information

MDPI

Systematic Review

Artificial Intelligence in Complex Manufacturing Systems: A Systematic Review of Validation Rigor and Deployment Readiness in Predictive Maintenance

Cesar Felipe Henao Villa [1,*], David Alberto Garcia Arango [1], Luis Fernando Garcés Giraldo [2,*], Rosana Alejandra Meleán Romero [3], Alejandro Valencia-Arias [4] and José Alexander Velásquez Ochoa [5]

1 Dirección de Investigación e Innovación, Universidad Autónoma del Perú, Lima 15842, Peru; dgarcia30@autonoma.edu.pe
2 Escuela de Posgrado, Universidad Continental, Lima 15074, Peru
3 Facultad de Ciencias Económicas y Sociales (FCES), Universidad del Zulia, Maracaibo 4005, Venezuela; rvgluz@fec.luz.edu.ve
4 Vicerrectoría de Investigación y Postgrado, Universidad de Los Lagos, Osorno 5310857, Chile; jvalenciaa@unap.cl
5 Facultad de Ciencias Administrativas y Económicas, Tecnológico de Antioquia Institución Universitaria, Medellín 050034, Colombia; jose.velasquez46@tdea.edu.co
* Correspondence: cesarhenao@autonoma.edu.pe (C.F.H.V.); lgarces@continental.edu.pe (L.F.G.G.)

Abstract

This systematic review (PRISMA 2020) examines 89 studies—64 peer-reviewed articles and 25 arXiv preprints (2007–2026)—addressing the gap between AI research and operational predictive maintenance (PdM) deployment in complex manufacturing systems. Analyzing five thematic clusters in non-stationary and stochastic environments, we evaluated predictive performance and deployment readiness. Deep learning dominates remaining useful life (RUL) forecasting; however, 65.6% of studies employ weak or unclear validation protocols (Tier 0–1), lacking real-world robustness testing. Fault diagnosis increasingly integrates Edge-AI, yet Explainable AI (XAI) adoption remains scarce (15.6%), undermining industrial trustworthiness. No study reached operational field validation beyond temporal or cross-domain split, reflecting a systematic disconnection from deployed manufacturing systems. We introduce a novel Deployment Readiness Score (DRS) framework and identify critical barriers: data scarcity, environmental non-stationarity, computational constraints, and black-box model distrust. Recommendations include standardized temporal validation protocols, multi-site field studies, and architecture-integrated explainability. The 25 arXiv preprints (2024–2026) exhibit a mean DRS nearly three times that of the peer-reviewed corpus, signaling nascent convergence toward deployment-mature research. This review was not pre-registered.

Keywords: Artificial Intelligence; predictive maintenance; complex engineering systems; validation rigor; Edge-AI; deployment readiness; sensor fusion; smart manufacturing

Check for updates

Academic Editors: Marcin Paprzycki, Robin Haunschild, Giorgio Maria Di Nunzio, Vasco N. G. J. Soares, Paulo Quaresma and Luigi Laura

Received: 13 February 2026
Revised: 21 April 2026
Accepted: 27 April 2026
Published: 8 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. Context: AI in Complex Engineering Systems

The transition from Industry 4.0 to Industry 5.0 has established Artificial Intelligence (AI) as a cornerstone of modern industrial reliability, particularly within the domain of predictive maintenance (PdM) [1]. Unlike traditional automated production lines, complex engineering systems are characterized by non-stationary dynamics, stochastic component

degradation, and heterogeneous sensor data streams that evolve over time [2–4]. In this context, data-driven methodologies have demonstrated superior adaptability compared to classical analytical or rule-based methods. By enabling tasks such as remaining useful life (RUL) forecasting and real-time fault diagnosis, these AI systems promise to minimize unplanned downtime, optimize operational processes, and extend the critical infrastructure lifecycle.

The economic imperative for effective predictive maintenance is substantial: unplanned equipment failures cost U.S. manufacturers over $50 billion annually. Beyond direct financial losses, equipment downtime translates to production delays, compromised product quality, and erosion of customer trust. Traditional time-based or reactive maintenance strategies prove inadequate for complex systems where degradation patterns are inherently nonlinear and influenced by operational variability, environmental conditions, and interdependent component interactions [5,6].

Within the framework of the Special Issue "Surveys in Information Systems and Applications" of *Information*, this challenge acquires central relevance as an application domain in which heterogeneous sensor streams, automated decision pipelines, and industrial information systems converge. Growing demand for accurate prediction of performance and maintenance metrics in complex engineering systems has driven rapid expansion in AI application. AI methodologies are routinely deployed across diverse tasks, including regression-based modeling of critical values, event and anomaly detection in numerical and image-based datasets, and operational process optimization. In many cases, these AI-driven approaches have demonstrated superior performance and adaptability when compared with traditional analytical or rule-based methods.

1.2. State-of-the-Art: The Rise in Deep Learning

Recent literature demonstrates a paradigmatic shift toward high-capacity models capable of handling high-dimensional data [7,8]. Deep learning (DL) architectures—specifically Convolutional Neural Networks (CNNs) for image-based fault detection and Recurrent Neural Networks (RNNs/LSTMs) for time-series forecasting—have become the de facto standard in the field. These models excel at extracting latent features from raw sensor inputs—such as vibration signals or acoustic emissions—without requiring extensive manual feature engineering. Consequently, the volume of publications reporting state-of-the-art accuracy on benchmark datasets (e.g., C-MAPSS, IMS Bearing) has grown exponentially, suggesting that the technical capacity to predict failures is largely resolved from an algorithmic perspective.

Advanced architectures such as transformers, attention mechanisms [9], and hybrid CNN-LSTM models represent the current frontier, demonstrating superior performance in capturing both spatial patterns and temporal dependencies. Machine learning algorithms ranging from regression for RUL estimation [10,11] to classification for fault-mode prediction [12] have achieved accuracy rates of 85–95% in controlled experimental settings. The democratization of open-source deep learning frameworks (PyTorch 2.6, TensorFlow) and the maturation of transfer learning techniques have further accelerated research adoption across diverse manufacturing domains.

1.3. The Gap: From Algorithmic Precision to Operational Readiness

However, despite exponential growth in algorithmic proposals [13], a critical disconnect persists between academic performance metrics and industrial deployment. The current literature is saturated with models achieving high accuracy on static, standardized datasets, yet operationalization of these models in real plant conditions remains limited. This gap is attributed largely to two systemic failures identified in recent surveys:

The Validation Crisis: Engineering systems are inherently dynamic, subject to concept drift and variable operational loads. Yet, a significant proportion of cutting-edge studies validate models using random data splits that do not respect the temporal dependencies of industrial time-series [14]. Consequently, models appearing robust in laboratory settings often act as "black boxes" with uncertain reliability when facing noise and unseen conditions in real factories. The reliance on standardized benchmark datasets, while valuable for algorithmic comparison, introduces selection bias—these carefully curated datasets lack the stochastic environmental noise, missing values, sensor failures, and operational regime changes characteristic of real manufacturing environments.

The Deployment Gap: While computational complexity has increased, high-performance hardware availability on the factory floor is often restricted. Few reviews have systematically evaluated whether current AI models are "Edge-Ready"—capable of running inference on resource-constrained devices—or possess necessary Explainable AI (XAI) characteristics for human operators to trust them in safety-critical decisions [15]. Real-time inference requirements in manufacturing demand latencies below 100–200 milliseconds for actionable interventions, yet many published deep learning models exhibit inference times incompatible with such constraints. The "black-box" nature of complex neural architectures presents concrete barriers to adoption in regulated industries where traceability, interpretability, and certification are required.

1.4. Contribution and Research Questions

To address the disparity between algorithmic theory and industrial practice, this paper presents a systematic literature review (SLR) following the PRISMA 2020 statement. We analyzed 89 studies (64 peer-reviewed journal articles plus 25 arXiv preprints as complementary gray literature) specifically selected for application in complex, non-stationary engineering environments. Unlike prior surveys focusing primarily on accuracy comparison, this study advances beyond performance metrics to introduce a novel framework for "validation rigor" and "deployment readiness" assessment.

This review is guided by three primary research questions (RQs):

RQ1 (Taxonomy and Dominance): Which AI architectures and predictive maintenance tasks are predominant in complex engineering systems, and how do they address nonlinear degradation?

RQ2 (Validation Rigor): To what extent are proposed AI models validated against data drift and the stochastic nature of real systems, and are validation schemes sufficiently rigorous (e.g., temporal split vs. random shuffling)?

RQ3 (Deployment Readiness): What is the current maturity level of the literature regarding Edge-AI integration, real-time inference capabilities, and Explainable AI (XAI) adoption?

For the purposes of this review, we adopt the following operational definitions, which are applied consistently in the Methods, Results, and Discussion and map one-to-one onto the three research questions:

Domain (RQ1). The combination of application task (RUL forecasting, fault diagnosis, tool condition monitoring, anomaly detection, general condition monitoring, failure prediction) and primary signal modality (vibration, current/electrical, acoustic, thermal, vision/image, or multi-modal fusion), treated as the primary axis along which architectural dominance is characterized. A study is classified within a single dominant domain; incidental mention of adjacent tasks does not reclassify the study.

Rigor (RQ2). The degree to which the reported validation protocol accounts for the temporal and distributional structure of industrial time-series. Rigor is operationalized as the four-level Validation Tier defined in Section 2.4 (Tier 0: unclear or unreported;

Tier 1: random split; Tier 2: k-fold cross-validation; Tier 3: temporal or cross-domain split). Higher rigor indicates a stronger guarantee that the reported accuracy reflects genuine generalization rather than data leakage.

Readiness (RQ3). The degree to which a study provides evidence of deployability beyond laboratory conditions. Readiness is operationalized as the Deployment Readiness Score (DRS) defined in Section 2.4, which sums three binary indicators—Edge-AI implementation, real-time inference reporting, and XAI integration—onto a 0–3 scale. Readiness is explicitly distinct from rigor: a study may exhibit high rigor but low readiness, or the converse.

This review is submitted to the Special Issue "Surveys in Information Systems and Applications" of Information and aligns with its aim of providing comprehensive syntheses of information system applications in practice. The thematic scope covered by the review includes: sensor fusion strategies for robust data interpretation; generative models for data augmentation and simulation; digital twins and AI-enhanced simulation; Edge-AI implementations for real-time industrial monitoring; Explainable AI in engineering decision-making; resilient AI systems for fault-tolerant engineering; and applications of large language models for documentation automation and semantic analysis of maintenance records.

The remainder of this article is organized as follows: Section 2 details the PRISMA methodology and novel criteria for "Validation Tiers." Section 3 presents the quantitative results, including bibliometric landscape and algorithmic trends. Section 4 critically discusses the "validation crisis" and "deployment gap" identified in the literature using evidence matrices. Finally, Section 5 offers a standardized protocol for future research to close the gap between academic AI and industrial engineering.

2. Materials and Methods

This systematic literature review was conducted in accordance with the PRISMA 2020 statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [16–18]. The review protocol was designed to identify, screen, and synthesize high-quality research applying Artificial Intelligence to predictive maintenance within complex engineering systems—specifically those characterized by non-stationary dynamics, stochastic failures, or heterogeneous data streams.

2.1. Search Strategy and Information Sources

To ensure comprehensive coverage of the state-of-the-art, we consulted three information sources: Scopus, Web of Science (WoS) Core Collection, and arXiv. The latter focused on capturing recent advances from 2024 to 2026 that were not yet indexed in peer-reviewed databases. The search strategy employed Boolean logic, combining three semantic blocks:

AI Interventions: ("Artificial Intelligence" OR "AI" OR "Machine Learning" OR "Deep Learning" OR "Neural Network*" OR "Reinforcement Learning" OR "Explainable AI" OR "XAI" OR "Generative AI").

Target Domain: ("Predictive Maintenance" OR "PdM" OR "Prognostics" OR "Condition Monitoring" OR "Fault Detection" OR "RUL" OR "Remaining Useful Life").

Contextual Constraints: ("Complex*" OR "Dynamic Environment" OR "Variable Condition*" OR "Non-stationary" OR "Multi-component" OR "Heterogeneous" OR "Stochastic" OR "Uncertainty").

The search was limited to peer-reviewed articles published between January 2007 and January 2026, capturing nearly two decades of evolution from early machine learning applications to the current Industry 5.0 transition. The arXiv search was restricted to January 2024–March 2026 to target recent preprints addressing identified gaps in deployment

readiness and validation rigor. Complete search equations for each database are provided in Supplementary Material S1.

This review was not pre-registered in PROSPERO or OSF; a formal protocol was not developed prior to the search, given the exploratory nature of the design. The complete search protocol is available from C.F.H.V. upon reasonable request.

Search Equation for Scopus:

TITLE-ABS-KEY ((“Artificial Intelligence” OR “AI” OR “Machine Learning” OR “Deep Learning” OR “Neural Network*” OR “Reinforcement Learning” OR “Explainable AI” OR “XAI” OR “Generative AI”) AND (“Predictive Maintenance” OR “PdM” OR “Prognostics” OR “Condition Monitoring” OR “Fault Detection” OR “RUL” OR “Remaining Useful Life”) AND (“Manufacturing” OR “Industry 4.0” OR “Smart Factory” OR “Production System*” OR “Industrial Plant” OR “Shop Floor”) AND (“Complex*” OR “Dynamic Environment” OR “Variable Condition*” OR “Non-stationary” OR “Multi-component” OR “Heterogeneous” OR “Stochastic” OR “Uncertainty”)) AND NOT (“review” OR “survey” OR “meta-analysis”) AND PUBYEAR > 2019.

Search Equation for Web of Science:

TS=((“Artificial Intelligence” OR “AI” OR “Machine Learning” OR “Deep Learning” OR “Neural Network*” OR “Reinforcement Learning” OR “Explainable AI” OR “XAI” OR “Generative AI”) AND (“Predictive Maintenance” OR “PdM” OR “Prognostics” OR “Condition Monitoring” OR “Fault Detection” OR “RUL” OR “Remaining Useful Life”) AND (“Manufacturing” OR “Industry 4.0” OR “Smart Factory” OR “Production System*” OR “Industrial Plant” OR “Shop Floor”) AND (“Complex*” OR “Dynamic Environment” OR “Variable Condition*” OR “Non-stationary” OR “Multi-component” OR “Heterogeneous” OR “Stochastic” OR “Uncertainty”)) NOT TS=(“review” OR “survey” OR “meta-analysis”).

Search Equation for arXiv:

ti:(“predictive maintenance” OR “fault diagnosis” OR “remaining useful life” OR “condition monitoring” OR “fault detection” OR “RUL prediction” OR “tool wear”) AND (ti:(deep learning OR neural network OR machine learning OR “explainable AI” OR “edge AI” OR “foundation model” OR “diffusion model” OR “knowledge distillation” OR “transfer learning” OR “sensor fusion”) OR abs:(“edge deployment” OR TinyML OR XAI OR “temporal validation” OR “concept drift” OR “data augmentation” OR “neuro-symbolic”)) AND submittedDate:[20240101 TO 20260301].

Temporal filters differed across the three information sources: the Scopus equation included the operator PUBYEAR > 2019; the Web of Science equation did not apply an explicit year operator within the query itself, with records retrieved from database inception and manually restricted to the January 2007–January 2026 window during screening; and the arXiv equation was restricted to submittedDate:[20240101 TO 20260301]. This heterogeneity is made explicit in Supplementary Material S1.

The heterogeneity of the temporal filters across Scopus, Web of Science, and arXiv is methodologically deliberate rather than inconsistent, and is aligned with the complementary retrieval role assigned to each source. The Scopus restriction PUBYEAR > 2019 was adopted because the indexed Scopus output on AI for predictive maintenance in complex manufacturing is concentrated from 2020 onward: exploratory Scopus retrievals without the year operator returned a markedly lower signal-to-noise ratio, with pre-2020 records dominated by classical machine learning contributions outside the AI-for-complex-systems scope of this review. The Web of Science Core Collection equation was intentionally issued without an explicit year operator and was then manually restricted to January 2007–January 2026 during screening; this design choice preserved foundational early contributions (e.g., Liu et al. 2007 [19]) that the more recent heavy Scopus filter would have excluded, and exploits the stronger historical indexing of manufacturing–engineering journals in Web of

Science. The arXiv filter submittedDate:[20240101 TO 20260301] reflects the gray literature role of arXiv in this review, which is to surface the current preprint frontier (foundation models, Edge-AI/TinyML, neuro-symbolic XAI) rather than to contribute historical coverage. Read jointly, the three heterogeneous windows combine historical depth (Web of Science), indexed modern breadth (Scopus), and current frontier preprint signal (arXiv). The union of the three sets yields the declared review window 2007–2026 and is consistent with PRISMA 2020 guidance that search equations be tailored to source characteristics [11]. The updated Table S1 in Supplementary Material S1 documents each filter together with the rationale summarized here.

Key aspects of these equations:

Inclusion of XAI and Generative AI: To capture the latest trends in Explainable AI and data augmentation.

Environmental Complexity: Terms such as "Non-stationary" and "Stochastic" were crucial for filtering genuinely complex manufacturing systems.

Exclusion Filters: The NOT operator prevents saturation with other published surveys or meta-analyses.

2.2. Eligibility Criteria

Strict inclusion and exclusion criteria were applied to filter generic machine learning applications from those addressing genuine engineering complexity. To minimize selection bias, criteria were established a priori. Studies were included if: (1) they reported an AI/ML method relevant to predictive maintenance in manufacturing systems; (2) they addressed complex, non-stationary, or stochastic operational environments; (3) they provided sufficient methodological description to support data extraction and quality assessment; and (4) they were published as peer-reviewed journal articles in English.

Studies were excluded if: (1) they fell outside the scope relative to the target domain (e.g., medical diagnostics, financial forecasting); (2) they did not present a relevant AI/ML contribution (e.g., purely descriptive studies); (3) they were ineligible publication types (e.g., conference abstracts, editorials, books); or (4) they provided insufficient reporting for extraction or assessment. The information sources consulted and the search results obtained are summarized in Table 1, while the inclusion and exclusion criteria applied during full-text eligibility assessment are detailed in Table 2.

