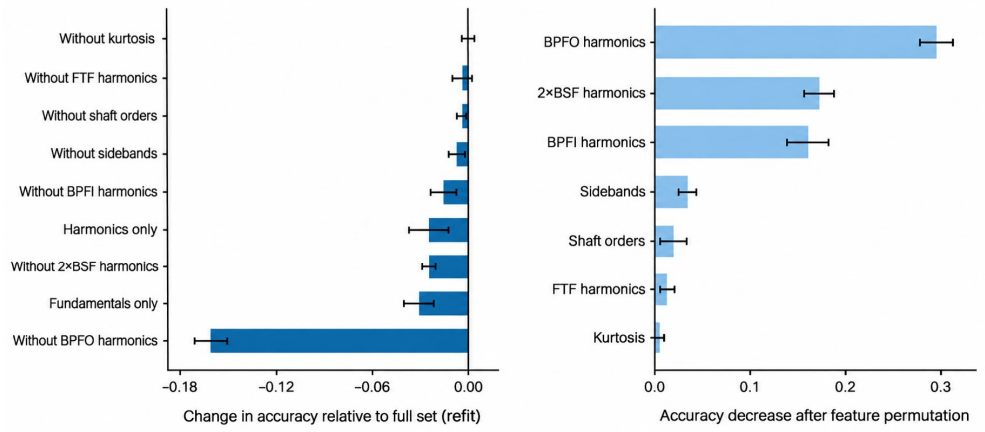



**Figure 9.** Effects of kinematic feature-group removal and permutation on cross-machine diagnostic accuracy

Note: BPFO = ball pass frequency of outer race; BPFI = ball pass frequency of inner race; BSF = ball spin frequency; FTF = fundamental train frequency. Error bars indicate standard deviations across five seeds.

The importance of individual feature groups was not uniform. Removing the BPFO harmonics reduced accuracy to $0.748 \pm 0.020$, a decrease of 0.160 relative to the full representation. Removing the harmonics associated with twice the ball-spin frequency reduced accuracy by 0.024, while excluding BPFI harmonics produced a reduction of 0.015. The effects of removing sidebands, shaft-order components, FTF harmonics, or envelope kurtosis were comparatively small.

Permutation analysis produced the same ordering among the most influential feature groups, although the magnitude of the accuracy reductions was greater. Permuting BPFO harmonics reduced accuracy by $0.295 \pm 0.017$, followed by twice the ball-spin frequency at $0.172 \pm 0.015$ and BPFI harmonics at $0.159 \pm 0.022$. The larger permutation effects arose because the trained classifier was evaluated with disrupted feature information, whereas the ablation experiments allowed the model to be refitted using the remaining features.

The pronounced contribution of BPFO harmonics is consistent with the signal-generation mechanism. In the reference simulator, outer-race faults were represented primarily by impulses at the characteristic BPFO order and its harmonics, without additional modulation sidebands. Removing these components therefore eliminated much

of the information used to identify outer-race faults. The relative importance of individual feature groups reflects the simulated fault mechanisms, while the comparison between full and reduced representations establishes the contribution of explicitly indexed fault-frequency information within the evaluated setting.

### 4.11 Class-Specific Diagnostic Performance and Unknown-Fault Classification

The aggregate accuracy results were examined further to determine how diagnostic errors were distributed among individual fault categories. Figure 10 presents the row-normalized confusion matrices obtained from the held-out machines, including the predicted labels assigned to the unknown mechanical-looseness condition.

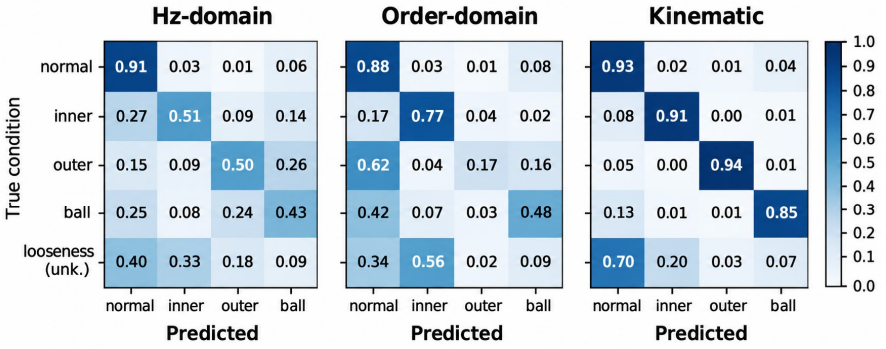



**Figure 10.** Row-normalized confusion matrices for held-out machines, including the unknown looseness condition

Note: unk. = unknown mechanical-looseness condition.

Geometry-anchored features achieved recall values of 0.93 for normal operation, 0.91 for inner-race faults, 0.94 for outer-race faults, and 0.85 for ball faults. Ball faults were the most difficult known condition, with approximately 13% of their windows classified as normal. Nevertheless, the representation maintained relatively balanced recognition across the four known classes.

The fixed-axis representations exhibited more pronounced class-dependent errors. With order-domain features, outer-race faults were correctly identified in only 17% of windows, while 63% were classified as normal. Ball-fault recall was 0.48. The frequency-domain representation produced recall values of approximately 0.51, 0.50, and 0.43 for inner-race, outer-race, and ball faults, respectively. These patterns are consistent with the ambiguity introduced when characteristic fault frequencies occupy different spectral positions across bearing configurations, or when similar positions correspond to different fault mechanisms.

The unknown mechanical-looseness condition revealed a separate difficulty. Even with geometry-anchored features, approximately 70% of unknown-fault windows were classified as normal. The remaining observations were assigned to inner-race faults (20%), outer-race faults (3%), and ball faults (7%). Because looseness was excluded from the training label space, these predictions reflected the tendency of the closed-set classifier to assign unfamiliar observations to the available categories.

The confusion analysis therefore distinguishes two aspects of diagnostic performance. Geometry-anchored features improved discrimination among the fault classes represented during training, but this improvement did not establish reliable recognition of an unknown fault. The novelty-detection results in Section 4.8 further showed that conventional confidence-based and conformal empty-set rules did not consistently resolve this distinction under the simulated machine shifts.

### 4.12 Class-Conditional, Conservative, and Adaptive Conformal Calibration

The preceding calibration results were based primarily on coverage averaged across machines and fault classes. To examine whether alternative calibration procedures could improve reliability for individual classes or atypical machines, pooled LMO-CP was compared with class-conditional, conservative, and adaptive-score variants. The results are summarized in Table 15 and Figure 11.

With geometry-anchored features, pooled LMO-CP produced class-specific coverage values of approximately 0.92 for normal operation, 0.90 for inner-race faults, 0.94 for outer-race faults, and 0.85 for ball faults. Although the overall coverage was close to the nominal target of 0.90, ball-fault coverage remained below that level. The class-conditional Mondrian procedure redistributed coverage across the classes, producing values of approximately 0.92, 0.89, 0.87, and 0.91, respectively. Mean prediction-set size increased only slightly, from 0.99 to 1.05 classes. However, worst-class coverage within individual machines decreased from 0.781 to 0.759, indicating that class-specific thresholds did not remove the variation associated with machine differences.

The adaptive prediction-set score also changed the balance between coverage and set size. With geometry-anchored features, it achieved overall coverage of $0.883 \pm 0.012$ and a mean set size of $2.05 \pm 0.03$ classes,

compared with $0.903 \pm 0.007$ and $0.99 \pm 0.03$ for the conventional pooled score. The larger sets did not produce a corresponding improvement in empirical coverage.

