

FROM PROGNOSTIC SIGNALS TO MAINTENANCE DECISIONS: A STRUCTURED EVIDENCE SYNTHESIS AND HUMAN-SUPERVISED AI-PDM FRAMEWORK FOR MANUFACTURING

Jha Keshav Gopal¹, Dr. Priyanka Jhavar², Dr. Sachin Baraskar³

Research Scholar, Department of Industrial Engineering, Sri Satya Sai University of Technology and
Medical Sciences¹

Associate Professor, Department of Industrial Engineering, Sri Satya Sai University of Technology
and Medical Sciences²

Associate Professor, Department of Industrial Engineering, Sri Satya Sai University of Technology
and Medical Sciences³

ABSTRACT

Manufacturing plants need maintenance policies that protect equipment reliability without sacrificing throughput, quality, or cost. This review examines past work on artificial intelligence (AI)-based predictive maintenance (PdM) and synthesizes evidence for a manufacturing framework connecting condition monitoring, failure diagnosis, remaining useful life (RUL), maintenance decisions, and production scheduling. A structured evidence synthesis covers foundational prognostics and health management, machine learning, deep learning, industrial IoT, digital twins, knowledge-based reasoning, and studies linking predictions to maintenance actions. The synthesis finds a consistent technical pattern: sensing and analytics are increasingly capable of detecting abnormal states and estimating degradation, yet evidence for sustained plant-level reliability and productivity gains remains less mature. Benchmark accuracy often dominates reporting, while calibration, lead time, false-alarm burden, cross-machine transfer, technician response, and production effects are reported inconsistently. Because equipment, datasets, failure definitions, horizons, and outcome measures differ, a pooled numerical effect estimate would be misleading; the paper therefore presents a transparent, paragraph-form descriptive meta-analysis rather than claiming a statistical effect-size meta-analysis. On this basis, it proposes a human-supervised, uncertainty-aware framework that fuses sensor, process, maintenance, and production data; deploys edge and cloud analytics; converts risk estimates into cost- and schedule-aware work orders; and learns from completed

interventions. Reliability and productivity are treated as joint outcomes measured through availability, mean time between failures, downtime, OEE, throughput, quality, and lifecycle cost. The paper identifies evidence gaps and proposes a validation ladder for future operational evaluation. Its contribution is an integrative synthesis and deployable architecture, not a new empirical performance claim.

Keywords: *Predictive maintenance¹; Artificial intelligence²; Manufacturing systems³; Reliability⁴; Productivity⁵; Prognostics and health management⁶; Meta-analysis⁷*

1. INTRODUCTION

1.1 Reliability and productivity as coupled manufacturing objectives

Manufacturing equipment is both a physical asset and a constraint on the flow of production. A failed spindle, pump, motor, robot, or compressor can interrupt a line, create downstream starvation, defer orders, and produce secondary quality losses. Reliability therefore cannot be treated as an engineering statistic detached from throughput. Mean time between failures (MTBF), failure rate, and availability describe asset behavior, whereas overall equipment effectiveness (OEE), throughput, schedule attainment, scrap, and energy intensity describe operating consequences. Maintenance choices influence both groups: repair may restore reliability but consume a production window, while delaying intervention may protect short-term output and increase the probability of an expensive breakdown. Machinery diagnostics and condition-based maintenance research established the basis for connecting observations of deterioration to planned action [1]–[3].

The maintenance policy has evolved from corrective repair, through calendar-based preventive maintenance, to condition-based and predictive strategies. Corrective work accepts failure before action; time-based work exchanges failure risk for periodic intervention; condition-based maintenance reacts to observed state; and PdM estimates a future state or failure risk to select when and how to intervene. The distinction matters because a forecast is not itself a maintenance outcome. It must also distinguish degradation that is actionable from harmless variation caused by load, recipe, ambient temperature, or normal process change. Reliability engineering and production planning therefore need a common decision horizon rather than separate optimization targets.

1.2 From condition monitoring to AI-enabled predictive maintenance

Prognostics and health management (PHM) organizes this chain as sensing, health assessment, diagnosis, prognosis, and decision support. Traditional condition monitoring uses vibration, temperature, current, pressure, acoustic emission, lubrication, speed, or process variables to detect changes in an asset. Statistical and physical models describe expected behavior, while data-driven artificial intelligence learns relationships between operating histories and failures or degradation. This shift has been enabled by connected sensors, programmable controllers, industrial networks, and storage for multivariate time series. Industry 4.0 reviews describe the resulting opportunity as the integration of intelligent sensing, analytics, and cyber-physical production rather than the isolated installation of a classifier [8], [11], [12], [28]. Standards provide useful lifecycle structure for condition-monitoring programs and their data processing interfaces [24], [25].

The AI toolbox spans supervised classification for fault labels, regression for degradation or RUL, unsupervised anomaly detection when labels are scarce, and sequential or deep models for complex temporal signals. Ensembles and multiple classifiers can represent differing alert horizons; convolutional networks learn signal patterns; recurrent networks model temporal dependencies; Bayesian approaches can express predictive uncertainty; and knowledge-based systems encode equipment context and maintenance rules. These methods answer different questions and should not be ranked by accuracy alone. Fault detection asks whether behavior is abnormal; diagnosis asks what is wrong; prognosis asks how condition may evolve; and prescriptive support asks which intervention best serves plant objectives. A robust PdM program links all four, exposes uncertainty, and permits operators to challenge recommendations.

1.3 Review rationale, questions, and contribution

Past reviews document rapid growth in learning algorithms, sensor infrastructure, and digital manufacturing, but also report uneven industrial validation and inconsistent evaluation practice [7], [9], [10], [13], [14], [17], [27]. This creates a practical evidence gap: an algorithm may classify a benchmark accurately yet fail to generalize across machines, shifts, tooling, loads, or factories; conversely, an interpretable model with modest accuracy may be more useful if it gives reliable lead time and supports safe work planning. The present review asks, in prose, which AI and PHM approaches have been applied to manufacturing maintenance, how prior work connects model outputs to reliability and productivity decisions, and what evidence is still required to substantiate operational benefit.