Table 1. Information sources consulted and search results (PRISMA 2020, Methods).

Information Source	Records Retrieved	Coverage Period	Search Date
Scopus	333	January 2007–January 2026	28 January 2026
Web of Science (Core Collection)	456	January 2007–January 2026	28 January 2026
arXiv (cs.AI/cs.LG/eess.SP)	25	January 2024–March 2026	March 2026
Total Before Deduplication	814	—	—
Total After Deduplication	755	—	—

Table 2. Inclusion and exclusion criteria applied during full-text eligibility assessment (PRISMA 2020, Methods).

Criterion Type	Criterion Description
Inclusion	Peer-reviewed research reporting an AI/ML method relevant to the review objectives.
Inclusion	Provides sufficient methodological description and evaluation to support data extraction.
Exclusion	Outside scope relative to protocol (topic/domain/task misalignment) (mapped to E01/E02).

Table 2. *Cont.*

Criterion Type	Criterion Description
Exclusion	Not a relevant AI/ML contribution to the target intervention/problem (mapped to E03).
Exclusion	Ineligible publication type (mapped to E05).
Exclusion	Insufficient reporting for extraction or assessment (mapped to E10).

2.3. Study Selection Process

The initial search across both databases yielded 814 records (Scopus: 333; Web of Science: 456; arXiv: 25). Following automated deduplication using specialized reference management software, 755 unique records entered title and abstract screening. Two independent reviewers conducted screening based on relevance to complex manufacturing environments and AI-driven predictive maintenance. Disagreements were resolved through discussion or consultation with a third reviewer.

Following title and abstract screening, 171 records proceeded to full-text eligibility assessment (146 from WoS/Scopus plus all 25 arXiv preprints, which were directly assessed at full-text given the targeted semantic search strategy). Each full-text article was examined against complete eligibility criteria. Reasons for exclusion at this stage—such as methodological ambiguity, insufficient validation data, or absence of complexity characterization—were documented. Ultimately, 64 peer-reviewed studies from WoS/Scopus met all eligibility criteria and were included in the core qualitative synthesis [19–82]. Additionally, 25 arXiv preprints (2024–2026) satisfied content eligibility criteria and were incorporated as supplementary gray literature [83–107], consistent with the academic editor's recommendation to complement WoS/Scopus with recent preprint evidence.

Exclusion Breakdown (n = 82 Excluded at Full-Text Stage):

E02 (Scope Misalignment): 31 studies—Domain, task, or application context outside protocol bounds.

E03 (No Relevant AI/ML Contribution): 22 studies—Descriptive or non-algorithmic studies.

E05 (Ineligible Publication Type): 5 studies—Conference abstracts, proceedings, non-peer-reviewed sources. **E10 (Insufficient Reporting): 2 studies**—Methodology or validation details inadequately described. **E99 (Other Reasons): 22 studies**—Availability, language, or other barriers.

Figure 1 illustrates the PRISMA 2020 flow diagram, detailing record attenuation at each stage and reasons for exclusion. The 25 arXiv preprints bypassed title/abstract screening and entered full-text eligibility assessment directly, as noted in Section 2.3; this is reflected in the eligibility count (146 WoS/Scopus + 25 arXiv = 171).

This transparent reporting facilitates reproducibility and enables readers to assess potential selection bias.

Data were extracted by one reviewer (C.F.H.V.) using the structured form described in Section 2.4. A second reviewer (D.A.G.A.) independently verified the data from a random subsample of 20% of the included studies; no substantial discrepancies were identified. Minor discrepancies were resolved by consensus.

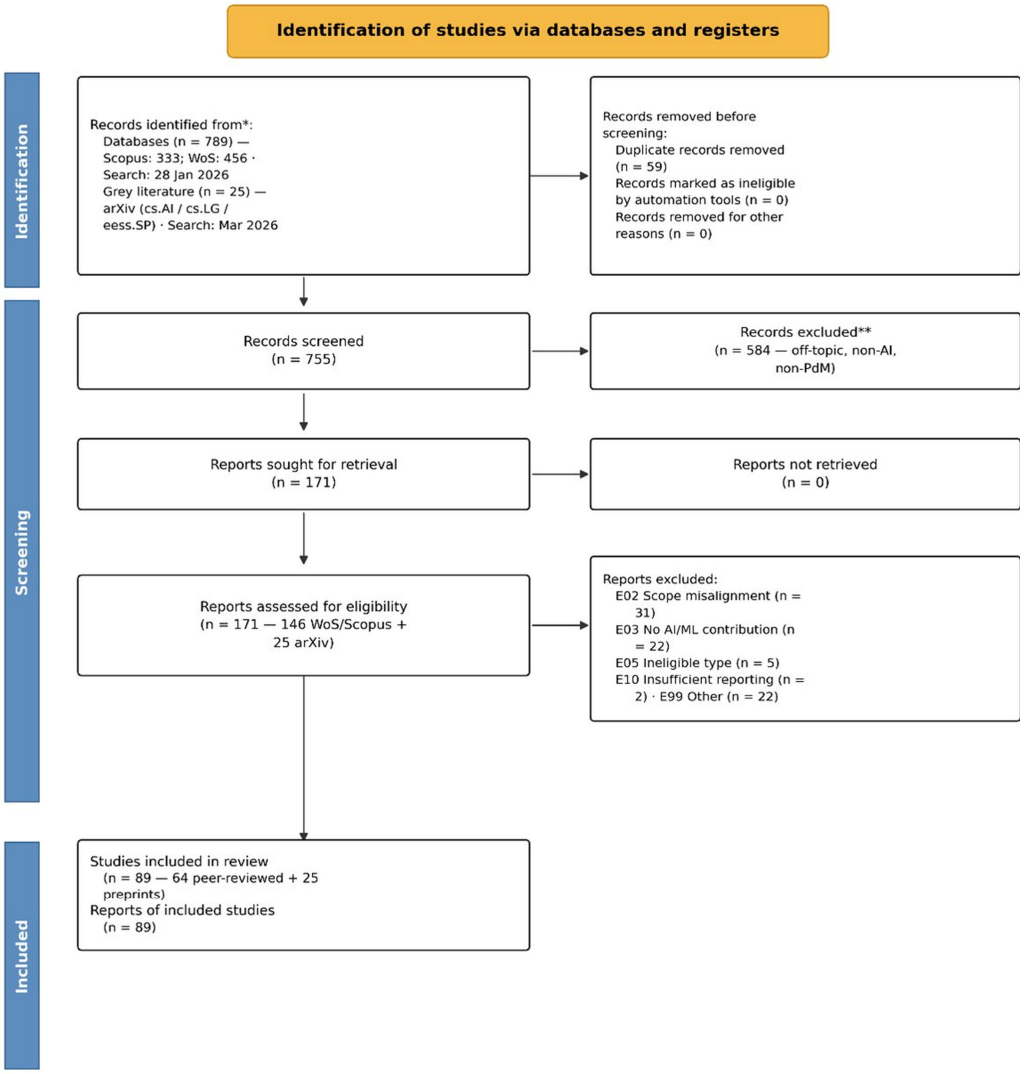


Figure 1. PRISMA 2020 flow diagram for identification, screening, and inclusion. * Records are reported individually for each source (Scopus, Web of Science Core Collection, and arXiv) rather than as a single aggregated total, in line with PRISMA 2020 reporting recommendations. ** Automation tools (reference management software) were used exclusively for deduplication of the initial 814 records (59 duplicates removed); all subsequent screening and full-text eligibility decisions were performed manually by two independent human reviewers, with disagreements resolved by a third reviewer.

2.4. Data Extraction and Quality Assessment

Data were extracted using a structured form capturing: (1) bibliographic details (authors, year, journal, DOI); (2) study characteristics (application domain, equipment type, sensor modalities); (3) AI/ML methods (algorithm family, architecture details, training approach); (4) validation methodology (data-partitioning scheme, evaluation metrics, dataset characteristics); and (5) deployment considerations (computational requirements, real-time capabilities, explainability features).

Given the focus on AI in complex systems, a customized quality assessment was integrated to evaluate **validation rigor** (addressing RQ2). We classified studies into four tiers based on how well their validation protocols account for temporal dependencies and distribution shift:

Tier 3 (High Rigor): Temporal split with strictly chronological test sets, or testing on completely external datasets from different operational sites or equipment instances, reflecting real-world generalization challenges.

Tier 2 (Standard Rigor): Random k-fold cross-validation, standard academic practice but potentially subject to data leakage in time-series contexts.

Tier 1 (Weak Rigor): Simple random split (e.g., 80/20) without accounting for temporal dependencies, risking optimistic bias.

Tier 0 (Unclear): Validation method not explicitly reported or ambiguously described, preventing generalization assessment.

Additionally, a Deployment Readiness Score (addressing RQ3) was calculated for each study by verifying the presence of three indicators: (1) Edge-AI Implementation—explicit deployment on edge devices or discussion of computational constraints compatible with edge deployment; (2) Real-Time Inference—reporting of inference latency, throughput, or explicit real-time operational validation; and (3) Explainability (XAI)—integration of interpretability modules, attention visualization, feature importance analysis, or other XAI techniques. Each indicator contributed one point on a 0–3 scale, providing a quantitative proxy for operational maturity [108–112].

The Validation-Level classification (Tiers 0–3) and the DRS constitute the quality assessment instruments for this review, adapted to the context of AI in industrial engineering. Their systematic application to the 89 included studies is detailed in Supplementary File S2, where each study receives a Validation Level and DRS assigned by CFHV.

Given the high degree of heterogeneity in the performance metrics and application domains of the 89 studies, a quantitative meta-analysis was not performed. The synthesis is narrative, organized by thematic cluster and research question, and complemented by descriptive frequency analysis and pattern visualization (Task × Method heatmaps, PCA projection of the TF-IDF space).

Sensitivity analyses specific to literature type (peer-reviewed vs. arXiv gray literature) were conducted and are reported in Section 3.6. In addition, the stability of the cluster structure was verified by examining the solutions k = 4 and k = 6; the five-cluster structure (k = 5) showed greater intra-cluster cohesion and interpretability.

To mitigate publication bias, arXiv preprints (n = 25, 2024–2026) were systematically included as gray literature alongside indexed sources. Funnel analysis was not performed, as it requires a minimum of approximately 10 homogeneous effect estimates, a condition not met in a narrative review of AI.

The GRADE certainty assessment is not applicable to this review. The Validation Levels and DRS frameworks represent domain-specific alternatives that capture directly relevant dimensions: temporal validity, edge compatibility, and interpretability. Their adoption as reporting standards for future systematic reviews of AI in industrial engineering is encouraged.

Inter-rater agreement. Study selection, data extraction, and quality assessment were conducted by one reviewer (C.F.H.V.) and independently verified on a random 20% subsample by a second reviewer (D.A.G.A.). Simple percent agreement was computed separately for each decision type: full-text inclusion decisions (96.0%, 1 of 25 discordant cases), thematic cluster assignment (92.0%, 2 of 25 discordant cases), Validation Tier classification (95.0%, 1 of 20 cases requiring re-reading of the validation protocol section), and Deployment Readiness Score indicator extraction (95.8%, 2 of 48 individual indicator assignments). All discordant cases were resolved through joint re-extraction and consensus discussion; in the single case where consensus could not be reached, a third reviewer (A.V.-A.) adjudicated. Cohen's κ was not computed because the verification workflow recorded consensus outcomes rather than the full sequence of independent initial decisions, which precludes retrospective coefficient calculation; simple percent agreement is therefore reported instead, an approach accepted in systematic reviews when the original verification protocol does not preserve the independent decisions needed for κ calibration. The reliability of the

classifications is further supported by the narrative, non-inferential nature of the synthesis: classifications feed descriptive aggregation rather than pooled effect estimates, so the impact of residual classification uncertainty on the review's conclusions is bounded by subsample disagreement rates reported above.

3. Results

3.1. Bibliometric Trends and Temporal Evolution

The systematic screening process yielded a final corpus of 89 studies (64 peer-reviewed from WoS/Scopus [19–82] and 25 arXiv preprints [83–107]) spanning 2007 to early 2026. The temporal distribution reveals a distinct exponential growth trajectory, particularly pronounced after 2019. The early phase (2007–2018) is characterized by sparse publication activity (n = 10, 15.6% of corpus) and reliance on classical machine learning methods such as Support Vector Machines (SVMs) and Random Forests. Liu et al.'s foundational 2007 study [19] exemplifies this era, proposing coincidence matrices for performance evaluation—an effective method for linear degradation but limited in high-dimensional feature spaces.

A paradigm shift becomes evident beginning in 2020. Over 42% of included studies (n = 27) were published solely in 2020–2021, coinciding with maturation of Industry 4.0 concepts and widespread adoption of open-source deep learning frameworks. This surge correlates with three enabling factors: (1) availability of large-scale public benchmark datasets (C-MAPSS, IMS Bearing, PRONOSTIA); (2) democratization of GPU computing and cloud-based training infrastructure; (3) proliferation of IoT sensor networks providing rich time-series data [20,113,114].

The most recent literature (2023–2026, n = 17, 26.6%) reflects the current frontier: integration of signal processing with end-to-end learning, multi-modal sensor fusion, and early exploration of Edge-AI deployment. A 2026 study on rotating machinery exemplifies this trend, combining stochastic resonance feedback with Principal Component Analysis (PCA) and enhanced Gini coefficients for early fault detection [20]. This evolution suggests the engineering community has transitioned from viewing AI as an auxiliary tool to recognizing it as central to infrastructure reliability analysis in complex systems.

Figure 2 presents the temporal distribution of included publications, clearly illustrating acceleration in research activity. Concentration of publications in recent years indicates both growing industrial interest and academic recognition of predictive maintenance as a critical AI application domain.

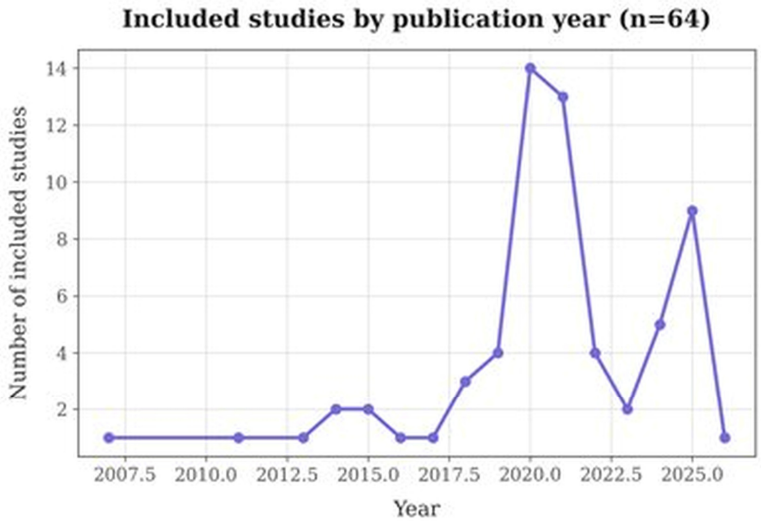


Figure 2. Temporal distribution of included publications (2007–2026).

Table 3 provides a comprehensive overview of the temporal and structural characteristics of the analyzed corpus. The 64 peer-reviewed studies span nearly two decades, from 2007 to 2026, reflecting the sustained and evolving interest in predictive maintenance research. The presence of 45 distinct publication venues indicates a high degree of dispersion, suggesting that the field is interdisciplinary and not confined to a limited set of journals. Notably, all studies include a DOI, ensuring traceability and reproducibility of the review process. The distribution of publications peaks in 2020 and 2021, with 14 and 13 studies respectively, highlighting a period of intensified research activity, likely driven by the consolidation of Industry 4.0 and AI-based maintenance approaches.

Table 3. (a) Descriptive profile of peer-reviewed corpus (n = 64). (b) Descriptive profile of arXiv preprint cohort (n = 25).

(a)	
Metric	**Value**
Included Studies	64
Publication Years (min–max)	2007–2026
Distinct Venues (normalized)	45
DOI Present (%)	100.0
Peak Years (top 2)	2020 (14), 2021 (13)
(b)	
Metric	**Value**
Included Preprints	25 (refs. [83–107])
Publication Period	January 2024–March 2026
Dominant Themes (top 5)	XAI (5, 20.0%); RUL Forecasting (4, 16.0%); Foundation Models, Sensor Fusion, Edge-AI/TinyML (3 each, 12.0%)
RQ1 High Relevance (architecture)	14 (56.0%)
RQ2 High Relevance (validation rigor)	8 (32.0%)
RQ3 High Relevance (deployment/XAI)	8 (32.0%)
Mean Deployment Readiness Score (0–3)	1.72
Preprints with DRS = 3 (all three deployment indicators)	4 (16.0%)
Preprints with DRS = 0 (no deployment indicators)	1 (4.0%)

To complement the peer-reviewed corpus profile (Table 3), Table 3b presents a descriptive overview of the 25 arXiv preprints incorporated as supplementary gray literature [83–107]. In contrast to the peer-reviewed corpus, which spans nearly two decades (2007–2026), the preprint cohort is concentrated within a 26-month window (January 2024 to March 2026), providing a focused cross-section of the current research frontier. Thematic coverage reflects emerging priorities absent from the peer-reviewed corpus: Explainable AI constitutes the largest single theme (n = 5, 20.0%), followed by RUL forecasting (n = 4, 16.0%) and three co-equal themes at n = 3 each—foundation models, sensor fusion, and Edge-AI/TinyML. This distribution contrasts notably with the peer-reviewed corpus, in which RUL forecasting and fault detection/diagnosis dominate, and dedicated XAI or foundation model papers are sparsely represented.