**Table 15.** Comparison of machine-level conformal calibration variants

| Features/Calibration Variant | Overall Coverage | Worst-Class Coverage | Worst-Machine Coverage | Machines $\geq 0.85$ | Mean Set Size |
|---|---|---|---|---|---|
| Order-domain/LMO-CP pooled | $0.928 \pm 0.009$ | 0.797 | 0.863 | 93% | $2.77 \pm 0.14$ |
| Order-domain/LMO-CP Mondrian | $0.922 \pm 0.004$ | 0.822 | 0.848 | 93% | $2.47 \pm 0.16$ |
| Order-domain/LMO-CP max | $0.977 \pm 0.007$ | 0.931 | 0.942 | 100% | $3.20 \pm 0.13$ |
| Order-domain/LMO-CP APS | $0.907 \pm 0.007$ | 0.741 | 0.819 | 90% | $2.89 \pm 0.05$ |
| Kinematic/LMO-CP pooled | $0.903 \pm 0.007$ | 0.781 | 0.751 | 80% | $0.99 \pm 0.03$ |
| Kinematic/LMO-CP Mondrian | $0.895 \pm 0.014$ | 0.759 | 0.748 | 70% | $1.05 \pm 0.08$ |
| Kinematic/LMO-CP max | $0.965 \pm 0.008$ | 0.915 | 0.866 | 95% | $1.50 \pm 0.30$ |
| Kinematic/LMO-CP APS | $0.883 \pm 0.012$ | 0.721 | 0.694 | 68% | $2.05 \pm 0.03$ |

Note: LMO-CP = leave-machine-out conformal prediction; APS = adaptive prediction sets.

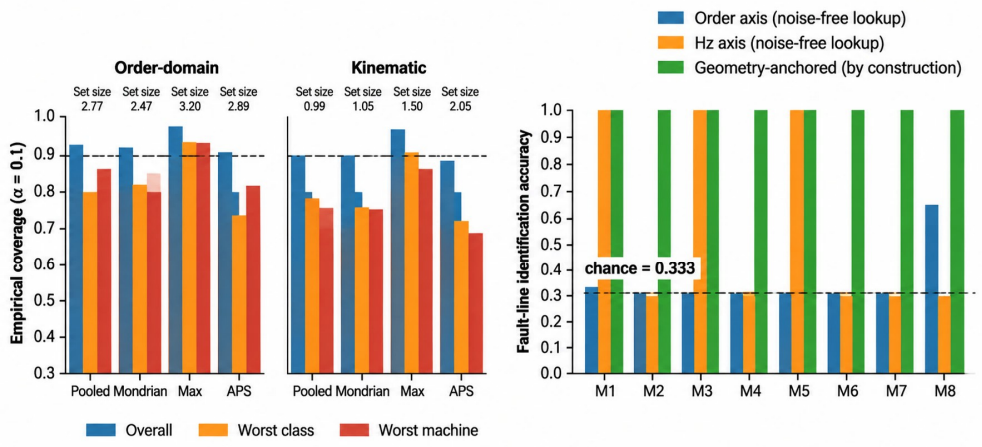



**Figure 11.** Overall, worst-class, and worst-machine coverage for four LMO-CP variants

Note: LMO-CP = leave-machine-out conformal prediction; APS = adaptive prediction sets.

The conservative LMO-max procedure produced a more substantial change. By selecting the largest machine-specific calibration quantile, it increased overall coverage to $0.965 \pm 0.008$. Worst-class coverage reached 0.915, and worst-machine coverage increased from 0.751 under pooled calibration to 0.866. The proportion of machine–seed cases achieving coverage of at least 0.85 increased from 80% to 95%. These improvements required larger prediction sets, with mean size increasing to $1.50 \pm 0.30$ classes.

The same conservative procedure was less informative when applied to order-domain features. Although overall coverage reached $0.977 \pm 0.007$ and worst-machine coverage increased to 0.942, the resulting sets contained $3.20 \pm 0.13$ of the four possible classes on average. The comparison demonstrates that a conservative threshold can improve coverage across heterogeneous machines, but its diagnostic value remains dependent on the underlying representation.

### 4.13 Robustness Under Kinematic Model Mismatch

The reference simulation generated fault-related impulses at characteristic orders calculated directly from the same bearing kinematic relationships used for feature extraction. To examine how departures from these relationships affected diagnostic performance, the stress test introduced additional slip, recording-level fault-order perturbations,

and within-window speed variation, as described in Section 3.8. Table 16 compares diagnostic accuracy, conformal coverage, prediction-set size, and worst-machine coverage before and after the kinematic model-mismatch stress test.

**Table 16.** Diagnostic performance under kinematic model mismatch (three seeds)

| Features | Accuracy | Naive Coverage | LMO-CP Coverage | LMO-CP Set Size | LMO-CP Worst-Machine Coverage |
|---|---|---|---|---|---|
| Hz-domain | 0.58 → 0.56 | 0.73 → 0.73 | 0.93 → 0.93 | 2.90 → 3.01 | 0.85 → 0.82 |
| Order-domain | 0.56 → 0.62 | 0.64 → 0.78 | 0.93 → 0.92 | 2.85 → 2.58 | 0.88 → 0.81 |
| Kinematic | 0.90 → 0.84 | 0.88 → 0.87 | 0.91 → 0.90 | 1.01 → 1.22 | 0.77 → 0.61 |

Note: LMO-CP = leave-machine-out conformal prediction.

Under the reference conditions for the same three seeds, geometry-anchored features achieved an accuracy of approximately 0.90. Accuracy decreased to 0.84 under the stress-test conditions. Despite this reduction, the geometry-anchored representation remained more accurate than both fixed-axis alternatives, which achieved 0.62 for order-domain features and 0.56 for frequency-domain features. The corresponding changes in the fixed-axis representations were comparatively small relative to their seed-to-seed variation.

The effect on conformal calibration was more complex. With geometry-anchored features, LMO-CP maintained average coverage close to 0.90, decreasing from approximately 0.91 under the reference conditions to 0.90 under model mismatch. Mean prediction-set size increased from 1.01 to 1.22 classes, indicating that the additional signal variation required less selective prediction sets to maintain aggregate coverage.

Machine-specific reliability was more strongly affected. Worst-machine coverage decreased from 0.77 to 0.61, while the proportion of machine–seed cases achieving coverage within 0.05 of the nominal target decreased from 83% to 79%. Thus, the average coverage remained close to nominal even though the least well-covered machine experienced a substantial deterioration.

The stress-test results show that geometry-anchored features retained a classification advantage under the specified departures from nominal bearing kinematics. They also establish that average conformal coverage and worst-machine coverage can respond differently to model mismatch. The experiment retained the underlying impulse-response structure of the reference simulator, so the observed robustness characterizes the evaluated perturbations rather than the full range of vibration conditions encountered in physical machinery.

### 4.14 Dependence on Classifier Selection

The reference experiments used histogram-based gradient boosting as the principal diagnostic classifier. To determine whether the observed differences between feature representations were specific to this model, the evaluation was repeated using standardized logistic regression and a random forest. Both classifiers were assessed using the same three random seeds and leave-one-machine-out protocol.

As shown in Table 17, geometry-anchored features achieved an accuracy of approximately 0.90 with both logistic regression and random forest. By comparison, order-domain features achieved 0.50 with logistic regression and 0.59 with random forest. The classification advantage of physically indexed fault features therefore persisted across the evaluated learning algorithms, indicating that the observed improvement was not restricted to the gradient-boosting classifier.

**Table 17.** Diagnostic performance across different classifiers (three seeds)

| Classifier/Features | Accuracy | Naive Coverage | LMO-CP Coverage | LMO-CP Set Size |
|---|---|---|---|---|
| Gradient boosting/order-domain | 0.56 | 0.64 | 0.93 | 2.85 |
| Logistic regression/order-domain | 0.50 | 0.73 | 0.92 | 3.14 |
| Random forest/order-domain | 0.59 | 0.65 | 0.90 | 2.84 |
| Gradient boosting/kinematic | 0.90 | 0.88 | 0.91 | 1.01 |
| Logistic regression/kinematic | 0.90 | 0.88 | 0.90 | 1.01 |
| Random forest/kinematic | 0.90 | 0.89 | 0.90 | 0.99 |

Note: LMO-CP = leave-machine-out conformal prediction.

The conformal results followed a similar pattern. For geometry-anchored features, naive calibration produced coverage of 0.88 with logistic regression and 0.89 with random forest. LMO-CP increased coverage to approximately

0.90 for both classifiers, while mean prediction-set size remained close to one class. With order-domain features, naive coverage was 0.73 and 0.65 for logistic regression and random forest, respectively. LMO-CP increased coverage to 0.92 and 0.90, but the resulting prediction sets contained approximately three classes on average.

