

Artificial Intelligence for Predictive Maintenance of Mission-Critical Engineered Systems: A Systematic Review of Methods, Frameworks, and Future Directions

Atharv Atmaram Jadhav*, Pradip M. Bhukan, Sakshi D. Dhembare, Bhagyashri R. Jadhav Tushar D. Kolhe

Navsahyadri Education Society's Group of Institutions, Pune, Maharashtra, India

Published October 2026

Abstract

Asset management in safety-critical industries is moving away from calendar-driven preventive schedules and toward autonomous, data-driven decision-making enabled by artificial intelligence (AI) and machine learning (ML). This work reports a PRISMA-compliant systematic review of the predictive maintenance (PdM) literature published between January 2018 and June 2026, covering domains such as aerospace, energy, and advanced manufacturing where component degradation must be tracked with high reliability. A structured search across four bibliographic databases (IEEE Xplore, Scopus, Web of Science, and the ACM Digital Library) produced an initial pool of 2,847 candidate records; after multi-stage screening, 127 peer-reviewed studies were retained for detailed synthesis. The retained literature is organized into six methodological threads: purely data-driven learning, physics-informed machine learning, digital-twin architectures, causal and explainable AI, uncertainty quantification, and human-AI collaboration. Although deep architectures achieve strong benchmark accuracy, the synthesis surfaces recurring shortcomings around out-of-distribution robustness, epistemic uncertainty handling, and model opacity. Eight recurring research gaps are catalogued, and a forward-looking agenda is proposed, culminating in a conceptual unifying framework, referred to here as NEXUS-AI, that combines physical priors, neuro-symbolic reasoning, and human oversight to support verifiable reliability in high-consequence settings.

Keywords: predictive maintenance; prognostics and health management; remaining useful life; physics-informed machine learning; digital twin; explainable AI; causal inference; uncertainty quantification; systematic review

1 Introduction

1.1 Background and motivation

Unplanned equipment failure remains one of the costliest operational risks faced by industry, with unscheduled downtime and its secondary effects estimated to cost roughly \$50 billion per year across sectors [1]. Maintenance philosophy has evolved through three broad stages: reactive, run-to-failure practices; scheduled preventive maintenance performed at fixed intervals; and, more recently, predictive and prescriptive maintenance (PdM), which times interventions to the actual health of an asset [2]. Systems that cannot tolerate failure — satellite constellations, aircraft propulsion units, and nuclear reactor coolant loops among them — are particularly poorly served by interval-based

maintenance, which either wastes useful component life or fails to catch anomalous wear patterns in time.

The spread of Industrial Internet of Things (IIoT) sensing, combined with the availability of large-scale compute, has made AI and ML the primary engines behind contemporary PdM research. More than 3,200 publications addressing data-driven prognostics appeared between 2018 and 2026. Even so, moving these methods from curated benchmark datasets into noisy, non-stationary plant environments has proven difficult. Models trained purely on historical data tend to produce unreliable outputs once operating conditions drift, and they offer no built-in guarantee of respecting the physical laws that actually govern mechanical degradation. This tension is pushing the field toward approaches that blend physical reasoning with statistical learning.

1.2 Scope and research questions

This review is organized around six research questions (RQs):

- **RQ1:** Which AI/ML architectures currently dominate PdM research, and how do they compare in predictive capability?
- **RQ2:** How are physics-based inductive biases being combined with purely data-driven neural models?
- **RQ3:** How are explainability and causal-inference methods being applied to interpret prognostic outputs?
- **RQ4:** What mathematical approaches are used to quantify aleatoric and epistemic uncertainty in remaining useful life (RUL) estimation?
- **RQ5:** What conceptual models describe collaboration between AI diagnostic agents and human reliability engineers?
- **RQ6:** What challenges remain unresolved, and what technical directions are likely to shape the next decade of PdM research?

1.3 Contributions

Earlier surveys have traced the rise of deep learning in fault diagnosis [4, 5], but the field has changed substantially since those accounts were written. This paper contributes in three ways: (1) it offers one of the broadest available syntheses of the current literature, drawing on 127 studies organized across six methodological categories; (2) it is, to the authors' knowledge, the first systematic review to critically assess the emerging role of large-scale time-series foundation models (e.g., TimeGPT, Lag-Llama) and causal AI in the context of physical-asset degradation; and (3) it formalizes a unified taxonomy of AI-driven PdM, identifies eight concrete capability gaps, and proposes a roadmap toward more deployable systems.

1.4 Paper organization

The review methodology section describes the PRISMA protocol and search strategy. The following section builds the methodological taxonomy spanning data-driven models, physics-informed architectures, digital twins, causal mechanisms, uncertainty quantification and human-AI collaboration. Benchmark datasets and state-of-the-art (SOTA) performance are then compiled and compared, adoption patterns are analyzed by application domain, and the identified research gaps and future research directions are presented. The paper closes with a discussion of implications and the conclusion.