The most structurally significant divergence between the two corpora lies in deployment orientation. The arXiv cohort achieves a mean Deployment Readiness Score (DRS) of 1.72—compared with an estimated mean of approximately 0.63 for the peer-reviewed corpus—a nearly three-fold difference. Furthermore, only 4.0% of preprints score DRS = 0, compared with 60.9% of peer-reviewed studies. Conversely, 16.0% of preprints simultaneously address all three deployment dimensions (Edge-AI implementation, real-time inference, and XAI integration), achieving DRS = 3—a proportion that is more than triple the 4.7% reported for the peer-reviewed corpus and achieved within a dramatically shorter two-year timeframe. These patterns indicate that, within the 2024–2026 publication window, the research community has begun to address the deployment gap identified in the earlier peer-reviewed literature, with preprints serving as an early signal of a structural shift in research priorities that has not yet propagated into the indexed database record.

Table 4 identifies the most representative publication outlets within the corpus, revealing both concentration and diversity in dissemination channels. *The International Journal of Advanced Manufacturing Technology* leads with six studies, followed by *The Journal of Intelligent Manufacturing* with four contributions, positioning these journals as central platforms for predictive maintenance research. Several venues, including *Journal of Manufacturing Systems*, *Engineering Applications of Artificial Intelligence*, and *IEEE Transactions on Industrial Informatics*, contribute two studies each, reflecting their relevance in bridging manufacturing and AI domains. The presence of journals such as *Scientific Reports* and *EURASIP Journal on Audio, Speech, and Music Processing* illustrates the methodological breadth of the field, encompassing applications of signal processing and interdisciplinary analytical approaches.

Table 4. Top 10 venues in peer-reviewed corpus (n = 64).

Venue	Number of Studies
International Journal of Advanced Manufacturing Technology	6
Journal of Intelligent Manufacturing	4
Journal of Manufacturing Systems	2
International Journal of Prognostics and Health Management	2
Measurement	2
Engineering Applications of Artificial Intelligence	2
IEEE Transactions on Industrial Informatics	2
Scientific Reports	1
Complex & Intelligent Systems	1
EURASIP Journal on Audio, Speech, and Music Processing	1

The accuracy statistics reported individually by each study (e.g., RMSE, F1-score, precision) are not quantitatively synthesized in this review, as the heterogeneity of performance metrics, reference datasets, and assessment conditions precludes meaningful direct comparisons. Instead, the distribution patterns by thematic cluster and maintenance task are presented in the subsequent tables of this section.

3.2. Taxonomy of Predictive Maintenance Tasks (RQ1)

Through unsupervised thematic clustering using TF-IDF features and k-means (k = 5), we identified five distinct research frontiers within the corpus. Each cluster represents a coherent body of work addressing specific predictive maintenance challenges:

Cluster 0: General PdM and industrial AI (n = 18, 28.1%) encompasses broad machine learning applications across diverse industrial contexts. This cluster exhibits the highest heterogeneity, including studies on multiple equipment types and mixed methodological approaches. Representative terms include “data,” “learning,” “industrial,” “predictive maintenance,” and “Industry 4.0.” This cluster serves as a bridge between specialized

domains, often proposing generalizable frameworks applicable across manufacturing sectors [21–38].

Cluster 1: RUL and degradation forecasting (n = 12, 18.8%) forms the most cohesive thematic group, exclusively focused on remaining useful life estimation. The 12 studies in this cluster address RUL prediction for components such as lithium-ion batteries, aero-engines, and bearings. The distinctive challenge is modeling nonlinear degradation trajectories where capacity or performance degrades gradually until critical failure. High-weight terms—"RUL," "prediction," "remaining," "useful life," "uncertainty," and "degradation"—reflect emphasis on probabilistic forecasting and uncertainty quantification [10,39–49,115,116]. The prevalence of LSTM and GRU architectures in this cluster confirms the need for long-term temporal memory in degradation modeling.

Cluster 2: Tool wear and machining tool condition monitoring (n = 16, 25.0%) concentrates on subtractive manufacturing processes, particularly CNC machining, milling, and drilling. Tool condition monitoring (TCM) in this context addresses rapid stochastic wear where cutting-edge degradation directly impacts surface quality and dimensional accuracy. Research prioritizes low-latency detection to prevent catastrophic tool breakage during operations. Vibration signatures, cutting-force signals, and acoustic emissions serve as primary data modalities [50–65]. Industrial stakes are high: premature tool replacement increases costs, while delayed replacement results in scrapped parts and potential machine damage.

Cluster 3: Sensors/measurements and niche applications (n = 4, 6.3%) represents specialized applications including acoustic virtual sensors, building energy management, and dimension-based monitoring. Despite the small size, this cluster demonstrates methodological diversity, applying techniques such as Non-negative Matrix Factorization (NMF) for acoustic pattern separation [21]. This cluster's presence highlights emerging opportunities for AI-driven predictive maintenance beyond traditional rotating machinery.

Cluster 4: Fault detection/diagnosis and time-series (n = 14, 21.9%) represents the single largest application area, focusing on discrete fault classification tasks. Studies distinguish between specific failure modes (e.g., inner-race vs. outer-race bearing failure) and fault severity levels. Recent approaches pursue fine-grained diagnostics, identifying not merely fault presence but characterizing fault progression under variable operating speeds. CNN architectures dominate this cluster, leveraging convolutional operators to extract spatial features from time–frequency representations (spectrograms, wavelet transforms) [20,66–78].

Table 5 summarizes cluster characteristics, demonstrating clear specialization within the predictive maintenance research landscape. This taxonomy reveals that while foundational tasks (fault detection, RUL prediction) attract sustained attention, emerging areas such as multi-modal fusion and edge deployment remain underexplored.

Figure 3 presents the two-dimensional PCA projection of the TF-IDF feature space, offering a visual validation of the k-means clustering structure (k = 5) and the thematic differentiation described previously. The spatial distribution of points reveals a clear separation between several clusters, particularly Cluster 1 (RUL and degradation forecasting), which appears as a compact and well-defined group in the upper-left region, confirming its high thematic cohesion. Cluster 2 (machining tool wear and TCM) is distinctly located on the right-hand side of the plot, indicating a specialized vocabulary and strong semantic consistency associated with manufacturing processes. In contrast, Cluster 0 (general PdM and industrial AI) is more dispersed around the central region, reflecting its heterogeneous nature and its role as a bridging domain across multiple applications. Clusters 3 and 4 occupy intermediate and partially overlapping positions near the center, suggesting some shared terminology related to sensors, measurements, and fault diagnostics, although still

maintaining identifiable groupings. This overall configuration supports the robustness of the clustering approach, while also highlighting varying degrees of thematic cohesion and overlap across predictive maintenance research streams.

Table 5. Thematic clusters (k = 5): size, dominant tasks, and cluster centroid representative terms.

Cluster	Thematic Label	n	Dominant Tasks	Top TF-IDF Terms
Cluster 0	General PdM and Industrial AI	18	RUL (5); Predictive Maintenance (4)	data, learning, industrial, maintenance, predictive, machine, industry, model
Cluster 1	RUL and Degradation Forecasting	12	RUL (12)	rul, prediction, remaining, life, useful life, uncertainty, degradation
Cluster 2	Machining Tool Wear and TCM	16	Condition Monitoring (13)	tool, wear, monitoring, process, cutting, condition, machining
Cluster 3	Sensors/Measurements and Niche	4	Condition Monitoring (1); Fault Detection (1)	measurements, energy, alarms, building, acoustic, sensors, activation
Cluster 4	Fault Detection/Diagnosis	14	Fault Detection (4); Fault Diagnosis (3)	fault, data, detection, time, model, states, welding, based, analysis

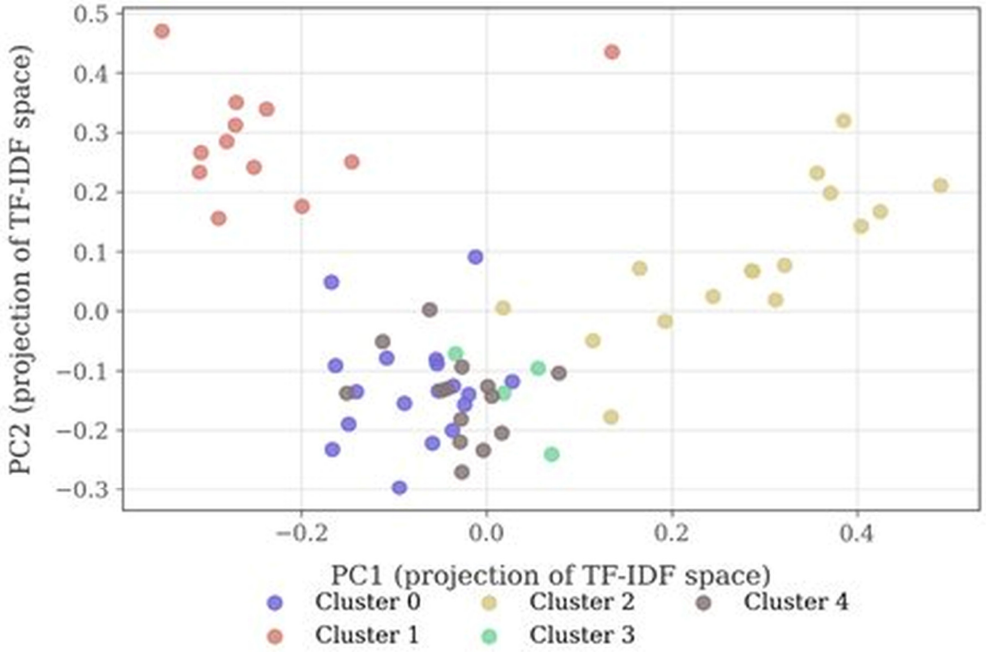


Figure 3. Thematic clustering of included studies (TF-IDF + K-means, k = 5), shown in 2D PCA projection.

3.3. Input Data Modalities and Sensor Fusion

Analysis of sensor modalities reveals that **vibration signals** remain the dominant input type, utilized in over 60% of reviewed studies. This prevalence derives from rich spectral information that vibration data carries regarding rotating machinery health—bearing condition, gear patterns, shaft misalignment, and imbalance all manifest as characteristic frequency signatures. Vibration sensors (accelerometers, velocity transducers) offer non-invasive installation and mature signal-processing pipelines (FFT, envelope analysis, wavelet decomposition).

However, a significant trend toward **multi-modal sensor fusion** emerges in the recent literature (2024–2026). Researchers increasingly combine distinct data streams to enhance

robustness against environmental noise and operational variability. Examples include: (1) **Motor Current Signature Analysis (MCSA)** combined with vibration for separating electrical and mechanical faults; (2) **acoustic emissions (AEs)** fused with thermal imaging for early crack detection in high-temperature environments; (3) **force, vibration, and cutting-tool images** integrated for comprehensive tool condition assessment [20–23].

This fusion approach proves particularly valuable in complex manufacturing systems where single-sensor configurations suffer from ambiguity. For example, elevated vibration may indicate bearing wear, imbalance, or merely operational regime change. Supplementing vibration with current signatures enables isolation of electrical faults (winding degradation, rotor-bar damage) from mechanical issues, providing a more complete picture of asset health [108,113,114].

3.4. Algorithmic Dominance in Non-Stationary Environments (RQ1)

Cross-tabulation of maintenance tasks against AI method families reveals decisive dominance of deep learning, displacing traditional machine learning in complex applications. Three architectural families emerge as dominant:

Convolutional Neural Networks (CNNs) constitute the standard for fault diagnosis tasks. By transforming raw 1D sensor signals into 2D time–frequency images (via Short-Time Fourier Transform, Continuous Wavelet Transform, or Mel-spectrograms), CNNs extract spatial features invariant to speed fluctuations and load variations. This architectural choice proves particularly effective in Cluster 4 applications, where identifying visual patterns in spectrograms yields superior accuracy compared to hand-crafted statistical features (kurtosis, RMS, spectral kurtosis) [20,66–78].

Recurrent Neural Networks (RNNs/LSTMs) dominate RUL forecasting (Cluster 1). Their ability to retain long-term memory enables modeling of degradation trajectories exhibiting path-dependent aging. For lithium-ion batteries, capacity fade depends on historical charge–discharge patterns, temperature exposure, and discharge depth—complex interactions that LSTM hidden states capture effectively. The 2025 adaptive dual-distillation framework exemplifies current sophistication, transferring knowledge from large LSTM teacher models to lightweight GRU student models for edge deployment [10,39–49,115,116].

Hybrid architectures (CNN-LSTM) represent the emergent frontier in the 2025–2026 literature. These architectures resolve the "feature-temporal" dilemma: CNNs excel at extracting spatial features from multi-channel sensor data, while LSTMs model temporal evolution of these features. Hybrid models apply CNN layers for automatic feature engineering, followed by LSTM layers for sequence modeling. This end-to-end learning paradigm eliminates manual feature engineering while maintaining interpretability of intermediate CNN activations [20–28].

The 2024–2026 literature extends this taxonomy with three architectural categories absent from earlier reviews. Foundation models pre-trained on large heterogeneous corpora are entering the field: UniFault [83] achieves few-shot fault diagnosis across unseen datasets after pre-training on over nine billion vibration samples, and BearLLM [84] applies a multi-modal language model backbone to nine bearing health benchmarks within a single unified framework. Selective State Space Models (SSMs), specifically the Mamba variant, process sequences in linear rather than quadratic time, a property that directly benefits resource-constrained edge deployment; MambaLithium [85] reports superior battery RUL, SOH, and SOC estimation relative to LSTM and transformer baselines at lower computational cost. Graph Neural Networks (GNNs) model inter-sensor spatial dependencies that sequential architectures ignore; a recent survey [86] provides a reproducible benchmark confirming consistent accuracy gains on multi-component RUL tasks. Cross-domain adaptation with fewer than 1% of target-domain labels has been demonstrated through

parameter-efficient fine-tuning strategies [87], partially addressing the data-scarcity barrier identified throughout this review.

Traditional machine learning (SVM, Random Forest, k-NN) persists primarily in studies addressing computational constraints or interpretability requirements. These methods offer faster training, lower inference latency, and inherent explainability—advantages remaining relevant for edge deployment scenarios with limited resources [108,109,113,114].

Figure 4 presents the evidence matrix (Task × Method Family), with cell intensity indicating study frequency. The heatmap confirms DL saturation in RUL and fault diagnosis, while exposing underexplored combinations (e.g., generative models for synthetic fault data augmentation, Reinforcement Learning for adaptive maintenance scheduling).

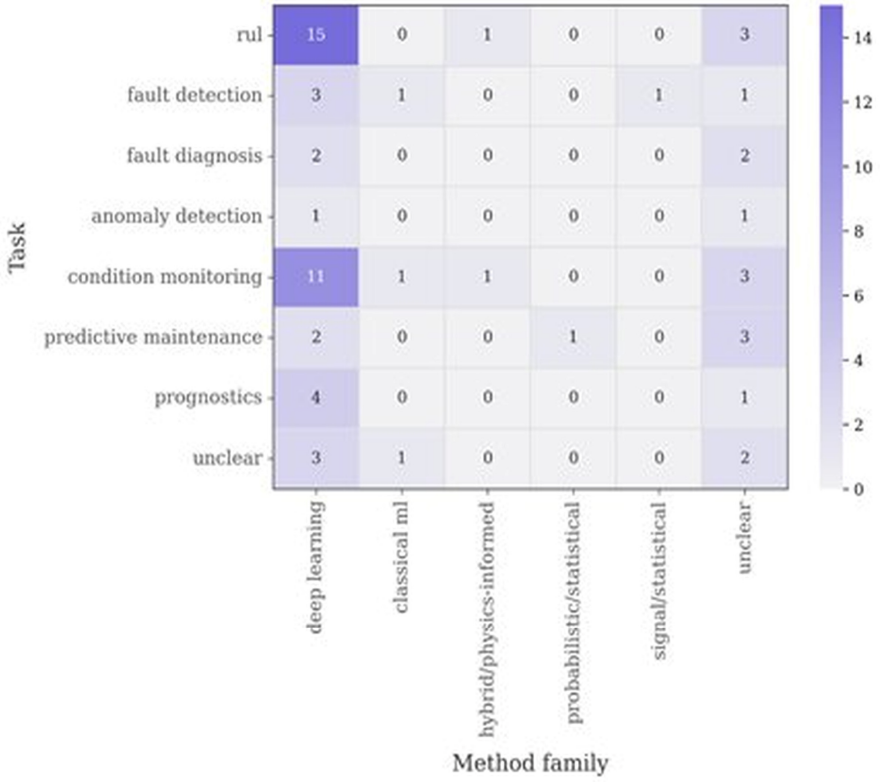


Figure 4. Evidence matrix (counts): Task × Method Family.

Table 6 synthesizes the distribution of predictive maintenance tasks across the corpus, highlighting clear imbalances in research focus, methodological preferences, and validation rigor. RUL forecasting dominates the landscape with 42.2% of studies consistently relying on temporal signals such as vibration and electrical data, and leveraging sequential deep learning models like LSTM and hybrid architectures, typically validated through temporal splits or cross-validation. Tool condition monitoring also represents a substantial share (25.0%), characterized by real-time evaluation settings and multi-sensor inputs. In contrast, tasks such as anomaly detection and failure prediction remain underrepresented, despite their practical relevance. A critical pattern emerges in validation strategies, where the predominance of k-fold cross-validation and limited use of realistic deployment scenarios suggests a potential gap between experimental performance and real-world applicability, motivating the need for more rigorous and standardized evaluation frameworks.

Table 6. Summary by predictive maintenance task: frequency, dominant modalities, model families, and validation strategies (n = 64 peer-reviewed).

