

A Review of IoT-Based Machine Health Monitoring and Predictive Maintenance Using Machine Learning: Architectures, Algorithms, and Open Challenges

Suwasini S. Chabukswar, Brijendra Gupta

Department of Information Technology, Siddhant College of Engineering (affiliated to Savitribai Phule Pune University),
Sudumbare, Pune, Maharashtra, India

Published October 2026

Abstract

Unplanned equipment failure remains one of the most expensive problems in modern manufacturing, and the maintenance strategy adopted by a plant has a direct bearing on its productivity, safety, and operating cost. Traditional reactive and time-based preventive approaches either allow machines to run to failure or replace healthy components too early, and neither uses the rich stream of data that industrial assets now generate. The convergence of the Industrial Internet of Things (IIoT) and machine learning has made a third option practical: predictive maintenance, in which the condition of a machine is monitored continuously and its future health is estimated from sensor data. This paper reviews the state of the art in IoT-based machine health monitoring and machine-learning-driven predictive maintenance. We examine the layered sensing-to-cloud architectures that underpin such systems, survey the feature-extraction and learning methods that have been applied to fault classification and remaining-useful-life estimation, and compare representative studies spanning support vector machines, ensemble methods, deep networks, image-based diagnosis, and survival analysis. Reported accuracies range from roughly 88% for a single-classifier failure detector to 96% for a vision-based gearbox diagnostic, while survival models achieve concordance indices above 0.97 on real industrial data. Drawing these threads together, we identify a set of recurring gaps: an over-reliance on single algorithms and balanced benchmark datasets, the fragmented treatment of fault classification and life estimation, the limited adoption of survival analysis in an Industry 4.0 setting, and unresolved questions around edge deployment, security, and interpretability. We close by outlining an integrated research direction that unifies real-time IoT sensing with a combined classification-and-prognostics pipeline.

Keywords: Predictive maintenance; Industrial Internet of Things; machine learning; condition-based monitoring; remaining useful life; fault diagnosis; Industry 4.0; prognostics and health management

1 Introduction

1.1 Background and motivation

Manufacturing plants, power stations, transport fleets, and process industries all depend on rotating and reciprocating machinery whose sudden failure can halt an entire production line. When a critical motor, pump, gearbox, or bearing fails without warning, the cost is rarely limited to the replacement part: lost production, idle labour, damaged downstream equipment, missed deliveries, and, in the worst cases, safety incidents all contribute to a bill that dwarfs the price of the component itself. Surveys of industrial operators have repeatedly placed unplanned downtime among the largest

controllable costs on the factory floor, which explains why maintenance has moved from a back-office function to a strategic concern for plant managers.

For most of the twentieth century, industry responded to this problem in one of two ways. The simplest was to run equipment until it broke and then repair it, an approach that keeps planning effort low but exposes the operation to exactly the catastrophic, poorly timed failures it should be avoiding. The more disciplined alternative was to service machines on a fixed calendar or usage schedule, replacing parts before they were statistically likely to fail. Time-based servicing certainly reduces surprise breakdowns, but it does so bluntly: components that still have useful life are discarded, maintenance crews are dispatched whether or not a machine needs attention, and the schedule cannot react to the fact that two nominally identical machines age at very different rates depending on load, environment, and duty cycle [29].

1.2 From reactive to predictive: The maintenance continuum

Predictive maintenance occupies a distinct place on this continuum. Rather than waiting for failure or trusting a calendar, it seeks to infer the actual health of each asset from measurements taken while the machine is running, and to intervene only when the evidence indicates that a fault is developing. The idea is not new; vibration analysts and reliability engineers have practised condition monitoring for decades using handheld instruments and periodic route-based inspections [24]. What has changed is the economics of sensing and computation. Low-cost micro-electromechanical sensors, cheap microcontrollers, ubiquitous wireless networking, and elastic cloud storage have made it feasible to instrument even modestly valued machines and to stream their data continuously rather than sampling it during an occasional inspection.

Once data are available in this volume, the analytical bottleneck shifts from measurement to interpretation. A human analyst can study a handful of vibration spectra, but no team can manually watch thousands of continuously reporting assets. This is precisely the task at which machine learning excels: given enough labelled examples, statistical models can learn the often subtle signatures that distinguish a healthy machine from one that is beginning to degrade, and can do so faster and more consistently than manual analysis. The marriage of pervasive sensing with data-driven modelling is therefore the technical foundation of contemporary predictive maintenance [8, 20].

1.3 The convergence of IoT and machine learning

The Industrial Internet of Things provides the nervous system of a predictive-maintenance deployment. Sensors attached to a machine measure quantities such as vibration, temperature, acoustic emission, motor current, rotational speed, and pressure; an embedded acquisition device digitises these signals; and a communication layer carries the data to where it can be processed, whether that is a nearby edge gateway or a remote cloud platform [12, 13, 30]. Machine learning supplies the interpretive layer that turns this raw telemetry into actionable estimates of machine health, classifying the current condition of an asset and, increasingly, forecasting how much useful life remains before intervention becomes necessary.