2 Review methodology

2.1 Search strategy

The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol [3]. Structured Boolean queries were run against IEEE Xplore, Scopus, Web of Science (Core Collection), and the ACM Digital Library, combining prognostics-related terms with computational-intelligence terms:

("predictive maintenance" OR "prognostics" OR "remaining useful life" OR "condition monitoring" OR "fault diagnosis") AND ("machine learning" OR "deep learning" OR "neural network" OR "digital twin" OR "foundation model" OR "causal inference")

The search window spanned January 1, 2018 to June 30, 2026, capturing both the maturation of deep learning methods and the more recent emergence of sequence-to-sequence foundation models.

2.2 Inclusion and exclusion criteria

Studies were included if they were: (i) peer-reviewed journal articles or high-impact conference papers; (ii) centered on algorithmic AI/ML development for mechanical, electrical, or structural PdM; (iii) written in English; and (iv) empirically validated on recognized benchmarks or industrial case data. Studies were excluded if they were secondary literature (reviews or surveys not serving as foundational anchors), purely theoretical without algorithmic implementation or experiments, or applied ML outside physical-asset engineering (e.g., software-defect prediction, clinical medicine).

2.3 Screening process

Database searches returned 2,847 records, supplemented by 143 records found through backward snowballing. Deduplication removed 891 records, leaving 2,099 for title/abstract screening by two independent reviewers. This step eliminated 1,787 studies that did not match the inclusion criteria, leaving 312 candidates for full-text review. Disagreements were resolved with a senior supervisor; inter-rater agreement was strong (Cohen's $\kappa = 0.83$). The process concluded with 127 papers retained for synthesis.

2.4 PRISMA flow summary

- **Identification:** Database records ($n = 2,847$) plus supplementary records ($n = 143$).
- **Screening:** Records remaining after deduplication ($n = 2,099$).
- **Eligibility:** Full-text articles assessed ($n = 312$); excluded with documented reasons ($n = 185$).
- **Included:** Studies retained for qualitative synthesis ($n = 127$).

2.5 Data extraction and quality assessment

Each retained study was logged in a standardized extraction sheet capturing architecture type, target physical system, dataset, evaluation metrics, stated limitations, and code availability. Study quality was assessed using a rubric adapted from PROBAST (Prediction model Risk Of Bias ASsessment Tool), favoring work with careful hyperparameter tuning, held-out validation, and baseline comparisons.

2.6 Bibliometric overview

Publication counts by year were: 2018 (8), 2019 (12), 2020 (18), 2021 (22), 2022 (24), 2023 (19), 2024 (14), 2025 (7), and partial-year 2026 (3), the later years reflecting stricter recency-based selection. By region, contributing institutions were distributed across China (34%), Europe (28%), the United States (22%), and India (8%). By application, the corpus concentrated on manufacturing (38%), aerospace propulsion (22%), energy generation (18%), and automotive/transportation (12%).

3 Taxonomy of AI methods for predictive maintenance

3.1 Data-driven methods

Most PdM literature assumes that degradation patterns can be learned directly from multivariate sensor time series without an explicit physical model.

Classical Machine Learning. Before deep learning became dominant, margin-based and ensemble classifiers — Random Forests, Support Vector Machines, and XGBoost — were the standard tools, valued for efficiency and interpretability [5]. Lei et al. [4] showed that these methods perform well on hand-engineered statistical features (kurtosis, crest factor, etc.) but plateau when applied directly to raw, high-dimensional vibration data, limiting end-to-end automation.

Deep Learning Architectures. Deep architectures remove the need for manual feature design through hierarchical representation learning. Early work converted 1-D vibration signals into 2-D spectrograms and applied convolutional neural networks (CNNs), effectively recasting fault diagnosis as an image-classification problem [9, 14]. Long Short-Term Memory (LSTM) networks [10–12] and attention-augmented bidirectional LSTMs [17] were subsequently used to capture longer-horizon, non-linear wear dynamics. Variational autoencoders (VAEs) [16] and generative adversarial networks (GANs) [15] have also been applied to mitigate class imbalance by synthesizing realistic fault signatures.

Transformer Architectures. Building on the self-attention mechanism [18], PdM sequence modeling is increasingly transformer-based. PatchTST [19] segments time series into patches to preserve local structure while reducing the quadratic cost of attention. TimesNet [20] models multi-periodic behavior via 2-D temporal representations, and the Temporal Fusion Transformer (TFT) [21] combines static operating context with dynamic telemetry while producing interpretable feature-importance weights.

Foundation Models. The newest direction involves zero-shot time-series foundation models trained on very large external corpora, such as TimeGPT [22], TimesFM [23], and Lag-Llama [24]. These models generalize impressively out of the box but are largely optimized for economic and retail forecasting; no foundation model in the reviewed literature was pretrained specifically on high-frequency, non-stationary mechanical telemetry such as rotating-machinery vibration.

