Available Online: 26 December 2025 DOI: 10.54254/2977-3903/2025.30893

AI-empowered maintenance: a review of challenges, technologies, and future perspectives

Nan Zhang

Department of Engineering Computer Science Building (K17), University of New South Wales, Kensington, Australia

z5610465@ad.unsw.edu.au

Abstract. Due to the rapid increase in the complexity and scale of industrial systems, mainstream maintenance models can no longer meet the demands of system operation. Condition monitoring and predictive maintenance based on sensor and artificial intelligence technologies have become the dominant approaches. Meanwhile, with the rapid development of technologies such as machine learning and digital twins, the technical framework of intelligent maintenance systems is also gradually evolving. To address the shortcomings of existing reviews, this paper aims to systematically examine the development trends, applicable tasks, and existing challenges of different methods from the perspectives of maintenance objectives and maintenance technologies, hoping to provide a guide for related research and industrial development.

Keywords: predictive maintenance, anomaly detection, fault diagnosis, remaining useful life

1. Introduction

Traditional industrial maintenance relies mainly on manual inspection and passive repair, which struggle to meet the growing industrial needs and has the disadvantages of high cost and lack of predictive ability. The introduction of condition monitoring marks the shift of maintenance methods to proactive maintenance. However, early condition monitoring systems relied on simple thresholds and manually set rules, which often failed to support real-time decision-making in real complex industrial environments. With the deep integration of the Internet of Things (IOT), cloud computing and artificial intelligence, intelligent maintenance has developed rapidly [1]. Especially in the past decade, the progress of machine learning and deep learning has further transformed the maintenance model [2]. By analysing data (including sensor readings, equipment fault records, and operation logs), subtle performance degradation trends can be detected before a fault actually occurs, thus laying the foundation for anomaly detection and fault diagnosis. This progress mainly relies on three factors. First, large-scale deployment of sensors enables continuous and detailed data acquisition. Second, the application of digital twin technology enables synchronisation between physical devices and virtual models, thereby achieving real-time simulation [3]. Third, the integration of edge computing brings data processing closer to sensor nodes, reducing latency and improving response speed. The fusion of these technologies forms the basis of intelligent maintenance systems, enabling traditional industrial production to progress towards a more optimised operation and maintenance mode. Predictive maintenance technology based on artificial intelligence is developing rapidly, but most existing reviews still fail to provide a comprehensive and systematic overview. To make up for this deficiency, this paper systematically reviews the current research status, focusing on two interrelated dimensions: the core tasks of predictive maintenance (such as fault detection, condition monitoring, and remaining useful life prediction) and the methods applied to these tasks.

Contribution:

1. This paper systematically reviews the latest advancements in predictive maintenance, summarizes key methodological breakthroughs, and identifies the main technical bottlenecks hindering the practical implementation of these technologies.

2. This paper provides a structured analysis of the field from three perspectives: anomaly detection, fault diagnosis, and remaining lifetime prediction, clarifying the conceptual boundaries of these core tasks and their inherent connections within a unified methodological framework.

2. Core tasks and technological evolution

From the perspective of research objectives and technical implementation, intelligent maintenance comprises three core tasks: anomaly detection, fault diagnosis, and lifespan prediction (as shown in Figure 1). Anomaly detection first identifies anomalous data from the raw input data; fault diagnosis, based on this, determines the type and location of the anomaly, providing a basis for subsequent health status assessments; lifespan prediction utilizes the health information obtained in the first two stages to infer future degradation trends and estimate remaining service life, supporting maintenance planning. Although these three tasks are progressive in their workflow, they are independent in their research paradigms: anomaly detection typically employs unsupervised or semi-supervised learning, fault diagnosis mainly relies on supervised methods for category identification, while lifespan prediction operates through model-driven or data-driven time series and degradation modelling methods.