It is the combination, rather than either technology alone, that creates value. Sensing without analytics produces dashboards that overwhelm operators with data they cannot act upon, while analytics without reliable, real-time data has nothing to reason about. When the two are integrated well, a plant gains the ability to schedule maintenance around production, to order spare parts before

they are urgently needed, and to avoid both the waste of premature replacement and the risk of run-to-failure operation. This review is concerned specifically with systems that bring these two capabilities together.

1.4 Scope and contributions of this review

A considerable body of literature now addresses predictive maintenance from different angles, and several excellent surveys already treat individual sub-topics such as deep learning for fault diagnosis or statistical remaining-useful-life estimation. The contribution of the present review is to bring the sensing side and the learning side into a single narrative, and to read a set of recent, concrete implementations through that combined lens. Specifically, this paper: (i) sets out the fundamental concepts and vocabulary of machine health monitoring and predictive maintenance; (ii) describes the layered IoT architecture common to practical deployments, from the sensing layer to the application layer; (iii) surveys the main families of machine-learning method applied to fault classification and life estimation, using recent studies as illustrative cases; (iv) compares those studies on a common footing of data, method, performance, and limitation; and (v) synthesises the recurring research gaps and proposes an integrated direction for future work that is aligned with an ongoing IoT-based predictive-maintenance project.

1.5 Organization of the paper

The remainder of the paper is organised as follows. The fundamental concepts of machine health monitoring, the Industrial Internet of Things and prognostics are introduced first, followed by the methodology used to select and review the literature. The IoT architecture that supports data collection and processing is then examined, and machine-learning approaches to predictive maintenance are surveyed by method family. A comparative analysis of representative studies follows, after which the research gaps and open challenges, a proposed research direction, and the conclusions and directions for future work are presented.

2 Fundamental concepts

2.1 Machine health monitoring and condition-based maintenance

Machine health monitoring is the practice of assessing the condition of equipment while it is in service, using physical measurements that correlate with the state of its components. Condition-based maintenance (CBM) is the maintenance policy built on top of such monitoring: instead of servicing at fixed intervals, work is triggered by evidence that a machine's condition has deteriorated beyond an acceptable threshold. The classical framing of CBM, articulated in influential reviews more than a decade ago, divides the problem into three steps — data acquisition, data processing, and maintenance decision-making — and this decomposition remains a useful way to reason about modern systems [6].

In a predictive-maintenance context, the condition of a machine is often expressed as a small number of discrete health states, such as healthy, warning, degraded, and faulty, or as a continuous health indicator that trends toward a failure threshold over time. Discrete states map naturally onto a classification problem, while a continuous indicator lends itself to regression and to the estimation of remaining useful life. Many practical systems maintain both views at once, using classification to raise alarms and regression to estimate how long an operator has to respond.

2.2 The Industrial Internet of Things

The Internet of Things refers to the networking of physical objects equipped with sensing, computation, and communication capabilities so that they can exchange data without human intervention [12]. When this paradigm is applied to factories, plants, and industrial assets, it is usually called the Industrial Internet of Things to emphasise the additional requirements that industrial settings impose: deterministic timing, robustness to harsh environments, long device lifetimes, and stringent expectations around reliability and security [30]. In a predictive-maintenance deployment, the IIoT is what turns an isolated machine into a continuously observable asset whose behaviour can be logged, analysed, and acted upon.

A distinguishing feature of industrial IoT data is its temporal and physical richness. Vibration signals, for example, may be sampled at tens of kilohertz to capture the high-frequency signatures of bearing defects, whereas temperature may be logged once per second. Reconciling these different time scales, aligning multiple sensor streams, and handling the inevitable gaps and noise are all part of the engineering that sits between raw sensing and useful analytics.

2.3 Prognostics, health management, and remaining useful life

Prognostics and Health Management (PHM) is the discipline concerned not only with detecting and diagnosing faults but with predicting their evolution. Its central quantity is the remaining useful life (RUL) of a component: the time, or number of operating cycles, from the present moment until the component can no longer perform its function. Reliable RUL estimation is the holy grail of predictive maintenance because it converts a vague warning into an actionable planning horizon, allowing maintenance to be scheduled at the most convenient and least costly moment [7, 10, 19].

Approaches to RUL estimation are commonly grouped into physics-based models, which encode the degradation mechanics of a component through equations; data-driven models, which learn degradation patterns from historical run-to-failure data; and hybrid schemes that combine the two. Data-driven methods have become dominant in the IoT era because they scale to fleets of machines and do not require a bespoke physical model for every asset, although they depend critically on the availability and quality of historical failure data [19, 26].

2.4 Machine learning paradigms in maintenance

The machine-learning methods used in predictive maintenance span the familiar paradigms. Supervised learning dominates, because fault classification and RUL regression both benefit from labelled examples of known conditions or known times-to-failure. Support vector machines, decision trees and their ensembles, and neural networks are the workhorses of the supervised setting [9, 14, 15, 16]. Unsupervised and semi-supervised techniques play an important role where labelled faults are scarce, using anomaly detection to flag deviations from a learned model of normal operation. More recently, survival analysis — a family of statistical methods originally developed in medicine to model time-to-event data — has been adapted to industrial life-cycle estimation, offering a principled way to handle the censored observations that arise when many machines have not yet failed [17, 18]. The remainder of this review returns repeatedly to how these paradigms have been instantiated in real systems.