The paper contributes a structured, paragraph-form meta-synthesis of past work and an integrated reference framework for manufacturing systems. It treats reliability and productivity as complementary outcomes, distinguishes predictive skill from decision utility, and makes explicit the barriers to numerical pooling across heterogeneous studies. The scope includes condition monitoring, fault diagnosis, RUL estimation, model uncertainty, digital twins, maintenance scheduling, and industrial deployment evidence published through June 2024. The review is not a new plant experiment, nor does it claim a pooled causal reduction in downtime. Instead, it organizes the literature into a traceable design logic, identifies where reported evidence is strongest or weakest, and proposes a validation pathway by which future deployments can connect sensor predictions to verifiable shop-floor outcomes.

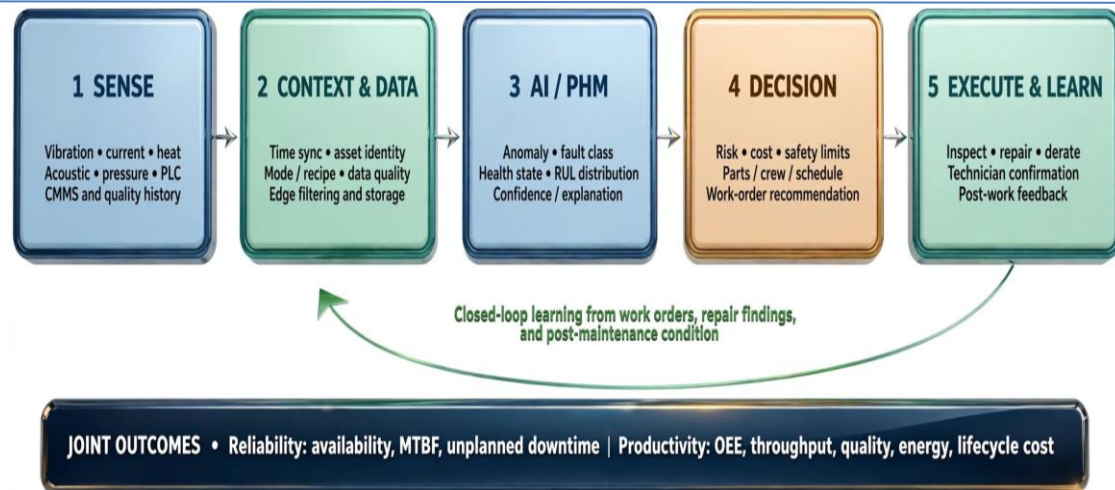


Figure 1: AI-PdM reference architecture linking sensing, PHM, decisions, and closed-loop maintenance

Figure 1 presents the proposed end-to-end architecture, emphasizing that model output is only one stage in maintenance value creation. Sensors and machine records must be synchronized with asset, product, and operating context before model inference. The analytics layer separates anomaly detection, fault diagnosis, and degradation prognosis, while uncertainty accompanies each estimate. A decision service then weighs failure risk against parts, labor, safety, and production constraints. Human-confirmed work orders close the loop by recording findings and verifying repairs. Reliability and productivity indicators are monitored together, preventing a high-scoring model from being mistaken for an operational improvement. The overall architecture is conceptual.

2. SURVEY OF PAST WORK

2.1 Foundations: condition monitoring, diagnostics, and prognostics

The conceptual roots of AI-based PdM lie in condition-based maintenance and machinery prognostics. Jardine, Lin, and Banjevic describe the diagnostic and prognostic reasoning needed to move from condition signals to maintenance decisions, while Heng and colleagues review rotating-machinery prognostics and emphasize degradation modeling, sensor choice, and uncertainty [1], [2]. Lee and colleagues organize PHM design into methods and applications for rotary systems, demonstrating that data acquisition, health indicators, fault isolation, and remaining-life assessment form a connected lifecycle rather than independent modeling tasks [3]. Statistical RUL research adds a formal treatment of censored observations, degradation processes, and population variability [4]. Later reviews extend the workflow from sensor acquisition and health-indicator construction to RUL prediction, underscoring that preprocessing and operating context can be as consequential as the model family [6].

Industrial monitoring programs also require engineering discipline around sensor placement, baselines, thresholds, and the interpretation of machine state. ISO 17359 provides general guidance for establishing a condition-monitoring program, and ISO 13374 describes functional stages for processing, communicating, and presenting diagnostic information [24], [25]. These standards are not AI recipes; their value is architectural. They encourage the plant to define the asset, failure modes, measurable parameters, and escalation logic before selecting a learning algorithm. A sensor stream without asset identity, operating mode, maintenance history, and quality checks is not

a usable health record. In synthesis, the strongest foundational studies thus frame PdM as a system-level activity: reliable measurement, meaningful health representation, decision thresholds, and feedback after maintenance.

Health indicators are constructed proxies, not direct measurements of damage; their usefulness depends on monotonicity, sensitivity, and robustness to operating change. Vibration features may reveal imbalance or bearing defects, while process variables can expose efficiency loss or abnormal cycle behavior. Reviews recommend joining signal features with asset engineering knowledge and operating regimes. A single threshold cannot represent every machine state, so context-aware baselines and repeated validation remain essential.

2.2 AI model families and predictive tasks

Classical machine learning made PdM practical for structured industrial data. Susto and colleagues use multiple classifiers with different prediction horizons to balance unexpected breakdowns against unnecessary early intervention in a semiconductor maintenance problem [5]. Reviews by Carvalho and colleagues and by Leukel and colleagues show broad use of support-vector machines, tree ensembles, neural networks, and feature engineering, while noting that datasets, preprocessing, and evaluation protocols differ markedly [7], [13]. Random forests and boosting can work well on tabular sensor summaries and offer useful feature attributions; one-class methods can learn normal operation when failures are rare; and calibrated classifiers can support risk thresholds. Their limits include dependence on handcrafted features, sensitivity to label definitions, and weak extrapolation beyond the training conditions.