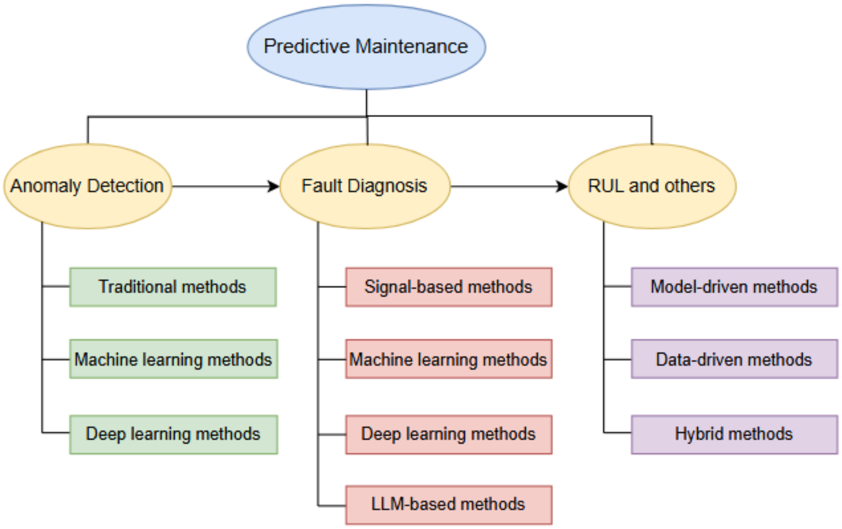

Figure 1. Structure of this review

2.1. Anomaly detection

Anomaly detection, as the first part of the workflow, aims to identify data that deviates from the normal operating mode, enabling the system to detect abnormal states early and maintain flexibility to different industrial scenarios. In actual industrial environments, fault samples are usually scarce and the cost of collection is high, so relying entirely on supervised learning is not feasible. Therefore, unsupervised or semi-supervised methods are more commonly used in engineering applications [4]. This technique is usually based on two key assumptions: first, normal data exhibits a continuous and clustered distribution in the feature space; second, the number of abnormal data is extremely small, and its distribution characteristics are significantly different from those of normal samples. Based on the above assumptions, most models focus more on learning the overall structure or boundary of normal behaviour rather than identifying specific fault modes. When a new sample deviates from the normal mode, it is considered an anomaly [5]. In other words, normal data forms a stable clustered distribution, while abnormal data is sparse and significantly different. Therefore, the model detects samples that deviate from this structure by characterising the normal distribution structure and makes anomaly judgments accordingly.

2.1.1. Technological evolution

2.1.1.1. Traditional statistics and signal processing

This established one of the early foundations of industrial anomaly detection. The basic idea is to use signal processing tools to extract statistical features with clear physical meaning, and then determine the system state based on thresholds set by experts.Time-domain: It is the simplest form of this approach. It looks directly at statistical measures of the monitored signals, such as the mean, variance, peak value, skewness, and kurtosis. Upper and lower limits are usually set by experience, and when any indicator goes beyond these limits, an alarm is raised. The method is easy to compute and works well in real time, but it is not very sensitive to early-stage faults [6] and has difficulty dealing with signals that vary widely with changing operating conditions [7].

Frequency-domain: It became possible as digital signal processing advanced, shifting the viewpoint from the time domain to the frequency domain. The Fourier Transform is the tool most commonly used for this purpose. It breaks a complicated time-domain waveform into a set of sine waves at different frequencies, which makes the spectral structure of the signal easier to see. By looking at how the amplitudes of certain frequency components change, periodic faults in rotating machines—such as imbalance or misalignment—can be identified with good reliability. However, the Fourier Transform is a global operation and does not indicate when a particular frequency component occurs in time. Because of this lack of time localisation, it has clear limitations when the signal is non-stationary or contains sudden changes [8].

Time–frequency: It was introduced into anomaly detection to address the limits of using the frequency domain alone. The Wavelet Transform uses a multi-scale way of looking at signals and can provide local information in both the time and frequency domains at the same time [9]. Because it can localize events in both domains, it is able to pull out weak, transient impulses even when they are buried in heavy noise [10], which makes it well suited for detecting non-stationary and sudden signals caused by early-stage equipment faults. Building on this, more accurate anomaly detection can be achieved by computing the energy or entropy of the wavelet coefficients and setting dynamic thresholds.