3.2 Physics-informed machine learning (PIML)

Purely data-driven networks can violate known conservation laws when extrapolating from sparse data.

Pure physics models. Early prognostic formulations relied on Paris' Law for crack growth or Arrhenius-type degradation equations [7, 8] — interpretable, but rigid.

Physics-informed neural networks (PINNs). Introduced by Raissi et al. [29] and extended by Karniadakis et al. [30], PINNs embed governing partial differential equations directly into the training loss:

$$L_{\text{Total}} = L_{\text{Data}} + \lambda L_{\text{PDE}} \quad (1)$$

Yucesan and Viana [41] applied this idea to wind-turbine bearing fatigue, constraining predictions to respect kinematic limits.

Hybrid architectures. Combining a mechanistic baseline with a data-driven residual learner [31, 32, 40] lets the physical model capture the nominal degradation trend while the network absorbs stochastic, asset-specific variation.

3.3 Digital twin frameworks

Digital twins (DTs) support continuous coupling between a physical asset and its computational counterpart [33, 34].

Static/nominal twins. Early implementations were largely high-fidelity CAD/CAE shadows with no feedback loop.

Adaptive/self-evolving twins. More recent work [37, 38] calls for bidirectional data flow. Thelen et al. [35] and VanDerHorn and Mahadevan [36] treat the DT as a dynamic probabilistic graphical model whose asset-specific parameters are continually updated via Bayesian particle filtering.

Federated twins. Qi et al. [39] proposed fleet-level learning in which distributed twins share operational insight without centralizing raw telemetry, preserving data sovereignty.

3.4 Explainable and causal AI

Opacity is a major obstacle to industrial certification of AI-based diagnostics.

Post-hoc explainability. SHAP [47] and LIME [48] are commonly used to attribute an RUL prediction to specific sensor channels (e.g., linking degradation to abnormal exhaust-gas temperature).

Causal discovery. Beyond correlation, Granger causality [43], the PC algorithm [44], and PCMCi [46] are increasingly used to build directed acyclic graphs that represent fault-propagation pathways in complex systems.

Counterfactual reasoning. Building on Pearl's do-calculus [45] and Wachter et al.'s counterfactual framework [49], some recent models support interventional queries such as, "how would RUL change if operating load were reduced by 15%?"

3.5 Uncertainty quantification (UQ)

Point predictions of failure time are inadequate for risk-averse scheduling; both aleatoric (data) and epistemic (model) uncertainty need to be captured.

Bayesian neural networks. Peng et al. [58] used variational inference over network weights to obtain native confidence bounds.

Ensembles and dropout. MC-Dropout [55] and deep ensembles [56] offer cheaper, non-Bayesian proxies for epistemic uncertainty.

Conformal prediction. Conformal methods [51, 52, 54] produce statistically valid prediction intervals with guaranteed marginal coverage, independent of the underlying model architecture, and are gaining traction in the PdM literature.

3.6 Human-AI collaborative maintenance

The end goal of PdM is to augment, not replace, human judgment.

Trust calibration. Lee and See [63] outline how trust between operators and autonomous systems should be dynamically allocated.

Selective prediction and deferral. Rejection-option algorithms [60–62] allow a model to abstain when predictive entropy exceeds a threshold, routing ambiguous cases to a human expert.

Interface design. Following the guidelines of Amershi et al. [64], effective PdM tools should present degradation trajectories and causal fault trees in a way that supports fast, confident decisions.

4 Benchmark datasets and evaluation

4.1 Benchmark dataset repository

Table 1 summarizes the open-source electromechanical datasets most frequently used in the reviewed literature.

Table 1. Benchmark datasets used in the reviewed PdM literature.

Dataset	Target Domain	Scale / Volume	Sensor Modalities	Primary Task	Studies Using It
C-MAPSS	Turbofan engines	709 units, 4 operating sub-conditions	21 sensors + 3 operational settings	RUL estimation	67
CWRU	Rolling bearings	12,000+ samples, varying loads	High-frequency vibration	Fault classification	43
FEMTO	Rolling bearings	17 run-to-failure experiments	Vibration, temperature	RUL estimation	28
NASA Battery	Li-ion cells	34 cells, accelerated cycling	Voltage, current, temperature	Capacity/SOH forecasting	19
Paderborn	Motor bearings	32 test trajectories	Vibration, motor current	Fault classification	15
IMS	Rolling bearings	4 channels, 3 run-to-failure tests	Vibration	RUL estimation	22
PHM08	Turbofan engines	Competition dataset	21 sensor streams	RUL estimation	12

4.2 Standardized evaluation metrics

Let N be the number of test samples, y_i the ground-truth RUL and $\hat{y}_i$ the predicted RUL, with error $e_i = \hat{y}_i - y_i$. The root mean square error (RMSE) is:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N e_i^2} \quad (2)$$

The mean absolute error (MAE) is:

$$MAE = \frac{1}{N} \sum_{i=1}^N |e_i| \quad (3)$$

The NASA asymmetric scoring function penalizes late predictions ($e_i > 0$) more severely than early ones ($e_i < 0$), reflecting the higher operational cost of an unanticipated failure relative to a premature maintenance action:

$$\text{Score} = \sum_{i=1}^N s_i, \quad s_i = \begin{cases} e^{-e_i/13} - 1, & e_i < 0 \\ e^{e_i/10} - 1, & e_i \geq 0 \end{cases} \quad (4)$$

For fault classification tasks, accuracy, macro F1-score, and diagnostic coverage are the dominant metrics.