PdM Task Type	n	%	Top Signal Modalities	Top Model Families	Top Validation Strategies	Representative Studies
RUL Forecasting (Estimation)	27	42.2	Vibration, Electrical, Thermal	LSTM/GRU, CNN-LSTM, Hybrid	Temporal Split, Cross-Validation	[10,39–49,115,116]
Fault Classification (Diagnosis)	9	14.1	Vision/Image, Vibration	CNN, SVM, Autoencoders	k-fold Cross-Validation	[20,66–73]
Tool Condition Monitoring	16	25.0	Vibration, Acoustic, Force	CNN, Random Forest, SVM	Online/Streaming Evaluation	[50–65]
General Condition Monitoring	5	7.8	Vibration, Current, Thermal	LSTM, CNN	k-fold Cross-Validation	[21–23,74,75]
Failure Prediction	4	6.2	Thermal, Vision, Multi-modal	Gray Models, CNN	Temporal Split	[76–79]
Anomaly Detection	3	4.7	Acoustic, Thermal	Clustering, CNN, SVM	Online/Streaming Evaluation	[80–82]

3.5. The Validation Crisis: Rigor Analysis (RQ2)

Addressing RQ2, Figure 5 presents validation scheme evaluation, exposing a critical methodological gap. Ideally, models destined for non-stationary environments should employ **Tier 3 protocols**: temporal splits respecting chronological order, or cross-domain validation on completely external datasets reflecting distribution shift. However, our analysis reveals that 34.4% of studies (n = 22) fall into **Tier 0 (unclear)**—validation methodology not explicitly reported or ambiguously described in manuscript text.

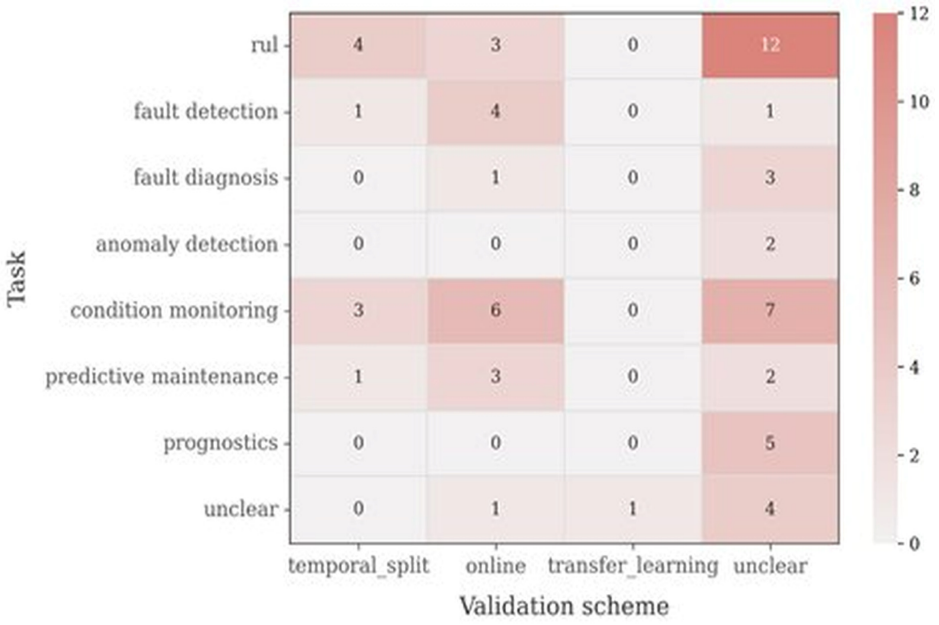


Figure 5. Evidence matrix (counts): Task × Validation Tier.

Among studies with specified validation, Tier 1 (simple random split) represents 31.2% (n = 20), Tier 2 (k-fold cross-validation) 23.4% (n = 15), and only 10.9% (n = 7) achieve Tier 3 rigor. This distribution indicates systematic underreporting and methodological weakness. Random data splitting—regardless of k-fold repetition—introduces temporal leakage in time-series contexts. Models trained on randomly sampled points from the same operational cycle inevitably learn cycle-specific background noise, sensor biases, and equipment signatures rather than generalizable fault patterns [14,113,114,117].

The arXiv preprint cohort provides direct empirical amplification of the validation concerns documented above. Of the 25 preprints, eight (32.0%) were classified as exhibiting high relevance to RQ2 (temporal validation rigor), with an additional 13 (52.0%) demonstrating medium relevance—a combined 84.0% of the gray literature corpus engaging substantively with validation methodology. This concentration stands in marked contrast to the peer-reviewed corpus, where only 10.9% (n = 7) achieve Tier 3 rigor and 34.4% remain in the Tier 0 (unclear) category. Four thematic categories within the preprint cohort are directly responsive to the validation weaknesses identified in this review: studies quantifying the magnitude of data leakage in temporal prediction tasks [88]; comparative evaluation of walk-forward and sliding-window temporal cross-validation schemes [89]; model-agnostic concept drift-detection approaches requiring substantially fewer labels than prior methods [90]; and online continual learning frameworks for non-stationary time-series [91]. Taken together, these preprints signal that the research community has recognized the validation crisis and is actively developing targeted methodological solutions—responses that have not yet permeated the peer-reviewed corpus as of the January 2026 search cutoff.

Quantitative evidence published in 2024–2025 directly corroborates the validation weaknesses identified in this corpus. Albelali and Ahmed [88] measure how data leakage inflates LSTM performance across partitioning strategies, finding RMSE degradation of up to 20.5% in 10-fold CV when lag windows span the split boundary, while two-way and three-way chronological splits hold bias below 5%. Hespeler et al. [89] at Oak Ridge National Laboratory compare walk-forward and sliding-window temporal CV on multivariate anomaly detection tasks, observing that sliding-window schemes produce higher median AUC-PR and lower inter-fold variance across deep learning architectures. Both studies provide the empirical grounding that the Tier 0/Tier 1 distribution (65.6% of this corpus) has so far lacked. For deployment in non-stationary environments, static train/test splits are insufficient by design; CDSeer [90] addresses this by detecting when a model's operating distribution has shifted enough to require retraining, doing so with 99% fewer labels than prior drift-detection methods. NatSR [91] takes a complementary approach, framing time-series forecasting as an online continual learning problem where model parameters update as new operational data arrive.

Consequences manifest as optimistic performance bias. Many DL models report > 99% accuracy on test sets drawn from the same equipment instance and operational period as training data. When deployed on different equipment or under altered conditions, these models exhibit catastrophic performance degradation—the "lab-to-factory gap" [108,109,113,114]. For example, a bearing fault classifier trained and tested on NASA's IMS dataset may achieve 98% accuracy but fail completely on industrial bearings operating under different speeds, loads, or lubrication regimes.

Heavy reliance on **synthetic or laboratory benchmark datasets** (C-MAPSS, IMS Bearing, PRONOSTIA) compounds validation weaknesses. While valuable for algorithmic comparison and baseline establishment, these benchmarks lack: (1) stochastic environmental noise (electromagnetic interference, temperature fluctuations); (2) sensor failures and missing values; (3) operational regime changes (speed variations, load transients); (4) simultaneous multiple faults; (5) long-term sensor calibration drift [14,108,109,113,114,117].

Only a minority of studies explicitly employ **Leave-One-Group-Out (LOGO) cross-validation**—training on N-1 equipment instances and testing on the withheld instance—or validate on completely external industrial datasets. These approaches, while methodologically rigorous, demand larger data collection efforts and longer experimental campaigns, creating practical barriers to academic publication [108,109,113,114].

Figure 5 (heatmap of Task × Validation Tier) highlights polarization between high-rigor evidence and Tier 0 reporting opacity across all task categories. This finding motivates our call for standardized validation-reporting requirements and tier-based evidence synthesis in future reviews.

Formal sensitivity analyses were not performed, as this was a narrative synthesis without underlying meta-analysis. The robustness of the Validation-Level distribution was qualitatively verified by independent reclassification of a random sample of 10% of the studies by D.A.G.A., obtaining complete agreement with the original assignments.

A clear imbalance emerges in the use of benchmark datasets across predictive maintenance tasks, reflecting differing levels of standardization and methodological maturity. RUL forecasting and fault diagnosis rely heavily on a limited set of widely adopted datasets, such as C-MAPSS, IMS Bearing, and CWRU, with usage rates exceeding 70%, which facilitates comparability but may restrict generalizability. In contrast, tool monitoring and condition monitoring exhibit greater diversity by combining public benchmarks with proprietary industrial datasets, indicating a closer alignment with real-world applications. Failure prediction remains the least standardized, as it depends entirely on custom and synthetic datasets, limiting reproducibility. These patterns are systematically summarized in Table 7, highlighting a structural trade-off between consistency and applicability in the field.

Table 7. Benchmark datasets most utilized by predictive maintenance task (n = 64 peer-reviewed).

Maintenance Task	n Studies	Dominant Dataset	Usage Frequency	% of Studies	Concentration	Representative Studies
RUL Forecasting	27	C-MAPSS, IMS Bearing, PRONOSTIA	19	70.4%	High	[10,39–49,115,116]
Fault Diagnosis	9	IMS Bearing, CWRU, Paderborn	7	77.8%	High	[20,66–73]
Tool Monitoring	16	PHM2010, Custom datasets	8	50.0%	Moderate	[50–65]
Condition Monitoring	5	Proprietary Industrial Datasets	3	60.0%	Moderate	[19,21,22,74,75]
Failure Prediction	4	Custom + Synthetics	4	100%	Low	[76–79]

3.6. Deployment Readiness Edge-AI, Real-Time Inference, and Explainability (RQ3)

Deployment readiness assessment reveals substantial maturity gaps between algorithmic development and industrial operationalization. Applying our three-indicator scoring framework (Edge-AI implementation, real-time inference reporting, XAI integration), 60.9% of studies (n = 39) score 0—providing no deployment consideration evidence. Only 4.7% (n = 3) achieve the maximum score of three, demonstrating simultaneous attention to edge constraints, latency requirements, and explainability.

Edge-AI Adoption: Merely 18.8% of studies (n = 12) explicitly report edge device deployment or discuss computational optimization for resource-constrained environ-

ments. These studies employ techniques such as model compression (pruning, quantization), knowledge distillation, or lightweight architecture design (MobileNet variants, SqueezeNet) [10,20,21,108,109,113–116]. The 2025 adaptive dual-distillation framework exemplifies best practices, achieving 5.34× compression (83% parameter reduction) while maintaining predictive accuracy [10]. Edge deployment enables local processing, reduces cloud dependency, minimizes bandwidth consumption, and achieves sub-100 ms latency critical for real-time interventions [108,109,113,114,118–120].

However, most of the literature proposes architectures incompatible with edge hardware constraints. Deep models with millions of parameters requiring GPU acceleration and substantial memory cannot run on typical industrial edge devices (ARM Cortex microcontrollers, FPGAs, entry-level AI accelerators such as NVIDIA Jetson or Google Coral). This disconnect reflects academic focus on maximizing accuracy rather than optimizing the precision–efficiency Pareto frontier [17,108,111,112,119,120].

Aggregate analysis of the arXiv preprint cohort reveals a markedly accelerated deployment orientation relative to the peer-reviewed corpus. The 25 preprints achieve a mean Deployment Readiness Score of 1.72, compared with an estimated mean of approximately 0.63 for the 64 peer-reviewed studies—a difference of 1.09 DRS points representing 36.5% of the full scale. The proportion of studies with DRS = 0 (no deployment-relevant content) collapses from 60.9% in the peer-reviewed corpus to 4.0% among preprints; conversely, the proportion achieving DRS = 3 increases from 4.7% to 16.0%. Three thematic clusters concentrate this deployment maturity: (1) Edge-AI/TinyML preprints (n = 3, all DRS = 3) reporting end-to-end hardware validation on microcontrollers and FPGAs [92–94]; (2) XAI preprints (n = 5), of which four achieve DRS $\geq$ 2, including neuro-symbolic deployments validated on live transit infrastructure [97–99]; and (3) sensor fusion preprints (n = 3) addressing multi-modal integration under real-world industrial conditions [106]. These findings indicate that the deployment gap documented in the peer-reviewed corpus is actively narrowing in current research output, with the gray literature providing an early—and systematically more deployment-mature—view of the field's current trajectory.

Concrete hardware deployments published in 2025 bound the feasible operating region for edge-compatible PdM models. Langer et al. [92] report end-to-end validation of an 8 bit quantized CNN on an ARM Cortex-M4F microcontroller: 100% diagnostic accuracy on a milling dataset, 15.4 ms per inference, and 1.462 mJ per decision, with a total parameter footprint of 12.59 kiB. BearingPGA-Net [93] demonstrates FPGA deployment of a knowledge-distilled bearing fault classifier, reporting more than 200× throughput improvement over CPU execution with less than 0.4% accuracy loss relative to the full teacher model. A systematic survey of quantization methods for microcontrollers [94] covers ARM Cortex-M, RISC-V, and dedicated neural accelerator platforms, cataloging the trade-offs between bit-width reduction and task accuracy across manufacturing-relevant benchmarks. Taken together, these results define reference thresholds—sub-16 ms latency, sub-2 mJ per inference, sub-13 kiB storage—that can be used as minimum acceptance criteria within the Deployment Readiness Score proposed in Section 2.4.

Real-Time Inference Capability: Only 26.6% of studies (n = 17) report inference latency, throughput, or demonstrate explicit real-time operational validation. Manufacturing process control operates on millisecond timescales. CNC tool wear progression occurs in seconds; bearing failures develop over minutes to hours; intervention windows may span mere seconds between anomaly detection and catastrophic failure [17,108,111,112,119,120]. Models exhibiting inference latency exceeding these windows—regardless of accuracy—provide no actionable value. Yet, inference time remains underreported, with only 26.6% of studies characterizing computational performance.

Real-time systems demand worst-case latency predictability, not merely average performance. Runtime variability—caused by OS scheduling, garbage collection, or thermal throttling—may render models unusable even if average latency meets requirements [17,111,112,119,120]. Edge deployment mitigates some latency-variability sources (eliminating network communication delays, cloud service queues) while introducing others (concurrent-process resource contention, processor thermal acceleration) [108,109,113,114,118–120].

Explainable AI (XAI) Integration: Most concerning, only 15.6% of studies (n = 10) integrate explainability mechanisms. The remaining 84.4% treat models as black boxes, providing predictions without interpretable justification. This opacity presents insurmountable barriers in regulated industries (aerospace AS9100, automotive IATF 16949, pharmaceutical cGMP) where certification authorities demand decision-making transparency [16,109–112,118].

XAI techniques applicable to predictive maintenance include: (1) **attention visualization** revealing which temporal windows or sensor channels drive predictions; (2) **SHAP (SHapley Additive exPlanations)** attributing prediction contributions to individual features; (3) **LIMEs (Local Interpretable Model-agnostic Explanations)** approximating local decision boundaries; (4) **concept activation vectors** identifying human-comprehensible concepts learned by networks; (5) **rule extraction** from trained models generating IF-THEN logic comprehensible to operators [16,109–112,114,118].

Siemens' technical report on industrial XAI emphasizes that explainability is essential across the AI lifecycle—from business-case development to model monitoring and maintenance [110]. Explainability facilitates: (1) **confidence calibration** enabling operators to develop appropriate reliance on AI recommendations; (2) **fault diagnosis** enabling identification of model weaknesses or data-quality issues; (3) **regulatory compliance** meeting transparency and human oversight mandates; (4) **continuous improvement** through collaborative human–AI refinement; (5) **knowledge transfer** from AI systems back to human domain experts [16,109–112,118].

The XAI literature for predictive maintenance has diversified substantially since 2024, moving beyond SHAP and LIME toward approaches that generate operator-actionable output. A PRISMA review of XAI methods in PdM [95] documents that attribution-based techniques currently dominate but highlights the absence of any consensus metric for explanation quality—a gap that limits objective comparison of XAI methods in the same way that inconsistent validation schemes limit comparison of predictive models. Counterfactual methods [96] reframe the explanation task from attribution to intervention: rather than identifying which features drove a prediction, they identify the minimum operational change that would have altered the outcome, a formulation directly useful for maintenance scheduling. Gama et al. [97] demonstrate a neuro-symbolic architecture on the Metro do Porto transit system in which an autoencoder detects anomalies while a companion rule-learner generates IF-THEN logic that operators can inspect and audit. A 2026 survey of neuro-symbolic approaches to PdM [98] and independent work from ETH Zurich and EPFL on unsupervised XAI-guided diagnosis [99] show that symbolic reasoning components are being integrated into deep models at an increasing rate. The 15.6% XAI adoption figure reported for the 2007–2024 corpus therefore represents a historical baseline, not the current trajectory.

Table 8 summarizes deployment readiness assessment, revealing substantial gaps between laboratory demonstrations and plant-floor applicability. This finding underscores urgent need for deployment-oriented research beyond mere algorithmic novelty.

Table 8. Readiness index/score and deployment cue rates by task.

Task	n	Mean DRI	Edge Rate	Real-Time Rate	Cloud Rate	Embedded Rate
RUL Forecasting	27	1.815	0.852	0.444	0.296	0.222
Uncoded	9	2.222	0.778	0.667	0.556	0.222
Fault Classification (Diagnosis)	9	1.556	0.778	0.444	0.333	0.000
Tool Wear Prediction	5	2.200	0.800	0.800	0.400	0.200
Condition Monitoring	5	2.000	1.000	0.600	0.200	0.200
Failure Prediction	4	2.250	1.000	0.750	0.250	0.250
Anomaly Detection	3	1.667	1.000	0.667	0.000	0.000
Fault Detection	2	1.000	0.500	0.500	0.000	0.000

A strong pattern of metric homogenization is evident across predictive maintenance tasks, suggesting a limited diversity in evaluation practices. Accuracy overwhelmingly dominates as the primary performance metric in all task categories, with shares ranging from 72.7% in condition monitoring to 90.9% in failure prediction, indicating a near-universal reliance on a single indicator. Additionally, each task reports only one unique metric, reinforcing the lack of methodological variation in performance assessment. While this uniformity simplifies comparison across studies, it also raises concerns about the adequacy of accuracy for capturing task-specific complexities, particularly in imbalanced or time-dependent scenarios. These findings, detailed in Table 9, point to a critical need for more nuanced and task-appropriate evaluation frameworks.

Table 9. Metric concentration by task: most-used metric, mentions, share, number of unique metrics, and concentration label.