The consistency of these results across gradient boosting, logistic regression, and random forest indicates that the principal differences between the representations arose from how fault-related information was encoded rather than from the choice of classifier alone. Machine-level calibration also produced similar changes in empirical coverage across the evaluated models. However, the diagnostic informativeness of the calibrated prediction sets remained strongly dependent on the feature representation. These findings support the use of physically indexed fault features as a common input to different classification models within the simulated cross-machine setting.

### 4.15 Structural Ambiguity of Fixed Frequency and Order Representations

The preceding experiments demonstrated that geometry-anchored features provided more consistent cross-machine diagnostic performance than fixed-frequency and order-domain representations. The structural fault-line audit described in Section 3.9 was conducted to examine the physical basis of these differences. Unlike the classification experiments, this analysis used only the bearing parameters listed in Table A1 and did not involve simulated noise, signal amplitudes, or trained diagnostic models.

Across the eight simulated machines, 24 characteristic fault lines were examined, comprising BPFO, BPFI, and twice the ball-spin frequency for each machine. On the order axis, 13 of these 24 lines (54%) lay within 3% of a characteristic line associated with a different fault on another machine, producing eight distinct cross-fault collision pairs (16 directed collision records). Increasing the tolerance to 5% raised the number of affected lines to 16 (67%). A comparable pattern was observed in the frequency domain, where 13 of the 24 lines occupied the same 10 Hz interval as a different-fault line from another machine.

Agreement between machines for the same fault was much less frequent. Among 168 directed same-fault, cross-machine line comparisons, only eight (4.8%) agreed within 3% on the order axis, while twelve (7.1%) occupied the same 10 Hz interval on the frequency axis. These results indicate that fixed spectral coordinates contained both substantial cross-fault ambiguity and limited consistency for the same fault mechanism across machines. The collision statistics are summarized in Table 18, while the machine-specific collision records, including the corresponding fault types, order values, and cross-machine matches, are provided in Table C1 (Appendix C).

**Table 18.** Structural audit of characteristic fault-line ambiguity

| Quantity | Order Axis | Frequency Axis |
|---|---|---|
| Fault lines colliding with a different-fault line on another machine (3% tolerance or same 10 Hz interval) | 13 of 24 (54%) | 13 of 24 (54%) |
| Distinct cross-fault collision pairs at 3% tolerance | 8 | Not applicable |
| Fault lines involved in collisions at 5% tolerance | 16 of 24 (67%) | Not applicable |
| Directed same-fault, cross-machine line comparisons agreeing within 3% or the same 10 Hz interval | 8 of 168 (4.8%) | 12 of 168 (7.1%) |
| Noise-free nearest-line accuracy under leave-one-machine-out evaluation (chance level = 0.333) | 0.333 | 0.583 |

Note: BPFO = ball pass frequency of outer race; BPFI = ball pass frequency of inner race; BSF = ball spin frequency.

The diagnostic consequences of this ambiguity were examined using the noise-free nearest-line classifier. Under leave-one-machine-out evaluation, the classifier achieved an accuracy of 0.333 on the order axis, corresponding to the chance level for three fault categories. The frequency-domain representation achieved 0.583, with machines M1, M3, and M5 classified correctly across all three fault lines. Although the frequency axis performed better in this particular configuration, neither fixed-axis representation consistently preserved the relationship between fault-line position and fault category across the eight machines.

The geometry-anchored representation provided a contrasting reference. Because its features were indexed by fault-specific characteristic orders calculated from each machine's geometry, the corresponding idealized identification accuracy was 1.0 by construction, as illustrated in Figure 12. This value represents the structural correspondence between feature coordinates and fault mechanisms, rather than the accuracy of a trained classifier on vibration signals.

The direction of these analytical results was consistent with the leave-one-machine-out classification results in Table 7, where order-domain and frequency-domain features achieved accuracies of 0.575 and 0.585, respectively. The numerical values are not directly comparable, however, because the trained classifiers also processed normal-condition windows, spectral amplitudes, noise, and interference components. The structural audit isolates the

ambiguity associated with fault-line positions, whereas the classification experiments measure the combined effect of the complete feature representations.

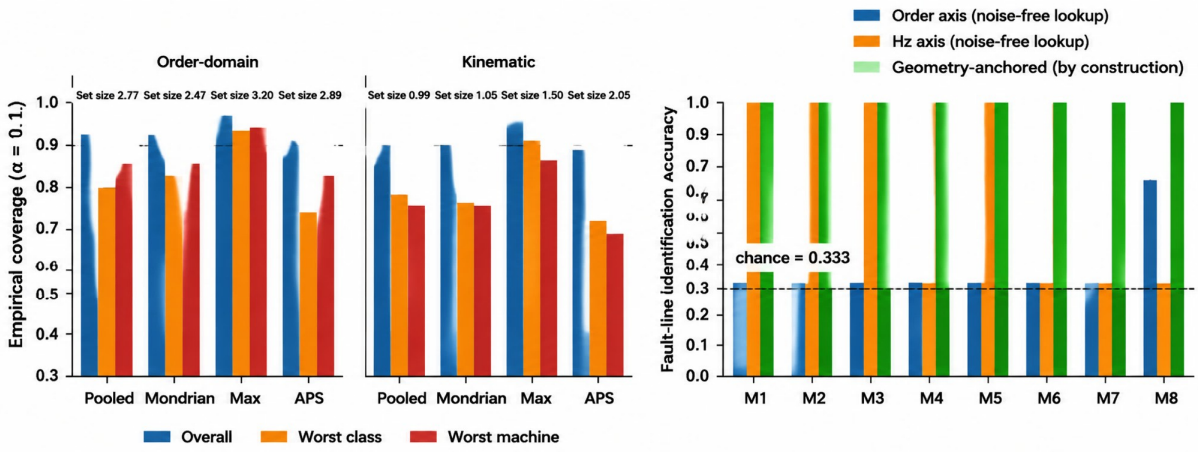



**Figure 12.** Noise-free fault-line identification under machine holdout

Note: BPFO = ball pass frequency of outer race; BPFI = ball pass frequency of inner race; BSF = ball spin frequency; APS = adaptive prediction sets; M1–M8 denote the eight simulated machines. Geometry-anchored accuracy is 1.0 by construction.

A further example occurred in machine M2, where the inner-race and ball-fault characteristic lines coincided at 4.8 shaft orders. In this configuration, the fundamental fault order alone could not distinguish the two mechanisms, making additional spectral information, including modulation sidebands, important for discrimination. M2 also exhibited the second-lowest geometry-anchored LMO-CP coverage, at 0.862, compared with 0.777 for M4 (Table 9). The coincident characteristic lines provide a possible physical explanation for the reduced coverage on M2, although the present audit does not isolate their contribution from other machine-dependent factors.

Taken together, the collision statistics and nearest-line analysis explain why a common frequency or order axis can produce conflicting fault information when bearing geometry changes. Geometry-anchored features address this structural difficulty by extracting spectral information at fault-specific locations determined for each machine. The analysis therefore provides a physical interpretation of the classification and calibration results obtained under machine-level distribution shift.

## 5 Synthesis of the Review and Simulation Findings

### 5.1 Assessment of the Research Hypotheses

The seven hypotheses developed from the literature synthesis were assessed against the controlled simulation results. The assessment considered the direction and consistency of the experimental findings, the paired statistical comparisons, and the analytical investigations reported in Section 4. A hypothesis was classified as supported when the simulated evidence consistently agreed with its proposed relationship, and as partly supported when the results depended on the feature representation, calibration procedure, or experimental condition.

As explained in Section 2.7, the hypotheses were formalized after the initial experiments had been completed. Their assessment therefore provides a structured interpretation of the simulation findings rather than a set of preregistered confirmatory tests. Table 19 summarizes the evidence associated with each hypothesis and the conditions under which the corresponding conclusions apply.

The hypothesis assessments reveal two closely connected findings. First, the choice of feature representation determined whether fault-related information remained discriminative when the target machine differed from those used for training. Second, the calibration procedure determined how uncertainty was expressed under that shift, but it could not compensate for a representation that failed to distinguish fault categories. This distinction explains why machine-level calibration produced prediction sets close to nominal coverage for all three representations, whereas only the geometry-anchored representation maintained small sets.