4.3 SOTA comparison on C-MAPSS

Table 2 tracks reported state-of-the-art RMSE and NASA score values on the four C-MAPSS sub-datasets (FD001–FD004) over time, illustrating the progression from classical models to attention-based and hybrid networks.

Table 2. Reported SOTA performance on C-MAPSS (RMSE, lower is better).

Method	Year	FD001	FD002	FD003	FD004
Standard SVM	2018	20.96	29.82	21.05	32.41
Vanilla LSTM	2019	16.14	24.49	16.18	28.17
DCNN	2020	12.61	22.36	12.64	23.31
BiLSTM + Attention	2021	11.35	19.42	11.89	21.16
Multi-Scale DCNN-LSTM	2022	10.92	18.66	10.74	20.35
Temporal Fusion Transformer	2023	10.15	17.51	10.42	18.91
Hybrid PINN	2023	11.02	15.34	11.15	16.42
PatchTST	2024	9.78	15.02	9.94	16.15
Causal Discovery GNN	2024	9.41	14.88	9.55	15.90
TimesNet variant	2025	9.15	14.21	9.02	15.11
Fed-DT Transformer	2026	8.89	13.94	8.76	14.65
Zero-shot FM (Lag-Llama)	2026	14.22	21.50	15.11	22.34

Notably, the zero-shot foundation model still trails domain-tuned supervised models on the more intricate thermodynamic dynamics captured by FD002 and FD004, underscoring that strong general-purpose forecasting ability does not automatically transfer to mechanical prognostics.

4.4 Cross-domain transfer behavior

Although within-dataset performance has largely saturated, transfer across datasets remains unreliable: models trained on CWRU bearing data commonly lose more than 40% of their accuracy when evaluated on Paderborn or FEMTO signals. This gap highlights the need for domain-adversarial training and physically grounded priors that generalize beyond a single test rig.

5 Analysis by application domain

Deployment realities differ sharply across sectors, shaped by regulation, sensing constraints, and the cost of failure.

5.1 Aerospace and aviation

Aerospace research concentrates on turbofan prognostics, structural health monitoring, and flight-control diagnostics. C-MAPSS underlies roughly 68% of aerospace RUL papers in the corpus, with attention-based and recurrent models driving recent accuracy gains [6, 12, 67]. Certification regimes such as DO-178C demand determinism and traceability that black-box deep models struggle to provide, pushing researchers toward physics-informed losses of the form in Equation (1) to keep predictions consistent with aerodynamic and thermodynamic constraints [29, 66].

5.2 Satellite and space systems

Because on-orbit repair is impossible, space-system PdM emphasizes early anomaly detection for solar-panel radiation degradation and reaction-wheel bearings [68]. Bandwidth limits and radiation-hardened, low-power processors motivate lightweight unsupervised methods, such as VAEs trained on thermal-vacuum ground test data and later deployed on-orbit [69]. Long-term radiation effects on electronics remain poorly modeled due to sparse failure history and coupled multi-physics degradation.

5.3 Manufacturing and industry

Manufacturing accounts for the largest share of the reviewed corpus, driven by the direct cost of unplanned downtime. Bearing and gear diagnostics using CWRU and Paderborn data dominate [4, 13], typically via wavelet-based time-frequency representations classified with CNNs [14]. Variable operating regimes (load, speed) create substantial domain shift, making domain-adaptation techniques such as maximum mean discrepancy or adversarial alignment standard practice [27, 28].

5.4 Energy systems

Energy-sector PdM targets wind turbines, gas turbines, and battery storage. Wind-turbine monitoring relies heavily on SCADA telemetry for early gearbox and generator fault warning [71, 72], complicated by severe class imbalance since catastrophic failures are rare, favoring semi- or unsupervised methods. Battery-health research increasingly combines electrochemical impedance data and partial-cycle information with electrochemical models to forecast capacity fade and optimize charging protocols [8, 70, 73].

5.5 Automotive and transportation

This sector splits between EV battery prognostics — which mirrors the energy domain but adds highly variable driver-dependent load profiles — and rail applications monitoring wheel-rail interfaces and bogie suspensions via onboard acoustic and vibration sensing [74]. High data volumes push processing toward the edge, motivating federated approaches that aggregate fleet-wide degradation patterns without centralizing proprietary vehicle data [73].

5.6 Cross-domain summary

Table 3 summarizes the dominant assets, data modalities, architectures, and bottlenecks per domain.