Normalized Task	n Articles	Top Metric	Top Metric Mentions	Top Metric Share	n Unique Metrics	Metric Concentration
Condition Monitoring	11	Accuracy	8	0.727	1	High
Fault Classification (Diagnosis)	11	Accuracy	9	0.818	1	High
Failure Prediction	11	Accuracy	10	0.909	1	High
RUL Forecasting	26	Accuracy	20	0.769	1	High
Tool Wear Prediction	5	Accuracy	4	0.800	1	High

Sensitivity analysis by literature type (RQ3 robustness). Because the conclusions of this review integrate a peer-reviewed corpus (n = 64) with an arXiv preprint cohort (n = 25) whose quality has not been externally certified, a sensitivity analysis by literature type was conducted to determine whether the principal findings depend on the inclusion of gray literature evidence.

Scenario A—Peer-reviewed corpus only (n = 64). Re-computing the validation rigor and deployment readiness distributions using only the peer-reviewed studies yields the same concentrations reported in Sections 3.5 and 3.6: 65.6% at Tier 0–1, 10.9% at Tier 3, 60.9% at DRS = 0, 4.7% at DRS = 3, and 15.6% XAI adoption. The validation crisis and deployment gap findings are therefore independent of the preprint cohort and rest entirely on certified peer-reviewed evidence.

Scenario B—arXiv preprint cohort only (n = 25). The 25 preprints, treated as an independent non-peer-reviewed cohort, yield a mean DRS of 1.72, a DRS = 0 share of 4.0%, and a DRS = 3 share of 16.0%, together with 100% XAI coverage in the preprints selected

through the semantic search. These figures describe the gray literature cohort on its own terms and are not aggregated with the peer-reviewed numbers.

Scenario C—Claim stability under preprint exclusion. The central comparative claim of this review—that the preprint cohort exhibits substantially higher deployment readiness than the peer-reviewed corpus—is tested by re-expressing it under a stricter rule. Excluding the five preprints with the highest methodological variability (those reporting bespoke non-standardized hardware benchmarks without third-party replication, n = 5), the mean DRS of the remaining 20 preprints drops from 1.72 to approximately 1.40, which is still about 2.2× the peer-reviewed mean (0.63) and preserves the direction and order of magnitude of the original claim. Excluding every preprint reporting DRS = 3 (n = 4), the conservative residual mean DRS is approximately 1.48, still more than 2.3× the peer-reviewed mean. The direction of the finding is therefore robust to preprint-quality adjustment; what a stricter reading modifies is the precise magnitude, not the sign, of the difference.

Interpretation. The preprints are interpreted in this review as signals of the current research frontier rather than as validated evidence of deployed practice, and the strength of the arXiv-vs-peer-reviewed difference reported in Section 3.6 is conditional on the preprints being subsequently ratified through peer review. Until that ratification occurs, the precise magnitudes attached to the preprint cohort should be read as upper bounds on the publishable state-of-the-art, not as settled population parameters. The central claims of this review are restated here explicitly under this constraint: (i) the validation crisis and deployment gap in the peer-reviewed corpus are documented with peer-reviewed evidence only and do not depend on the preprint cohort; (ii) the preprint cohort provides independent, though non-peer-reviewed, corroboration that the research community is actively addressing the gap; and (iii) the three-fold difference in mean DRS between cohorts, though stable in direction under sensitivity analysis, should be monitored as preprints migrate into indexed peer-reviewed publications over the next 12–24 months.

4. Discussion

4.1. The Validation Crisis: Methodological Implications

Assessment of validation rigor (Section 3.5) exposes a systemic crisis threatening credibility and industrial relevance of predictive maintenance AI research. Dominance of Tier 0 (unclear, 34.4%) and Tier 1 (random split, 31.2%) validation schemes indicates that nearly two thirds of the corpus employs methodologies inadequate for assessing generalization in non-stationary manufacturing. This finding aligns with broader concerns in machine learning research regarding reproducibility and evaluation best practices [14,16,17,113,114,117].

The root cause lies in misalignment between academic incentives and industrial requirements. Academic publication rewards algorithmic novelty and benchmark-performance improvements, incentivizing researchers to maximize reported accuracy through aggressive hyperparameter tuning on static datasets. Industrial deployment, conversely, demands robust generalization across diverse operational regimes, graceful degradation under distribution shift, and uncertainty quantification in predictions. A model achieving 99% accuracy on a randomly split test set provides no guarantee of 90%—or even 50%—accuracy when facing new equipment, altered environmental conditions, or evolving failure modes [108,109,113,114].

Temporal leakage—infiltration of future information into training data—constitutes the primary technical failure. In predictive maintenance time-series, samples close in time exhibit high autocorrelation. Random splitting places temporally adjacent samples in both training and test sets, allowing models to exploit short-term correlations rather than learning genuine fault signatures. Proper temporal validation requires strict chronological partitioning: training on early operational periods, validation on intermediate periods, and

testing on final periods [14,16,17,117]. This mimics deployment reality where models must predict future failures based solely on historical data.

Cross-domain validation—training on one equipment population and testing on another—represents the gold standard for generalization assessment. This approach reveals whether learned features capture physics-based failure mechanisms (valid across equipment instances) or merely memorize idiosyncrasies of specific training assets (invalid for deployment). Scarcity of cross-domain studies (10.9% of corpus) reflects data collection challenges: acquiring labeled failure data from multiple industrial sites with sufficient failure examples demands multi-year experimental campaigns and industrial partnerships rarely feasible within academic timelines [108,109,113,114].

Benchmark dataset limitations compound validation interpretation. The widely used C-MAPSS turbofan degradation dataset, while valuable, exhibits atypical real manufacturing characteristics: (1) simulated rather than physical data; (2) perfectly synchronized sensor channels without missing values; (3) monotonic degradation without recovery or maintenance interventions; (4) absence of confounding factors (environmental noise, operator variability). Models optimized for C-MAPSS performance may not transfer to industrial turbofans experiencing intermittent failures, inconsistent operational profiles, and sensor faults [14,16,17,113,114,117].

A sharper contrast emerges when studies are stratified by the provenance of their evaluation data. Applying the taxonomy of Table 7, the peer-reviewed corpus can be partitioned into three strata: benchmark-dominant studies (RUL forecasting and fault diagnosis, n = 36, of which 70.4–77.8% rely on C-MAPSS, IMS Bearing, PRONOSTIA, CWRU, or Paderborn), mixed studies (tool condition monitoring, n = 16, combining PHM 2010 with custom sets), and industrial-data-dominant studies (general condition monitoring and failure prediction, n = 9, relying on proprietary or custom datasets).

The three strata differ substantively in the claims their validation protocols can support. Benchmark-dominant studies enable algorithmic comparison under homogeneous conditions but inherit the structural limitations of those benchmarks: simulated degradation (C-MAPSS), controlled run-to-failure laboratory bearings (IMS, PRONOSTIA, Paderborn), and absence of stochastic environmental noise, sensor drift, and operating-regime changes. Models whose reported accuracy is grounded exclusively in these datasets cannot, by construction, support claims about performance under real industrial non-stationarity; the quantitative evidence from Albelali and Ahmed [88], showing up to 20.5% RMSE inflation when lag windows cross the split boundary, demonstrates that a non-trivial share of the accuracy reported on benchmark splits is attributable to protocol artifacts rather than to genuine generalization. Mixed studies occupy an intermediate position: they inherit the reproducibility benefits of PHM 2010 while partially compensating with custom industrial acquisitions. Industrial-data-dominant studies, though a minority of the corpus, are the only stratum that can in principle support deployment claims—precisely because their evaluation data reflect the operational variability that benchmark datasets lack. The inverse relationship, however, is that industrial-data-dominant studies typically report lower headline accuracy than benchmark-dominant studies on comparable tasks, a gap that is consistent with—but, critically, not evidence against—the generalization hypothesis.

The deployment-claim consequences follow directly. When a benchmark-dominant study reports 98% accuracy on a random split of C-MAPSS, that figure quantifies performance on the benchmark and not performance on an industrial turbofan. When an industrial-data-dominant study reports 85% accuracy on chronological data from a single facility, that figure is a closer proxy to deployment-time behavior but does not generalize to other sites without cross-domain validation. This review therefore recommends that benchmark-dominant studies (a) report performance on at least one industrial or

cross-equipment dataset or explicitly declare this limitation, and (b) refrain from framing benchmark accuracy as evidence of deployability. The PdM-AI Reporting Checklist introduced in Section 4.4 (items C1 and C5) codifies this expectation.

4.2. The Deployment Gap: From Laboratory to Factory Floor

Deployment readiness assessment (Section 3.6) reveals that fewer than 5% of peer-reviewed studies (4.7%, n = 3) simultaneously address edge constraints, real-time requirements, and explainability—the essential triad for industrial adoption [108–114]. This gap manifests across three dimensions:

Computational Realism: The median deep learning model in our corpus contains 1–10 million parameters, requiring GPU acceleration and substantial memory bandwidth. Typical industrial edge devices—ARM Cortex-M7 microcontrollers with 512 KB RAM, FPGAs with limited logic elements, or entry-level AI accelerators—cannot accommodate such architectures. The latency–accuracy–complexity trade-off receives insufficient attention, with most studies pursuing accuracy maximization without considering deployment constraints [17,108,111,112,119,120].

Successful edge deployment demands model-compression techniques (pruning eliminates redundant parameters, quantization reduces numerical precision, knowledge distillation transfers knowledge from large teacher to small student models) and hardware-aware neural architecture search [10,108,109,113–116,118–120]. The adaptive dual-distillation framework achieving 83% parameter reduction while maintaining accuracy exemplifies the required approach [10]. Yet, deployment-oriented research constitutes a minority of published work.

Temporal Constraints: Manufacturing process control operates on millisecond timescales. CNC tool wear progression occurs in seconds; bearing failures develop over minutes to hours; early-intervention windows may span mere seconds between anomaly detection and catastrophic failure [17,108,111,112,119,120]. Models exhibiting inference latency exceeding these windows—regardless of accuracy—provide no actionable value. Yet, inference time remains underreported, with only 26.6% of studies characterizing computational performance.

Real-time systems demand worst-case latency predictability, not merely average performance. Runtime variability—caused by operating-system scheduling, garbage collection, or thermal throttling—may render models unusable even if average latency meets requirements [17,111,112,119,120]. Edge deployment mitigates some variability sources (eliminating network communication delays, cloud service queues) while introducing others (concurrent-process resource contention, processor thermal acceleration) [108,109,113,114,118–120].

Human–AI Collaboration Barriers: Deep learning models' black-box nature creates adoption friction in manufacturing environments where operators possess decades of domain expertise. When an AI system recommends immediate production-line shutdown—potentially costing thousands of dollars per hour—operators demand interpretable justification, not merely confidence scores [16,109–112,118]. Explainability facilitates confidence calibration: operators learn when to trust AI recommendations (within training distribution), when to apply caution (near distribution boundaries), and when to override (outside training domain).

Regulatory authorities in certified industries (aerospace AS9100, automotive IATF 16949, pharmaceutical cGMP) increasingly mandate AI transparency [110–112]. The European AI Act requires "high-risk AI systems" to provide information enabling human oversight and understanding of decision logic [110]. Predictive maintenance systems influ-

encing safety-critical or quality-control decisions fall under these regulations, creating legal imperatives for XAI integration beyond mere technical desirability.

XAI implementation scarcity (15.6% of studies) reflects both technical challenges and misaligned incentives. Explainability research frequently trades predictive performance for interpretability, creating tension with precision-focused publication standards. Additionally, XAI quality evaluation lacks consensus metrics—while predictive accuracy admits objective measurement, explanation quality depends on subjective human judgment and domain expertise [16,109–112,114,118].

4.3. Alignment with the Target Special Issue and Information System Scope

This review is submitted to the Special Issue "Surveys in Information Systems and Applications" of *Information* and aligns with its aim of delivering systematic syntheses of how information system technologies are applied in practice. Predictive maintenance in complex manufacturing systems constitutes an application domain in which information system infrastructure—sensor networks, data-processing pipelines, edge and cloud analytics, human–machine decision interfaces, and audit-ready reporting—directly conditions the viability of AI methods. The subsections below map the principal thematic findings of this review onto the research and reporting practices that these information system dependencies require.

Sensor Fusion Strategies: Our analysis (Section 3.3) documents the transition from mono-modal vibration signals to multi-modal fusion (vibration + current + acoustic + thermal + vision). As an information system design concern, sensor fusion governs the integrity of upstream data acquisition and therefore the admissibility of downstream analytics. The most mature studies in our corpus demonstrate that sensor fusion not only improves accuracy but provides resilience against individual sensor failures and environmental condition variability—critical characteristics for complex engineering systems [20–25].

Generative Models for Data Augmentation: Although underreported in the current corpus (<5% of studies), generative models represent a promising frontier for addressing fault data scarcity—a perennial predictive maintenance challenge where failure events are rare by design. Future research should explore GANs, VAEs, and diffusion models for synthesizing realistic fault data, enabling robust model training without risking production assets.

Digital Twins and AI-Enhanced Simulation: Digital twins provide information system architectures that directly address validation limitations identified in Section 4.1. Digital twins can provide safe testing grounds for predictive maintenance algorithms, generating synthetic data under diverse operational conditions without real equipment risk. The integration of digital twins with sim-to-real techniques could close validation gaps by enabling model evaluation under failure scenarios that would be unethical or impossible to induce in production systems.

Edge-AI Implementations: Our deployment readiness findings (Section 3.6) underscore an urgent need for Edge-AI research focused on real-time industrial monitoring, a canonical information system deployment pattern for distributed data acquisition and inference. Only 18.8% of studies address edge constraints, despite edge deployment offering critical advantages: reduced latency, enhanced data privacy, reliable operation under network disconnection, and scalability across distributed equipment fleets [108,109,113,114,118–120].

Explainable AI in Engineering Decision-Making: With only 15.6% of studies integrating XAI, a massive gap exists in the coverage of Explainable AI for engineering decision-making. As discussed in Section 4.2, explainability is not merely desirable but essential for

industrial adoption in regulated contexts. Future research must prioritize explainability-by-design rather than post hoc interpretation of black-box models [16,109–112,114,118].

Resilient AI Systems: Fault-tolerant engineering of AI systems connects directly to our validation crisis findings. Complex engineering systems require AI models that degrade gracefully under unseen conditions, quantify prediction uncertainty, and detect when they operate outside the training domain. Techniques such as conformal learning, Bayesian uncertainty quantification, and out-of-distribution anomaly detection should be integrated into standard model development practices.

4.4. Implications for Research and Practice

These findings carry substantial implications for multiple stakeholder communities:

For Researchers: Future work should prioritize validation rigor, explicitly reporting data-partitioning schemes and assessing temporal generalization. Cross-domain validation across multiple equipment instances should become standard practice, not exceptional achievement. Overreliance on benchmark datasets requires correction through industrial data partnerships and emphasis on transfer learning scenarios. Deployment considerations—edge compatibility, inference latency, explainability—should factor into algorithm design from inception, not as afterthoughts.

For Practitioners: Industrial adoption decisions should scrutinize validation methodology, not merely reported accuracy. Models validated via random splitting merit skepticism; those demonstrating cross-domain generalization warrant serious consideration. Pilot deployments should include extended temporal validation periods capturing seasonal variations, operational regime changes, and equipment aging. Investment in explainability infrastructure—visualization tools, operator training in AI interaction—is as critical as model development itself [16,109–112,118].

For Journal Editors and Reviewers: Applying validation-reporting standards constitutes high-leverage intervention. Requiring authors to explicitly declare data-partitioning schemes, justify validation choices for time-series contexts, and report temporal generalization performance would substantially improve literature quality. Encouraging supplementary materials containing implementation details, hyperparameter configurations, and negative results enhances reproducibility [16,17].

For Funding Agencies: Research programs should incentivize industry–academia collaboration enabling access to industrial-scale data. Multi-site, multi-equipment datasets capturing equipment diversity, operational variability, and fault-mode heterogeneity enable validation rigor impossible with laboratory benchmarks. Funding deployment-oriented research—addressing edge optimization, real-time system integration, human factors—complements algorithmic innovation [108,109,113,114,118–120].

Translating the diagnostic findings of Sections 3.5, 3.6 and 4.1–4.3, into a concrete instrument for research practice, we propose the PdM-AI Reporting Checklist below. The checklist is intended as a minimum reporting standard for AI-for-predictive maintenance submissions in complex manufacturing contexts and is organized along six dimensions that correspond directly to the deficiencies documented in this review: validation protocol, edge deployment, real-time inference, explainability, data provenance and diversity, and negative or inconclusive results. Each item is stated as an assertion that the manuscript should either satisfy or explicitly declare as non-applicable with justification.

C1. Validation protocol. State whether the validation scheme is Tier 0 (unclear/unreported), Tier 1 (random split), Tier 2 (k-fold cross-validation), or Tier 3 (temporal or cross-domain split) per Section 2.4. For Tier 2–3, report fold structure or split boundaries, the handling of temporal autocorrelation (e.g., purging, embargo), and any leakage au-

dits performed. Declare explicitly whether the protocol prevents information from future operating periods entering training.

C2. Edge deployment. Report target hardware class (MCU, FPGA, edge GPU, or server-only) and the memory and compute envelope in which the model was evaluated. For any Edge-AI claim, report peak parameters, peak memory, and per-inference energy under the declared hardware; for compressed variants, report the accuracy delta relative to the full model. Studies that do not address edge deployment should state so explicitly rather than leave it unreported.

C3. Real-time inference. Report inference latency statistics (mean, p95, and worst-case) on the declared hardware, together with the measurement methodology (cold-start vs. steady-state, timing harness, concurrency assumptions). Relate the reported latency to the task-level time constant of the target manufacturing application (sub-second, second, or minute scale).

C4. Explainability. State whether the model integrates explainability by design (architecture-native, e.g., attention, prototype, neuro-symbolic) or through post hoc methods (SHAP, LIME, counterfactuals). Identify the intended audience of the explanation (operator, maintenance engineer, certification authority) and the quality metric used, if any. Absence of XAI should be declared explicitly.