The results also distinguish aggregate calibration from reliability on individual machines. Although pooled LMO-CP achieved average coverage close to the nominal level, the worst-machine results varied substantially. Conservative calibration reduced this variation at the cost of larger prediction sets. These observations motivate the analytical examination of machine-level coverage in the following section.

### 5.2 Analytical Interpretation of Machine-Level Coverage

The experiments in Sections 4.4 and 4.12 showed that average coverage across machines could approach the nominal level while coverage on individual machines remained substantially lower. A rank-based argument provides

a complementary interpretation of the conservative LMO-max threshold and clarifies the role of the number of available source machines.

**Table 19.** Assessment of the seven research hypotheses

| ID | Assessment | Principal Evidence | Scope of Interpretation |
|---|---|---|---|
| H1 | Supported | Random-window accuracy of 0.834 and 0.866 decreased to 0.585 and 0.575 under machine holdout; recording-level accuracy remained 0.847 for order features (Section 4.1). | Eight simulated machines. |
| H2 | Supported | Kinematic accuracy of 0.908 versus 0.585 and 0.575; paired gain of 0.333 over order features (95% CI: 0.313–0.353); fundamental features reached 0.877; 13 of 24 fault lines collided (Sections 4.2, 4.9, 4.10, and 4.15). | Feature construction shares the simulator's kinematic relationships; accuracy decreased to 0.84 under model mismatch. |
| H3 | Partly supported | Order-domain coverage increased from 0.642 to 0.928 with LMO-CP; kinematic naive coverage of 0.879 was not statistically distinguishable from nominal (Sections 4.3 and 4.9). | Stronger evidence for machine-dependent representations. |
| H4 | Supported | Kinematic prediction sets averaged 0.99 classes, compared with 2.77–2.96 for fixed-axis features; mean set size increased to 1.89 at low SNR (Sections 4.3 and 4.7). | Evaluated signal-generation and noise conditions. |
| H5 | Supported | Worst-machine coverage increased from 0.751 with pooled LMO-CP to 0.866 with LMO-max, at an additional 0.51 classes per set; Mondrian calibration did not improve worst-class coverage (Sections 4.4 and 4.12). | Five seeds and 192 known-class windows per machine. |
| H6 | Supported within the evaluated configurations | Weighted effective sample size was 4%–14%; kinematic softmax AUROC was 0.831; empty sets flagged 8.7% of unknown faults; Bayes cost was 0.289 versus 0.30 for universal inspection (Sections 4.3, 4.8, and 4.11). | One unknown-fault condition and one cost matrix. |
| H7 | Partly supported | Under model mismatch, kinematic accuracy decreased from 0.90 to 0.84 and worst-machine coverage from 0.77 to 0.61, while average coverage remained near 0.90; logistic regression and random forest reproduced the main representation effect (Sections 4.13 and 4.14). | Three seeds and the evaluated classifier families. |

Note: ID = Identifier; LMO-CP = leave-machine-out conformal prediction; CI = confidence interval; SNR = signal-to-noise ratio; AUROC = area under the receiver operating characteristic curve.

Let $q_j$ denote the population $(1-\alpha)$-quantile of the nested nonconformity scores associated with machine $j$. Consider $K$ source machines and one target machine, indexed by $j = K+1$. The conservative calibration threshold, $q_{\max}$, is the largest of the source-machine quantiles.

Proposition 1 (idealized machine-level quantile ordering). Assume that (A1) the $K+1$ machine-level quantiles are exchangeable and have no ties; (A2) each machine contributes its full nonconformity-score distribution, so that the corresponding quantile is a population quantity; and (A3) the score distribution of the deployed predictor on the target machine agrees with the distribution represented by its nested calibration scores. Under these assumptions,

$$P\left(q_{K+1} \leq q_{\max}\right) = \frac{K}{K+1} \tag{1}$$

Whenever the target-machine quantile does not exceed the conservative threshold, prediction sets constructed using that threshold cover at least a proportion $1-\alpha$ of the target-machine score distribution.

The result follows from the ranks of exchangeable machine-level quantiles. Under the stated assumptions, the target-machine quantile is equally likely to occupy any of the $K+1$ ranks. It exceeds every source-machine quantile only when it occupies the largest rank, an event with probability $1/(K+1)$. In all other cases, the conservative threshold is at least as large as the target-machine population quantile, establishing the stated coverage relationship.

For source-machine counts from two to seven, the corresponding probabilities are 0.667, 0.750, 0.800, 0.833, 0.857, and 0.875. Under the idealized assumptions, achieving a probability of at least $1 - \delta$ requires $K \geq \frac{1-\delta}{\delta}$. Accordingly, nine source machines are required for a target probability of 0.90, and nineteen for 0.95. With seven source machines, the corresponding value is 0.875.

This argument concerns the relative ranks of population-level machine quantiles. It does not establish a finite-sample coverage theorem for the implemented LMO-CP procedure, in which each machine contributes a finite collection of potentially dependent windows and the final classifier differs from the nested calibration models. The pooled calibration threshold also differs from the conservative threshold because it combines machine-level scores rather than selecting the largest machine-specific quantile.

The empirical results were consistent with the direction of the rank argument. With geometry-anchored features, LMO-max increased worst-machine coverage from 0.751 to 0.866, while mean prediction-set size increased from 0.99 to 1.50 classes (Table 15). In addition, 95% of machine–seed cases achieved coverage of at least 0.85. This empirical criterion differs from the nominal 0.90 coverage level considered in Proposition 1, so the result should not be interpreted as a direct verification of the proposition. Rather, the analysis explains why conservative machine-level calibration can improve coverage across heterogeneous machines and why the number of independent source machines matters when interpreting reliability at the machine level.

### 5.3 Relationship to Previous Research

The findings can be understood in relation to several established results in cross-domain diagnosis and predictive uncertainty. The substantial reduction in fixed-axis classification accuracy under machine holdout resembles the cross-dataset performance deterioration reported for bearing diagnosis between CWRU and SEU [10]. Similarly, the loss of empirical coverage under machine shift is consistent with broader observations that predictive uncertainty becomes less reliable as the test distribution departs from the training distribution [33].

The behavior of weighted conformal calibration provides a further connection to domain-adaptation theory. In the experiments, the domain classifier distinguished source and target machines with high accuracy, leaving normalized effective sample sizes of only 4%–14% after weighting. This behavior is consistent with the difficulty of transferring predictions when source and target feature distributions differ substantially, as reflected in theoretical analyses of domain adaptation [29]. The observed variation in coverage between individual machines is also consistent with the limitations of distribution-free conditional coverage [39].

The present study extends these observations by examining feature representation and calibration within the same machine-level evaluation framework. Geometry-anchored features preserved the physical association between feature coordinates and bearing fault mechanisms, while nested machine holdouts produced calibration scores under a transfer setting closer to the target evaluation. Their joint assessment showed that improved empirical coverage was most informative when the representation also maintained diagnostic discrimination. The resulting distinction between coverage and prediction-set size provides a basis for evaluating uncertainty-aware diagnosis beyond classification accuracy alone.

### 5.4 Reporting Requirements for Cross-Machine Diagnosis

The methodological and experimental findings were consolidated into a reporting checklist for studies evaluating bearing diagnosis on previously unseen machines. The twelve reporting requirements listed in Table 20 address the principal sources of variation identified in the review and examined in the controlled experiments. Its purpose is to make the evaluation setting, calibration assumptions, and operational interpretation of diagnostic predictions explicit.