Table 3. Cross-domain summary of PdM adoption patterns.

Domain	Primary Assets	Dominant Data Modality	Prevalent Architectures	Primary Bottleneck
Aerospace	Turbofans, structures	Thermodynamic, strain	Transformers, PINNs	Regulatory certification (DO-178C)
Space	Reaction wheels, panels	Telemetry, thermal	VAEs, lightweight CNNs	Zero-repair constraint, compute limits
Manufacturing	Bearings, CNC tools	Vibration, acoustics	CNNs, domain adaptation	Cross-machine operational variability
Energy	Wind turbines, batteries	SCADA, voltage/current	Graph NNs, LSTMs	Extreme class imbalance, rare failures
Transportation	EV batteries, rail bogies	Kinematic, electrical	Federated learning, edge ML	Dynamic user behavior, data privacy

6 Critical analysis and research gaps

Across the 127 reviewed studies, eight structural gaps recur that limit translation from academic benchmarks to industrial deployment.

6.1 Gap 1: Domain Silos

About 94% of the reviewed work targets a single asset class in isolation. Techniques developed for bearing diagnostics are rarely tested against, say, battery degradation, despite shared temporal structure, leading to duplicated effort and limited transfer.

6.2 Gap 2: Largely Static Digital Twins

Digital twins appear in 42% of the corpus, but 89% of these implementations are static surrogates that do not update as the asset ages [33, 36, 37]. Genuine self-correction via recursive Bayesian updating or online parameter estimation is rare, so twin calibration degrades over time.

6.3 Gap 3: Correlation-Based Explainability

Only about 5.5% of the reviewed studies use formal causal-inference tools; most rely on post-hoc, correlation-based methods such as SHAP or LIME [47, 48], which flag salient features without distinguishing cause from symptom (e.g., a temperature spike that accompanies, rather than causes, degradation).

6.4 Gap 4: Uncalibrated Uncertainty

Roughly 78% of studies output deterministic point predictions only. Where UQ is present, it typically relies on costly Bayesian networks or simplified Gaussian assumptions that miss heteroscedastic degradation behavior [57, 58]; only three studies in the corpus applied conformal prediction to obtain distribution-free, finite-sample interval guarantees [59].

6.5 Gap 5: Absent Human-AI Frameworks

Most reviewed work implicitly treats PdM output as directly actionable, with no human step in between. Formal deferral or dynamic trust-calibration frameworks, common in medical and HCI research [61, 62, 65], are almost entirely absent from the engineering PdM literature.

6.6 Gap 6: Limited Edge Deployment

Fewer than 15% of studies address quantization, pruning, or distillation for edge inference, leaving many state-of-the-art transformer models impractical for latency- or power-constrained settings such as vehicles or remote wind farms.

6.7 Gap 7: Data Scarcity and Class Imbalance

Because engineered systems are designed not to fail, run-to-failure data is inherently scarce. GAN-based augmentation is common but can generate physically implausible samples, and few-shot approaches for multivariate time series remain underdeveloped relative to the field's reliance on large labeled datasets.

6.8 Gap 8: Fragmented End-to-End Frameworks

Fewer than 2% of studies connect anomaly detection, RUL estimation, and maintenance scheduling into a single pipeline with consistent uncertainty propagation across stages.

Table 4 summarizes coverage, severity, and proposed direction for each gap.

Table 4. Summary of identified research gaps.

Gap	% Addressing	Severity	Proposed Direction
Domain silos	6%	High	Cross-domain foundation models
Static digital twins	11%	Critical	Self-evolving, auto-calibrating twins
Correlation-based explainability	5.5%	High	Causal discovery and counterfactuals
Uncalibrated uncertainty	22%	Critical	Conformal prediction for time series
Absent human-AI frameworks	<1%	High	Algorithmic deferral, trust metrics
Limited edge deployment	14%	Medium	Model compression, federated ML
Data scarcity	31%	High	Physics-guided few-shot learning
Fragmented frameworks	2%	Critical	Integrated end-to-end architectures

7 Future research directions

7.1 Cross-Domain Foundation Models

As with large language models, progress may depend on domain-agnostic tokenization that maps heterogeneous sensor streams — irrespective of sampling rate or physical unit — into a shared latent representation, enabling pretraining across industries and zero- or few-shot transfer to new asset types [25, 26].

7.2 Self-Evolving Digital Twins

Closing the sim-to-real gap likely requires digital twins that update continuously as new data arrives, effectively performing online system identification. Coupling neural ordinary differential equations [42] with recursive Bayesian filtering is one promising route to continuous-time models that adapt to wear without drifting into physically implausible states.

7.3 Causal AI for Root-Cause Diagnosis

Automatically constructing causal graphs from observational sensor data — for example with PCMCI [46, 50] — would let systems be modeled as structural causal models, supporting counterfactual queries such as “what would RUL be if load had been reduced 20% three days ago?” This is central to producing explanations that are physically, not just statistically, grounded.

