POSITION: WHY AI FOR PHM MAY NEVER HAVE ITS "IMAGENET MOMENT" (AND WHY IT PROBABLY SHOULDN'T)

Anonymous authors
Paper under double-blind review

ABSTRACT

The "ImageNet moment"—a historical convergence in which a unified benchmark triggers methodological unification and discipline-wide reorganization—has become a default aspiration for ML entering new domains. We argue that the data-driven machine learning paradigm within prognostics and health management (PHM) may never, and should not, experience such a moment. Structural constraints including data inaccessibility, irreducible modality heterogeneity, and the absence of a dominant predictive architecture make the preconditions for such convergence structurally absent; an ImageNet-style leaderboard is therefore neither achievable nor desirable for physical-world industrial tasks. Instead, PHM requires small-scale, diagnostic, scenario-specific evaluation protocols that respect domain specificity over generic ranking. This position has implications beyond PHM: as ML expands into AI for Science and engineering, the assumption that universal benchmarks are the natural endpoint of methodological maturity must be re-examined.

1 INTRODUCTION

One of the main assumptions in mainstream data-driven machine learning, that universal benchmarks are the natural endpoint of methodological maturity, destabilizes when applied to the physical world. We argue that data-driven prognostics and health management (PHM) will never have its "ImageNet moment"—not merely an ImageNet-style benchmark, but the historical convergence of methodology, data, and community around a single shared target—and the field must stop waiting for one. Historically, such convergence required three ingredients to coincide: a single, well-defined central task; a homogeneous input modality; and abundant public data. We argue that in PHM the first is absent and the latter two are structurally out of reach—and that the field is better off for it. This mismatch is not a symptom of backwardness but an early warning for the AI4S/AI4E community as it expands into industrial environments.

The waiting is understandable. ImageNet did not merely rank algorithms; it reorganized an entire discipline around a shared target, and the fields that inherited this template—vision, then language—now enjoy what PHM conspicuously lacks: common tasks, common splits, common leaderboards, and a common sense of progress (Deng et al., 2009; Russakovsky et al., 2015). Meanwhile, a discouraging fraction of PHM publications still follow the same ritual: take one of a handful of aging public datasets, report a small improvement over a weakly-tuned baseline, and declare victory (Smith & Randall, 2015; Lei et al., 2020). When a field's dominant evaluation practice consists of squeezing decimal points out of laboratory datasets collected under conditions no factory would recognize, the desire for a great unifying benchmark is not just understandable—it is almost a call for help.

Recent developments make the aspiration look newly plausible. Time-series classification, the methodological backbone of vibration-based diagnostics, has seen a visible wave of new benchmark suites within just the past few years (Dau et al., 2019; Bagnall et al., 2018; Middlehurst et al., 2024), and PHM-adjacent venues have begun to publish explicit benchmarking efforts rather than yet another architecture paper (Hendriks et al., 2022; Zhao et al., 2024). The field appears to be pivoting, slowly, from method-driven to evaluation-driven research. If one extrapolates this trend naively, the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

destination seems obvious: eventually someone will assemble the definitive PHM benchmark, the community will gather towards it, and progress will become legible.

We write for ML and AI4S/AI4E researchers who may encounter PHM as an application domain; no prior PHM background is assumed. PHM encompasses both physics-based degradation modeling and data-driven predictive methods. These traditions have fundamentally different evaluation logics: physics-based methods are validated against mechanistic consistency, whereas data-driven methods are evaluated against held-out test data. This paper addresses the benchmarking challenges of the latter; physics-based modeling remains essential to PHM, but its validation framework is orthogonal to the benchmark culture crisis we analyze. Whether PHM is classified as AI4S, AI4E, or AIOps is immaterial to our argument; the challenges we identify are shared across all physical-world AI domains.

This paper argues that the extrapolation is negative, in both directions. Section 2 argues the descriptive half: the two structural conditions that made ImageNet possible—abundant public data of adequate quality, and a single homogeneous modality—do not exist in PHM and are not on any credible roadmap. Section 3 argues the normative half: even if such a benchmark could be built, converging on it would actively harm the field, because leaderboard incentives reward precisely the properties that industrial deployment punishes. Section 4 sketches what evaluation looks like without an ImageNet moment: a pluralistic ecology of small, failure-mode-driven diagnostic tests with efficiency-aware metrics, presented not as a deployment specification but as the natural shape of progress when single-score convergence is structurally impossible. Section 5 argues that PHM's predicament is a preview of the evaluation crisis facing AI for the physical world at large, not by claiming all physical-world AI is identical, but by identifying the structural gradient along which different domains sit. Section 6 concludes, and the Alternative Views section at the end of the paper answers the strongest objections we have encountered.