**Table 20.** Reporting checklist for cross-machine bearing fault diagnosis

| No. | Reporting Requirement | Supporting Evidence |
|---|---|---|
| 1 | Identify the evaluation unit excluded from training: window, recording, operating condition, machine, or dataset. | H1; Table 1. |
| 2 | Report accuracy under both random splitting and machine-level holdout using the same classifier. | H1; Section 4.1. |
| 3 | Specify the physical and operational information required by each feature representation, including bearing geometry and shaft speed. | H2; Sections 4.6 and 4.13. |
| 4 | Describe the origin of calibration observations and distinguish known-machine, nested machine-holdout, and labelled target-machine calibration. | H3; Section 4.3. |
| 5 | Report coverage for individual held-out machines and identify the worst-machine result. | H5; Section 4.4. |

| | | |
|---|---|---|
| 6 | Report mean prediction-set size alongside empirical coverage. | H4; Section 4.3. |
| 7 | State the number of independent source machines and discuss the implications of the idealized $K/(K+1)$ rank relationship. | H5; Section 5.2. |
| 8 | Report class-specific coverage, particularly for fault categories associated with substantial maintenance consequences. | H5; Section 4.12. |
| 9 | Include at least one unknown-fault condition and report the proportion incorrectly classified as normal. | H6; Section 4.11. |
| 10 | Specify the decision-cost matrix, examine its influence, and compare diagnostic policies with universal inspection. | H6; Section 4.8. |
| 11 | Evaluate sensitivity to departures from the physical assumptions used in feature construction. | H7; Section 4.13. |
| 12 | Provide the implementation code, random seeds, and aggregated experimental results. | Data availability. |

Note: H1–H7 = research hypotheses 1–7.

These reporting items distinguish model performance under a specified data partition from evidence of transfer to unfamiliar equipment. They also encourage separate reporting of classification accuracy, prediction-set coverage, and the diagnostic information retained within the prediction sets. For industrial applications, this distinction is particularly important because a statistically well-covered prediction set may still contain too many candidate faults to support a specific maintenance decision.

## 6 Discussion

### 6.1 Principal Findings

The experiments revealed a clear distinction between diagnostic discrimination and prediction reliability. Geometry-anchored features reduced the dependence of fault classification on machine-specific spectral coordinates, whereas machine-level calibration addressed the uncertainty associated with transferring predictions to unseen machines. The evaluation protocol provided a separate means of determining whether these effects persisted when complete machines were excluded from model development. Together, these three elements defined the experimental assessment of cross-machine diagnostic reliability.

The difference between random splitting and machine-level holdout was particularly pronounced for fixed-axis representations. Random window splitting overestimated diagnostic accuracy by as much as 0.29, while recording-level separation did not fully expose the deterioration associated with an unseen machine. Geometry-anchored features, by contrast, maintained substantially higher accuracy across the evaluation protocols. The ablation experiments in Section 4.10 showed that the characteristic fault-order amplitudes accounted for much of this advantage, while the structural audit in Section 4.15 identified the ambiguity introduced when similar frequency or order coordinates corresponded to different fault mechanisms across machines.

The calibration results revealed a complementary relationship. For machine-dependent representations, calibration using observations from previously encountered machines produced prediction sets with coverage below the nominal target. LMO-CP improved average empirical coverage by generating calibration scores from models that had not encountered the corresponding calibration machine. However, the improvement in coverage did not necessarily produce informative diagnostic predictions. With fixed-axis representations, the calibrated sets contained nearly three of the four known classes on average, whereas geometry-anchored features produced sets containing approximately one class. This distinction indicates that the effectiveness of uncertainty calibration depends partly on whether the underlying representation preserves fault discrimination across machines.

Machine-specific coverage introduced a further consideration. Although pooled LMO-CP achieved average coverage close to the nominal level, individual machines exhibited substantial differences. Conservative calibration increased worst-machine coverage but required larger prediction sets. The analytical argument in Section 5.2 also demonstrated how the number of independent source machines constrains machine-level coverage statements under idealized exchangeability assumptions. These findings distinguish empirical average coverage from reliability on a particular machine and show why both quantities should be reported when assessing cross-machine diagnosis.

The unknown-fault and decision analyses identified a separate challenge. Approximately 70% of the mechanical-looseness windows were classified as normal, despite the improved discrimination of known faults achieved by geometry-anchored features. The evaluated novelty-detection and abstention procedures did not consistently reduce operational costs under the specified decision assumptions. Thus, reliable classification among known fault categories and recognition of previously unseen fault mechanisms represent distinct requirements of an industrial diagnostic system.

The combined findings indicate that geometry-anchored features and machine-level conformal calibration address complementary aspects of cross-machine diagnosis. Physically indexed features support fault discrimination across bearing configurations, while machine-level calibration provides an empirical means of assessing prediction-set coverage under machine shift without labelled target-machine observations. Their joint evaluation offers a more complete account of diagnostic performance than classification accuracy alone. Table 21 distinguishes the findings established within the controlled simulation from the corresponding requirements for experimental validation.

**Table 21.** Principal simulation findings and requirements for experimental validation

| Finding | Evidence From the Controlled Simulation | Requirements for Experimental Validation |
|---|---|---|
| Evaluation protocols influence apparent cross-machine accuracy | Random splitting overestimated accuracy by 0.25 and 0.29 for the fixed-axis representations (Section 4.1). | Quantify the corresponding differences using independent experimental bearing recordings. |
| Geometry-anchored features support cross-machine fault discrimination | Accuracy reached 0.908, compared with approximately 0.58 for fixed-axis features; the advantage persisted under model mismatch and classifier substitution. | Examine real slip, load-dependent contact angles, structural transmission paths, and sensor characteristics. |
| Machine-level calibration improves average empirical coverage | LMO-CP achieved coverage of approximately 0.90–0.93, compared with 0.64–0.88 for naive calibration. | Assess coverage across independently sampled machines with different operating and measurement conditions. |
| Machine-specific coverage varies despite adequate average coverage | Worst-machine coverage was 0.751 with pooled LMO-CP and 0.866 with conservative calibration. | Evaluate machine-level reliability using a larger population of independent machines. |
| Prediction-set informativeness depends on feature representation | Mean set size was 0.99 classes for geometry-anchored features, compared with 2.77–2.96 for fixed-axis representations. | Examine prediction-set size under real noise, variable loading, and heterogeneous sensor configurations. |
| Unknown-fault recognition and cost-sensitive decisions require separate evaluation | Empty sets flagged 8.7% of unknown-fault windows, approximately 70% were classified as normal, and cost-sensitive automation showed no consistent cost reduction. | Include additional unknown-fault categories, realistic fault prevalence, and maintenance cost structures. |

Note: LMO-CP = leave-machine-out conformal prediction.

### 6.2 Implications for Industrial Condition Monitoring

The results have several implications for the development and evaluation of intelligent bearing-monitoring systems. The first concerns the selection of evaluation data. When a diagnostic model is intended for equipment not represented during training, random window splitting or recording-level separation alone does not establish cross-machine generalization. Machine-level holdout provides a more appropriate assessment because it separates the equipment used for model development from that used for testing. This distinction is particularly important when bearing geometry, sensor placement, and structural vibration transmission differ between source and target machines.

The second implication concerns feature construction. Where bearing geometry and rotational speed are available, characteristic fault frequencies can be incorporated directly into the representation rather than leaving their relationship to fault categories for a classifier to infer from a limited number of machines. The sensitivity experiments in Section 4.6 showed that the accuracy of this representation depended on the quality of the geometric and speed information supplied to the feature extractor. Consequently, the availability and precision of these parameters should be considered part of the diagnostic system specification.

The third implication concerns the interpretation of uncertainty. Prediction-set coverage should be reported together with set size and machine-specific performance. A set containing several fault classes may satisfy an aggregate coverage target while providing limited guidance for maintenance decisions. Similarly, average coverage close to the nominal level may conceal substantially lower coverage on individual machines. The choice between pooled and conservative machine-level calibration should therefore reflect the relative importance of diagnostic

specificity and coverage on atypical equipment. In applications where missed faults carry substantial consequences, a conservative threshold may be preferable, provided that the resulting increase in prediction-set size is acceptable.

The computational requirements of the evaluated calibration procedure were modest at the scale of the simulated experiment. For $K$ source machines, LMO-CP required $K + 1$ classifier fits, including the final model trained on all source machines. On a single CPU core, one gradient-boosting fit required approximately 0.24 s with geometry-anchored features and 0.64 s with order-domain features. The complete calibration procedure required approximately 2 s and 5 s, respectively, for the evaluated eight-machine configuration, compared with one or two model fits for conventional split calibration. These measurements indicate that the additional fitting cost was manageable for the selected classifiers and dataset size, although computational requirements will depend on model complexity and the number of machines in larger applications.