7.4 Conformal Prediction for Prognostics

Rigorous, distribution-free uncertainty quantification via conformal prediction can wrap any black-box RUL estimator to yield intervals with guaranteed marginal coverage,

$$P(Y_{t+k} \in \widehat{C}(X_t)) \geq 1 - \alpha \quad (5)$$

though extending this to the covariate shift typical of degradation processes will require non-conformity scores that adapt to the asset’s underlying hazard rate [53, 54].

7.5 Human-AI Collaborative Maintenance

“Learning to defer” frameworks, in which a model routes high-uncertainty cases to a human expert [61, 62], along with adaptive explanation interfaces that tailor detail to the audience (executive summary vs. sensor-level causal graph), would help calibrate operator trust to actual model competence [64, 65].

7.6 Unified End-to-End Frameworks

We outline a conceptual architecture, NEXUS-AI (Networked Ecosystem for eXplainable and Unified System-health), intended to connect data ingestion, causal diagnostics, conformal prognostics, and reinforcement-learning-based scheduling within one pipeline, with differentiable uncertainty propagation across stages to avoid silent failure cascades.

7.7 Emerging Directions

Large language models grounded in telemetry and maintenance manuals could serve as natural-language “maintenance copilots.” Federated learning is likely to remain important for fleet-wide learning without centralizing sensitive data, and early-stage quantum optimization may eventually help with the NP-hard scheduling problems that arise across large fleets.

8 Discussion

8.1 Key Findings

Deep learning now appears in roughly 72% of recent publications, having displaced classical ML as the default approach. Since 2022, transformer-based models have increasingly displaced LSTMs for complex forecasting tasks, echoing trends seen in natural language processing [19, 75]. This progress, however, coexists with persistent weaknesses: sparse uncertainty quantification, reliance on correlation-based explanations, and limited attention to the human factors that determine whether a system is actually trusted and used.

8.2 Implications for Practitioners

Benchmark accuracy alone is a poor predictor of deployment success. Practitioners should prioritize data-pipeline quality over chasing the newest architecture, insist on calibrated uncertainty rather than point estimates, and require some combination of physics-informed constraints or human-in-the-loop review before AI outputs drive maintenance decisions.

8.3 Implications for Researchers

Continued incremental gains on saturated benchmarks such as C-MAPSS or CWRU offer diminishing returns. Higher-value research directions include domain adaptation, causal discovery, and distribution-free uncertainty quantification on messier, less curated industrial data. Sharing pretrained weights and full data-curation methodology, not just code, would meaningfully accelerate progress toward industrial-scale foundation models.

8.4 Limitations of This Review

The search was restricted to English-language records in three major databases, which may have excluded relevant work published in regional venues or proprietary industry reports. Reported SOTA benchmarks are likely to age quickly given the field's pace. Finally, assigning individual papers to a single application domain sometimes required subjective judgment, since many methodological papers are validated across multiple datasets.

9 Conclusion

AI has materially advanced what is achievable in predictive maintenance for mission-critical systems, but this review's synthesis of 127 studies indicates that enterprise-scale deployment is still constrained by siloed development, weak causal grounding, and inconsistent uncertainty quantification. Addressing these constraints will likely require a shift away from purely data-driven curve fitting toward physics-grounded, human-collaborative systems. Advances in time-series foundation models, self-evolving and causally aware digital twins, and conformal prediction offer a plausible path toward AI-based maintenance solutions that are robust, transparent, and scalable enough for the next generation of industrial and autonomous infrastructure.

Acknowledgment

The authors thank Navsahyadri Education Society's Group of Institutions for institutional support during the preparation of this review.

References

- [1] Mobley RK. *An Introduction to Predictive Maintenance*. 2nd ed. Amsterdam, The Netherlands: Elsevier; 2022.
- [2] Jardine AKS, Lin D, Banjevic D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech Syst Signal Process*. 2006;20(7):1483–1510.
- [3] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.
- [4] Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi AK. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mech Syst Signal Process*. 2020;138:106587.
- [5] Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX. Deep learning and its applications to machine health monitoring. *Mech Syst Signal Process*. 2019;115:213–237.