2.5 Evaluating predictive-maintenance models

Because this review compares studies that report their success in different ways, it is worth setting out the metrics on which such judgements rest. For the classification task — deciding which health state a machine occupies — the confusion matrix is the fundamental object, tabulating correct and incorrect predictions for each class. From it are derived accuracy, the fraction of predictions that are correct; precision, the fraction of positive predictions that are truly positive; recall, or sensitivity, the fraction of true faults that are detected; and the F1-score, the harmonic mean of precision and recall. In a maintenance setting, recall on the fault classes often matters more than raw accuracy, because a missed fault is far more costly than a false alarm, and a headline accuracy figure can be dangerously flattering when healthy examples vastly outnumber faulty ones [8, 9].

For the estimation of a continuous quantity such as remaining useful life, regression metrics take over: the mean absolute error and the root-mean-square error express, in the units of the target itself, how far predictions fall from reality, with the latter penalising large misses more heavily. Survival models are judged differently again. The concordance index measures how well a model orders machines by risk — informally, the probability that of a random pair the one predicted to fail sooner actually does — while the Brier score and its integrated form assess how well the predicted survival probabilities are calibrated over time [5, 18, 19]. Appreciating these distinctions is essential to reading the comparative analysis without being misled by headline numbers reported under different measures.

3 Review methodology

This review was assembled by combining a focused reading of recent primary studies on IoT-based predictive maintenance with a broader base of established survey and foundational literature. Primary studies were selected to represent the principal method families in current use rather than to exhaustively enumerate every published system, since the field is now too large for a single review to catalogue completely. Preference was given to work published between 2020 and 2025 that reported a concrete implementation, described its data and evaluation protocol, and stated quantitative results, because such studies allow a meaningful comparison of methods on a common footing.

Four recent primary studies form the analytical core of the review. They were chosen because, taken together, they span the main axes along which predictive-maintenance systems differ: the learning algorithm, the nature of the sensor data, the system architecture, and the prediction target. One study applies a single classical classifier to a benchmark failure dataset; a second performs vision-based fault diagnosis on gearbox vibration signals; a third describes an end-to-end IoT architecture that integrates edge and cloud processing; and a fourth applies survival analysis to estimate component life on real industrial data. These primary studies are read against a wider body of surveys and foundational papers that supply the historical and methodological context [6, 7, 8, 10, 11, 20, 21, 22]. Where a primary study exhibited limitations in its own referencing or reporting, those limitations are noted rather than propagated, so that the comparative analysis rests on the substance of each study rather than on its self-description.

4 IoT architecture for machine health monitoring

Practical predictive-maintenance systems are almost always organised as a layered architecture in which data flow upward from the physical machine toward the analytical application, and control or

insight flows back down toward the operator. Although different authors draw the boundaries slightly differently, a five-layer view captures the structure well: a perception layer that senses the machine, a network layer that transports the data, an edge layer that performs local processing, a cloud layer that provides storage and heavyweight computation, and an application layer that delivers insight to users [4]. The following subsections consider each in turn.

4.1 Perception and sensing layer

At the base of the architecture sits the instrumentation that converts physical phenomena into electrical signals. The choice of sensor is dictated by the failure modes of interest. Accelerometers capture the vibration signatures that reveal bearing wear, imbalance, misalignment, and gear-tooth defects; thermocouples and infrared sensors track the temperature rises that accompany friction and electrical faults; current and voltage transducers expose electrical asymmetries in motors; tachometers or encoders record rotational speed; and pressure and flow sensors monitor hydraulic and pneumatic subsystems [3, 4]. Because a single fault often manifests across several of these channels, multi-sensor deployments are the norm, and the fusion of heterogeneous signals is itself an active area of study.

Vibration deserves particular mention because it is the richest and most widely used channel for rotating machinery. The high sampling rates it demands — commonly several kilohertz and sometimes tens of kilohertz — make it the most data-intensive signal in a typical deployment, and much of the feature-extraction effort described below is aimed specifically at condensing vibration waveforms into compact, informative descriptors.

4.2 Data acquisition and edge layer

Between the raw sensor and the network lies the data-acquisition device, which digitises analogue signals, timestamps them, and often performs a first pass of filtering and buffering. Low-cost microcontrollers and single-board computers — the Arduino and ESP32 families, the Raspberry Pi, and purpose-built data-acquisition modules — have made this layer inexpensive enough to justify instrumenting machines that would never have warranted a traditional condition-monitoring installation.