Deep learning reduces some feature-engineering burden by learning representations from raw or transformed signals. Convolutional networks capture local patterns in vibration or current signals; recurrent models, including long short-term memory networks, represent temporal sequence; and hybrid architectures combine feature extraction with temporal prediction. Reviews compare these families across anomaly detection, fault classification, and RUL tasks but caution that benchmark gains do not establish plant generality [17], [18]. The LSTM RUL approach of Zheng and colleagues and convolutional RUL work by Li and colleagues illustrate sequence and signal-learning strategies often evaluated on run-to-failure benchmarks [29], [30]. Motor-focused synthesis reports increasing deep-learning use in induction-motor diagnostics, alongside continuing reliance on public datasets and a need for more industrial records [20].

The task definition determines what evidence a model can supply. Anomaly detection flags deviation but may not identify a cause; classification assigns a known fault label but may miss unseen failures; RUL regression offers a time estimate but can create false precision if uncertainty is ignored. Supervised failure prediction can be useful when failure events and exposure windows are consistently logged. Unsupervised or self-supervised models reduce label dependence but require careful validation of alert meaning. Reviews covering manufacturing and broader industrial use cases emphasize that the choice among statistical, physics-informed, classical ML, and deep learning should follow data availability, risk, interpretability, and computational constraints [9], [10], [11], [14], [21]. A hybrid approach is often more credible than a universal algorithm because physical knowledge constrains implausible outputs, and machine-learning components capture nonlinear residual patterns.

Method comparison is complicated by unequal labels and prediction objectives. Windows may surround known failures, while anomaly models learn nominal periods and RUL studies use run-to-failure trajectories. These tasks have different sampling units and utility functions. A fault classifier cannot be compared with an RUL regressor. Appropriate baselines include maintenance rules, statistical models, and classifiers, not deep networks. Transparent ablations establish whether added complexity earns its deployment cost.

2.3 Industrial data, sensing, and deployment

Manufacturing PdM draws on vibration, electrical current, temperature, acoustics, pressure, oil condition, controller alarms, cycle time, tool load, quality inspection, and work-order data. Intelligent sensors and edge devices can detect or summarize signals near the machine; cloud or plant platforms can support fleet analysis and model management [12], [28]. The value of multi-sensor fusion is conditional: sensors must be synchronized, calibrated, associated with the correct asset and operating mode, and tied to failure records. IoT-enabled production-line studies illustrate the potential for real-time data streams, while systematic visual-analytics reviews highlight the role of dashboards and human interpretation in anomaly review and maintenance planning [15], [19]. These systems need reliable network behavior, secure interfaces, data governance, and a plan for legacy equipment with limited connectivity.

Evidence maturity is constrained by the difference between controlled benchmarks and production conditions. Public bearing or engine datasets make method comparison possible, but often represent a narrow asset type, controlled degradation, fixed sensing, and selected operating regimes. Real factories add maintenance interventions that truncate degradation histories, sensor replacement, changing recipes, nonstationary loads, and failure labels recorded for business rather than research purposes. Data may be imbalanced because failures are rare, and random sample splits can leak information from adjacent time windows into both training and test sets. Reviews of industrial ML adoption repeatedly call for stronger out-of-sample evaluation, clear reporting of data preparation, and validation under changing operating conditions [7], [13], [14], [20]. Therefore, model design and data engineering must be evaluated together.

Data integration changes the governance problem. Sensor records are high-frequency and sensitive, maintenance logs are often free text, and production systems may encode equipment differently. Linking them requires a stable asset hierarchy, timestamp discipline, consistent failure taxonomy, and access controls. Edge processing can lower latency and bandwidth, while central fleet learning supports longer-term comparisons. It must define data location, label authority, and model rollback after drift or faults.

2.4 Connecting prognosis to maintenance and production decisions

A central transition in past work is from forecasting asset health to deciding what the plant should do. Multi-model PHM surveys call for integration of diagnosis and prognosis, while production-oriented studies show how operation-specific health estimates can inform schedules [9], [16]. Zhai, Gehring, and Reinhart explicitly connect a generative health-indicator approach to integrated production scheduling and test it using both a benchmark and machining-center data, making the connection between a predicted state and a production plan visible [16]. Knowledge-based systems add equipment rules and semantic context to analytics; digital-twin reviews describe

how a virtual representation may combine current state, history, and simulation for maintenance support [22], [23]. These approaches advance from dashboard alerts toward a decision layer with actionable context.

Uncertainty-aware prognosis is especially important when a false alarm can cause a costly shutdown or when a missed failure threatens safety or major delivery commitments. Zhuang, Xu, and Wang integrate Bayesian deep-learning RUL distributions with maintenance and spare-parts decisions, illustrating how uncertainty may flow into planning rather than being discarded after point prediction [26]. Recent PHM synthesis distinguishes life-consumption monitoring from system-health monitoring and explains how health information can support PdM scheduling [27]. The decision should consider failure probability, consequence, intervention duration, component lead time, production urgency, crew availability, and the possibility of continued safe operation. Accordingly, a predicted RUL is not a maintenance date by itself. It is one input to a constrained decision that may recommend inspection, derating, planned repair, spare-part reservation, or monitored deferral.

Across this literature, the evidence is strongest for technical feasibility and weakest for sustained causal effects on plant-wide productivity. Many papers report detection accuracy, F1, RMSE, or a benchmark score; fewer report calibrated risk, warning time, false-alert workload, work-order acceptance, repeat failures, or net production effects. Visual analytics research notes that maintenance personnel are often involved in data exploration but less consistently in model construction and deployment [19]. Reviews of digital twins and AI-based maintenance likewise identify implementation, interoperability, and validation as open issues [10], [22], [27]. This gap motivates a framework that treats technician feedback, work-order closure, and post-maintenance condition as model inputs, and treats reliability, production, and cost as measured outcomes rather than assumed consequences of algorithmic accuracy.