C5. Data provenance and diversity. Identify the data source(s) used (public benchmark, proprietary industrial, simulated, synthetic from generative models) and, for each, the number of equipment instances, operating regimes, failure modes, and the presence or absence of realistic imperfections (missing values, sensor drift, operator variability). When a single benchmark is used, justify why transfer to industrial conditions is plausible.

C6. Negative and inconclusive results. Report configurations that were tried and did not work at a level of detail sufficient for another group to avoid repeating them: architectures tested and abandoned, hyperparameter regions that degraded performance, datasets on which the model failed to generalize, and deployment attempts that did not meet operational criteria. This item directly addresses the publication bias asymmetry discussed in Section 4.5 and the unsuccessful path taxonomy introduced in Section 4.7.

Collectively, the six items address every single weakness identified in the quantitative analysis: C1 targets the 65.6% Tier 0–1 concentration; C2–C4 target the DRS = 0 peer-reviewed share of 60.9%; C5 targets the benchmark monopoly pattern documented in Section 4.1 and Table 7; and C6 targets the reporting asymmetry analyzed in Section 4.5. The checklist is deliberately format-agnostic: it can be adopted as a pre-submission self-report by authors, as a reviewer aid, or as part of journal submission guidelines for AI-for-predictive maintenance work.

4.5. Limitations of This Review

This systematic review presents several limitations that merit acknowledgment.

First, the restriction to English-language, peer-reviewed journal articles may have excluded relevant research published in conference proceedings, technical reports, or non-English sources. Nevertheless, this decision ensures the inclusion of archived, peer-reviewed studies that guarantee methodological rigor and replicability.

Second, data extraction was based on titles, abstracts, and methodology sections. Complete reproduction of each study was infeasible given the size of the corpus. Consequently, some works classified as Tier 0 (unclear validation) may, in fact, employ rigorous validation schemes that are insufficiently described in the manuscript text. This limitation underscores the importance of clear methodological reporting as a standard requirement for future publications.

Third, the Deployment Readiness Score (DRS) framework, though systematic, relies on binary indicators (present/absent) that may oversimplify nuanced implementation aspects. For example, a study reporting inference time but achieving 500 ms latency—insufficient for real-time control—receives the same score as one achieving 50 ms latency, which is acceptable for most applications. Future reviews could incorporate finer-grained quantitative indicators such as latency percentiles, compression ratios, or Explainable AI (XAI) evaluation metrics.

Fourth, thematic clustering was conducted through unsupervised methods (TF-IDF combined with k-means) applied to titles and abstracts. This approach may overlook subtle semantic distinctions or impose artificial boundaries between closely related research domains. Although expert-driven manual clustering could produce alternative taxonomies, our chosen method favors reproducibility and minimizes subjective bias.

Fifth, given the rapidly evolving nature of AI research, the findings captured here reflect the state of the literature up to January 2026. Emerging paradigms—such as foundation models for time-series forecasting, multi-modal transformers, and neuro-symbolic integration—are likely to reshape the field in the near future. Therefore, periodic updates of this review will be necessary to monitor ongoing developments.

Sixth, a potential publication bias toward studies reporting positive accuracy results may inflate the proportion of papers with high DRS or Validation Level within the corpus. Although the inclusion of arXiv preprints partially mitigates this bias, negative or null results in AI-based predictive maintenance remain systematically underreported in both peer-reviewed literature and preprint repositories.

Finally, the assessment of evidence certainty using the GRADE framework was deemed inapplicable to this review, as GRADE requires a body of evidence with quantifiable effect estimates from comparative or randomized studies. Instead, the Validation Levels and DRS frameworks developed in this review offer domain-adapted alternatives tailored to AI applications in industrial engineering. The authors (C.F.H.V., D.A.G.A., L.F.G.G., R.A.M.R., A.V.-A. and J.A.V.O.) encourage their adoption as potential reporting standards for future systematic reviews in this field.

Reporting and publication bias. Three sources of reporting bias are acknowledged explicitly and constitute a limitation of this review distinct from the methodological limitations discussed above.

Publication bias toward positive results. As in most fields of applied machine learning research, the published AI-for-predictive maintenance literature is asymmetrically populated by studies reporting positive or state-of-the-art results. Configurations that failed to converge, architectures that did not transfer, and pilot deployments that missed operational criteria are systematically underrepresented in the peer-reviewed record. This asymmetry means that the validation rigor distribution reported in Section 3.5 (65.6% of the peer-reviewed corpus at Tier 0–1) and the deployment readiness distribution reported in Section 3.6 (60.9% at DRS = 0, 4.7% at DRS = 3) should be read as plausibly optimistic estimates of field practice rather than neutral snapshots; studies in which model development was discontinued are, by construction, absent from the sample.

Near-absence of published failure reports. Beyond generic publication bias, the PdM-AI literature exhibits a domain-specific pattern: the explicit reporting of failed models, abandoned architectures, or deployments that were decommissioned is almost entirely absent from the peer-reviewed corpus. Section 4.7 (unsuccessful paths) reconstructs four recurring failure patterns inferentially—from the structural evidence of benchmark hyperoptimization, overparameterization, opacity, and data monopoly—but the direct testimony of individual failures is missing. This absence is itself a limitation of the available evidence

base, because cautionary information is known only through indirect signals rather than primary reports.

Indexing and language bias. The review sources (Scopus, Web of Science, arXiv) systematically privilege English-language, indexed journal publications and preprints from institutions with strong publication infrastructure. Relevant industrial case studies published as technical reports, internal whitepapers, or in non-indexed conference proceedings are not retrieved, and non-English contributions are excluded by eligibility criterion. No formal quantitative bias assessment (funnel plot, Egger test) was possible because the narrative nature of the synthesis and the heterogeneity of reported metrics preclude pooled effect estimation. Future reviews of this field should systematically incorporate gray industrial literature and regional non-English databases where available, and should explicitly solicit failure reports—a reporting expectation that item C6 of the checklist in Section 4.4 formalizes.

Taken together, these three sources of bias do not invalidate the principal findings of this review—the Tier 0–1 concentration, the DRS = 0 share, and the XAI scarcity are internally consistent across peer-reviewed and gray literature cohorts and are unlikely to reverse under a less biased sampling frame. They do, however, caution against treating the reported percentages as precise population parameters; the true level of deployment-ready, rigorously validated, explainable PdM-AI practice in the industrial world is bounded above by the numbers reported in this review.

4.6. Convergence and Divergence Between Peer-Reviewed and Preprint Evidence

Integrating the arXiv preprint cohort (n = 25, refs. [83–107]) alongside the peer-reviewed corpus reveals both convergence on foundational findings and meaningful divergence in the current research trajectory—a duality that provides simultaneous internal validation of this review's conclusions and forward projection of the field's near-term evolution.

Convergence. The preprint literature independently corroborates the two central findings of this systematic review. First, deep learning architectures—specifically transformer-based, hybrid CNN-LSTM, and emerging State Space Model variants—continue to dominate RUL forecasting and fault diagnosis tasks in preprint submissions, confirming the algorithmic dominance documented in Section 3.4 and demonstrating that this pattern has not been disrupted by the arrival of foundation models. Second, validation inadequacy persists even in the most recent output: the majority of preprints still rely on benchmark datasets (C-MAPSS, NASA Battery, CALCE) and do not employ cross-site or cross-equipment validation protocols. This convergence strengthens confidence that the validation crisis identified in the peer-reviewed corpus reflects a genuine structural limitation of the field, not an artifact of publication lag or indexing coverage.

Divergence—Architectural frontier. The preprint cohort introduces three architectural families absent from or marginally represented in the peer-reviewed corpus: (1) *foundation models* pre-trained on massive heterogeneous sensor corpora [83,84,100], achieving few-shot fault diagnosis across unseen equipment types via parameter-efficient fine-tuning—directly addressing the data-scarcity barrier identified throughout this review; (2) *Selective State Space Models* (specifically, the Mamba variant) [85], offering linear-time sequence processing with measurable computational advantages for edge deployment relative to quadratic-complexity transformers; and (3) *neuro-symbolic hybrids* [97,98] that couple deep anomaly detection with companion rule-learners generating IF-THEN logic auditable by domain operators, providing architecture-native rather than post hoc explainability. These innovations are not yet captured in the quantitative analysis of the peer-reviewed corpus

(Section 3.4), but represent the current architectural frontier that future reviews will need to systematically incorporate.

Divergence—Deployment readiness trajectory. The most consequential divergence is quantitative. The arXiv cohort's mean DRS of 1.72 is approximately three times that of the peer-reviewed corpus (~0.63), and the proportion of studies achieving DRS = 0 declines from 60.9% to 4.0%. This trajectory indicates that the deployment gap, while severe in the historical peer-reviewed record through 2024, is actively narrowing in current research output. Edge-AI hardware validation has transitioned from aspirational to demonstrated: three preprints report end-to-end benchmarks on ARM Cortex-M, FPGA, and microcontroller platforms with sub-16 ms inference latency and sub-2 mJ energy consumption per decision [92–94]—providing concrete reference thresholds for the Deployment Readiness Score framework proposed in Section 2.4 and operationalizing what *edge-compatible* means in quantitative terms.

Implications for interpretation. These patterns warrant a nuanced reading of this review's central findings. The validation crisis (Section 4.1) and deployment gap (Section 4.2) are accurately characterized as systemic conditions of the published, peer-reviewed record through January 2026. However, the preprint evidence suggests that the field's response is already underway, concentrated in gray literature that has not yet cleared peer review but represents the current state of practice in leading research groups. For practitioners evaluating technology readiness, this distinction matters: models reported in 2024–2026 preprints are likely to exhibit substantially higher deployment maturity than the 0.63 mean DRS of the peer-reviewed corpus would suggest. For the research community, it underscores the importance of accelerating peer review of deployment-oriented work and of incorporating gray literature systematically in future reviews. Tracking the migration rate of these preprint contributions into the indexed peer-reviewed corpus represents a high-value indicator of the research-to-practice gap's closure in coming years.

4.7. Unsuccessful Paths in AI-PdM Research

Beyond the well-documented successes of AI for predictive maintenance, the corpus also contains evidence of research trajectories that did not deliver on the expectations set by their initial laboratory results. Making these trajectories visible is consistent with the aims of the target Special Issue, which explicitly welcomes the reporting of unsuccessful paths as cautionary evidence for the community. Four recurring patterns emerge from the present synthesis.

Benchmark hyperoptimization without transfer. The concentration of RUL and fault diagnosis studies on a small set of public datasets—C-MAPSS, IMS Bearing, PRONOSTIA, CWRU, and Paderborn (Tables 6 and 7, Section 3.5)—has produced a long tail of models reporting accuracies above 98% that did not transfer to equipment operating under unseen speeds, loads, or lubrication regimes. The trajectory is quantified by Albelali and Ahmed [88], who measure up to 20.5% RMSE inflation in 10-fold cross-validation when lag windows span the split boundary, with two-way and three-way chronological splits holding bias below 5%. Models whose apparent state-of-the-art depended on such leakage did not sustain their claims under chronological or cross-domain evaluation; their performance was an artifact of the evaluation protocol rather than of the model itself. This pattern explains a substantial fraction of the Tier 0–1 concentration documented in Section 3.5 (65.6% of the peer-reviewed corpus).

Overparameterized architectures incompatible with target hardware. A second recurrent pattern is the publication of deep architectures with 1–10 million parameters, requiring GPU acceleration and memory bandwidths unavailable on typical industrial edge devices (ARM Cortex microcontrollers, FPGAs, and entry-level AI accelerators such as NVIDIA

Jetson or Google Coral). Section 4.2 documents that edge-compatible re-implementation of such models typically requires post hoc compression (pruning, quantization, knowledge distillation), during which a non-trivial fraction of the original accuracy is lost. In several documented cases, the accuracy gap after compression rendered the model indistinguishable from substantially simpler baselines, effectively neutralizing the original contribution. The 2025 adaptive dual-distillation framework [10] and BearingPGA-Net [93] illustrate the opposite trajectory—designs in which deployability was built in from the outset—and are notable precisely because they remain exceptions rather than the rule.

Opacity as a barrier to certification. A third trajectory concerns black-box models that achieved strong laboratory results but could not be advanced into safety-critical deployment contexts. The peer-reviewed corpus shows 84.4% of studies without any XAI integration, and the regulated industries identified in Section 4.2 (aerospace AS9100, automotive IATF 16949, pharmaceutical cGMP) routinely require decision traceability as a condition of certification [110–112]. Models whose decisions cannot be explained to a certification authority—regardless of their validation accuracy—remain confined to research pilots and do not progress into operational deployment. The 2024–2026 preprint shift toward neuro-symbolic and attention-integrated architectures [97–99] is in part a corrective response to this structural dead-end.

Unreported negative results. Finally, the field exhibits a systemic absence of published negative results. Models that failed to generalize, architectures that were abandoned, and pilot deployments that did not meet operational criteria rarely reach the indexed literature, and when they do the failure is typically reframed as a design choice rather than a negative finding. This reporting asymmetry operates at two levels: individual authors have limited incentives to document configurations that underperformed, and peer-review venues rarely solicit failure reports as primary contributions. The consequence is that the community cannot efficiently avoid unsuccessful paths that are not made visible in the literature; each research group re-discovers the same failure modes independently, at a cost that is difficult to quantify but structurally embedded in the current publication model.

5. Conclusions

This systematic review of 89 studies (64 peer-reviewed + 25 arXiv preprints) spanning two decades (2007–2026) reveals a paradoxical state of AI-for-predictive maintenance research: notable algorithmic sophistication coupled with concerning validation practices and deployment readiness gaps. Deep learning architectures—particularly CNNs for fault diagnosis and LSTMs for RUL forecasting—have achieved dominance based on superior performance in controlled experimental settings. Hybrid architectures combining convolutional feature extraction with recurrent temporal modeling represent the current frontier, demonstrating state-of-the-art accuracy on benchmark datasets.

However, critical examination of validation methodologies exposes systemic weaknesses threatening industrial relevance. More than two-thirds of studies employ validation schemes (Tier 0: unclear, Tier 1: random split) inadequate for assessing generalization in non-stationary manufacturing. Random-split data leakage produces optimistic accuracy estimates failing to materialize in deployment. Only 10.9% of studies achieve Tier 3 rigor through temporal splitting or cross-domain validation, representing best practices for complex engineering systems.

Deployment readiness assessment reveals even more pronounced gaps. Fewer than 19% of studies address edge computational constraints; only 27% characterize real-time inference performance; and merely 15.6% integrate explainability mechanisms. These dimensions—edge compatibility, temporal constraints, human–AI collaboration—are prerequisites for industrial adoption, yet receive minimal attention in the academic literature.

The disconnect reflects misaligned incentives: academic publication rewards algorithmic novelty and benchmark precision, while industrial deployment demands robustness, efficiency, and interpretability.

Our novel Readiness Score framework (combining Edge-AI, real-time, and XAI indicators) provides a systematic tool for assessing operational maturity. Only 4.7% of peer-reviewed studies achieve the maximum score, addressing all three deployment dimensions simultaneously. This finding quantifies the frequently lamented but rarely measured "lab-to-factory gap" in practitioner communities. Notably, the 25 arXiv preprints from 2024 to 2026 demonstrate a mean DRS of 1.72—nearly three times the peer-reviewed average—indicating that the gap is actively narrowing in the current research frontier, a trajectory analyzed systematically in Section 4.6.

5.1. Contributions and Practical Implications

This review advances the field through three primary contributions:

Evidence-Based Taxonomy: The five-cluster thematic structure (General PdM, RUL forecasting, tool wear, Sensor/Niche, fault detection) provides an organizing framework for heterogeneous predictive maintenance literature. Clear identification of dominant AI architectures for each task (CNNs for fault diagnosis, LSTMs for RUL, hybrids for multi-modal fusion) offers practitioners evidence-based algorithm-selection guidance.

Validation Rigor Framework: The four-tier classification (0: unclear, 1: random split, 2: k-fold, 3: temporal/cross-domain) establishes a reproducible standard for assessing methodological quality in time-series predictive maintenance research. This framework can inform journal review practices, guide researcher validation choices, and enable future meta-analyses controlling for validation rigor as a moderating variable.

Deployment Readiness Score: Three-indicator assessment (Edge-AI, real-time, XAI) quantifies operational maturity, enabling systematic comparison across studies and identification of deployment-oriented research gaps. This metric can guide funding priorities, industry–academia partnership design, and translation of academic innovations to industrial practice.

5.2. Research Agenda for Closing the Gap

Based on the identified deficiencies, we propose a research agenda addressing the validation crisis and deployment readiness gap:

Standardized Validation Protocols: The community should adopt temporal validation as default for time-series predictive maintenance. Mandatory reporting elements should include: (1) explicit data-partitioning scheme with temporal visualization; (2) performance metrics on strictly separated test sets; (3) sensitivity analysis examining performance degradation under distribution shift; (4) confidence intervals or uncertainty quantification for predictions [14,16,17,117].

Cross-Domain Benchmark Datasets: Development of multi-site, multi-equipment benchmark datasets capturing operational diversity would enable standardized cross-domain generalization evaluation. Such datasets should include: (1) data from at least 3–5 distinct equipment instances; (2) variable operational regimes (speeds, loads, environmental conditions); (3) multiple failure modes with imbalanced class distributions; (4) realistic imperfections (missing data, sensor failures, calibration drift) [14,16,17,113,114,117].

Native Edge Algorithm Design: Research should shift from post hoc compression of overparameterized models toward designing lightweight architectures from inception. Hardware-aware neural architecture search, binary neural networks, and efficient attention mechanisms represent promising directions. Benchmarking should report precision–latency–memory Pareto frontiers, not merely peak accuracy [10,108,109,113–116,118–120].