### 6.3 Applicability to Industrial Bearing Systems

The geometry-anchored representation requires information about bearing geometry and shaft rotational speed. The geometric parameters include the number of rolling elements, the ratio of rolling-element diameter to pitch diameter, and the contact angle. In industrial equipment, these parameters may be obtained from bearing designations, manufacturer catalogues, maintenance records, or machine drawings. Once the bearing configuration is identified, the characteristic fault orders can be calculated using the kinematic relationships described in Section 3.2 [13].

Rotational speed may be measured using a tachometer or keyphasor, obtained from the operating information of a variable-speed drive, or estimated from motor operating parameters. Where direct measurements are unavailable, shaft speed may also be estimated from the vibration spectrum. The accuracy of these inputs becomes particularly important when bearings have been replaced without updated maintenance records, when the installed bearing type is unknown, or when mechanical transmission elements separate the motor from the monitored bearing.

The parameter-sensitivity experiments provide an indication of the required precision. A 1% speed error produced an accuracy of 0.867, close to the 0.870 obtained with exact inputs, whereas a 2% error reduced accuracy to 0.834. Errors in the bearing diameter ratio had a greater effect: accuracy decreased to 0.823 with a 3% error and to 0.807 with a 6% error. The largest geometric perturbation, combining a 10% diameter-ratio error with a $5^\circ$ contact-angle error, reduced accuracy to 0.762. Across these conditions, average LMO-CP coverage remained close to 0.90, while mean prediction-set size increased from 1.10 to as much as 1.61 classes (Table 10).

These results indicate that geometry-anchored diagnosis is most suitable when the installed bearing can be identified, its geometric parameters are documented, and shaft speed is measured or estimated with sufficient precision. The approach also requires vibration measurements in which characteristic bearing fault components can be distinguished from other mechanical sources. Gear-mesh frequencies, belt-related vibration, structural resonances, and overlapping spectral components may complicate the interpretation of the envelope spectrum.

Real bearing systems introduce additional effects, including load-dependent contact angles, rolling-element slip, manufacturing tolerances, and speed variation within the analysis window. The model-mismatch experiment in Section 4.13 examined controlled departures from nominal fault orders and showed that geometry-anchored features retained a classification advantage under the specified perturbations. These results provide an initial indication of robustness to kinematic mismatch, while the behavior under independently measured industrial vibration remains a separate validation requirement.

When bearing geometry is unavailable, the characteristic fault orders cannot be calculated with the information required by the present feature extractor. In such cases, alternative representations or methods for estimating the missing physical parameters are necessary. The experiments also showed that machine-level calibration maintained average empirical coverage under parameter perturbations by producing larger prediction sets. Thus, calibration and feature construction should be treated as complementary components rather than interchangeable solutions to inaccurate physical information.

### 6.4 Study Limitations

The findings should be interpreted in relation to the controlled simulation design and the assumptions used to generate the vibration signals. The simulator incorporated established bearing fault-frequency relationships, and the geometry-anchored features were constructed using the same underlying kinematic principles. This correspondence allowed the influence of machine geometry to be examined directly, but it also favored representations that explicitly used those relationships. The geometry-perturbation and model-mismatch experiments assessed departures from the nominal assumptions while retaining the principal impulse-response structure of the simulated signals.

The evaluation involved eight simulated machines, with seven source machines available in each leave-one-machine-out experiment. Although this design permitted direct comparisons across machine configurations, the number of independent machine-level units was limited. The reported LMO-CP coverage values therefore describe empirical performance under the evaluated simulation conditions. They do not constitute a general finite-sample

guarantee for individual machines or establish the exchangeability of physical machines encountered in industrial applications.

The literature synthesis was structured around selected methodological themes rather than a registered systematic-review protocol. Its purpose was to identify the relationships between evaluation practice, feature representation, distribution shift, and uncertainty calibration. The seven research hypotheses were formalized after the initial experiments, and their assessment should consequently be understood as an exploratory interpretation of the evidence.

The classifier comparison covered histogram-based gradient boosting, logistic regression, random forest, and a small multilayer perceptron. These models provided a basis for examining whether the observed representation effects depended on the classifier family. However, the experiments did not include convolutional networks, transformer-based architectures, or adversarial domain-adaptation models under a common machine-level evaluation protocol. Model-specific hyperparameter optimization was also outside the experimental design.

The unknown-fault analysis considered mechanical looseness as the sole fault category excluded from training. The cost-sensitive decision analysis used one specified cost matrix, while the robustness experiments involved a limited number of random seeds. The resulting observations characterize the selected fault condition, decision assumptions, and perturbation settings rather than the full range of industrial fault mechanisms and maintenance requirements.

Dependence between observations is another consideration. Windows originating from the same recording shared operating and signal-generation parameters, so the number of windows should not be interpreted as the number of independent machine-level observations. The bootstrap intervals and statistical comparisons quantified variation within the simulated evaluation design, but their interpretation remains tied to the available machines and random realizations. Similarly, the fault-line collision audit examined selected bearing geometries under deterministic conditions, while the model-mismatch experiments retained the principal structure of the simulator.

These considerations define the contribution as a controlled methodological investigation of geometry-anchored features, machine-level conformal calibration, and diagnostic performance under specified forms of machine shift. They also establish the experimental conditions that should be addressed when extending the framework to independently collected bearing measurements.

### 6.5 Future Research Directions

The findings provide a basis for extending the framework from controlled simulation toward experimental and industrial bearing monitoring. Five directions are particularly relevant.

(1) Experimental validation and cross-dataset benchmarking

The first priority is to evaluate the proposed framework using measured bearing vibration signals from datasets such as CWRU, Paderborn, and XJTU-SY. Evaluation should distinguish complete-machine holdout from operating-condition transfer and maintain separation between training and testing recordings. Where available, bearing geometry and operating-condition metadata should be incorporated into feature construction. Comparisons with conventional classifiers, deep-learning models, and representative domain-adaptation methods would allow the observed effects of physically indexed features and machine-level calibration to be examined under measured signal variability.

(2) Evaluation of machine-level coverage across larger equipment populations

The number-of-machines analysis should be extended to include a larger and more heterogeneous population of independent machines. In addition to average coverage, the evaluation should report the distribution of machine-specific coverage, the proportion of machines reaching the nominal target, and the relationship between source-machine diversity and prediction-set size. Such experiments would provide a stronger empirical basis for interpreting conservative calibration and the influence of the number of available machines.

(3) Cost-sensitive conformal decision-making

Future work should investigate the integration of machine-level calibration with conformal risk control [40], allowing diagnostic decisions to be evaluated through operational losses rather than prediction-set coverage alone. Relevant outcomes include missed faults, unnecessary inspections, and maintenance delays. Comparisons across alternative cost matrices and fault prevalences would clarify the circumstances under which uncertainty-aware classification provides a measurable operational advantage.

(4) Unknown-fault recognition under machine shift

The open-set evaluation should be extended beyond mechanical looseness to include additional fault mechanisms and combinations of faults absent from training. Dedicated novelty-detection methods could be examined alongside conformal prediction sets and selective classification. The evaluation should distinguish distribution shift caused by unfamiliar equipment from changes caused by previously unseen fault categories, since both may influence diagnostic confidence.

(5) Hierarchical and group-aware conformal calibration

Further research should examine calibration procedures that explicitly account for the hierarchical organization of vibration observations within recordings and machines. Machine-stratified, hierarchical, and group-conditional methods could be compared with LMO-CP in terms of average coverage, machine-specific reliability, calibration-data requirements, and sensitivity to heterogeneous operating conditions. Particular attention should be given to the assumptions required for finite-sample guarantees when the number of independent machines is small.

Together, these directions establish a progression from experimental validation and larger machine-level evaluations toward uncertainty-aware maintenance decisions and calibration procedures that account for the structure of industrial monitoring data.

## 7 Conclusions

This study investigated cross-machine bearing fault diagnosis by combining geometry-anchored feature representation with machine-level conformal calibration. A structured literature review identified five methodological gaps and seven research hypotheses, which were examined using controlled simulations of eight machines with different bearing geometries, operating speeds, structural resonances, and noise characteristics.