Decision policies use expected loss: failure probability multiplied by consequence, plus intervention, production interruption, inventory, and quality costs. Actions are selected under crew, parts, safety, and schedule constraints, with threshold rules reserved for critical hazards. This clarifies why one RUL threshold is inadequate across assets. It also enables comparison of AI assistance with current practice using work orders, observed repairs, and attributable changes in equipment and production outcomes.

3. METHODOLOGY

This review uses a structured, paragraph-form evidence synthesis focused on AI-enabled predictive maintenance in discrete and process manufacturing. The scope covers peer-reviewed research and authoritative standards published from 2003 through June 2024, with earlier foundational ideas considered through seminal works. Targeted discovery combined terms for predictive maintenance, condition monitoring, prognostics and health management, machine learning, deep learning, remaining useful life, manufacturing, reliability, productivity, digital twins, and industrial IoT. Publisher records, scholarly-index results, and citation chaining were used to identify relevant sources. The resulting reference set contains 30 sources cited in this paper; it is a purposive core corpus, not a claim that every eligible publication worldwide was retrieved. The review is consequently positioned as a transparent narrative meta-synthesis rather than a registered PRISMA systematic review. The concepts were

iteratively refined to include variants such as failure prediction, health indicators, condition-based maintenance, and maintenance scheduling. Searches were bounded by the review's manufacturing focus.

Sources were retained when they contributed at least one of four forms of evidence: a manufacturing or machinery PdM application; a review or systematic synthesis of AI, PHM, or RUL methods; a framework linking health estimates to maintenance or production decisions; or a standard relevant to monitoring architecture. Studies focused only on unrelated clinical prediction, generic AI without equipment-health relevance, or nontechnical commentary without a traceable argument were outside scope. The unit of synthesis was the source's reported task and evidence, not an assumed independent effect size. For each source, the review coded its asset or context, sensor or data type, task, model family, treatment of uncertainty, validation setting, maintenance linkage, and reported reliability or productivity outcomes. Review articles were used to characterize the field and identify recurring limitations; they were not counted as independent experiments. Each source was appraised for relevance, methodological transparency, validation design, and directness of its maintenance link. No shared numeric quality score was imposed because reviews, empirical studies, and standards do not use comparable reporting units.

Synthesis proceeded by comparing like with like conceptually across monitoring, detection, diagnosis, prognosis, decision support, and deployment. Findings were grouped as technical capability, data and validation quality, actionability, and operational outcome. The paper reports recurring patterns in prose and uses original figures to visualize the framework and evidence logic; it does not present unsupported frequency estimates or a fabricated study-flow count. A statistical meta-analysis would require comparable populations, failure definitions, prediction horizons, baselines, and extractable effect sizes, which the cited sources do not provide consistently. In particular, benchmark accuracy cannot be pooled meaningfully with plant-level downtime or OEE changes. No pooled percentage benefit or causal effect is therefore asserted. The synthesis is limited by language and publication bias, a purposeful rather than exhaustive corpus, and uneven reporting in the underlying literature; conclusions should guide design and evaluation, not substitute for local validation. The coding also recorded whether validation used simulation, laboratory experiments, public benchmarks, industrial data, or field deployment, and whether reported benefits were measured or inferred. This distinction informed the strength assigned to each synthesis claim.

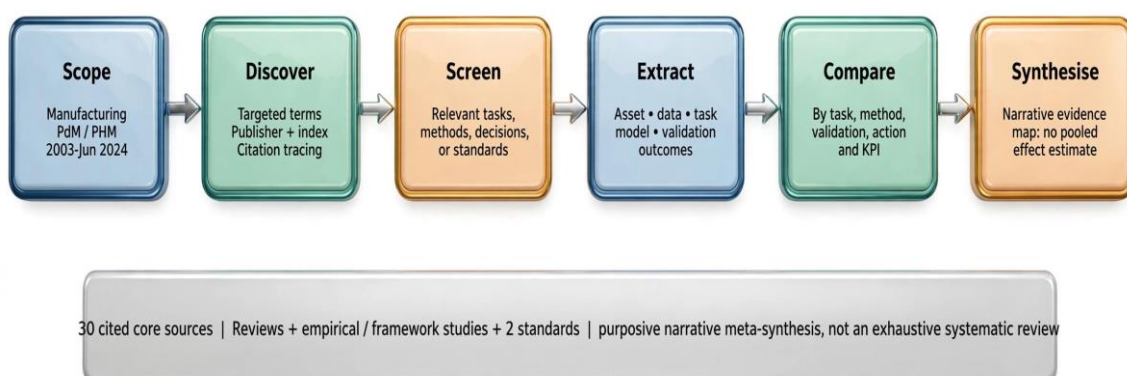


Figure 2: Review evidence logic and coding protocol; the diagram intentionally reports no unsupported screening totals

Figure 2 shows the review's evidence logic without inventing database hit or exclusion totals. Search and citation tracing identify relevant work; scope screening retains manufacturing maintenance, machinery PHM, decision integration, and standards evidence. The 30-source core corpus is then coded against dimensions, including task, model, data, validation, intervention, and outcome. Synthesis compares patterns across evidence types rather than averaging incompatible performance numbers. This diagram is not a PRISMA flow chart and does not imply exhaustive retrieval. It makes the method more auditable by separating discovery, eligibility, extraction, and interpretation, and by identifying the point at which heterogeneity prevents numerical pooling.