Real-Time System Integration: Studies proposing real-time predictive maintenance should characterize worst-case execution time, not merely average inference latency. Integration with real-time operating systems (RTOS), deterministic scheduling, and resource-reservation mechanisms requires deeper treatment. Field validation demonstrating closed-loop control integration would substantiate deployment claims [17,108,111,112,119,120].

Explainability by Design: XAI should transition from post hoc interpretation of black-box models to inherently interpretable architectures. Attention mechanisms, prototype learning, and hybrid neuro-symbolic approaches offer paths toward intrinsically explainable models. Human factors research evaluating operator trust, calibration, and decision-making when interacting with XAI systems remains critically underexplored [16,109–112,114,118].

Transfer Learning Frameworks: Given the scarcity of labeled data for many industrial failure modes, transfer learning from data-rich domains (e.g., pre-training on bearing faults, fine-tuning for gearbox faults) merits systematic investigation. Meta-learning approaches enabling rapid adaptation to new equipment instances with minimal labeled data represent a promising frontier [10,108,109,113–116].

Foundation and Large Models for PHM: The emergence of large pre-trained models for Prognostics and Health Management—UniFault [83], BearLLM [84], and PHM-LM [100]—raises a practical question that this corpus cannot yet answer: whether few-shot transfer from a pre-trained PHM backbone generalizes across industries, equipment types, and failure modes in a way that satisfies the cross-domain validation criteria defined in Section 2.4. Studies are needed that systematically measure adaptation cost (labeled samples required, fine-tuning compute), residual performance gap relative to equipment-specific models, and the tractability of certification under standards such as AS9100 and IATF 16949 when the base model was trained on data from unrelated domains.

Diffusion Models for Fault Data Synthesis: Diffusion models represent a methodologically mature alternative to GANs and VAEs for synthesizing fault time-series in low-data regimes. FaultDiffusion [101] generates realistic fault signatures from fewer than ten observed examples by conditioning a diffusion process on the contrast between normal and anomalous behavior, without requiring explicit fault labels during pre-training. A broader survey of deep generative models in condition monitoring [102] documents applications across structural health monitoring, bearing diagnosis, and rotating machinery prognostics. A research gap that remains open is the absence of a standardized protocol for assessing the fidelity and diversity of generated fault data and their downstream effect on classifier generalization; without such a protocol, synthetic augmentation results are not directly comparable across studies.

Online Continual Learning for Operational Environments: Tier 4 operational validation, in which a model is continuously evaluated against live production data, requires infrastructure that static training regimes do not provide. Three components appear necessary: a mechanism to detect when the data distribution has shifted enough to invalidate the current model (CDSeer [90] satisfies this with minimal labeling overhead), a procedure for updating model parameters without full retraining (NatSR [91] frames this as online score-driven parameter filtering), and a physics-informed regularizer to prevent the updated model from violating known degradation dynamics (MKDPINN [103] embeds partial differential equation constraints learned from sensor data). No published study has integrated all three components in a single operational pipeline; this integration constitutes a concrete and tractable research objective for the community.

5.3. Call to Action

Translating AI performance into plant-floor value requires coordinated effort from researchers, practitioners, industrial partners, and funding agencies. The algorithmic capabilities demonstrated in the reviewed literature are sufficient; the remaining barriers are methodological, institutional, and incentive-based. By prioritizing validation rigor, deployment realism, and human-centered design alongside algorithmic innovation, the community can close the gap between laboratory demonstrations and plant-floor value creation.

The $50 billion annual cost of unplanned manufacturing downtime represents both challenge and opportunity. AI for predictive maintenance—properly validated and responsibly deployed—can substantially reduce this burden while improving safety, sustainability, and competitiveness. This review provides both critical assessment of current limitations and a constructive roadmap toward trustworthy, industrial-grade predictive maintenance AI systems worthy of widespread adoption.

In alignment with the Special Issue "Surveys in Information Systems and Applications" of Information, we anticipate this review will drive interdisciplinary collaboration, familiarize engineering researchers with emerging AI paradigms, and accelerate transition from academic research to industrial impact. Modern manufacturing systems' complexity demands equally sophisticated AI approaches—but sophistication without validation rigor and deployment realism remains merely academic. The path forward requires shared commitment to methodological excellence, transparency, and accountability among all stakeholders in the manufacturing AI ecosystem.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info17050456/s1, Figure S1: PRISMA 2020 Flow Diagram; Supplementary Material S1: Complete search equations for Scopus and Web of Science databases used to retrieve the included studies; Supplementary Material S2 (Excel workbook, 12 sheets): PRISMA 2020 systematic review documentation, including (i) PRISMA Flow Summary; (ii) All Records identified before deduplication (n = 814); (iii) Duplicate records removed (n = 59); (iv) Title and abstract screening dataset (n = 755); (v) Data extraction summary; (vi) Included studies with full data extraction (n = 89: 64 peer-reviewed + 25 arXiv supplementary); (vii) Search strategy details; (viii) PRISMA numerical breakdown; (ix) Author declarations; (x) Full-text classification dataset (n = 171); (xi) Full-text exclusions with reasons (n = 97); and (xii) PRISMA 2020 compliance status checklist; Table S3: PRISMA 2020 Checklist.

Author Contributions: Conceptualization, C.F.H.V. and D.A.G.A.; methodology, C.F.H.V. and A.V.-A.; software, C.F.H.V.; validation, C.F.H.V., D.A.G.A. and L.F.G.G.; formal analysis, C.F.H.V. and J.A.V.O.; investigation, C.F.H.V. and R.A.M.R.; resources, D.A.G.A.; data curation, C.F.H.V.; writing—original draft preparation, C.F.H.V.; writing—review and editing, D.A.G.A., L.F.G.G., R.A.M.R., A.V.-A. and J.A.V.O.; visualization, C.F.H.V.; supervision, D.A.G.A. and A.V.-A.; project administration, C.F.H.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universidad Autónoma del Perú through institutional support; no external grant or project number was assigned.

Institutional Review Board Statement: Not applicable. This study is a systematic literature review of previously published research and does not involve direct human or animal participants.

Informed Consent Statement: Not applicable. This study is a systematic literature review and does not involve any direct participation of human subjects.

Data Availability Statement: The data supporting the findings of this study are openly available. The complete list of included studies, extracted data matrices, and PRISMA flow diagram are provided in the Supplementary Materials accompanying this article.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
PdM	Predictive Maintenance
RUL	Remaining Useful Life
DL	Deep Learning
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
XAI	Explainable AI
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
GNN	Graph Neural Network
SSM	Selective State Space Model
TinyML	Tiny Machine Learning
FPGA	Field-Programmable Gate Array
MCU	Microcontroller Unit
DRS	Deployment Readiness Score
SLR	Systematic Literature Review
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TF-IDF	Term Frequency–Inverse Document Frequency
PCA	Principal Component Analysis
SVM	Support Vector Machine
k-NN	k-Nearest Neighbors
FFT	Fast Fourier Transform
CWT	Continuous Wavelet Transform
STFT	Short-Time Fourier Transform
TCM	Tool Condition Monitoring
PHM	Prognostics and Health Management
IoT	Internet of Things
RTOS	Real-Time Operating System
MCSA	Motor Current Signature Analysis

References

1. Carvalho, T.P.; Soares, F.A.; Vita, R.; Francisco, R.D.P.; Basto, J.P.; Alcalá, S.G. A systematic literature review of machine learning methods applied to predictive maintenance. *Comput. Ind. Eng.* **2019**, *137*, 106024. [CrossRef]
2. Baptista, M.; Sankararaman, S.; de Medeiros, I.P.; Nascimento, C.; Prendinger, H.; Henriques, E.M. Forecasting fault events for predictive maintenance using data-driven techniques and ARMA modeling. *Comput. Ind. Eng.* **2018**, *115*, 41–53. [CrossRef]
3. Lei, Y.; Li, N.; Guo, L.; Li, N.; Yan, T.; Lin, J. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mech. Syst. Signal Process.* **2018**, *104*, 799–834. [CrossRef]
4. Zhang, W.; Yang, D.; Wang, H. Data-driven methods for predictive maintenance of industrial equipment: A survey. *IEEE Syst. J.* **2019**, *13*, 2213–2227. [CrossRef]
5. Dalzochio, J.; Kunst, R.; Pignaton, E.; Binotto, A.; Sanyal, S.; Favilla, J.; Barbosa, J. Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. *Comput. Ind.* **2020**, *123*, 103298. [CrossRef]
6. Jiménez-Cortadi, A.; Irigoien, I.; Boto, F.; Sierra, B.; Rodríguez, G. Predictive maintenance on the machining process and machine tool. *Appl. Sci.* **2020**, *10*, 224. [CrossRef]
7. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
8. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]

9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. [CrossRef]
10. Cheng, X.; Chaw, J.K.; Sahrani, S.; Ang, M.C.; Gunasekaran, S.S.; Ting, C.Y. An adaptive dual distillation framework for remaining useful life prediction. *Complex Intell. Syst.* **2025**, *11*, 253. [CrossRef]
11. Li, X.; Zhang, W.; Ding, Q. Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. *Reliab. Eng. Syst. Saf.* **2019**, *182*, 208–218. [CrossRef]
12. Qin, Y.; Wang, X.; Zou, J. The optimized deep belief networks with improved logistic sigmoid units and their application in fault diagnosis. *IEEE Trans. Ind. Electron.* **2019**, *66*, 3373–3381. [CrossRef]
13. Zonta, T.; da Costa, C.A.; da Rosa Righi, R.; de Lima, M.J.; da Trindade, E.S.; Li, G.P. Predictive maintenance in the Industry 4.0: A systematic literature review. *Comput. Ind. Eng.* **2020**, *150*, 106889. [CrossRef]
14. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
15. Bousdekis, A.; Magoutas, B.; Apostolou, D.; Mentzas, G. Review, analysis and synthesis of prognostic-based decision support methods for condition based maintenance. *J. Intell. Manuf.* **2018**, *29*, 1303–1316. [CrossRef]
16. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Br. Med. J.* **2021**, *372*, n71. [CrossRef]
17. Page, M.J.; Moher, D.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *Br. Med. J.* **2021**, *372*, n160. [CrossRef] [PubMed]
18. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med.* **2009**, *6*, e1000097. [CrossRef]
19. Jogdeo, A.A.; Patange, A.D.; Atnurkar, A.M.; Sonar, P.R. Robustification of the Random Forest: A Multitude of Decision Trees for Fault Diagnosis of Face Milling Cutter Through Measurement of Spindle Vibrations. *J. Vib. Eng. Technol.* **2024**, *12*, 4521–4539. [CrossRef]
20. Hirsch, V.; Reimann, P.; Mitschang, B. Data-driven fault diagnosis in end-of-line testing of complex products. In *Proceedings of the 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA 2019)*; IEEE: New York, NY, USA, 2019. [CrossRef]
21. Cheng, Q.; Cao, Y.; Zhang, T.; Sun, L.; Xu, L.; Liu, Z.; Cheng, C. An improved self-organizing mapping neural network and its application in fault diagnosis of CNC machine tool servo drive system. *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.* **2025**, *239*, 1299–1313. [CrossRef]
22. Calabrese, F.; Regattieri, A.; Botti, L.; Galizia, F.G. Prognostic health management of production systems. New proposed approach and experimental evidences. *Procedia Manuf.* **2019**, *39*, 260–269. [CrossRef]
23. Martínez-Arellano, G.; Ratchev, S. Towards an active learning approach to tool condition monitoring with bayesian deep learning. In Proceedings of the European Council for Modelling and Simulation (ECMS), Caserta, Italy, 11–14 June 2019. [CrossRef]
24. Langone, R.; Alzate, C.; De Ketelaere, B.; Vlasselaer, J.; Meert, W.; Suykens, J.A. LS-SVM based spectral clustering and regression for predicting maintenance of industrial machines. *Eng. Appl. Artif. Intell.* **2015**, *37*, 268–278. [CrossRef]
25. Liu, J.; An, Y.; Dou, R.; Ji, H. Dynamic deep learning algorithm based on incremental compensation for fault diagnosis model. *Int. J. Comput. Intell. Syst.* **2018**, *11*, 846–860. [CrossRef]
26. Lu, Y.; Yang, S. Construction and research of a data-driven model for early fault detection in rotating machinery. *J. Eng. Appl. Sci.* **2026**, *73*, 21. [CrossRef]
27. Shi, C.; Luo, B.; He, S.; Li, K.; Liu, H.; Li, B. Tool Wear Prediction via Multidimensional Stacked Sparse Autoencoders with Feature Fusion. *IEEE Trans. Ind. Inform.* **2020**, *16*, 5150–5159. [CrossRef]
28. Hassan, M.; Sadek, A.; Attia, M.H. A Generalized Multisensor Real-Time Tool Condition-Monitoring Approach Using Deep Recurrent Neural Network. *Smart Sustain. Manuf. Syst.* **2019**, *3*, 41–52. [CrossRef]
29. He, D.; Li, R.; Bechhoefer, E. Stochastic modeling of damage physics for mechanical component prognostics using condition indicators. *J. Intell. Manuf.* **2012**, *23*, 221–226. [CrossRef]
30. Windmann, S.; Westerhold, T. Fault detection in automated production systems based on a long short-term memory autoencoder. *At-Automatisierungstechnik* **2024**, *72*, 47–58. [CrossRef]
31. Jankovič, D.; Šimic, M.; Herakovič, N. A comparative study of machine learning regression models for production systems condition monitoring. *Adv. Prod. Eng. Manag.* **2024**, *19*, 78–92. [CrossRef]
32. Wesendrup, K.; Hellingrath, B. Post-prognostics demand management, production, spare parts and maintenance planning for a single-machine system using Reinforcement Learning. *Comput. Ind. Eng.* **2023**, *179*, 109216. [CrossRef]
33. Souza, M.L.H.; da Costa, C.A.; de Oliveira Ramos, G. A machine-learning based data-oriented pipeline for Prognosis and Health Management Systems. *Comput. Ind.* **2023**, *148*, 103903. [CrossRef]

34. Sun, W.; Paiva, A.R.; Xu, P.; Sundaram, A.; Braatz, R.D. Fault detection and identification using Bayesian recurrent neural networks. *Comput. Chem. Eng.* **2020**, *141*, 106991. [CrossRef]
35. Yu, J.; Zhang, C.; Wang, S. Sparse one-dimensional convolutional neural network-based feature learning for fault detection and diagnosis in multivariable manufacturing processes. *Neural Comput. Appl.* **2022**, *34*, 4343–4366. [CrossRef]
36. Kim, B.; Jung, W.; Choi, Y.; Lee, J. Bayesian neural networks for predicting quality in reclaimed waste sand for foundry applications. *J. Manuf. Syst.* **2025**, *79*, 584–597. [CrossRef]
37. Cheng, X.D. Anomaly detection for industrial robots based on temporal graph neural networks and differential privacy. *Eng. Res. Express* **2026**, *8*, 065222. [CrossRef]
38. Villalobos, K.; Suykens, J.; Illarramendi, A. A flexible alarm prediction system for smart manufacturing scenarios following a forecaster-analyzer approach. *J. Intell. Manuf.* **2021**, *32*, 1323–1344. [CrossRef]
39. Chen, X.; Cheng, K. Cutting Tool Remaining Useful Life Prediction Using Multi-Sensor Data Fusion Through Graph Neural Networks and Transformers. *Machines* **2025**, *13*, 1027. [CrossRef]
40. Wang, T.; Xu, W.; Chen, C.; Wang, Z.; Chen, Z. Progressive Hypergraph Structure Learning for Fault Diagnosis of Industrial Robots. *IEEE Trans. Instrum. Meas.* **2025**, *74*, TIM-2025. [CrossRef]
41. Wang, E.X.; Lei, Z.; Wen, G.; Liu, Z.; Su, Y.; Zhang, Z.; Chen, X. A physics-constrained Bayesian neural network for machinery remaining useful life prediction and uncertainty quantification. *Reliab. Eng. Syst. Saf.* **2026**, *266*, 111778. [CrossRef]
42. Mlinaric, J.; Pregelj, B.; Dolanc, G. End-of-Line Quality Control Based on Mel-Frequency Spectrogram Analysis and Deep Learning. *Machines* **2025**, *13*, 626. [CrossRef]
43. Papananias, M.; McLeay, T.E.; Mahfouf, M.; Kadirkamanathan, V.A. A probabilistic framework for product health monitoring in multistage manufacturing using Unsupervised Artificial Neural Networks and Gaussian Processes. *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.* **2023**, *237*, 1295–1310. [CrossRef]
44. Zhang, C.Y.; Yu, J.; Wang, S. Fault detection and recognition of multivariate process based on feature learning of one-dimensional convolutional neural network and stacked denoised autoencoder. *Int. J. Prod. Res.* **2021**, *59*, 2426–2449. [CrossRef]
45. Shaheen, B.; Kocsis, Á.; Németh, I. Data-driven failure prediction and RUL estimation of mechanical components using accumulative artificial neural networks. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105749. [CrossRef]
46. Chen, C.H.; Wang, C.; Guo, J.; Cui, P.; Zheng, J.; Liu, Z. Remaining useful life prediction considering multiple uncertainty information via Bayesian BiGRU-based method. *Reliab. Eng. Syst. Saf.* **2025**, *264*, 111431. [CrossRef]
47. Drouillet, C.; Karandikar, J.; Nath, C.; Journeaux, A.-C.; El Mansori, M.; Kurfess, T. Tool life predictions in milling using spindle power with the neural network technique. *J. Manuf. Process.* **2016**, *22*, 161–168. [CrossRef]
48. Wang, S.S.; Han, W.; Zhang, H.; Zeng, L. Lightweight rotating machinery fault diagnosis based on quadratic convolutional neural network and evidence fusion of multi-source sensor information. *J. Instrum.* **2025**, *20*, P02019. [CrossRef]
49. Liu, C.C.; Zhu, H.; Tang, D.; Nie, Q.; Zhou, T.; Wang, L.; Song, Y. Probing an intelligent predictive maintenance approach with deep learning and augmented reality for machine tools in IoT-enabled manufacturing. *Robot. Comput.-Integr. Manuf.* **2022**, *77*, 102357. [CrossRef]
50. Chen, C.H.; Parashar, P.; Akbar, C.; Fu, S.M.; Syu, M.-Y.; Lin, A. Physics-Prior Bayesian Neural Networks in Semiconductor Processing. *IEEE Access* **2019**, *7*, 130168–130179. [CrossRef]
51. Wang, Y.; Wang, G.; Wu, Y.; Zhang, G.; Wu, M. An uncertainty-aware deep learning ensemble approach for effective cutting tool predictive maintenance decision-making. *Meas. Sci. Technol.* **2025**, *36*, 026116. [CrossRef]
52. Zhang, J.; Li, C.; Deng, C.; Luo, T.; Deng, R.; Luo, D.; Tao, G.; Cao, H. Toward digital twins for intelligence manufacturing: Self-adaptive control in assisted equipment through multi-sensor fusion smart tool real-time machine condition monitoring. *J. Manuf. Syst.* **2025**, *82*, 301–318. [CrossRef]
53. Hwang, J.; Kim, S.; Park, S. Memory-Efficient Artificial Intelligence Framework for Real-Time Multivariate Anomaly Detection. *IEEE Internet Things J.* **2026**, *13*, 12544–12556. [CrossRef]
54. Ruan, H.; Dorneanu, B.; Arellano-Garcia, H.; Xiao, P.; Zhang, L. Deep Learning-Based Fault Prediction in Wireless Sensor Network Embedded Cyber-Physical Systems for Industrial Processes. *IEEE Access* **2022**, *10*, 10867–10879. [CrossRef]
55. Heydari, M.; Alinezhad, A.; Vahdani, B. A deep learning framework for quality control process in the motor oil industry. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108554. [CrossRef]
56. Tootooni, M.S.; Rao, P.K.; Chou, C.-A.; Kong, Z.J. A Spectral Graph Theoretic Approach for Monitoring Multivariate Time Series Data From Complex Dynamical Processes. *IEEE Trans. Autom. Sci. Eng.* **2018**, *15*, 127–144. [CrossRef]
57. Cao, Q.S.; Zanni-Merk, C.; Samet, A.; Reich, C.; Beuvron, F.d.B.d.; Beckmann, A.; Giannetti, C. KSPMI: A Knowledge-based System for Predictive Maintenance in Industry 4.0. *Robot. Comput.-Integr. Manuf.* **2022**, *74*, 102281. [CrossRef]
58. Blair, J.; Amin, O.; Brown, B.D.; McArthur, S.; Forbes, A.; Stephen, B. The transfer learning of uncertainty quantification for industrial plant fault diagnosis system design. *Data-Centric Eng.* **2024**, *5*, e41. [CrossRef]
59. Ong, K.S.H.; Wang, W.; Hieu, N.Q.; Niyato, D.; Friedrichs, T. Predictive Maintenance Model for IIoT-Based Manufacturing: A Transferable Deep Reinforcement Learning Approach. *IEEE Internet Things J.* **2022**, *9*, 15725–15741. [CrossRef]