Increasingly, this layer does more than acquire data: it performs edge computing, running lightweight analytics close to the machine so that only summarised results or alerts need to travel upstream. Edge processing reduces the bandwidth consumed by high-rate vibration data, lowers the latency between a developing fault and its detection, and allows a system to keep functioning when connectivity to the cloud is intermittent [3, 30]. The trade-off is that edge devices are constrained in memory and compute, which limits the complexity of the models they can host and motivates research into compact, efficient learning algorithms.

4.3 Network and communication layer

Data move from edge to cloud over a communication layer whose technology is chosen to balance range, bandwidth, power, and cost. Within a facility, Wi-Fi and wired Ethernet offer high bandwidth suitable for streaming raw signals, while lightweight publish-subscribe protocols such as MQTT are popular for their low overhead and tolerance of unreliable links [3]. Where machines are dispersed across a wide area or where power is scarce, low-power wide-area technologies such as LoRa trade bandwidth for range and battery life, transmitting compact feature vectors rather than raw

waveforms. The protocol stack chosen at this layer has a direct effect on what analytics are feasible downstream, since it determines whether the cloud sees full signals or only pre-computed features.

4.4 Cloud and application layer

The cloud layer provides the durable storage and elastic computation that training and large-scale analytics require. Historical data accumulated in the cloud become the training corpus from which models learn degradation patterns, and the cloud's scalability makes it the natural home for computationally heavy tasks such as retraining models across a fleet. Sitting above it, the application layer translates model outputs into forms that people can act on: dashboards that visualise current health, alerts that fire when a threshold is crossed, and maintenance recommendations that feed into work-order systems. Lightweight web dashboards, including those built with frameworks such as Streamlit, have become a common way to present real-time health indicators and RUL estimates to operators and managers [5].

4.5 Data pre-processing and feature extraction

No matter how the layers are arranged, raw sensor data must be conditioned before a model can use them. Pre-processing typically includes filtering to remove noise and drift, normalisation to place different channels on comparable scales, segmentation of continuous streams into windows, and the handling of missing or corrupt samples. Feature extraction then reduces each window to a set of descriptors that summarise its content. Time-domain features such as root-mean-square amplitude, kurtosis, crest factor, and peak value are cheap to compute and sensitive to many mechanical faults, while frequency-domain features derived from the fast Fourier transform expose the characteristic frequencies at which bearing and gear defects announce themselves [10, 24]. The quality of these engineered features often matters more to the final result than the choice of classifier, a point to which the comparative analysis returns.

A notable recent trend is the partial replacement of hand-crafted features by representations learned automatically from the data. Deep networks can, in principle, discover informative features directly from raw or lightly processed signals, and some systems side-step conventional feature engineering entirely by transforming signals into images and applying computer-vision techniques, an approach examined under vision- and image-based fault diagnosis [11, 25].

5 Machine-learning approaches for predictive maintenance

This section surveys the principal families of machine-learning method used for fault classification and life estimation, drawing on recent studies to ground each family in a concrete implementation. The families are presented roughly in order of increasing model complexity, from classical margin-based classifiers to deep and specialised statistical models.

5.1 Support vector machines for fault classification

The support vector machine (SVM) remains one of the most reliable classifiers for predictive maintenance, especially when the amount of labelled data is modest. By finding the hyperplane that maximally separates classes in a transformed feature space, and by using kernels to handle non-linear boundaries, the SVM often generalises well without the large data appetite of deep networks [14]. A recent study exemplifies its use: applying an SVM to the widely used AI4I 2020 predictive-maintenance dataset, the authors performed a grid search over the penalty parameter and the radial-

basis-function kernel width and selected the configuration that balanced training and test performance [2]. Their best model reached about 98% accuracy on the training data and 88% on unseen test data, for an overall accuracy in the region of 90%.

The study is instructive both for what it achieves and for what it acknowledges. The gap between training and test accuracy is a reminder that a classifier tuned aggressively on one split can overstate its real-world performance, and the authors are candid that they evaluated only a single algorithm on a single subset of the data. They explicitly recommend extending the comparison to random forests, neural networks, and logistic regression — an admission that a single-classifier result, however respectable, is a starting point rather than a conclusion. This limitation recurs across much of the literature and is taken up again under research gaps and open challenges [2].

5.2 Ensemble and tree-based methods

Decision trees partition the feature space with a sequence of simple threshold tests, and while a single tree is prone to over-fitting, ensembles of trees are among the most dependable methods in the entire predictive-maintenance toolkit. The random forest builds many decorrelated trees on bootstrapped samples and aggregates their votes, yielding a classifier that is robust, resistant to over-fitting, and largely insensitive to feature scaling [15]. Gradient-boosting methods instead grow trees sequentially, each correcting the errors of its predecessors, and frequently achieve the strongest accuracy of any tree-based approach on tabular sensor data. A further practical advantage of tree ensembles is that they yield a natural measure of feature importance, which helps engineers understand which sensor channels drive a prediction — a form of interpretability that pure black-box models lack [8, 9].

In IoT-based systems, ensemble methods are frequently the default choice for the fault-classification stage. A recent architectural study of an industrial predictive-maintenance system, for example, adopts random forest alongside SVM and a recurrent network as the candidate learners for classifying machine condition from temperature, vibration, pressure, and speed data, reflecting the standing of the random forest as a strong, low-maintenance baseline against which more elaborate models must justify themselves [3].