Figure 3: Qualitative evidence map of task coverage across method families; coverage is not a frequency count

Figure 3 maps prior approaches to the maintenance task they support. Statistical and physics-based methods anchor condition and degradation modeling; classical machine learning supports structured detection and classification; deep learning learns signal and sequence representations; hybrid or knowledge-based designs link analytics to context and decisions. Filled cells indicate documented use in the cited literature, not comparative superiority or study frequency. RUL differs from diagnosis; models may span cells. The map highlights a research imbalance: sensing and prediction are richly studied, whereas intervention planning, production integration, and post-maintenance learning receive less consistent treatment. It is a conceptual map, not a count.

4. CRITICAL ANALYSIS OF PAST WORK

A first strength of past work is the maturation of the technical pipeline. Foundational PHM research clarified the importance of health indicators, diagnostic isolation, degradation modeling, and RUL estimation [1]–[4], [6]. Subsequent reviews map classical ML and deep learning to a wider range of machinery and show how sensing and computing have expanded the feasible analysis [7], [11], [14], [17], [20]. Multiple classifiers and Bayesian prognosis improve the conceptual link between prediction horizon, uncertainty, and action [5], [26]. These advances mean that manufacturers can often construct a credible proof of concept for a defined asset, fault family, and sensor set. The critical limitation is that technical sophistication is frequently mistaken for operational

maturity. A strong model score answers whether test examples were separated or estimated well under one protocol; it does not establish that a plant can act early, safely, economically, and repeatedly.

The most persistent weakness is external validity. Laboratory rigs and public datasets offer clean signals and reproducible labels, but industrial assets operate under changing loads, products, maintenance histories, and sensor configurations. Reviews repeatedly identify a shortage of industrial datasets and limited transfer across assets or working conditions [10], [13], [14], [20], [21]. A random train-test split can overstate generalization when nearby time windows, the same asset, or overlapping run-to-failure trajectories appear on both sides. A defensible evaluation should split by time, asset, production line, or site according to the intended deployment claim; report exposure time and failure prevalence; and compare against simple rules and existing preventive schedules. Few reported studies support a sufficiently broad transfer claim. The evidence therefore supports feasibility more strongly than universal portability.

A second limitation is metric mismatch. Accuracy and ROC area can look persuasive on imbalanced data while concealing missed failures or an intolerable alert rate. F1 and precision-recall measures improve visibility of rare positives but omit warning time, repair opportunity, and failure consequence. RUL metrics such as MAE or RMSE compress asymmetric decision risk into an average error and can reward a forecast that arrives too late for the maintenance window. Plant evaluation should report calibration, lead-time distributions, false alerts per asset-month, missed-event cost, decision coverage, and uncertainty intervals, alongside MTBF, availability, downtime, MTTR, OEE, throughput, quality, and maintenance cost. Without a common denominator, such as asset-hours at risk or production orders exposed, results cannot be compared reliably. This heterogeneity blocks a responsible numerical meta-analysis and cautions against transferring percentage claims between sectors.

A third issue is the separation of prediction from intervention. Several studies develop classifiers or RUL models, but plant benefit depends on whether a recommendation reaches a competent technician, whether parts and access are available, and whether the intervention fits the schedule. Integrated scheduling work is a notable advance because it recognizes that a machine's health and its assigned operations affect one another [16]. Knowledge-based and digital-twin approaches also bring context and rule constraints into decision support [22], [23]. Yet these architectures need evidence that the model changes action appropriately, rather than merely adding another dashboard. Maintenance history is also endogenous: a predicted fault can trigger replacement and prevent the very failure that would have confirmed the prediction. Evaluation must treat intervention, censoring, and counterfactual outcomes explicitly. Otherwise, a model may be penalized for a successful preventive action or rewarded for a warning that operators ignore.

Finally, the human, organizational, and lifecycle dimensions remain underdeveloped relative to algorithm design. Operators possess tacit knowledge about sound, vibration, process transitions, and the plausibility of an alert; maintenance teams also determine whether a warning is timely and actionable. Visual analytics reviews identify an incomplete role for maintenance personnel beyond exploration [19]. Digital infrastructure creates further concerns about cyber-security, interface interoperability, model drift, sensor outages, access rights, and support responsibility [10], [22], [24], [25]. These are not secondary implementation details: they determine data quality, trust, safe fallback, and ongoing cost. The reviewed corpus also contains publication and selection bias toward

successful demonstrations, while negative results and long-term service costs are rarely visible. Overall, past work demonstrates a compelling engineering opportunity, but not a single transferable return-on-investment coefficient. Strong future studies should use prospective, multi-asset evaluations, prespecified endpoints, transparent baselines, and an audit trail from prediction to work order and post-repair verification.

An additional methodological concern is dependence among publications. Several surveys review overlapping primary studies, so treating each citation as an independent confirmation would inflate support. This paper therefore uses reviews to locate patterns and applications to illustrate mechanisms, without adding their conclusions as separate effect observations. Evidence strength is judged by design, validation realism, and outcome proximity, not by citation frequency. The same caution applies to success narratives: a single pilot cannot establish a typical benefit across sectors. Future evidence repositories should identify assets, operating regimes, event definitions, exposure, interventions, and denominators, allowing replication and pooling when outcomes become comparable.