2 Why will not have an ImageNet moment

2.1 Data bottleneck is structural, not temporary

The quantity problem. Industrial data is not scarce because the community has been lazy; it is scarce because the parties who own it have rational, permanent reasons not to release it. Condition-monitoring streams sit inside proprietary maintenance records whose disclosure implies disclosure of failure histories, operating envelopes, and—by extension—trade secrets and liability exposure (Jardine et al., 2006; Lei et al., 2018). The predictable consequence is that the public data ecology of PHM is dominated by laboratory test rigs: a small set of bearing and gearbox datasets, collected under a handful of controlled operating conditions, with clean sensors, clean labels, and failure modes induced on purpose (Smith & Randall, 2015; Hendriks et al., 2022). These datasets are valuable teaching material. They are not, and cannot be inflated into, an ImageNet-scale resource. ImageNet drew on a pre-existing public ocean of natural images; there is no equivalent ocean of industrial run-to-failure data, and no plausible mechanism that would create one (Deng et al., 2009; Lei et al., 2020).

The quality problem. Worse, even the data that is public often turns out to be unsuitable for meaningful benchmarking once examined closely. A recent, unpublished attempt to construct a reproducible benchmark from a public aviation-maintenance dataset implemented and re-audited a suite of mainstream time-series methods end-to-end—including a reproducibility fix in one method's official code that had not been previously reported by other users of the dataset. The result was sobering: The audit (Appendix A) covers the binary fault-detection task, and the F1 scores of seven distinct methods collapsed into a band of roughly 0.63–0.76, with the top two methods separated by less than 0.01 F1 (Table A.1). The only axis along which methods separated cleanly was inference latency, which spans three orders of magnitude (0.1 ms to 100+ ms) with negligible F1 gaps. Also, according to the dataset authors, the original multi-class severity task, classifiers collapse to the majority classes over 95% of the time—performance is dominated by class imbalance rather than by any methodological distinction Yang & Desell (2022). We do not believe that this dataset is an outlier; we believe that it is a illustrative case of what "public PHM data" actually look like once you try to use it as a measuring instrument rather than as an infrastructure. A benchmark built on data with no discriminative power does not measure algorithms, it measures noise, hardware, and luck. A compact audit table is provided in Appendix A.

Conclusion. When both the quantity and the quality of public data are in doubt, a "universal benchmark" is not an engineering challenge awaiting a heroic effort; it is a category error. You cannot standardize measurement on an instrument that does not discriminate.

2.2 Modality heterogeneity is an intrinsic property, not an engineering debt

Vision and language could be unified under single leaderboards because each field possesses a homogeneous representational support: a vision benchmark can plausibly claim to cover "vision," because its inputs are all images; a language benchmark, texts. The shared modality is what makes a single ranking coherent in the first place.

PHM has no such support and none appears on any credible roadmap. A realistic condition-monitoring deployment is natively multimodal: high-frequency vibration signals, periodic current and voltage waveforms, low-frequency temperature and pressure traces, free-text maintenance logs, and tabular operating parameters—differing in sampling rate by orders of magnitude, in physical units, in noise characteristics, and in the very meaning of a "label" (Lei et al., 2018; 2020). This heterogeneity forces any would-be universal benchmark into a dilemma with no good exit. Restrict the benchmark to one modality (say, vibration) amputates most of what created PHM—the result is a generic time-series competition that happens to take place in an industrial setting, not a genuine PHM evaluation. Force multimodal fusion into the task, and you immediately face an unsolved and rarely-acknowledged methodological problem: there exists no accepted protocol for fairly comparing contributions across modalities with incommensurable information content and acquisition cost.

We therefore reject the framing of modality heterogeneity as a temporary inconvenience awaiting a clever architecture. It is a constitutive feature of the problem domain, a direct reflection of the fact that physical degradation expresses itself through many channels at once. Any evaluation methodology that treats this as a bug will produce numbers that are precise, comparable, and meaningless.

Even if future data-sharing policies improve, modality heterogeneity implies that cross-modal pretraining–transfer is effectively impossible in PHM prediction. A pretraining weight learned on bearing vibration signals has no transferable shared representation with gearbox current signals, let alone with maintenance logs or thermal traces. This is not a data-volume problem; it is an ontological gap between physically distinct measurement channels. The ImageNet pretraining narrative—train once on a generic dataset, then finetune everywhere—assumes a shared representational space that multi-modal physics in PHM simply does not provide.