The geometry-anchored representation achieved a leave-one-machine-out classification accuracy of 0.908, compared with 0.585 and 0.575 for frequency-domain and order-domain features. Random window splitting overestimated the accuracy of the fixed-axis representations by as much as 0.29. Feature ablation and the structural fault-line audit showed that the improvement was associated with extracting spectral information at fault-specific kinematic orders. Across the simulated machines, 13 of 24 characteristic fault lines overlapped with lines associated with different faults, explaining an important source of ambiguity in fixed-axis representations.

Machine-level conformal calibration improved average empirical coverage under machine shift. At a nominal coverage level of 0.90, LMO-CP achieved average coverage between approximately 0.90 and 0.93 across the evaluated representations. However, prediction-set informativeness depended strongly on feature construction. Geometry-anchored features produced sets containing approximately one class on average, compared with nearly three classes for the fixed-axis alternatives.

Coverage also varied between individual machines. With geometry-anchored features, worst-machine coverage was 0.751 under pooled LMO-CP and increased to 0.866 under conservative calibration, accompanied by larger prediction sets. The idealized rank argument established the relationship between the number of source machines and the probability of attaining nominal machine-level coverage under its stated assumptions. The robustness experiments further showed that the geometry-anchored representation retained its classification advantage under controlled kinematic mismatch and across alternative classifiers, although worst-machine coverage decreased under more demanding signal conditions.

The unknown-fault and decision analyses identified separate requirements for industrial application. Approximately 70% of unknown mechanical-looseness windows were classified as normal, while the evaluated novelty-detection and abstention procedures did not produce consistent operational improvements under the selected conditions. These findings distinguish reliable classification among known faults from the broader requirements of unknown-fault recognition and maintenance decision-making.

The study establishes a controlled evaluation framework in which diagnostic accuracy, prediction-set coverage, machine-specific reliability, and prediction-set size are assessed together. Its principal contribution is the joint examination of physically indexed fault features and calibration based on machine-level holdouts. The resulting evidence supports the use of machine-level evaluation and physically meaningful feature construction when developing diagnostic models for unfamiliar equipment. Experimental validation using independent bearing measurements remains necessary to establish how these findings translate to industrial operating conditions.

### Author Contributions

Conceptualization, A. Karkadakattil and A. Kalath; methodology, A. Karkadakattil and A. Kalath; software, A. Karkadakattil; validation, A. Karkadakattil and A. Kalath; formal analysis, A. Karkadakattil and A. Kalath; investigation, A. Karkadakattil; data curation, A. Karkadakattil; writing—original draft preparation, A. Karkadakattil; writing—review and editing, A. Karkadakattil and A. Kalath; visualization, A. Karkadakattil; supervision, A. Kalath. All authors have read and agreed to the published version of the manuscript.

### Data Availability

The data used to support the research findings are available from the corresponding author upon request.

### Conflicts of Interest

The authors declare no conflicts of interest.

## Declaration on the Use of Generative AI and AI-assisted Technologies

AI-assisted tools were used to support language editing and improve manuscript clarity. The authors reviewed and verified the manuscript and take responsibility for its final content.

## References


[1] 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," *Mech. Syst. Signal Process.*, vol. 138, p. 106587, 2020. https://doi.org/10.1016/j.ymssp.2019.106587

[2] S. Zhang, S. Zhang, B. Wang, and T. G. Habetler, "Deep learning algorithms for bearing fault diagnostics—A comprehensive review," *IEEE Access*, vol. 8, pp. 29 857–29 881, 2020. https://doi.org/10.1109/ACCESS.2020.2972859

[3] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," *Mech. Syst. Signal Process.*, vol. 64–65, pp. 100–131, 2015. https://doi.org/10.1016/j.ymssp.2015.04.021

[4] J. Hendriks, P. Dumond, and D. A. Knox, "Towards better benchmarking using the CWRU bearing fault dataset," *Mech. Syst. Signal Process.*, vol. 169, p. 108732, 2022. https://doi.org/10.1016/j.ymssp.2021.108732

[5] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World (1st ed.)*. New York, NY, USA: Springer, 2005. https://doi.org/10.1007/b106715

[6] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494–591, 2023. https://doi.org/10.1561/2200000101

[7] R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas, "Conformal prediction under covariate shift," in *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2019.

[8] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *Ann. Stat.*, vol. 51, no. 2, pp. 816–845, 2023. https://doi.org/10.1214/23-AOS2276

[9] Z. Zhao, Q. Zhang, X. Yu, C. Sun, S. Wang, R. Yan, and X. Chen, "Applications of unsupervised deep transfer learning to intelligent fault diagnosis: A survey and comparative study," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–28, 2021. https://doi.org/10.1109/TIM.2021.3116309

[10] M. Mei, Y. Li, Y. Qian, and Z. Jia, "Calibrated prediction set in fault detection with risk guarantees via significance tests," *arXiv Preprint*, p. arXiv:2508.01208, 2025. https://doi.org/10.48550/arXiv.2508.01208

[11] V. Vovk, "Cross-conformal predictors," *Ann. Math. Artif. Intell.*, vol. 74, pp. 9–28, 2015. https://doi.org/10.1007/s10472-013-9368-4

[12] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," *Ann. Stat.*, vol. 49, no. 1, pp. 486–507, 2021. https://doi.org/10.1214/20-AOS1965

[13] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics—A tutorial," *Mech. Syst. Signal Process.*, vol. 25, no. 2, pp. 485–520, 2011. https://doi.org/10.1016/j.ymssp.2010.07.017

[14] J. Antoni, "The spectral kurtosis: A useful tool for characterising non-stationary signals," *Mech. Syst. Signal Process.*, vol. 20, no. 2, pp. 282–307, 2006. https://doi.org/10.1016/j.ymssp.2004.09.001

[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.

[16] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Computer Vision—ECCV 2016 Workshops*. Cham, Switzerland: Springer, 2016, pp. 443–450. https://doi.org/10.1007/978-3-319-49409-8_35

[17] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, "Adaptive batch normalization for practical domain adaptation," *Pattern Recognit.*, vol. 80, pp. 109–117, 2018. https://doi.org/10.1016/j.patcog.2018.03.005

[18] Z. Chen, A. Mauricio, W. Li, and K. Gryllias, "A deep learning method for bearing fault diagnosis based on cyclic spectral coherence and convolutional neural networks," *Mech. Syst. Signal Process.*, vol. 140, p. 106683, 2020. https://doi.org/10.1016/j.ymssp.2020.106683

[19] K. Xu, X. Kong, Q. Wang, S. Yang, N. Huang, and J. Wang, "A bearing fault diagnosis method without fault data in new working condition combined dynamic model with deep learning," *Adv. Eng. Inform.*, vol. 54, p. 101795, 2022. https://doi.org/10.1016/j.aei.2022.101795

[20] T. E. Alam, M. M. Ahsan, and S. Raman, "Physics-informed multimodal bearing fault classification under variable operating conditions using transfer learning," *arXiv Preprint*, p. arXiv:2508.07536, 2025. https://doi.org/10.48550/arXiv.2508.07536

[21] X. M. Liu, R. M. Zhang, J. P. Li, Y. F. Xu, and K. Li, "A motor bearing fault diagnosis model based on multi-adversarial domain adaptation," *Sci. Rep.*, vol. 14, p. 29078, 2024. https://doi.org/10.1038/s41598-024-80743-1

[22] J. Che, L. Fang, Q. Ma, G. Yu, X. Sun, and X. Zhu, "An intelligent bearing fault transfer diagnosis method based on improved domain adaption," *Entropy*, vol. 27, no. 11, p. 1178, 2025. https://doi.org/10.3390/e27111178

[23] I. Gibbs, J. J. Cherian, and E. J. Candès, "Conformal prediction with conditional guarantees," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 87, no. 4, pp. 1100–1126, 2025. https://doi.org/10.1093/jrsssb/qkaf008

[24] K. Kasa, Z. Zhang, H. Yang, and G. W. Taylor, "Adapting prediction sets to distribution shifts without labels," in *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence*, Rio de Janeiro, Brazil, 2025, pp. 1990–2010.