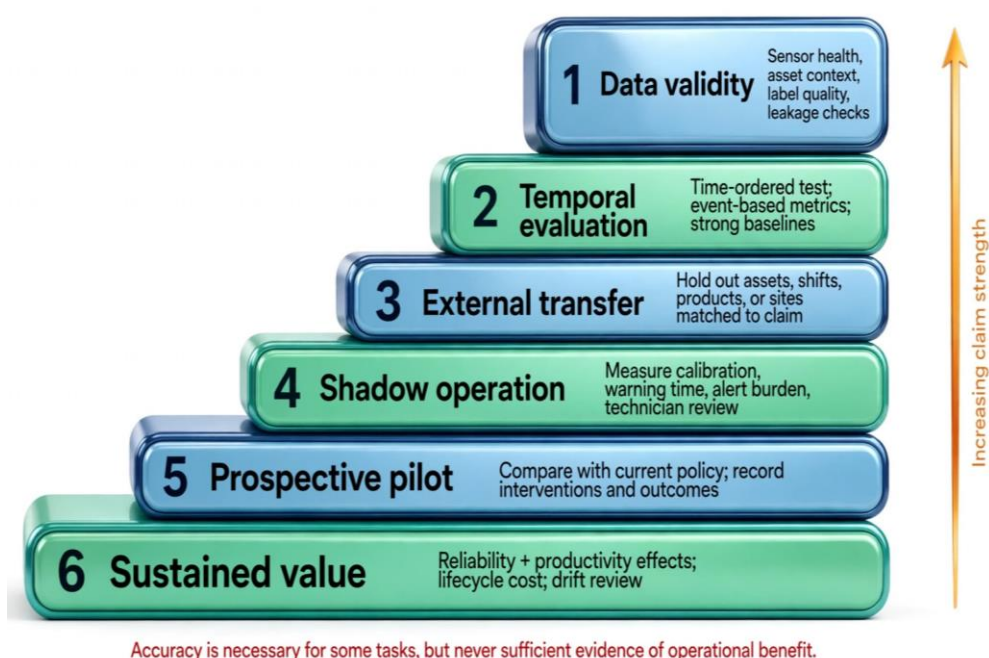


Figure 4: Validation ladder from data validity to sustained operational impact

Figure 4 shows a validation ladder for evidence as claims approach operations. Sensor and label checks precede time-separated testing; cross-asset evaluation tests transfer; shadow deployment measures alerts without changing maintenance; prospective pilots assess whether decisions improve outcomes. Each rung poses different questions, so strong benchmark performance cannot substitute for a field trial. The ladder also emphasizes calibration, false alerts, warning time, and technician workload alongside discrimination scores. At the highest levels, reliability and productivity effects require suitable comparisons and an auditable intervention record. The figure is a recommended evaluation pathway, not a report that every reviewed study completed each stage.

5. DISCUSSION

The synthesis supports a layered framework in which sensing, analytics, and maintenance actions are designed together. Sensor and controller data first pass through asset-context mapping, clock alignment, quality checks, and edge filtering. A health layer then combines interpretable baselines, fault classifiers, anomaly detection, and degradation models according to asset criticality and label availability. A prognosis layer should return a probability distribution or risk band over relevant horizons, not just a single RUL value. The decision layer combines that information with consequence of failure, safety limits, crew and spare-part capacity, maintenance duration, and production schedule. This design is consistent with PHM standards and with work integrating predictive maintenance, production scheduling, Bayesian uncertainty, and knowledge-based reasoning [16], [23]–[27].

The operational loop should preserve human authority. The system can rank assets, explain the signal or operating context behind an alert, estimate time-to-intervention, and generate a draft work order, while a technician or reliability engineer confirms the diagnosis and risk. The work order should capture the decision, action, parts, time, and findings; post-repair measurements then indicate whether the health state reset, remained abnormal, or degraded again. Where safety risk exceeds a defined threshold, established protection systems remain authoritative and the AI must not replace interlocks. Edge inference can support low-latency alarms; centralized services can manage model versions, fleet comparisons, and governance. A staged rollout begins with shadow monitoring, then advisory use, then tightly bounded automation after prospective evidence and hazard review.

Evaluation should test technical validity and economic utility at the same time. Before deployment, teams should specify the failure modes, prediction horizons, allowable false-alert burden, action windows, and relevant outcome denominators. Offline testing should use temporal and asset-level holdouts; shadow trials should measure calibration and workload; controlled pilots should compare the new policy with existing practice. Results should include reliability, production, quality, energy, labor, and lifecycle cost, with uncertainty and subgroup differences across machines or shifts. Benefits should be credited only when intervention and comparison conditions support attribution. Future research should publish reusable industrial protocols, negative results, drift and cybersecurity monitoring, and human-factors measures. This moves the field from model-centric demonstrations toward dependable maintenance services that earn trust through transparent performance.

Implementation should begin with a business case for a small set of critical assets rather than a plant-wide promise. Teams should map failure modes, establish baseline exposure and maintenance practice, and quantify the cost of false alarms, missed failures, planned stops, and spare-part delays. They can then select sensors and models matched to those failure modes, define ownership for data and alerts, and train technicians to record evidence consistently. A pilot should preserve an untreated or historical comparison where feasible, document production mix and maintenance changes, and track outcomes long enough to capture both avoided failures and added interventions. The scaling decision should consider integration effort, cybersecurity, lifecycle support, and workforce acceptance as well as model scores. This staged approach limits risk, tests local assumptions, and prevents technology procurement from preceding maintenance-process redesign. Named plant owners should review all relevant evidence.