2.3 The absence of a dominant architecture

ImageNet was a dual event. On the surface, it was a dataset; underneath it was a methodological convergence. A single architecture family (CNN, ResNet, then ViT) achieved dominance on the shared benchmark, triggering a paradigm shift that restructured the entire field around end-to-end feature learning. In PHM's core predictive tasks—time-series signal modeling, fault diagnosis, RUL estimation—this convergence has never happened.

The methodological landscape remains fragmented. Empirical reviews of the vibration-acoustic PHM literature indicate that classical machine learning methods (SVM, decision trees, random forests) continue to account for a substantial share of deployed applications, often outperforming or matching deep learning in low-data regimes (Lei et al., 2020). More tellingly, the winning solutions in recent PHM data competitions have relied on techniques that the ML mainstream would consider "pre-deep-learning": kernel regression with statistical heuristics for RUL prediction, similarity-based matching with physics-informed features for fault diagnosis, and ensemble methods with explicit feature engineering (Zhao et al., 2024). These are not retrograde holdouts; they are the empirical optimum under the constraints that PHM data actually impose. Large language models and agentic systems (associated with knowledge-based methods or not) have recently entered industrial maintenance, but their current deployments remain confined to downstream decision support—documentation, scheduling, and compliance querying—rather than the core predictive tasks (fault diagnosis, RUL estimation) that define the methodological backbone of PHM. They have not produced a convergent predictive architecture; if anything, they introduce yet another modality with its own evaluation logic. We discuss their implications for PHM benchmarking culture in Section 3.

Recent surveys of foundation models in PHM acknowledge the field's lack of consensus on how to respond to the large-model paradigm, and emphasize roadmapping over deployed convergence (Li et al., 2024).

The broader time-series evidence. Outside PHM, the time-series classification and forecasting communities have independently reached the same conclusion. MiniRocket—a method that uses fixed random convolutional kernels followed by a linear classifier, with *no learnable parameters*—achieves accuracy statistically indistinguishable from deep models such as InceptionTime on the UCR archive, while operating orders of magnitude faster (Dempster et al., 2021). On the forecasting side, recent large-scale benchmarks have found that simple statistical methods (VAR, linear regression) outperform recently proposed state-of-the-art deep-learning methods on a non-trivial subset of datasets (Qiu et al., 2024). Theoretically, Transformer architectures have been shown to fail to outperform linear models on autoregressive time-series prediction tasks, with chain-of-thought reasoning causing predictions to collapse exponentially to the mean (Zhou et al., 2025).

3 WHY SHOULD NOT HAVE AN IMAGENET MOMENT

Suppose we are wrong about Section 2, and a large, clean, unified PHM benchmark could somehow be conjured into existence. We argue the field should decline the invitation.

3.1 LEADERBOARDS OPTIMIZE THE WRONG OBJECTIVE

The sociology of benchmark-centric fields is by now well documented and should be required reading for anyone proposing a new one. Once a leaderboard acquires authority, the community's objective function silently mutates: the goal is no longer to solve the underlying problem but to ascend the table (Dehghani et al., 2021). The symptoms are familiar from vision and language: methods selected for their friendliness to the benchmark's distribution rather than for robustness in deployment; architectures effectively fitted to the test set through thousands of adaptive evaluations; and a near-total neglect of inference cost, memory footprint, and engineering viability, because the scoreboard does not ask (Recht et al., 2019; Roelofs et al., 2019).

The inadequacy of performance-centric ranking for safety-critical engineering ML has been documented in domain tutorials (Nemani et al., 2023), which argue for uncertainty quantification, calibration, and decision-risk awareness as explicit evaluation components—precisely the dimensions that leaderboard culture systematically suppresses in favor of accuracy rankings. The convergence is not ours alone. Salinas-Camus et al. (2025) review data-driven prognostic models along four axes—uncertainty, robustness, interpretability, and feasibility—and propose a user-oriented, non-compensatory framework in which feasibility concerns can veto otherwise accurate models. That a methods review elevates feasibility to a first-class criterion corroborates our central claim: industrial prognostic evaluation is non-compensatory, and the field already knows it. Industrial fault detection operates under constraints that are not secondary considerations but the product: bounded compute and memory on edge hardware, hard limits on false-negative rates because a missed fault is a safety event, strict stability requirements across operating conditions, interpretability sufficient to convince a maintenance engineer, and an explicit return-on-investment calculus (Jardine et al., 2006; Lei et al., 2020). An algorithm that tops an accuracy leaderboard but requires a GPU cluster and a three-minute inference window is not a slightly inconvenient champion; on the factory floor it is worthless. Mainstream leaderboard culture, inherited from CV and NLP, has never treated efficiency and reliability as co-primary with accuracy—at best they appear as an afterthought column that nobody's ranking depends on (Schwartz et al., 2020; Mattson et al., 2020). Importing that culture into a field whose entire raison d'être is the performance–efficiency–reliability trade-off would not elevate PHM's standards; it would overwrite them.