[25] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification," in *Proceedings of the 3rd European Conference of the Prognostics and Health Management Society*, Bilbao, Spain, 2016. https://doi.org/10.36001/phme.2016.v3i1.1577

[26] 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. https://doi.org/10.1109/TR.2018.2882682

[27] C. Li, S. Zhang, Y. Qin, and E. Estupinan, "A systematic review of deep transfer learning for machinery fault diagnosis," *Neurocomputing*, vol. 407, pp. 121–135, 2020. https://doi.org/10.1016/j.neucom.2020.04.045

[28] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Trans. Ind. Inform.*, vol. 15, no. 4, pp. 2446–2455, 2019. https://doi.org/10.1109/TII.2018.2864759

[29] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, pp. 151–175, 2010. https://doi.org/10.1007/s10994-009-5152-4

[30] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, 2023. https://doi.org/10.1109/TPAMI.2022.3195549

[31] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *arXiv Preprint*, p. arXiv:2007.01434, 2020. https://arxiv.org/abs/2007.01434

[32] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 1321–1330.

[33] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2019.

[34] M. Sadinle, J. Lei, and L. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," *J. Am. Stat. Assoc.*, vol. 114, no. 525, pp. 223–234, 2019. https://doi.org/10.1080/01621459.2017.1395341

[35] Y. Romano, M. Sesia, and E. Candès, "Classification with valid and adaptive coverage," in *Advances in Neural Information Processing Systems*, Virtual Event, 2020.

[36] V. Vovk, D. Lindsay, I. Nouretdinov, and A. Gammerman, "Mondrian confidence machine," On-line Compression Modelling Project, Working Paper 4, 2003. https://alrw.cs.rhul.ac.uk/old/04.pdf

[37] A. Podkopaev and A. Ramdas, "Distribution-free uncertainty quantification for classification under label shift," in *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, Virtual Event, 2021, pp. 844–853.

[38] I. Gibbs and E. J. Candès, "Adaptive conformal inference under distribution shift," in *Advances in Neural Information Processing Systems*, Virtual Event, 2021, pp. 1660–1672.

[39] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *J. Am. Stat. Assoc.*, vol. 113, no. 523, pp. 1094–1111, 2018. https://doi.org/10.1080/01621459.2017.1307116

[40] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," in *Proceedings of the International Conference on Learning Representations*, Vienna, Austria, 2024.

[41] S. Farouq, S. Byttner, M. R. Bouguelia, and H. Gadd, "A conformal anomaly detection based industrial fleet monitoring framework: A case study in district heating," *Expert Syst. Appl.*, vol. 201, p. 116864, 2022. https://doi.org/10.1016/j.eswa.2022.116864

[42] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv Preprint*, p. arXiv:1610.02136, 2016. https://arxiv.org/abs/1610.02136

[43] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, Montreal, Canada, 2018.

[44] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 4878–4887.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer,

R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

## Appendix

### Appendix A. Simulated Machines

**Table A1.** Simulated machine parameters and characteristic fault orders

| Machine | Balls $n$ | $d/D$ | $\varphi$ (deg) | Shaft (Hz) | $f_s$ (Hz) | Resonance (Hz) | BPFO | BPFI | $2 \times$ BSF | FTF |
|---|---|---|---|---|---|---|---|---|---|---|
| M1 | 9 | 0.25 | 0 | 25 | 12,000 | 3,200 | 3.375 | 5.625 | 3.750 | 0.375 |
| M2 | 8 | 0.20 | 0 | 30 | 12,000 | 2,800 | 3.200 | 4.800 | 4.800 | 0.400 |
| M3 | 12 | 0.15 | 0 | 20 | 25,600 | 6,000 | 5.100 | 6.900 | 6.517 | 0.425 |
| M4 | 10 | 0.30 | 0 | 35 | 16,384 | 4,000 | 3.500 | 6.500 | 3.033 | 0.350 |
| M5 | 7 | 0.28 | 15 | 40 | 20,000 | 5,000 | 2.553 | 4.447 | 3.310 | 0.365 |
| M6 | 11 | 0.18 | 10 | 28 | 16,384 | 3,500 | 4.525 | 6.475 | 5.381 | 0.411 |
| M7 | 8 | 0.35 | 0 | 45 | 12,000 | 2,500 | 2.600 | 5.400 | 2.507 | 0.325 |
| M8 | 9 | 0.22 | 20 | 22 | 25,600 | 7,000 | 3.570 | 5.430 | 4.351 | 0.397 |

Note: BPFO = ball pass frequency of outer race; BPFI = ball pass frequency of inner race; BSF = ball spin frequency; FTF = fundamental train frequency.

### Appendix B. Implementation Notes

Window length 16,384 samples. Candidate resonance bands: 14 centres between 6% and 42% of the sampling rate, each 8% wide. Per-machine noise standard deviation 0.9 to 1.5 (arbitrary units), gain 0.7 to 1.2, interference rate 6 to 15 impulses per second. Recording severity is $\exp(U(\ln 0.9, \ln 2.6))$ divided by a difficulty factor (0.5 in the reference experiments, 0.4 and 0.7 for the noise sensitivity runs), multiplied by noise level and gain. Speed fluctuation across recordings is 4% (SD). Gradient-boosting classifiers use random_state equal to the seed (nested models add the held-out machine index). The domain classifier for weighted conformal prediction is class-balanced logistic regression with $C = 0.5$. The number of experiments is small enough to run on a single central processing unit (CPU) core in under ten minutes per seed. The experiments were implemented in Python using scikit-learn. The Python and scikit-learn versions, together with the remaining software dependencies, are documented in the supplementary implementation files.

### Appendix C. Order-Collision Audit Details

**Table C1.** Cross-machine fault-order collisions within a 3% relative tolerance

| Machine | Fault Line | Order | Number of Collisions | Colliding Lines (Machine, Fault, Order) |
|---|---|---|---|---|
| M1 | BPFO | 3.375 | 1 | M5 $2 \times$ BSF (3.310) |
| M1 | BPFI | 5.625 | 0 | none |
| M1 | $2 \times$ BSF | 3.750 | 0 | none |
| M2 | BPFO | 3.200 | 0 | none |
| M2 | BPFI | 4.800 | 0 | none |
| M2 | $2 \times$ BSF | 4.800 | 0 | none |
| M3 | BPFO | 5.100 | 0 | none |
| M3 | BPFI | 6.900 | 0 | none |
| M3 | $2 \times$ BSF | 6.517 | 2 | M4 BPFI (6.500); M6 BPFI (6.475) |
| M4 | BPFO | 3.500 | 0 | none |
| M4 | BPFI | 6.500 | 1 | M3 $2 \times$ BSF (6.517) |
| M4 | $2 \times$ BSF | 3.033 | 0 | none |
| M5 | BPFO | 2.553 | 1 | M7 $2 \times$ BSF (2.507) |
| M5 | BPFI | 4.447 | 2 | M6 BPFO (4.525); M8 $2 \times$ BSF (4.351) |
| M5 | $2 \times$ BSF | 3.310 | 1 | M1 BPFO (3.375) |
| M6 | BPFO | 4.525 | 1 | M5 BPFI (4.447) |
| M6 | BPFI | 6.475 | 1 | M3 $2 \times$ BSF (6.517) |
| M6 | $2 \times$ BSF | 5.381 | 2 | M7 BPFI (5.400); M8 BPFI (5.430) |
| M7 | BPFO | 2.600 | 0 | none |
| M7 | BPFI | 5.400 | 1 | M6 $2 \times$ BSF (5.381) |
| M7 | $2 \times$ BSF | 2.507 | 1 | M5 BPFO (2.553) |

| | | | | |
|---|---|---|---|---|
| M8 | BPFO | 3.570 | 0 | none |
| M8 | BPFI | 5.430 | 1 | M6 $2 \times$ BSF (5.381) |
| M8 | $2 \times$ BSF | 4.351 | 1 | M5 BPFI (4.447) |

Note: BPFO = ball pass frequency of outer race; BPFI = ball pass frequency of inner race; BSF = ball spin frequency. The within-machine coincidence of BPFI and $2 \times$ BSF for M2 is excluded. The table lists directed collision records; each distinct cross-machine collision pair appears twice.