5.3 Artificial neural networks and deep learning

Artificial neural networks approximate complex non-linear mappings by composing layers of weighted connections, and their deep variants have transformed machine health monitoring over the last decade. For the sequential telemetry that machines produce, recurrent architectures and in particular the long short-term memory (LSTM) network are attractive because they model temporal dependencies explicitly, learning how a signal's history bears on its future [16]. This makes them well suited to RUL estimation, where the trajectory of a health indicator, not merely its instantaneous value, carries the predictive information [11, 26].

The strengths of deep learning come with costs that matter in an industrial setting. Deep models require large volumes of labelled data to train well, they are computationally demanding relative to the constrained edge devices on which real-time inference must sometimes run, and their internal reasoning is opaque, which complicates the task of convincing a plant engineer to trust a maintenance recommendation. These tensions — between accuracy and data appetite, between model complexity and edge deployment, and between predictive power and interpretability — are a running theme of the literature and are revisited among the open challenges [22, 23].

5.4 Vision- and image-based fault diagnosis

A distinctive and increasingly popular strategy converts one-dimensional vibration signals into two-dimensional images and then applies image-processing and computer-vision techniques to diagnose faults. The appeal is that a mature toolkit of image operators — edge detection, pooling, and spatial interpolation — becomes available, and that spatial patterns invisible in a raw waveform can become salient once the signal is rendered as an image. A recent study of gearbox fault diagnosis demonstrates the approach on an industrial machine-learning platform [1]. Working from a dataset of five hundred samples spanning five gear conditions — a healthy gear and four fault types comprising a missing tooth, a crack, spalling, and a chipped tooth tip — sampled at twenty kilohertz, the authors transformed the vibration signals into images and applied a pipeline of edge detection, max pooling, and inverse-distance-weighted interpolation with a seven-nearest-neighbour scheme.

Trained on ninety percent of the data and tested on the remaining tenth, the resulting diagnostic achieved about 96% accuracy across the five conditions, a strong result that illustrates the promise of image-based methods for the fine-grained discrimination of fault types. It should be read, however, alongside the modest size of the dataset: five hundred samples is small by the standards of image classification, and the extent to which such a model generalises to other gearboxes, operating conditions, and fault severities is precisely the kind of question that a single controlled study cannot settle [1].

5.5 Survival analysis and data-driven RUL estimation

The methods discussed so far mainly answer the question "what is wrong with this machine?" Survival analysis addresses the complementary and arguably more valuable question "how long until this machine fails?" Originating in biomedical statistics, survival models are designed to estimate time-to-event data while correctly handling censoring — the common situation in which many units have not yet failed by the end of the observation window, so that their true failure times are unknown. This is exactly the situation faced by a maintenance analyst studying a fleet in which most machines are still running, and it is a situation that conventional regression handles awkwardly [17].

A recent study brings this toolkit to bear on component life-cycle extension, comparing the classical Cox proportional-hazards model with three machine-learning survival methods — random survival forests, gradient-boosting survival analysis, and a fast survival support vector machine — on two datasets: telemetry from a spot-welding robot in a real appliance-manufacturing plant and a widely used public predictive-maintenance dataset [5, 18]. On the real industrial data, the Cox model achieved a concordance index of about 0.97, with the fast survival SVM reaching roughly 0.87 and the two remaining methods close to 0.79, where the concordance index measures how well a model ranks machines by their risk of failure. The models were evaluated not only by concordance but by the Brier score and its integrated form, and the system was wrapped in a web dashboard so that estimates could be surfaced to operators in real time.

Beyond its numerical results, the study is valuable for a candid observation about the field. The authors report that searching the scholarly literature for survival analysis in the context of Industry 4.0 returns very little, even though the same techniques are routine in medicine, and they argue that this represents a missed opportunity for industrial prognostics [5]. That observation, corroborated by the relative scarcity of survival-based work in the broader surveys, directly motivates one of the research gaps identified below.

5.6 Hybrid, transfer, and emerging directions

Beyond the established method families, several emerging directions are reshaping how predictive-maintenance models are built and deployed. Hybrid approaches that pair physics-based degradation models with data-driven learners aim to combine the interpretability and extrapolation of the former with the flexibility of the latter, and they are especially attractive where run-to-failure data are too scarce for a purely data-driven model to learn from [7, 19]. The digital twin — a continuously updated virtual replica of a physical asset — extends this idea, providing a simulation environment in which a model can be exercised and in which what-if scenarios can be explored before any intervention is made on the real machine [25].

Two further trends respond directly to the data and privacy constraints noted elsewhere in this review. Transfer learning seeks to carry knowledge learned on one machine, or on abundant simulated data, across to a different machine for which labelled faults are scarce, mitigating the data-scarcity problem that limits so many deployments [11, 23]. Federated learning goes a step further, training a shared model across several sites without moving raw data off-premises, so that a fleet can benefit from collective experience while each operator retains control of its own sensitive telemetry — an appealing prospect wherever the security of industrial data is a first-order concern [22, 30]. These approaches are less mature than the classification and survival methods surveyed above, but they indicate the direction in which the field is moving and are likely to feature prominently in the predictive-maintenance systems of the coming years.