Leaderboard chasing encodes a compensatory decision rule, while industrial decision is non-compensatory. In leaderboard culture, performance is the objective and efficiency is an adaptation constraint: a method is expected to retain its score under a resource budget, and being faster or cheaper counts as an improvement only if accuracy does not drop. Industrial deployment inverts this. Metrics (performance, efficiency, reliability, .etc) are not one objective plus other constraints; they are co-equal dimensions of a selection problem with hard thresholds. No method can theoretically dominate on all dimensions, so the evaluation must expose the trade-off surface rather than

hide it behind a weighted sum. A model that misses the false-negative limit or the edge-latency budget is not ranked lower; it is inadmissible. The leaderboard question is "which model scores highest?" The industrial question is "which models are admissible, and among those, which is the best trade-off for this asset, this operating regime, or this failure cost?" Collapsing the second into the first discards precisely the dimensions industrial decision count on, and produces a systematic misallocation of research effort: the community optimizes what is measurable on the benchmark, while industry ignores what the benchmark measures.

3.2 INDUSTRY PAYS FOR SPECIFIC RELIABILITY, NOT GENERAL INTELLIGENCE

There is a deeper mismatch, one that no metric redesign can fix, because it concerns what is being purchased. The implicit promise of an ImageNet-style benchmark is generality: a single number summarizing competence across a broad task distribution. That promise matches what CV and NLP users want. It does not match what industrial users need.

What an industrial user needs is not "competent prediction for arbitrary equipment under arbitrary conditions." It is reliable prediction, on *their* specific asset, under *their* specific operating regime, with *their* specific failure costs (Jardine et al., 2006; Lei et al., 2018). These are different optimization targets, and the gap between them is not a detail. A model that ranks first on a composite leaderboard but was never stress-tested on the user's fault modes is, for procurement purposes, an unknown quantity; a model that ranks seventh but has demonstrated stability on the user's exact scenario is the correct choice. Composite rankings flatten exactly the dimension—scenario specificity—along which industrial value is created.

A medical analogy makes the point concrete. Hospitals do not procure one general-purpose diagnostic machine that performs "adequately across all diseases"; they procure specialized instruments, each validated for a department, a protocol, a patient population. Nobody considers this a failure of medical-device benchmarking—it is what fitness-for-purpose looks like. PHM is in the same position, and a field whose evaluation infrastructure ignores this will find, as it deserves to, that industry ignores the infrastructure.

3.3 THE EMERGING LLM FRONTIER.

Large language models and agentic systems are beginning to enter industrial maintenance workflows, initially for documentation, scheduling, and compliance querying, but increasingly for diagnostic reasoning over multimodal sensor data. We are optimistic about this trajectory: domain-specific language models trained on maintenance logs and technical corpora offer genuine capabilities that traditional signal-processing pipelines lack. However, we caution against importing the NLP community's benchmarking template wholesale. These capabilities operate on the *textual sediment* of PHM—logs, reports, and API documentation—not on the physical signals that generate them. An LLM that tops a generic question-answering leaderboard on synthetic PHM queries is not a PHM solution; it is a language model that has been fitted to a benchmark distribution.

Recent agentic benchmarks illustrate this category error: they evaluate LLM orchestration of maintenance workflows using Pass@1 and Pass-all-3 metrics, without latency bounds, cost accounting, or stability guarantees (Das & Patel, 2026). In safety-critical industrial deployment, a diagnostic agent that produces correct answers stochastically but unpredictably is not a viable decision support system; it is a risk that procurement teams cannot accept. Industrial deployment requires validation against physical constraints, regulatory thresholds, and edge-case robustness—dimensions that NLP-style leaderboards systematically discard. The data-knowledge-driven paradigm emerging in some recent work points toward a more promising path, but only if evaluation remains anchored to industrial utility rather than generic accuracy.