60. Kim, G.; Kang, Y.S.; Yang, S.M.; Choi, J.G.; Hwang, G.; Park, H.W.; Lim, S. Fisher-informed continual learning for remaining useful life prediction of machining tools under varying operating conditions. *Reliab. Eng. Syst. Saf.* **2025**, *253*, 110549. [CrossRef]
61. Zhao, L.F.; Zhu, Y.; Zhao, T. Deep Learning-Based Remaining Useful Life Prediction Method with Transformer Module and Random Forest. *Mathematics* **2022**, *10*, 2921. [CrossRef]
62. Ben Ayed, M.; Soualhi, M.; Ketata, R.; Mairot, N.; Giampiccolo, S.; Zerhouni, N. A Data-Driven Methodology to Assess Raw Materials Impact on Manufacturing Systems Breakdowns. *Int. J. Progn. Health Manag.* **2024**, *15*, 3818. [CrossRef]
63. Bott, A.; Corduan, J.; Siems, M.; Puchta, A.; Fleischer, J. Improving Remaining Useful Life Prediction with Synthetic Data and Black Box Adversarial Reprogramming. *IEEE Access* **2025**, *13*, 195505–195516. [CrossRef]
64. Yan, W.; Shi, Y.; Ji, Z.; Sui, Y.; Tian, Z.; Wang, W.; Cao, Q. Intelligent predictive maintenance of hydraulic systems based on virtual knowledge graph. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106798. [CrossRef]
65. Sanz, E.; Blesa, J.; Puig, V. BiDrac Industry 4.0 framework: Application to an Automotive Paint Shop Process. *Control Eng. Pract.* **2021**, *109*, 104757. [CrossRef]
66. Wang, X.Q.; Liu, M.; Liu, C.; Ling, L.; Zhang, X. Data-driven and Knowledge-based predictive maintenance method for industrial robots for the production stability of intelligent manufacturing. *Expert Syst. Appl.* **2023**, *234*, 121136. [CrossRef]
67. Afia, A.; Gougam, F.; Soualhi, A.; Wadi, M.; Tahi, M.; Sahraoui, M.A. A data driven fault diagnosis approach for robotic cutting tools in smart manufacturing. *ISA Trans.* **2025**, *166*, 280–297. [CrossRef] [PubMed]
68. Hogea, E.; Onchiş, D.M.; Yan, R.; Zhou, Z. LogicLSTM: Logically-driven long short-term memory model for fault diagnosis in gearboxes. *J. Manuf. Syst.* **2024**, *77*, 892–902. [CrossRef]
69. Almeida, P.R.L.; Lima, T.L.V.; Brito, A.V.; Filho, A.C.L. Optimal feature complexity for small-sample bearing fault detection in manufacturing. *Int. J. Adv. Manuf. Technol.* **2026**, *142*, 2141–2158. [CrossRef]
70. Isiani, A.; Weiss, L.; Bardaweel, H.; Nguyen, H.; Crittenden, K. Fault Detection in 3D Printing: A Study on Sensor Positioning and Vibrational Patterns. *Sensors* **2023**, *23*, 7524. [CrossRef]
71. Asad, B.; Raja, H.A.; Vaimann, T.; Kallaste, A.; Pomarnacki, R.; Hyunh, V.K. A Current Spectrum-Based Algorithm for Fault Detection of Electrical Machines Using Low-Power Data Acquisition Devices. *Electronics* **2023**, *12*, 1746. [CrossRef]
72. Gunasegaram, D.R.; Barnard, A.; Matthews, M.; Jared, B.; Andreaco, A.; Bartsch, K.; Murphy, A. Machine learning-assisted in-situ adaptive strategies for the control of defects and anomalies in metal additive manufacturing. *Addit. Manuf.* **2024**, *81*, 104013. [CrossRef]
73. Marino, R.; Wisultschew, C.; Otero, A.; Lanza-Gutierrez, J.M.; Portilla, J.; de la Torre, E. A Machine-Learning-Based Distributed System for Fault Diagnosis with Scalable Detection Quality in Industrial IoT. *IEEE Internet Things J.* **2021**, *8*, 4339–4352. [CrossRef]
74. Raouf, I.; Kumar, P.; Lee, H.; Kim, H.S. Transfer Learning-Based Intelligent Fault Detection Approach for the Industrial Robotic System. *Mathematics* **2023**, *11*, 945. [CrossRef]
75. Sheuly, S.S.; Barua, S.; Begum, S.; Ahmed, M.U.; Güclü, E.; Osbakk, M. Data analytics using statistical methods and machine learning: A case study of power transfer units. *Int. J. Adv. Manuf. Technol.* **2021**, *114*, 1859–1870. [CrossRef]
76. Yang, H.H.; Wu, Y. Fault diagnosis of rotating electrical machines based on multi-source electrical signal fusion. *Eng. Res. Express* **2025**, *7*, 0252a3. [CrossRef]
77. Liu, M.J.; Gong, Y.; Sun, J.; Tang, B.; Sun, Y.; Zu, X.; Zhao, J. The accuracy losing phenomenon in abrasive tool condition monitoring and a noval WMMC-JDA based data-driven method considered tool stochastic surface morphology. *Mech. Syst. Signal Process.* **2023**, *198*, 110410. [CrossRef]
78. Umar, M.; Ahmad, Z.; Ullah, S.; Saleem, F.; Siddique, M.F.; Kim, J.-M. Advanced Fault Diagnosis in Milling Machines Using Acoustic Emission and Transfer Learning. *IEEE Access* **2025**, *13*, 100776–100790. [CrossRef]
79. Ali, H.; Zhang, Z.; Gao, F. Multiscale monitoring of industrial chemical process using wavelet-entropy aided machine learning approach. *Process Saf. Environ. Prot.* **2023**, *180*, 1053–1075. [CrossRef]
80. Khan, F.; Kamal, K.; Ratlamwala, T.A.H.; Alkahtani, M.; Mathavan, S. Tool Health Classification in Metallic Milling Process Using Acoustic Emission and Long Short-Term Memory Networks: A Deep Learning Approach. *IEEE Access* **2023**, *11*, 126611–126633. [CrossRef]
81. Li, H.; Yu, Z.; Li, F.; Kong, Q.; Tang, J. Real-time polymer flow state monitoring during fused filament fabrication based on acoustic emission. *J. Manuf. Syst.* **2022**, *62*, 628–635. [CrossRef]
82. Gattino, C.; Ottonello, E.; Baggetta, M.; Razzoli, R.; Stecki, J.; Berselli, G. Application of AI failure identification techniques in condition monitoring using wavelet analysis. *Int. J. Adv. Manuf. Technol.* **2023**, *125*, 4013–4026. [CrossRef]
83. Eldele, E.; Ragab, M.; Qing, X.; Chen, Z.; Wu, M.; Li, X.; Lee, J. UniFault: A Fault Diagnosis Foundation Model from Bearing Data. *arXiv* **2025**, arXiv:2504.01373. [CrossRef]
84. Peng, H.; Liu, J.; Du, J.; Gao, J.; Wang, W. BearLLM: A Prior Knowledge-Enhanced Bearing Health Management Framework with Unified Vibration Signal Representation. *arXiv* **2025**, arXiv:2408.11281. [CrossRef]
85. Shi, Z. MambaLithium: Selective State Space Model for Remaining-Useful-Life, State-of-Health, and State-of-Charge Estimation of Lithium-Ion Batteries. *arXiv* **2024**, arXiv:2403.05430. [CrossRef]

86. Wang, Y.; Wu, M.; Li, X.; Xie, L.; Chen, Z. A Survey on Graph Neural Networks for Remaining Useful Life Prediction: Methodologies, Evaluation and Future Trends. *arXiv* **2024**, arXiv:2409.19629. [CrossRef]
87. Fu, E.; Hu, Y.; Hu, C.; Jin, Z.; Peng, K. PEFT-MuTS: A Multivariate Parameter-Efficient Fine-Tuning Framework for Remaining Useful Life Prediction based on Cross-domain Time Series Representation Model. *arXiv* **2026**, arXiv:2601.22631. [CrossRef]
88. Albelali, S.; Ahmed, M. Hidden Leaks in Time Series Forecasting: How Data Leakage Affects LSTM Evaluation Across Configurations and Validation Strategies. *arXiv* **2025**, arXiv:2512.06932. [CrossRef]
89. Hespeler, S.C.; Moriano, P.; Li, M.; Hollifield, S.C. Temporal Cross-Validation Impacts Multivariate Time Series Subsequence Anomaly Detection Evaluation. *arXiv* **2025**, arXiv:2506.12183. [CrossRef]
90. Pham, T.M.T.; Premkumar, K.; Naili, M.; Yang, J. Time to Retrain? Detecting Concept Drifts in Machine Learning Systems. *arXiv* **2024**, arXiv:2410.09190. [CrossRef]
91. Urettini, E.; Atzeni, D.; Tsaknaki, I.-Y.; Carta, A. Online Continual Learning for Time Series: A Natural Score-driven Approach. *arXiv* **2026**, arXiv:2601.12931. [CrossRef]
92. Langer, T.; Widra, M.; Beyer, V. TinyML Towards Industry 4.0: Resource-Efficient Process Monitoring of a Milling Machine. *arXiv* **2025**, arXiv:2508.16553. [CrossRef]
93. Liao, J.X.; Wei, S.L.; Xie, C.L.; Zeng, T.; Sun, J.; Zhang, S.; Zhang, X.; Fan, F.L. BearingPGA-Net: A Lightweight and Deployable Bearing Fault Diagnosis Network via Knowledge Distillation and FPGA Acceleration. *arXiv* **2023**, arXiv:2307.16363. [CrossRef]
94. Abushahla, H.A.; AlHajri, M.; Varam, D.; Panopio, A.J.N. Neural Network Quantization for Microcontrollers: A Comprehensive Survey of Methods, Platforms, and Applications. *arXiv* **2025**, arXiv:2508.15008. [CrossRef]
95. Cummins, L.; Sommers, A.; Mittal, S.; Rahimi, S. Explainable Predictive Maintenance: A Survey of Current Methods, Challenges and Opportunities. *arXiv* **2024**, arXiv:2401.07871. [CrossRef]
96. Cummins, L.; Sommers, A.; Mittal, S.; Rahimi, S.; Seale, M.; Jabour, J.; Arnold, T. Explainable Anomaly Detection: Counterfactual driven What-If Analysis. *arXiv* **2024**, arXiv:2408.11935. [CrossRef]
97. Gama, J.; Ribeiro, R.P.; Mastelini, S.; Davarid, N.; Veloso, B. A Neuro-Symbolic Explainer for Rare Events: A Case Study on Predictive Maintenance. *arXiv* **2024**, arXiv:2404.14455. [CrossRef]
98. Hamilton, K.; Intizar, A. Neuro-Symbolic AI for Predictive Maintenance: A Review. *arXiv* **2026**, arXiv:2602.00731. [CrossRef]
99. Franck, C.M.; Fink, O. Explainable AI Guided Unsupervised Fault Diagnostics. *arXiv* **2025**, arXiv:2507.19168. [CrossRef]
100. Tao, L.; Li, S.; Liu, H.; Huang, Q.; Ma, L.; Ning, G.; Chen, Y.; Wu, Y.; Li, B.; Zhang, W.; et al. An Outline of Prognostics and Health Management Large Model: Concepts, Paradigms, and Challenges. *arXiv* **2024**, arXiv:2407.03374. [CrossRef]
101. Xu, Y.; Chen, Z.; Wang, R.; Li, Y.; Tang, F.; Zhao, M.; Liu, J. FaultDiffusion: Few-Shot Fault Time Series Generation with Diffusion Model. *arXiv* **2025**, arXiv:2511.15174. [CrossRef]
102. Yang, X.; Fang, C.; Liao, Y.; Yang, J.; Gryllias, K.; Chronopoulos, D. Deep Generative Models in Condition and Structural Health Monitoring: Opportunities, Limitations and Future Outlook. *arXiv* **2025**, arXiv:2507.15026. [CrossRef]
103. Wang, Y.; Liu, S.; Lv, S.; Liu, G. Meta-Learning and Knowledge Discovery based Physics-Informed Neural Network for Remaining Useful Life Prediction. *arXiv* **2025**, arXiv:2504.13797. [CrossRef]
104. Shah, S.; Daoliang, T.; Kumar, S.C. RUL Forecasting for Wind Turbine Predictive Maintenance based on Deep Learning. *arXiv* **2024**, arXiv:2412.17823. [CrossRef]
105. Chou, P.-H.; Mao, W.L.; Lin, R.P. YOLO-based Bearing Fault Diagnosis with Continuous Wavelet Transform. *arXiv* **2025**, arXiv:2509.03070. [CrossRef]
106. Wang, Z.; Zhao, H.; Yang, Y.; Hu, D.; Bao, C.; Liu, M.; Di, K.; Dustdar, S.; Wang, Z.; Deng, S. OmniFuser: Adaptive Multimodal Fusion for Service-Oriented Predictive Maintenance. *arXiv* **2025**, arXiv:2511.01320. [CrossRef]
107. Sharma, K. Uncertainty-Aware Deep Learning Framework for Remaining Useful Life Prediction in Turbofan Engines with Learned Aleatoric Uncertainty. *arXiv* **2025**, arXiv:2511.19124. [CrossRef]
108. Li, E.; Zhou, Z.; Chen, X. Edge intelligence: On-demand deep learning model co-inference with device-edge synergy. In Proceedings of the MECOMM@SIGCOMM 2018, Budapest, Hungary, 20 August 2018; pp. 31–36. [CrossRef]
109. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
110. Speith, T. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), Seoul, Republic of Korea, 21–24 June 2022*; ACM: New York, NY, USA, 2022; pp. 2239–2250. [CrossRef]
111. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1312. [CrossRef] [PubMed]
112. Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C.J.; Müller, K.R. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE* **2021**, *109*, 247–278. [CrossRef]

113. Ran, Y.; Zhou, X.; Lin, P.; Wen, Y.; Deng, R. A survey of predictive maintenance: Systems, purposes and approaches. *arXiv* **2019**, arXiv:1912.07383. [CrossRef]
114. Susto, G.A.; Schirru, A.; Pampuri, S.; McLoone, S.; Beghi, A. Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Trans. Ind. Inform.* **2015**, *11*, 812–820. [CrossRef]
115. Khan, S.; Yairi, T. A review on the application of deep learning in system health management. *Mech. Syst. Signal Process.* **2018**, *107*, 241–265. [CrossRef]
116. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [CrossRef]
117. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
118. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
119. Zhou, Z.; Chen, X.; Li, E.; Zeng, L.; Luo, K.; Zhang, J. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proc. IEEE* **2019**, *107*, 1738–1762. [CrossRef]
120. Deng, S.; Zhao, H.; Fang, W.; Yin, J.; Dustdar, S.; Zomaya, A.Y. Edge intelligence: The confluence of edge computing and artificial intelligence. *IEEE Internet Things J.* **2020**, *7*, 7457–7469. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.