6 Comparative analysis of reviewed studies

Table 1 summarises the four primary studies that anchor this review, placing them side by side in terms of the learning method used, the data on which they were evaluated, their prediction target, the performance reported, and their principal limitation. Reading the studies together, rather than in isolation, brings the trade-offs of the field into focus.

Table 1. Comparative summary of the primary studies reviewed.

Study	Method / Model	Data & Sensors	Target	Reported Result	Key Limitation
Assagaf et al., 2023 [2]	Support vector machine (RBF kernel, grid-searched)	AI4I 2020 benchmark; process variables (temperature, speed, torque, tool wear)	Binary failure classification	≈90% overall (98% train / 88% test)	Single algorithm on one data subset; train-test gap; authors urge broader comparison
Horváth & Zelei, 2024 [4]	Image-based diagnosis: edge detection + max pooling + inverse-distance weighting	500 samples, 5 gear conditions; vibration at 20 kHz rendered as images	Multi-class gear fault diagnosis	≈96% accuracy (90/10 split)	Small dataset; generalisation to other gearboxes and conditions untested
Hem-anth et al., 2025 [3]	Random forest, SVM, LSTM within a 5-layer IoT architecture	Temperature, vibration, pressure, RPM; MQTT / Wi-Fi / LoRa; edge + cloud	Condition classification & alerting	Architecture and framework (system-level)	Emphasis on architecture over rigorous, reproducible evaluation
Moreira et al., 2024 [5]	Survival analysis: Cox PH, random survival forest, gradient-boosting survival, fast survival SVM	Real spot-welding robot data + public PdM dataset; Streamlit dashboard	Remaining useful life / risk ranking	C-index up to ≈0.97 (Cox) on real data	Survival methods still rare in Industry 4.0; results dataset-dependent

Several patterns emerge from the comparison. First, the highest headline accuracies belong to the most narrowly scoped studies: the 96% gearbox diagnostic is trained and tested on a single, small, carefully curated dataset, whereas the more broadly applicable failure classifier reports a more modest 88% on held-out data. This inverse relationship between scope and reported accuracy is a caution against comparing headline numbers across studies as if they were commensurable; a figure means little without the context of the data and the evaluation protocol that produced it.

Second, the studies target different questions. Classification-oriented work asks whether a machine is faulty and, if so, how; survival analysis asks how long a machine will last. Both are needed in practice, yet the two strands are pursued largely in isolation, with few systems attempting to deliver classification and life estimation from a single integrated pipeline. Third, the architecture-focused study demonstrates a complete sensing-to-cloud pipeline but places comparatively little weight on rigorous, reproducible evaluation of its models, whereas the algorithm-focused studies evaluate their models carefully but on data divorced from a live IoT deployment. The field, in other words, tends to be strong on either the systems side or the analytics side, but rarely on both at once.

7 Research gaps and open challenges

Synthesising the studies reviewed above with the broader survey literature reveals a set of recurring gaps. These are not criticisms of any single paper — each of the studies discussed makes a genuine contribution — but rather structural weaknesses in the field that define where further work is most needed.

7.1 Over-reliance on single algorithms

A large share of studies evaluate a single learning algorithm and report its performance without a systematic comparison against alternatives on the same data. The SVM study reviewed here is explicit about this limitation and calls for exactly such a comparison [2]. Where multiple algorithms are named, they are often listed as design options rather than benchmarked head-to-head under identical conditions. The consequence is that the community accumulates many isolated accuracy figures that cannot be reliably compared, making it difficult to say which method is genuinely best for a given class of problem.

7.2 Data scarcity, imbalance, and the synthetic–real gap

Predictive maintenance is haunted by a data problem: healthy operation is abundant while genuine failures are, by design, rare. This produces severe class imbalance, and many studies sidestep it by using balanced benchmark or synthetic datasets that do not reflect the skewed distributions of a real plant. Models validated only on such data may perform far worse when deployed. The small size of some fault datasets — five hundred samples in one of the studies reviewed — compounds the problem, leaving open the question of how well reported accuracies survive contact with the messier, larger, and more imbalanced data of an operational facility [1, 8].

7.3 Fragmented treatment of classification and life estimation

Fault classification and remaining-useful-life estimation are treated as largely separate problems, addressed by different communities using different methods and datasets. Yet an operator ultimately wants both answers from the same system: what is wrong, and how long do I have. The scarcity of integrated frameworks that deliver classification and prognostics together, from a common sensor

stream and in real time, is one of the clearest gaps in the literature and one of the most direct opportunities for contribution.