4 WHAT EVALUATION LOOKS LIKE WITHOUT AN IMAGENET MOMENT

If ImageNet moment is structurally impossible, what does methodological progress look like in PHM? We sketch the contours of an evaluation ecology that is naturally compatible with the constraints identified in Sections 2–3. The four components below are not a deployment specification;

they are an existence proof that evaluation-driven research is possible without a single-score convergence event.

(1) Physics-based synthetic and semi-synthetic data. Since the fundamental barrier is that real industrial run-to-failure data cannot be public at scale, the realistic path is not to wait for it but to manufacture its stand-ins: degradation models, physics-based simulators, and generative approaches that produce data whose failure phenomenology mirrors the field even when its provenance is synthetic (Karniadakis et al., 2021). Crucially, the target here is not volume but controlled coverage—the ability to dial specific fault modes, severities, and noise regimes that real datasets cannot offer at will.

(2) Non-compensatory evaluation protocols. Performance must be reported jointly with the resources required to obtain it, but the reporting protocol itself must reflect the non-compensatory logic of industrial deployment. We advocate a two-stage protocol as an example: first, a feasibility screen: a model must satisfy hard constraints—edge latency budget, false-negative bound, cross-regime stability threshold—or be eliminated regardless of its accuracy. Second, a Pareto comparison only among admissible models, displaying performance–efficiency trade-offs in two dimensions rather than collapsing them into a scalar ranking (Schwartz et al., 2020; Mattson et al., 2020). This is not a cosmetic change: it makes the knockout logic that industrial procurement actually uses structurally visible, and it removes the free lunch that leaderboard culture currently grants to compute-hungry methods.

(3) Small, specific, task-targeted benchmarks. Replace the monolith with a menu: many small benchmarks, each scoped to a concrete, well-defined scenario (a fault type, an asset class, an operating regime), from which industrial users select the subset matching their deployment context. Comparability is preserved where it matters—within a scenario—and abandoned where it never meant anything—across incommensurable scenarios (Dehghani et al., 2021).

(4) Failure-mode-driven diagnostic tests and probes. The most valuable unit of evaluation is not the dataset but the question: a small, deliberately adversarial probe targeting a specific known failure mode—sensitivity to operating-condition shift, to sensor dropout, to label noise, to class imbalance—designed so that failing it tells you something actionable (Ribeiro et al., 2020). Diagnostics of this kind are cheap to build, cheap to run, and orthogonal to one another, which makes them naturally composable into scenario-specific evaluation suites.

The strategic virtue of this ecology deserves emphasis. A leaderboard culture requires a number to chase; this design offers none. There is no global ranking to overfit, no single split to adaptively mine, no prestige attached to a scalar. Researchers are forced to improve specific, named capabilities—which happen to be precisely the capabilities engineering practice cares about. And because the required data are small-to-medium, carefully designed synthetic or semi-synthetic collections rather than a mythical public ocean, the proposal is buildable now, with the resources the community actually has. The alternative—waiting for perfect large-scale public data that will never arrive—is not rigor; it is procrastination dressed as rigor. This ecology resists Goodhart effects structurally, not procedurally. By eliminating a global scalar score, it removes the single optimization target that drives adaptive tuning. Selective reporting across scenarios is mitigated by requiring full Pareto disclosure: a method cannot claim victory on one metric while hiding failure on another. Fragmentation into incomparable bespoke tests is a feature, not a bug—industrial procurement already operates on scenario-specific validation; our proposal merely makes this practice explicit and auditable.

5 PHM AS A PREVIEW OF EVALUATION CRISIS IN AI FOR SCIENCE

A remark we heard at a recent summer school presentation related to AI distills this point better than any formalism: the physical world is the ultimate AI benchmark.[1] We endorse the sentiment, but we draw from it a sharper consequence than it is usually given. If the ultimate benchmark is the physical world itself, then every offline benchmark is at best a proxy—and a proxy earns trust only by demonstrating fidelity to specific deployment conditions, not by accumulating scale or prestige. A leaderboard, however large, cannot be the world.

[1] Personal communication, E.T. Matsubara (2026).

This suggests a broader hypothesis: the universal-benchmark paradigm may be an artifact of evaluating AI on human-made data—images and text, artifacts with homogeneous substrates and inexhaustible public supply. As AI moves from describing human artifacts to reasoning about the physical world, the paradigm's preconditions quietly disappear, and our evaluation philosophy must change with them. PHM, which hit this wall early and visibly, may thus serve as a small but useful reference point for how AI4Science could evaluate itself: not by converging on one number, but by pluralizing into many honest ones.