- [6] Saxena A, Goebel K, Simon D, Eklund N. Damage propagation modeling for aircraft engine run-to-failure simulation. In: *Proc. Int. Conf. Prognostics Health Manage. (PHM)*; 2008. p. 1–9.
- [7] Si X, Wang W, Hu C, Zhou D. Remaining useful life estimation—A review on the statistical data driven approaches. *Eur J Oper Res*. 2011;213(1):1–14.
- [8] Goebel K, Saha B, Saxena A, Celaya JR, Christophersen J. Prognostics in battery health management. *IEEE Instrum Meas Mag*. 2008;11(4):33–40.
- [9] Sateesh Babu G, Zhao P, Li X. Deep convolutional neural network based regression approach for estimation of remaining useful life. In: *Proc. DASFAA*; 2016. p. 214–228.
- [10] Zheng S, Ristovski K, Farahat A, Gupta C. Long short-term memory network for remaining useful life estimation. In: *Proc. IEEE ICPHM*; 2017. p. 88–95.
- [11] Zhang J, Wang P, Yan R, Gao RX. Long short-term memory for machine remaining life prediction. *J Manuf Syst*. 2018;48:78–86.
- [12] Li X, Ding Q, Sun J. Remaining useful life estimation in prognostics using deep convolution neural networks. *Rel Eng Syst Saf*. 2018;172:1–11.
- [13] Wen L, Gao L, Li X. A new deep transfer learning based on sparse auto-encoder for fault diagnosis. *IEEE Trans Syst, Man, Cybern, Syst*. 2019;49(1):136–144.
- [14] Ince T, Kiranyaz S, Eren L, Askar M, Gabbouj M. Real-time motor fault detection by 1-D convolutional neural networks. *IEEE Trans Ind Electron*. 2016;63(11):7067–7075.
- [15] Shao S, Wang P, Yan R. Generative adversarial networks for data augmentation in machine fault diagnosis. *Comput Ind*. 2019;106:85–93.
- [16] Chen Z, Mauricio A, Li W, Gryllias K. A deep generative model for fault diagnosis of rolling element bearings. *Mech Syst Signal Process*. 2021;150:107335.
- [17] Song Y, Gao S, Li Y, Jia L, Li Q, Pang F. Distributed attention-based temporal convolutional network for remaining useful life prediction. *IEEE Internet Things J*. 2021;8(12):9594–9602.
- [18] Vaswani A, et al. Attention is all you need. In: *Proc. NeurIPS*; 2017.
- [19] Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A time series is worth 64 words: Long-term forecasting with transformers. In: *Proc. ICLR*; 2023.
- [20] Wu H, Hu T, Liu Y, Zhou H, Wang J, Long M. TimesNet: Temporal 2D-variation modeling for general time series analysis. In: *Proc. ICLR*; 2023.
- [21] Lim B, Arık SO, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int J Forecast*. 2021;37(4):1748–1764.
- [22] Garza A, Mergenthaler-Canseco M. TimeGPT-1. *arXiv:2310.03589* [Preprint]. 2023.
- [23] Das A, et al. A decoder-only foundation model for time-series forecasting. In: *Proc. ICML*; 2024.
- [24] Rasul K, et al. Lag-Llama: Towards foundation models for probabilistic time series forecasting. *arXiv:2310.08278* [Preprint]. 2024.
- [25] Yue Z, et al. TS2Vec: Towards universal representation of time series. In: *Proc. AAAI*; 2022.
- [26] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: *Proc. ICML*; 2020.
- [27] Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;22(10):1345–1359.
- [28] Lu W, Liang P, Li Y, Liu L, Li M, Zhang J. Transfer learning for remaining useful life prediction of bearings under different working conditions. *Rel Eng Syst Saf*. 2022;225:108620.
- [29] Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear PDEs. *J Comput Phys*. 2019;378:686–707.
- [30] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. *Nat Rev Phys*. 2021;3(6):422–440.
- [31] Karpatne A, et al. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Trans Knowl Data Eng*. 2017;29(10):2318–2331.
- [32] Willard J, Jia X, Xu S, Steinbach M, Kumar V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Comput Surv*. 2022;55(4):1–37.