7.4 Under-utilization of survival analysis in Industry 4.0

Although survival analysis is tailor-made for time-to-failure problems and handles censored data naturally, it remains conspicuously rare in the industrial setting. The observation that a literature search for survival analysis in Industry 4.0 returns almost nothing, despite its ubiquity in medicine, points to a substantial untapped opportunity [5]. Bringing survival models into mainstream predictive maintenance, and combining them with the IoT data pipelines that classification systems already use, is a promising and under-explored direction.

7.5 Edge deployment, latency, and security

Moving analytics from the cloud toward the edge reduces latency and bandwidth and improves resilience, but edge devices are resource-constrained, which limits the models they can host and forces trade-offs between accuracy and footprint. At the same time, connecting industrial machinery to networks enlarges the attack surface, and the security and privacy of the data and control paths in an IIoT deployment are frequently acknowledged as concerns yet seldom treated in depth within predictive-maintenance studies [22, 30]. The synopsis-stage project that motivates this review explicitly lists data security alongside efficiency as an objective, reflecting a recognition that a predictive-maintenance system is only as trustworthy as the pipeline that carries its data.

7.6 Interpretability and operator trust

The most accurate models are often the least interpretable, and a maintenance recommendation that cannot be explained is unlikely to be trusted or acted upon by the engineers responsible for a machine. Tree-based methods offer feature-importance measures and survival models offer interpretable hazard relationships, but the deep networks that excel at raw-signal analysis remain largely opaque. Bridging the gap between predictive power and explainability, so that a system can justify its alerts in terms a human can verify, is essential for adoption and remains only partially solved [23].

7.7 Standardization, benchmarking, and generalization

Finally, the absence of shared, realistic benchmarks and standard evaluation protocols makes cross-study comparison hazardous, as the comparative analysis above illustrated. Results are reported on disparate datasets with different metrics, and models tuned and tested on one machine rarely demonstrate that they transfer to another. Progress toward common benchmarks, agreed metrics, and explicit tests of cross-machine generalisation would let the field measure its own advance more reliably [20, 21].

8 Proposed research direction

The gaps identified above point toward an integrated system rather than another isolated model, and they align closely with the direction of an ongoing project on IoT-based machine health monitoring and predictive maintenance. The central proposal is to unify real-time IoT sensing with a two-stage learning pipeline that delivers both fault classification and remaining-useful-life estimation from the same data stream, closing the fragmentation identified under fragmented treatment of classification and life estimation.

In such a design, a suite of sensors measuring vibration, temperature, motor current, rotational speed, and pressure would feed a low-cost acquisition device that performs pre-processing and time- and frequency-domain feature extraction at the edge. A first learning stage would classify the machine's condition into health states such as healthy, warning, degraded, and faulty, using an ensemble classifier chosen for its robustness and its interpretable feature-importance output. A second stage would estimate remaining useful life, and here the under-used survival-analysis toolkit is a natural candidate, allowing the system to rank machines by failure risk and to produce a planning horizon while correctly handling the censored data of a fleet that is mostly still running. Rather than committing to a single algorithm, the framework would benchmark several classifiers and several survival models against one another on the same operational data, addressing the single-algorithm limitation directly.

Attention to the practical concerns raised under research gaps and open challenges would distinguish such a system from a purely academic model. Class imbalance would be confronted explicitly through resampling or cost-sensitive learning and through validation on realistically skewed data rather than balanced benchmarks. Security would be treated as a first-class objective of the data pipeline rather than an afterthought. Results would be surfaced through a real-time dashboard so that classification states and life estimates reach operators in an actionable form, and the choice of interpretable models at the classification stage would help engineers trust and verify the system's recommendations. The value of the proposal lies less in any single novel algorithm than in the disciplined integration of sensing, dual-target learning, and deployment concerns into one coherent system — precisely the combination that the current literature addresses only in fragments.

9 Conclusion and future scope

Predictive maintenance has moved from an aspiration to a practical reality on the strength of two converging technologies: the Industrial Internet of Things, which makes continuous machine monitoring affordable, and machine learning, which turns the resulting data into estimates of machine health and remaining life. This review has traced that convergence, describing the layered architecture that carries data from sensor to application and surveying the main families of learning method through the lens of four recent studies. Those studies show that classical support vector machines, tree ensembles, deep networks, image-based diagnostics, and survival models each have a role to play, with reported performance ranging from around 88% test accuracy for a general failure classifier to 96% for a focused vision-based gearbox diagnostic and concordance indices above 0.97 for survival-based life estimation on real industrial data.

The comparison also exposed the field's structural weaknesses. Too many studies evaluate a single algorithm on balanced or synthetic data; classification and life estimation are pursued in isolation; survival analysis remains strangely rare in industry despite its suitability; and the practical matters of edge deployment, security, interpretability, and standardised benchmarking are acknowledged more often than they are solved. These gaps define a clear agenda. The most valuable future work is likely to be integrative rather than narrowly algorithmic: systems that combine real-time IoT sensing with a pipeline delivering both fault classification and remaining-useful-life estimation, that benchmark multiple models honestly on realistically imbalanced operational data, that treat security and interpretability as design goals, and that present their conclusions to operators in a form they can trust and use. Pursuing that agenda would move predictive maintenance closer to its promise of

keeping industrial machinery healthy, safe, and productive while spending maintenance effort only where and when it is genuinely needed.