2.1.1.2. Traditional machine learning

Compared with statistical methods that rely on a single threshold, traditional machine learning approaches introduce a multi-dimensional feature space and employ geometric relationships among samples to establish more refined decision mechanisms. According to their underlying assumptions, these approaches can be grouped into several categories.Distance- and density-based Methods are built on the idea that normal samples tend to appear in regions of high density, whereas anomalous samples usually lie in sparse areas or maintain large distances from their neighbours. K-Nearest Neighbours (KNN) is the most straightforward distance-based approach. It uses the average distance between a test sample and its K nearest neighbours as its anomaly score. Gao [11] refined KNN technique incorporating a soft incremental three-way decision strategy was proposed for anomaly detection in network intrusion scenarios. However, in high-dimensional spaces, KNN faces the difficulty that a single global distance threshold cannot adapt well to data with uneven density distributions. To address this issue, Lu [12] introduced an algorithm called RkCNN. This method constructs multiple kCNN classifiers by randomly selecting a subset of features from high-dimensional data, and then combines them using a weighted scheme based on separability. This design helps reduce the performance loss of traditional KNN in high-dimensional settings due to noisy features and reduced effectiveness of distance metrics.

Clustering-based Mechanisms: These methods rely on the assumption that normal data forms clear cluster structures in the feature space, while outliers either fall outside the main clusters or appear only as very small micro-clusters. K-Means and its Variants are among the most commonly used benchmark algorithms. By applying iterative optimisation, the data are divided into K clusters, and during testing, the distance from a sample to the nearest cluster centroid is used as a measure of its anomaly level. Arin [13] integrated K-Means with digital twin techniques to support anomaly detection and clustering in the context of industrial equipment performance enhancement. Although K-Means offers high computational efficiency and is suitable for large data sets, it relies on strong assumptions about cluster geometry, typically assuming spherical and convex structures. As a result, it has difficulty modelling the non-convex or manifold-like distributions that frequently appear in industrial data. In addition, outliers may shift the cluster centroids during training, which can reduce the accuracy of anomaly detection [14].

Boundary- and isolation-based Methods It does not estimate data densities directly. Instead, they aim to construct a decision boundary that separates normal instances from anomalous ones in an explicit manner.

One-Class Support Vector Machine (OCSVM): It applies a kernel function to project data into a high-dimensional space and searches for a minimum-volume hypersphere, or an optimal separating hyperplane, that encloses the normal samples. Any point falling outside this boundary is regarded as an anomaly. Li et al. [15] applied OCSVM to the detection of post-earthquake building damage in high-resolution remote sensing imagery, achieving one-class anomaly detection requiring only normal training samples by fusing spectral and spatial features. Additionally, Bountzis [16] also integrated autoencoders with OCSVM, where the decision scores of OCSVM were used to strengthen the autoencoder's ability to detect deviations from normal behaviour. They further introduced a heuristic approach for tuning the OCSVM parameters based solely on one-class data, and showed that this method performs well against previously unseen attacks in network intrusion detection.

Isolation Forest: This method identifies outliers by gradually "isolating" samples through random binary trees. It has linear time complexity and is suitable for high-dimensional data. Wang et al. [17] proposed an improved Isolation Forest combined with binary particle swarm optimisation for detecting defects and hidden dangers in power transmission lines, which significantly improved Area Under the Curve (AUC) and anti-interference ability. Xiang et al. [18] proposed an anomaly detection method based on Isolation Forest under the federated learning framework, which effectively alleviated the contradiction between data privacy and collaborative detection in edge IoT scenarios. However, Monemizadeh et al. [19] identified the existence of "ghost clusters" in traditional Isolation Forest and proposed a method involving the random rotation of data to eliminate this bias, thereby improving anomaly detection accuracy.

2.1.1.3. Deep learning

With the rapid increase in the dimensionality of industrial big data, the shortcomings of traditional methods that depend on manual feature extraction have become more evident. Deep learning, with its strong ability for end-to-end automatic feature learning, has gradually become a leading approach for anomaly detection. According to how the models capture the underlying structure of the data, these approaches can be divided into three main paradigms.Reconstruction-based Paradigm: This is the most mature framework currently used in deep anomaly detection. It is built on the assumption that a model trained solely on normal data learns to compress and reconstruct normal patterns, and thus produces noticeably larger reconstruction errors when encountering anomalous patterns.