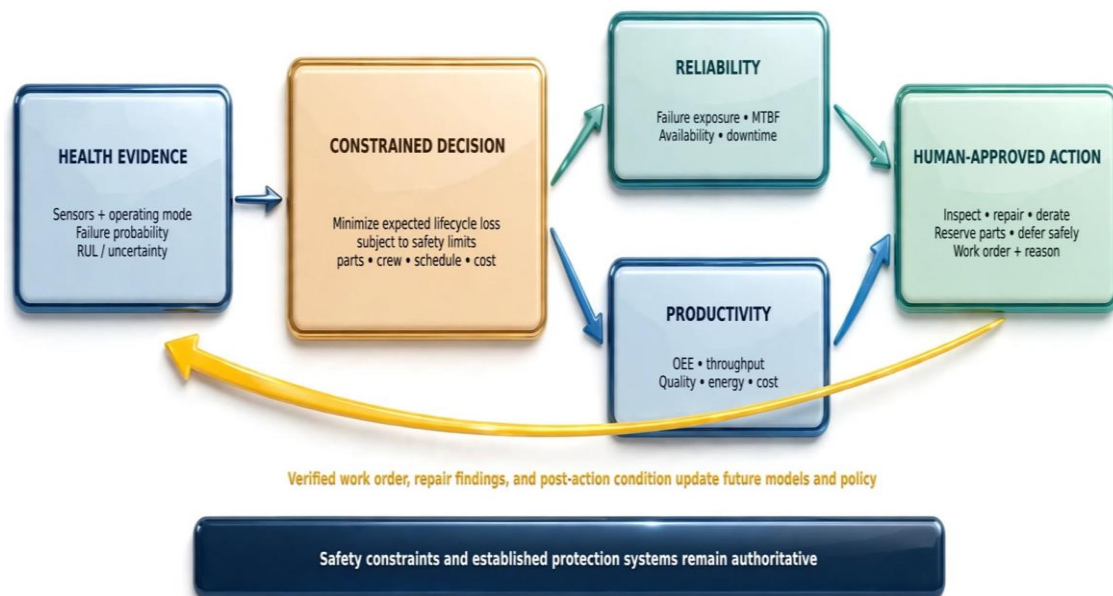


Figure 5: Constrained decision loop balancing reliability, productivity, safety, and human approval

Figure 5 presents the proposed reliability and productivity decision loop. The health model supplies risk and RUL distributions; the decision engine considers consequence, schedule, resources, and safety before recommending inspection, repair, derating, or monitored deferral. Technician confirmation and work-order completion return evidence to the system. Reliability outcomes include failure exposure, MTBF, availability, and downtime; productivity outcomes include OEE, throughput, quality, and resource use. The conceptual objective is to minimize expected lifecycle loss rather than maximize predictive accuracy in isolation. A plant may set different weights by asset criticality, but safety constraints remain non-negotiable. Feedback holds models accountable for measured consequences.

6. CONCLUSION

This review synthesized past work on AI-based predictive maintenance for manufacturing systems, focusing on the path from sensor evidence to reliability and productivity outcomes. The literature establishes mature building blocks for condition monitoring, fault detection, diagnosis, degradation assessment, and RUL prediction. Classical machine learning, deep learning, Bayesian methods, knowledge-based systems, and digital twins each contribute useful capabilities, but none alone constitutes an effective maintenance strategy. Operational value emerges only when estimates are calibrated, contextualized, converted into feasible work, coordinated with production, and verified after intervention.

The principal gap is the distance between benchmark prediction scores and sustained plant-level evidence. Dataset shift, rare failures, inconsistent labels, metric heterogeneity, uncertain warning time, and incomplete maintenance feedback limit transfer and block defensible pooled effect estimates. Accordingly, this paper presents a descriptive meta-synthesis and does not claim a numerical reduction in downtime or cost. Its proposed framework connects sensing and data quality, AI-based health assessment, uncertainty-aware decisions, human review, production scheduling, and closed-loop learning. Future studies should validate that chain prospectively across assets and shifts, compare against existing policies, report reliability and productivity together, and document workload,

safety, and lifecycle cost. AI-PdM can improve manufacturing performance when it is treated not as a model installation, but as an accountable socio-technical maintenance process. Industrial researchers should report negative trials, forecast horizons, asset-hour exposure, and intervention details, then repeat evaluation across lines and operating regimes. Such reporting will make replication possible, strengthen future quantitative reviews, and help plants judge whether claimed benefits are transferable and sustainable.

7. REFERENCES

- [1] A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical Systems and Signal Processing*, vol. 20, no. 7, pp. 1483–1510, 2006, doi: 10.1016/j.ymssp.2005.09.012.
- [2] A. Heng, S. Zhang, A. C. C. Tan, and J. Mathew, "Rotating machinery prognostics: State of the art, challenges and opportunities," *Mechanical Systems and Signal Processing*, vol. 23, no. 3, pp. 724–739, 2009, doi: 10.1016/j.ymssp.2008.06.009.
- [3] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems Reviews, methodology and applications," *Mechanical Systems and Signal Processing*, vol. 42, nos. 1–2, pp. 314–334, 2014, doi: 10.1016/j.ymssp.2013.06.004.
- [4] X. S. Si, W. Wang, C. H. Hu, and D. H. Zhou, "Remaining useful life estimation A review on the statistical data driven approaches," *European Journal of Operational Research*, vol. 213, no. 1, pp. 1–14, 2011, doi: 10.1016/j.ejor.2010.11.018.
- [5] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: A multiple classifier approach," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 812–820, 2015, doi: 10.1109/TII.2014.2349359.
- [6] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mechanical Systems and Signal Processing*, vol. 104, pp. 799–834, 2018, doi: 10.1016/j.ymssp.2017.11.016.
- [7] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, R. da P. Francisco, J. P. Basto, and S. G. S. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," *Computers & Industrial Engineering*, vol. 137, Art. no. 106024, 2019, doi: 10.1016/j.cie.2019.106024.
- [8] T. Zonta, C. A. da Costa, R. da Rosa Righi, M. J. de Lima, E. S. da Trindade, and G.-P. Li, "Predictive maintenance in the Industry 4.0: A systematic literature review," *Computers & Industrial Engineering*, vol. 150, Art. no. 106889, 2020, doi: 10.1016/j.cie.2020.106889.
- [9] J. J. M. Jimenez, S. Schwartz, R. Vingerhoeds, B. Grabot, and M. Salaün, "Towards multi-model approaches to predictive maintenance: A systematic literature survey on diagnostics and prognostics," *Journal of Manufacturing Systems*, vol. 56, pp. 539–557, 2020, doi: 10.1016/j.jmsy.2020.07.008.
- [10] J. Dalzochio, R. Kunst, E. Pignaton, A. Binotto, S. Sanyal, J. Favilla, and J. Barbosa, "Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges," *Computers in Industry*, vol. 123, Art. no. 103298, 2020, doi: 10.1016/j.compind.2020.103298.