-
- [33] Grieves M. *Digital twin: Manufacturing excellence through virtual factory replication*. White paper; 2014.
 - [34] Tao F, Zhang H, Liu A, Nee AY. Digital twin in industry: State-of-the-art. *IEEE Trans Ind Informat*. 2019;15(4):2405–2415.
 - [35] Thelen A, et al. A comprehensive review of digital twin—part 1: Modeling and twinning enabling technologies. *Struct Multidiscip Optim*. 2022;65:354.
 - [36] VanDerHorn E, Mahadevan S. Digital twin: Generalization, characterization and implementation. *Decis Support Syst*. 2021;145:113524.
 - [37] Jones D, Snider C, Nassehi A, Yon J, Hicks B. Characterising the digital twin: A systematic literature review. *CIRP J Manuf Sci Technol*. 2020;29:36–52.
 - [38] Liu Y, et al. A novel cloud-based framework for the elderly healthcare services using digital twin. *IEEE Access*. 2019;7:49088–49101.
 - [39] Qi Q, et al. Enabling technologies and tools for digital twin. *J Manuf Syst*. 2021;58:3–21.
 - [40] Chao MA, Kulp C, Yeo CSK, Rocco CM, Zio E. Hybrid deep learning approach for RUL prediction. *Rel Eng Syst Saf*. 2022;222:108420.
 - [41] Yucesan E, Viana FA. *A physics-informed neural network for wind turbine main bearing fatigue*. Int J Prognostics Health Manage. 2020;11(1).
 - [42] Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud DK. Neural ordinary differential equations. In: *Proc. NeurIPS*; 2018.
 - [43] Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*. 1969;37(3):424–438.
 - [44] Spirtes P, Glymour CN, Scheines R. *Causation, Prediction, and Search*. Cambridge, MA, USA: MIT Press; 2000.
 - [45] Pearl J. *Causality*. Cambridge, U.K.: Cambridge Univ. Press; 2009.
 - [46] Runge J, Nowack P, Kretschmer M, Flaxman S, Sejdinovic D. *Detecting and quantifying causal associations in large nonlinear time series datasets*. Sci Adv. 2019;5(11):eaau4996.
 - [47] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Proc. NeurIPS*; 2017.
 - [48] Ribeiro MT, Singh S, Guestrin C. ‘Why should I trust you?’: Explaining the predictions of any classifier. In: *Proc. KDD*; 2016.
 - [49] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv J Law Technol*. 2017;31:841.
 - [50] Nauta M, Bucur D, Seifert C. Causal discovery with attention-based convolutional neural networks. *Mach Learn Knowl Extr*. 2019;1(1):312–340.
 - [51] Vovk V, Gammerman A, Shafer G. *Algorithmic Learning in a Random World*. Berlin, Germany: Springer; 2005.
 - [52] Romano Y, Patterson E, Candes E. Conformalized quantile regression. In: *Proc. NeurIPS*; 2019.
 - [53] Tibshirani RJ, Foygel Barber R, Candes E, Ramdas A. Conformal prediction under covariate shift. In: *Proc. NeurIPS*; 2019.
 - [54] Barber RF, Candes EJ, Ramdas A, Tibshirani RJ. Conformal prediction beyond exchangeability. *Ann Statist*. 2023;51(2):816–845.
 - [55] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: *Proc. ICML*; 2016.
 - [56] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Proc. NeurIPS*; 2017.
 - [57] Abdar M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf Fusion*. 2021;76:243–297.
 - [58] Peng W, Ye Y, Chen N. Bayesian deep learning for remaining useful life prediction of engineering systems. *Rel Eng Syst Saf*. 2019;191:106544.
 - [59] De Pater I, Mitici M. Conformal prediction for remaining useful life estimation of aircraft components. *Rel Eng Syst Saf*. 2023;233:109094.
 - [60] Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In: *Proc. NeurIPS*; 2017.
-

-
- [61] Madras D, Pitassi T, Zemel R. Predict responsibly: Improving fairness and accuracy by learning to defer. In: *Proc. NeurIPS*; 2018.
 - [62] Mozannar H, Sontag D. Consistent estimators for learning to defer to an expert. In: *Proc. ICML*; 2020.
 - [63] Lee JD, See KA. Trust in automation: Designing for appropriate reliance. *Hum Factors*. 2004;46(1):50–80.
 - [64] Amershi S, et al. Guidelines for human-AI interaction. In: *Proc. CHI*; 2019.
 - [65] Okamura K, Yamada S. Adaptive trust calibration for human-AI collaboration. *Int J Hum-Comput Stud*. 2023;172:102985.
 - [66] Ellefsen AL, Bjørlykhaug E, Æsøy V, Ushakov S, Zhang H. Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture. *Rel Eng Syst Saf*. 2019;183:240–251.
 - [67] Mo Y, Wu Q, Li X, Huang B. Remaining useful life estimation via transformer-based architectures with multi-head attention. *Rel Eng Syst Saf*. 2021;210:107532.
 - [68] Tipaldi M, D'Arrigo PS, Macchia BC. Satellite anomaly detection using machine learning techniques. *Adv Space Res*. 2015;55(12):2884–2895.
 - [69] Ibrahim SK, Hameed MSA, Hameed KA. Machine learning approaches for spacecraft telemetry monitoring. In: *Proc. IEEE Aerosp. Conf.*; 2019. p. 1–10.
 - [70] Severson KA, et al. Data-driven prediction of battery cycle life before capacity degradation. *Nat Energy*. 2019;4:383–391.
 - [71] Stetco A, et al. Machine learning methods for wind turbine condition monitoring: A review. *Renew Energy*. 2019;133:620–638.
 - [72] Tautz-Weinert S, Watson SJ. Using SCADA data for wind turbine condition monitoring—A review. *IET Renew Power Gener*. 2017;11(4):382–394.
 - [73] Hu X, Xu L, Lin X, Pecht M. Battery lifecycle prognosis for electric vehicles. *J Power Sources*. 2020;448:227443.
 - [74] Fink O, Wang Q, Svensen M, Dersin P, Lee WJ, Ducoffe M. Potential, challenges and future directions for deep learning in prognostics and health management applications. *Eng Appl Artif Intell*. 2020;92:103678.
 - [75] Zerveas G, Jayaraman S, Patel D, Bhamidipaty A, Eickhoff C. A transformer-based framework for multivariate time series representation learning. In: *Proc. KDD*; 2021.