Autoencoders and their Variants: Zhou [20] applied unsupervised autoencoders to fuse features from multi-modal data for aircraft sensor prediction. Denoising Autoencoders (DAE) are also used, where noise is added to the inputs during training so that the model learns representations that remain stable under noise, which helps prevent it from learning a trivial identity mapping. Qiu [21] constructed a deep learning framework focused on DAE, which effectively reduced downtime and operational costs in practical applications. Furthermore, Variational Autoencoders (VAE) introduce probability distribution constraints in the latent space, enabling the model to learn the underlying manifold structure of the data rather than merely direct spatiotemporal point-to-point mappings. Chen et al. [22] introduced a hybrid architecture that combines Bayesian Long Short-Term Memory (LSTM) with deep generative models for multi-level anomaly detection and fault prediction in diesel engine lubrication systems. Liu [23] developed a hybrid model that integrates VAE with LSTM for sewage treatment systems, enabling the model to capture both the latent data distribution and temporal dependencies and thereby achieve accurate anomaly detection. To address the issue that conventional autoencoders may still reconstruct anomalous samples relatively well, recent studies have proposed memory-augmented autoencoders, in which the model is constrained during training to reconstruct inputs only from prototype features of normal samples stored in a memory bank. This design substantially strengthens the model's ability to distinguish anomalous data. Chang et al. [24] incorporated a memory-augmented module into a lightweight transformer architecture and applied it to video anomaly detection in industrial vision or surveillance environments, achieving efficient and precise anomaly identification.

Prediction-based Paradigm: Methods in this category use prediction error as an anomaly indicator by learning the temporal dynamics of normal data. Early work mainly relied on LSTM and GRU architectures, while in recent years, transformers have become widely used due to their parallel computation capability and their strength in modelling long-range dependencies. Liu et al. [25] introduced a transformer model that incorporates global attention and reconstruction-trend analysis for multivariate time-series anomaly detection. Zhang et al. [26] combined time-series decomposition with a patch-based transformer to model normal patterns in complex time-series data and detect anomalies. Owing to the self-attention mechanism, the transformer can capture long-range temporal correlations across the sequence and shows strong performance in trend prediction under complex operating conditions.

Generative model-based Paradigm: Approaches in this category detect anomalies by modelling the distribution of normal data. GANs learn the boundary of this distribution through the adversarial interaction between a generator and a discriminator, whereas diffusion models—owing to their stepwise denoising and precise distribution modelling—have shown strong potential in image and signal anomaly detection. Riaz et al. [27] employed an industrial anomaly detection framework that incorporates a Wasserstein GAN to identify minority classes within highly imbalanced IIoT data. Beizaee [28] enhanced the diffusion model by treating anomalies as noise in the latent space and introducing a selective region transformation mechanism, making anomalous areas easier to distinguish and improving performance in multi-class image anomaly detection tasks.

2.1.2. Challenges

Significant breakthroughs have been achieved in anomaly detection technology, but its large-scale application in real-world industrial environments still faces several key challenges that require further overcoming.

Insufficient Generalisation Ability under Complex Conditions: Industrial environments are dynamic; fluctuations in factors such as load, speed, and temperature can cause data distribution shifts. Traditional methods, relying primarily on fixed thresholds and static features, often perform poorly in such situations. While deep models can extract features, their performance depends on the data distribution during training and testing. When encountering unknown operating conditions, they may misclassify normal, unknown conditions as anomalies.

Severe Sample Imbalance: Equipment operates normally most of the time, resulting in scarce and incomplete fault samples. Unsupervised methods lack prior fault knowledge, making it difficult to identify normal but rare behaviours. Start-up and shutdown processes or slight load changes are often mistaken for faults.

2.2. Fault diagnosis

Fault diagnosis seeks to recognise known fault patterns and is generally carried out using supervised learning, in which a model learns the mapping from input signals to fault categories based on a large amount of labelled data. When labels are sufficient, such approaches can achieve high accuracy. To improve adaptability across different operating conditions and equipment types, transfer learning has been widely employed; for example, Li [29] provides a systematic overview of the theories, applications, and challenges of deep transfer learning in industrial fault diagnosis. Few-shot learning addresses situations where labelled samples are extremely limited. Liu [30] used Fourier features together with temporal convolutional networks to perform few-shot bearing fault identification. More recently, self-supervised learning has become an important research direction. Song [31] proposed pretraining models using contrastive learning or reconstruction tasks, followed by fine-tuning with a small amount of labelled data, which improves diagnostic performance under sparse annotation and varying operating conditions.