References

- [1] Matzka S. Explainable artificial intelligence for predictive maintenance applications. In: *Proc. 3rd Int. Conf. Artificial Intelligence for Industries (AI4I)*, Irvine, CA, USA. 2020. p. 69–74.
- [2] Assagaf I, Lomi Ga J, Sukandi A, Abdillah AA, Arifin S. Machine predictive maintenance by using support vector machines. *Recent in Engineering Science and Technology (RIESTech)*. 2023;1(1). <https://doi.org/10.59511/riestech.v1i01.6>
- [3] Hemanth N, Mahesh Babu B, Kalyan Raj K. IoT-based predictive maintenance system for industrial machinery. *International Journal of Research Publication and Reviews (IJRPR)*. 2025;6(2):1047–1053.
- [4] Horváth K, Zelei A. Gearbox fault diagnosis using industrial machine learning techniques. *Engineering Proceedings*. 2024;79(1). <https://doi.org/10.3390/engproc2024079036>
- [5] Moreira M, Pereira E, Gonçalves G. Data-driven predictive maintenance for component life-cycle extension. In: *Proc. 21st Int. Conf. Informatics in Control, Automation and Robotics (ICINCO)*. 2024. p. 126–136. <https://doi.org/10.5220/0013014200003822>
- [6] Jardine AKS, Lin D, Banjevic D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*. 2006;20(7):1483–1510.
- [7] Lee J, Wu F, Zhao W, Ghaffari M, Liao L, Siegel D. Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. *Mechanical Systems and Signal Processing*. 2014;42(1–2):314–334.
- [8] Carvalho TP, Soares FAAMN, Vita R, Francisco RdP, Basto JP, Alcalá SGS. A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*. 2019;137.
- [9] Susto GA, Schirru A, Pampuri S, McLoone S, Beghi A. Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*. 2015;11(3):812–820.
- [10] Lei Y, Li N, Guo L, Li N, Yan T, Lin J. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*. 2018;104:799–834.
- [11] Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX. Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*. 2019;115:213–237.
- [12] Atzori L, Iera A, Morabito G. The Internet of Things: A survey. *Computer Networks*. 2010;54(15):2787–2805.
- [13] Gubbi J, Buyya R, Marusic S, Palaniswami M. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*. 2013;29(7):1645–1660.
- [14] Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995;20(3):273–297.
- [15] Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32.
- [16] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*. 1997;9(8):1735–1780.
- [17] Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*. 1972;34(2):187–202.
- [18] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *The Annals of Applied Statistics*. 2008;2(3):841–860.
- [19] Si XS, Wang W, Hu CH, Zhou DH. Remaining useful life estimation—A review on the statistical data driven approaches. *European Journal of Operational Research*. 2011;213(1):1–14.
- [20] Zonta T, da Costa CA, da Rosa Righi R, de Lima MJ, da Trindade ES, Li GP. *Predictive maintenance in the Industry 4.0: A systematic literature review*. *Computers & Industrial Engineering*. 2020;150.
- [21] Çınar ZM, Abdussalam Nuhu A, Zeeshan Q, Korhan O, Asmael M, Safaei B. *Machine learning in predictive maintenance towards sustainable smart manufacturing in Industry 4.0*. *Sustainability*. 2020;12(19).
- [22] Zhang W, Yang D, Wang H. Data-driven methods for predictive maintenance of industrial equipment: A survey. *IEEE Systems Journal*. 2019;13(3):2213–2227.
- [23] Saufi SR, Ahmad ZAB, Leong MS, Lim MH. Challenges and opportunities of deep learning models for machinery fault detection and diagnosis: A review. *IEEE Access*. 2019;7:122644–122662.

-
- [24] Randall RB, Antoni J. Rolling element bearing diagnostics—A tutorial. *Mechanical Systems and Signal Processing*. 2011;25(2):485–520.
 - [25] Wang J, Ma Y, Zhang L, Gao RX, Wu D. Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*. 2018;48:144–156.
 - [26] Nguyen KTP, Medjaher K. A new dynamic predictive maintenance framework using deep learning for failure prognostics. *Reliability Engineering & System Safety*. 2019;188:251–262.
 - [27] Ran Y, Zhou X, Lin P, Wen Y, Deng R. A survey of predictive maintenance: Systems, purposes and approaches. *arXiv:1912.07383 [Preprint]*. 2019.
 - [28] Schwabacher M, Goebel K. A survey of artificial intelligence for prognostics. In: *Proc. AAAI Fall Symp.* 2007. p. 108–115.
 - [29] Mobley RK. *An introduction to predictive maintenance*. 2nd ed. Amsterdam: Butterworth-Heinemann; 2002.
 - [30] Sisinni E, Saifullah A, Han S, Jennehag U, Gidlund M. Industrial Internet of Things: Challenges, opportunities, and directions. *IEEE Transactions on Industrial Informatics*. 2018;14(11):4724–4734.