Similar misalignment have been documented in biological dynamical systems (Martinelli, 2025), where physics-inspired benchmarks are criticized for creating 'a false sense of generality.' The convergence is striking: whether the domain is industrial prognostics or biological simulation, the importation of generic ML benchmarking templates into physical-world settings produces the same pathology.

6 Conclusion

We have argued a two-part position. PHM will not have an ImageNet moment, because the prerequisites—abundant high-quality public data and a homogeneous modality—are structurally absent: public industrial data is scarce for permanent economic reasons, the data that does exist often lacks the discriminative power to rank anything, and multimodality is constitutive of the problem rather than a solvable inconvenience. And PHM should not want one, because leaderboard incentives reward precisely what industrial deployment punishes, while industrial value lives in scenario-specific reliability that no composite ranking can capture. The constructive half of the position is that this is liberating rather than depressing. A field that stops waiting for a monolith can start building what is actually buildable: synthetic data with controlled failure coverage, efficiency-aware protocols, small scenario-scoped benchmarks, and failure-mode probes—an evaluation ecology that is structurally resistant to gaming because it offers no single score to chase.

7 Alternative Views

We close by addressing the strongest objections we have encountered to this position.

The most fundamental objection holds that PHM does not need benchmarks at all, that industrial PHM is inherently domain-specific, each asset operates under unique physical constraints and degradation mechanisms, and the traditional PHM community has functioned for decades without benchmarks, relying on physical modeling and case-by-case validation. We agree that PHM's specificity makes *generic* benchmarks inappropriate—this is precisely the argument of Section 2. However, the absence of a generic benchmark has not produced healthy pluralism; it has produced *evaluative anarchy*. Fewer than 1% of surveyed PHM papers made both code and data publicly available (von Hahn & Mechefske, 2023), and without shared, auditable evaluation infrastructure, progress becomes unverifiable: a paper claiming 95% accuracy on a private dataset cannot be compared to another claiming 93% on a different private dataset, leaving the industrial practitioner with no basis for selecting between them. The question is therefore not *whether* to benchmark, but *what kind* of evaluation is historically viable for physical-world AI. Recent efforts to build PHM evaluation infrastructure confirm that the demand is real; their failure mode lies not in their *existence* but in their *practical utility*, as some replicate the accuracy-first logic of vision and language benchmarks without asking whether those metrics matter on the factory floor.

A second objection concedes the predictive core but sees hope in the decision layer: large language models equipped with tool use and domain knowledge could plausibly converge on a unified architecture for maintenance scheduling, cost optimization, and safety compliance. This view conflates two distinct layers of PHM value creation. The industrial significance of PHM lies in its **predictive and diagnostic core**—the ability to infer equipment health from sensor data and anticipate failures—while decision support is a downstream application that consumes PHM outputs rather than producing them. These systems operate on the *textual sediment* of PHM—logs, reports, and API documentation—not on the physical signals that generate them; an LLM has no embodied experience of physical degradation, and its "knowledge" is the statistical distribution of tokens describing degradation, not degradation itself. Moreover, industrial decisions are bound by physical and regulatory constraints (IEC 61508, FAA thresholds, OSHA requirements) that vary

across sectors and cannot be captured by a generic decision benchmark. Even within decision support, LLMs face structural mismatches: hallucination of regulatory thresholds, latency incompatible with emergency intervention windows, and operational costs that may exceed the decision value. LLMs may improve PHM's user interfaces, but they do not address the foundational evaluation challenges of physical-world prediction.

A third suggestion is that PHM should simply adopt the vigorous benchmark infrastructure that the time-series community has already built—the UCR Archive, the Monash Repository, and recent large-scale suites such as GIFT-Eval and TFB—rather than fragmenting efforts. This proposal overlooks a fundamental mismatch between *generic* time-series data and *industrial prognostic* time-series data. PHM data—high-frequency vibration, current harmonics, acoustic emissions, thermal transients—exhibit characteristics that break the assumptions underlying generic benchmarks: extreme non-stationarity (degradation shifts the distribution over time), severe class imbalance (healthy states dominate; failure samples are rare), multi-rate sampling (vibration at kHz, temperature at Hz, logs at event frequency), and physics-coupled anomalies (fault signatures manifest as specific frequency components invisible to generic metrics). A model that tops the UCR classification leaderboard may fail catastrophically on a bearing fault diagnosis task where the critical information lies in a narrow frequency band masked by operational noise. PHM needs benchmarks designed *from* industrial sensor physics, not benchmarks borrowed *from* electricity meters or stock tickers.