2.2.1. Technological evolution path

2.2.1.1. Signal processing and mechanisms

Before artificial intelligence became widely used, fault diagnosis depended mainly on physics-based mechanistic models and signal processing techniques. These approaches applied mathematical transformations to convert time-domain signals into the frequency or time–frequency domains in order to extract and match specific fault features. Spectral analysis is the most fundamental technique [31] and is used to diagnose imbalance or misalignment in rotating machinery by examining amplitude changes in the fundamental frequency and its harmonics. For modulated impulse signals produced by components such as bearings, envelope analysis [32] and cepstrum analysis are widely employed because they can demodulate fault-related frequencies that are hidden within resonance bands.

2.2.1.2. Classical machine learning methods

During the development of fault diagnosis, classical machine learning methods played a pivotal role in the transition phase from signal analysis to intelligent identification. The core framework of this phase consists of two components: first, relying on expert knowledge to manually extract features from raw signals (e.g., time-domain statistics, frequency-domain energy, or wavelet coefficients); and second, feeding these engineered feature vectors into a supervised classifier for training. Support Vector Machine (SVM): SVM achieves classification by finding a hyperplane that maximizes the inter-class margin. It can flexibly handle nonlinear feature relationships through kernel function mechanism. This method is particularly suitable for scenarios with limited sample size but high feature discrimination [33].

Random Forests: As an ensemble learning method, it completes classification by constructing multiple decision trees and using a voting mechanism. It has strong noise resistance and insensitivity to high-dimensional features, and has high robustness under complex working conditions or multi-source sensor data. In addition, it can evaluate the importance of features and provide a basis for feature selection [34].

In addition, algorithms such as k-nearest neighbours, Naive Bayes and Extreme Gradient Boosting (XGBoost) have also been widely used. In general, these traditional methods have laid the foundation for data-driven fault diagnosis; however, their fundamental limitation is that the diagnostic effect is highly dependent on the quality of extracted features, and the over-reliance on expert experience [35] and the inability to accurately represent complex systems [36] are areas that need improvement.

2.2.1.3. Deep learning methods

Supervised learning-based deep learning methods eliminate the need for manual feature extraction, as neural networks can directly learn useful features from the raw signal. This directly improves the ability to model complex systems.

Convolutional Neural Networks (CNNs): It uses convolution and pooling operations to extract local spatial patterns from signals or from time–frequency representations such as spectrograms. By learning these patterns directly from raw vibration data, CNNs have led to marked improvements in the accuracy of rolling bearing fault diagnosis [37].

Recurrent Neural Networks: It aims to model how the features of a time series evolve over time. Such models are well suited to long-duration signals that describe equipment operating conditions, including measurements such as temperature, rotational speed, and energy consumption. Iqbal [38] introduced an LSTM-based variant that maintains good diagnostic performance even when only a small number of labelled samples are available, and is particularly designed for data collected under varying operating conditions.

In addition, many studies have explored combining different methods to build hybrid architectures. Fu [39] designed a parallel CNN–LSTM network in which the CNN extracts spatial features from time–frequency representations, while the LSTM models temporal correlations in one-dimensional signals. This combination improves both the accuracy and the robustness of rolling-bearing fault classification. Other work [40] further improves diagnostic performance under non-stationary vibration signals by incorporating attention mechanisms to highlight important segments of the signal or by using residual connections to construct deeper and more stable networks [41].

2.2.1.4. LLM-based fault diagnosis

Fault diagnosis based on Large Language Models (LLM) represents a new approach for the evolution of data-driven methods [42]. Its core idea lies in leveraging the powerful language understanding and reasoning capabilities of the model. LLM is used for pre-training and learning prior knowledge in relevant domains to convert measurement data into textual descriptions during the diagnostic process, and these descriptions are then correlated with fault logs to ultimately generate diagnostic results that are easier for humans to understand.