- [11] Z. M. Çinar, A. A. Nuhu, Q. Zeeshan, O. Korhan, M. Asmael, and B. Safaei, "Machine learning in predictive maintenance towards sustainable smart manufacturing in Industry 4.0," *Sustainability*, vol. 12, no. 19, Art. no. 8211, 2020, doi: 10.3390/su12198211.
- [12] M. Pech, J. Vrchota, and J. Bednář, "Predictive maintenance and intelligent sensors in smart factory: Review," *Sensors*, vol. 21, no. 4, Art. no. 1470, 2021, doi: 10.3390/s21041470.
- [13] J. Leukel, J. González, and M. Riekert, "Adoption of machine learning technology for failure prediction in industrial maintenance: A systematic review," *Journal of Manufacturing Systems*, vol. 61, pp. 87–96, 2021, doi: 10.1016/j.jmsy.2021.08.012.
- [14] M. Nacchia, F. Fruggiero, A. Lambiase, and K. Bruton, "A systematic mapping of the advancing use of machine learning techniques for predictive maintenance in the manufacturing sector," *Applied Sciences*, vol. 11, no. 6, Art. no. 2546, 2021, doi: 10.3390/app11062546.
- [15] S. Ayvaz and K. Alpay, "Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real-time," *Expert Systems with Applications*, vol. 173, Art. no. 114598, 2021, doi: 10.1016/j.eswa.2021.114598.
- [16] S. Zhai, B. Gehring, and G. Reinhart, "Enabling predictive maintenance integrated production scheduling by operation-specific health prognostics with generative deep learning," *Journal of Manufacturing Systems*, vol. 61, pp. 830–855, 2021, doi: 10.1016/j.jmsy.2021.02.006.
- [17] O. Serradilla, E. Zugasti, J. Rodriguez, and U. Zurutuza, "Deep learning models for predictive maintenance: A survey, comparison, challenges and prospects," *Applied Intelligence*, vol. 52, pp. 10934–10964, 2022, doi: 10.1007/s10489-021-03004-y.
- [18] C. Ferreira and G. Gonçalves, "Remaining useful life prediction and challenges: A literature review on the use of machine learning methods," *Journal of Manufacturing Systems*, vol. 63, pp. 550–562, 2022, doi: 10.1016/j.jmsy.2022.05.010.
- [19] X. Cheng, J. K. Chaw, K. M. Goh, T. T. Ting, S. Sahrani, M. N. Ahmad, R. Abdul Kadir, and M. C. Ang, "Systematic literature review on visual analytics of predictive maintenance in the manufacturing industry," *Sensors*, vol. 22, no. 17, Art. no. 6321, 2022, doi: 10.3390/s22176321.
- [20] M. Drakaki, Y. L. Karnavas, I. A. Tzifettas, V. Linardos, and P. Tzionas, "Machine learning and deep learning based methods toward Industry 4.0 predictive maintenance in induction motors: State of the art survey," *Journal of Industrial Engineering and Management*, vol. 15, no. 1, pp. 31–57, 2022, doi: 10.3926/jiem.3597.
- [21] M. Fernandes, J. M. Corchado, and G. Marreiros, "Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: A systematic literature review," *Applied Intelligence*, vol. 52, pp. 14246–14280, 2022, doi: 10.1007/s10489-022-03344-3.
- [22] B. van Dinter, B. Tekinerdogan, and C. Catal, "Predictive maintenance using digital twins: A systematic literature review," *Information and Software Technology*, vol. 151, Art. no. 107008, 2022, doi: 10.1016/j.infsof.2022.107008.

- [23] Q. Cao, C. Zanni-Merk, A. Samet, C. Reich, F. de Bertrand de Beuvron, A. Beckmann, and C. Giannetti, "KSPMI: A knowledge-based system for predictive maintenance in Industry 4.0," *Robotics and Computer-Integrated Manufacturing*, vol. 74, Art. no. 102281, 2022, doi: 10.1016/j.rcim.2021.102281.
- [24] ISO, ISO 17359:2018, Condition Monitoring and Diagnostics of Machines General Guidelines, 3rd ed. Geneva, Switzerland: International Organization for Standardization, 2018.
- [25] ISO, ISO 13374-1:2003, Condition Monitoring and Diagnostics of Machines Data Processing, Communication and Presentation Part 1: General Guidelines. Geneva, Switzerland: International Organization for Standardization, 2003.
- [26] L. Zhuang, A. Xu, and X.-L. Wang, "A prognostic driven predictive maintenance framework based on Bayesian deep learning," *Reliability Engineering & System Safety*, vol. 234, Art. no. 109181, 2023, doi: 10.1016/j.ress.2023.109181.
- [27] C. Huang, S. Bu, H. H. Lee, C. H. Chan, S. W. Kong, and W. K. C. Yung, "Prognostics and health management for predictive maintenance: A review," *Journal of Manufacturing Systems*, vol. 75, pp. 78–101, 2024, doi: 10.1016/j.jmsy.2024.05.021.
- [28] J. Lee, H. Davari, J. Singh, and V. Pandhare, "Industrial artificial intelligence for Industry 4.0-based manufacturing systems," *Manufacturing Letters*, vol. 18, pp. 20–23, 2018, doi: 10.1016/j.mfglet.2018.09.002.
- [29] S. Zheng, K. R. Ristovski, A. Farahat, and C. Gupta, "Long short-term memory network for remaining useful life estimation," in *Proc. IEEE International Conference on Prognostics and Health Management (ICPHM)*, Dallas, TX, USA, 2017, pp. 88–95, doi: 10.1109/ICPHM.2017.7998311.
- [30] X. Li, Q. Ding, and J.-Q. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," *Reliability Engineering & System Safety*, vol. 172, pp. 1–11, 2018, doi: 10.1016/j.ress.2017.11.021.