A fourth objection grants the problems but denies their relevance: data scarcity, modality heterogeneity, and the absence of a dominant architecture are, on this view, merely symptoms of PHM's immaturity as an ML application domain—growing pains for the PHM community to solve internally. This view mistakes symptom for cause. PHM's evaluation difficulties arise not *despite* ML's involvement but, in part, *because of* the uncritical importation of ML benchmarking templates into physical-world domains: the PHM community has resisted benchmarks not out of stubbornness but because the available templates—accuracy-first, latency-agnostic, cost-oblivious—do not map onto industrial requirements, and when ML researchers enter PHM and construct benchmarks, they reproduce the same ImageNet-derived evaluation logic that this paper critiques. The relevance to ML is not charitable concern for PHM's welfare but **methodological self-interest**. AI4S is one of the fastest-growing frontiers in ML, and physical-world domains share PHM's structural properties: data scarcity, physical constraints, multi-modal sensing, and safety-critical deployment. If ML continues to assume that its standard evaluation playbook transfers automatically, it will waste enormous resources on benchmarks that measure the wrong things. PHM is not a backwater asking for ML's attention; it is an early warning system for the evaluation challenges that every physical-world AI domain will face.

Finally, physics-informed methods (e.g., PINNs) are sometimes offered as bridging the gap between data-driven and physics-based approaches, supposedly making our critique obsolete. Such methods are a promising research direction, but they are not related to the challenge we identify. The "ImageNet moment" is a data-driven phenomenon: it presupposes large-scale, public, homogeneous data and convergence of data-driven methods on that data. Physics-informed methods are a different paradigm, whose training and validation are shaped by physical constraints rather than by dataset scale. Whether they have their own evaluation debates is not in question here; they are simply outside the scope of an argument about the evaluation infrastructure for data-driven prediction. Our claim concerns what happens when data-driven predictors meet leaderboard culture, not what happens when physical equations are encoded into neural networks.

AI USE STATEMENT

Generative AI tools were used in this work for brainstorming and developing the paper's core arguments, refining its conceptual framing and hypotheses, and providing feedback on the overall research methodology. They were not used for generating synthetic data sets, implementing methods, designing or conducting experiments, formulating mathematical claims, or assisting with proofs; translation and qualitative data analysis are not applicable to this work. AI tools were additionally used for scoping the research topic, outlining and structuring the paper, drafting all sections of the manuscript, identifying and locating relevant literature, and editing the text for readability.

The manuscript was produced in a human-in-the-loop process: the author developed and revised the argument and outline through dialogue with AI tools; every AI-drafted section was read, fact-checked, and edited by the author; all cited references were verified against the original sources; and every AI-suggested change in the final revision pass was individually reviewed by the author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work studies evaluation methodology and proposes no new system or application. It uses only the publicly released NGAFID dataset and involves no human subjects. Our critique targets evaluation practices and incentive structures rather than individuals; where specific prior work is discussed, we restrict ourselves to methodological claims supported by public artifacts. The authors have no conflicts of interest or external funding to disclose.

Reproducibility statement

The position advanced in this paper rests on two publicly verifiable sources of evidence: the methodological critique, whose claims about specific benchmarks can be checked against their released artifacts, and the NGAFID reproducibility audit summarized in Appendix A, which is based on a public dataset and publicly available analysis code. The paper introduces no new experimental pipeline of its own.

References

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive. *arXiv preprint arXiv:1811.00075*, 2018.

Ayan Das and Dhaval Patel. PHMForge: A scenario-driven agentic benchmark for industrial asset lifecycle maintenance. *arXiv preprint arXiv:2604.01532*, 2026.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayeve Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

Mostafa Dehghani, Yi Tay, Anurag Arnab, Lucas Beyer, and Ashish Vaswani. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Angus Dempster, Daniel F Schmidt, and Geoffrey I Webb. MiniRocket: A very fast (almost) deterministic transform for time series classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(3):1–34, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

J Hendriks, P Dumond, and DAK Knox. Towards better benchmarking using the CWRU bearing fault dataset. *Mechanical Systems and Signal Processing*, 169:108732, 2022.

Andrew KS Jardine, Daming Lin, and Dragan Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7):1483–1510, 2006.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Yaguo Lei, Na Li, Liang Guo, Ning Li, Tao Yan, and Jing Lin. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, 104:799–834, 2018.

Yaguo Lei, Bibo Yang, Xiang Jiang, Feng Jia, Ning Li, and Asoke K Nandi. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing*, 138:106587, 2020.

Yan-Fu Li, Huan Wang, and Muxia Sun. Chatgpt-like large-scale foundation models for prognostics and health management: A survey and roadmaps. *Reliability Engineering & System Safety*, 243: 109850, 2024. doi: 10.1016/j.ress.2023.109850.

Julien Martinelli. Position: Biology is the challenge physics-informed ML needs to evolve. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Position Paper Track.

Peter Mattson, Vijay Janapa Reddi, Christine Cheng, Cody Coleman, Greg Diamos, David Kanter, Paulius Micikevicius, David Patterson, Guenther Schmuelling, Han Tang, et al. MLPerf inference benchmark. In *Proceedings of the ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, pp. 446–459, 2020.

Matthew Middlehurst, Patrick Schäfer, and Anthony Bagnall. Bake off redux: A review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery*, 38:1958–2031, 2024.

Venkat Nemani, Luca Biggio, Xun Huan, Zhen Hu, Olga Fink, Anh Tran, Yan Wang, Xiaoge Zhang, and Chao Hu. Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. *Mechanical Systems and Signal Processing*, 205:110796, 2023. doi: 10.1016/j.ymssp.2023.110796.

Xiangfei Qiu, Jingtian Hu, Jingwei Li, Lekui Zhou, Xingjian Wu, Junyang Du, Bin Zhang, Chenjuan Guo, Aoying Zhou, Christian S Jensen, Zhenli Sheng, et al. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. *arXiv preprint arXiv:2403.20150*, 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 5389–5400, 2019.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4902–4912, 2020.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

Mariana Salinas-Camus, Kai Goebel, and Nick Eleftheroglou. A comprehensive review and evaluation framework for data-driven prognostics: Uncertainty, robustness, interpretability, and feasibility. *Mechanical Systems and Signal Processing*, 237:113015, 2025. ISSN 0888-3270. doi: 10.1016/j.ymssp.2025.113015.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, 2020.

Wade A Smith and Robert B Randall. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mechanical Systems and Signal Processing*, 64–65: 100–131, 2015.

Tim von Hahn and Chris K. Mechefske. Computational reproducibility within prognostics and health management. *Journal of Dynamics, Monitoring and Diagnostics*, 2(1):42–50, 2023. doi: 10.37965/jdmd.2023.141.

Hong Yang and Travis Desell. A large-scale annotated multivariate time series aviation maintenance dataset from the NGAFID. *arXiv preprint arXiv:2210.07317*, 2022.

Chen Zhao, Enrico Zio, and Weiming Shen. Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study. *Reliability Engineering & System Safety*, 245:109964, 2024.

Yufa Zhou, Yixiao Wang, Surbhi Goel, and Anru R Zhang. Why do transformers fail to forecast time series in-context? *arXiv preprint arXiv:2510.09776*, 2025.

A NGAFID AUDIT: METHODS, RESULTS, AND REPRODUCIBILITY NOTE

The audit was conducted as an informal reproducibility study on the NGAFID aviation maintenance dataset (Yang & Desell, 2022). Table A.1 summarizes the protocol and results.

Table A.1: NGAFID audit (5-fold CV, flight-level, no augmentation).

Method	F1	Latency (ms)
ConvMHSA	0.756 ± 0.009	3.1 ± 0.1
InceptionTime	0.748 ± 0.009	2.8 ± 0.0
MiniRocket (GPU)	0.723 ± 0.016	22.4 ± 0.2
MiniRocket (CPU)	0.715 ± 0.009	81.7 ± 22.1
Fusion (MR+stats)	0.712 ± 0.008	105.2
CNN+BiMamba	0.682 ± 0.112	6.1 ± 0.5
LogReg (stats)	0.629 ± 0.008	$0.1^{\dagger}$

†Excl. feature extraction.

The official MiniRocket demonstration passes untransposed dimensions $(N, 4096, 23)$ to `MiniRocketFeatures`, causing a $\sim$13-point accuracy drop. The corrected implementation (reported above) lands at $\sim$0.72 F1, not $\sim$0.60. Full protocol details and analysis code will be made available upon acceptance.