Currently, LLM-based diagnostic research mainly focuses on two directions: knowledge assistance and semantic enhancement. On the one hand, LLM is used to build intelligent maintenance assistants, enabling technicians to query fault causes or obtain maintenance steps through natural language. Liu et al. [43] proposed a maintenance decision framework based on Retrieval-Enhanced Generation (RAG), in which LLM improves the accuracy of question answering and can more effectively extract relevant maintenance strategies from technical manuals. On the other hand, LLM can also be integrated with traditional deep models to achieve knowledge-enhanced diagnosis. In such methods, the output of CNNs can be associated with semantic knowledge bases, thereby improving the clarity and reliability of diagnostic results.

In the future, LLM hold greater application potential in multimodal data fusion, industrial knowledge graph construction [44] , and human-machine collaborative maintenance.

2.2.2. Challenges

While significant breakthroughs have been achieved in fault diagnosis technologies, numerous challenges remain in real-world industrial scenarios.

Difficulty in Distinguishing Similar Faults: Many fault types (e.g., inner ring spalling versus minor outer ring cracks) produce signals with minimal differences. These differences become even more difficult to detect in noisy and rapidly changing environments. Models relying solely on automatic feature extraction often struggle to effectively differentiate such faults.

Limited Ability to Identify Unknown Faults: New fault types constantly emerge in real-world applications. However, most models are trained on a fixed set of known categories, and due to a lack of explicit judgment mechanisms, they often categorize unknown faults into these existing categories. Effective diagnostic systems need to update their judgment criteria for unknown fault patterns based on constantly changing operating conditions and data.

Limitations of LLM in Industrial Diagnostics: Despite the powerful reasoning capabilities of large language models, their application in industrial diagnostics faces numerous limitations due to their requirement for large amounts of industrial data, their susceptibility to AI illusions, and the stringent data security requirements of industry. Therefore, LLM as a standalone diagnostic system still faces significant challenges.

2.3. Remaining Useful Life prediction

The purpose of Remaining Useful Life (RUL) prediction is to estimate the remaining time before a component or system fails, based on its current degradation state. This information is important and fundamental for the planning and allocation of maintenance resources.

RUL prediction methods are generally classified into three categories:

Model-driven Methods: These methods build degradation models based on physical formulas. They offer high interpretability, but due to the complexity of some systems' mechanisms, the physical models are difficult to describe explicitly analytically, thus presenting a significant limitation.

Data-driven Methods: These methods learn degradation patterns directly from historical observation data through statistical learning or deep learning. They are well-suited for high-dimensional and multimodal data, but heavily rely on sufficient labelled samples and typically have limited interpretability.

Hybrid Methods: These methods combine elements of model-driven and data-driven frameworks—for example, embedding physical constraints into data-driven models or using neural networks to correct errors in mechanistic models. They tend to perform well under complex operating conditions and with limited data.

2.3.1. Technological evolution

2.3.1.1. Model-driven methods

The development of model-driven approaches moved from analytical formulas to simulation. Early research often relied on analytical degradation models, such as using Paris's law to describe fatigue crack propagation. These formulas provide deterministic mathematical expressions that reflect the underlying physical mechanisms of degradation. Recently, Jia [45] introduced a modified version of the deterministic Paris model and applied it to fatigue failure analysis in concrete structures. As ideas from control theory entered the field, state-space modelling and filtering techniques became widely used. By defining state-transition equations and applying algorithms such as Kalman or particle filtering, these methods enable dynamic tracking of equipment health and probabilistic forecasting of its future degradation path.

2.3.1.2. Data-driven methods

Data-driven approaches have progressed from shallow statistical techniques to deep representation learning. Early work mainly used traditional machine learning models such as Support Vector Machines (SVMs) and Random Forests, which depended on handcrafted features for regression-based RUL estimation. With the rise of deep learning, architectures including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and transformers made it possible to learn multi-scale features directly from raw signals, reducing the need for manual feature design. Xu [46] introduced a transformer-based hybrid model that achieved strong performance in aero-engine RUL prediction, obtaining very low Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). Zhao [47] proposed an interpretable Kernel Function Convolutional-Former (KFC-Former), which combines phase-space reconstruction with a Convolutional-Transformer structure and demonstrated high accuracy in RUL prediction tasks. More recently, research has begun to address the challenges posed by data silos and distribution differences. Reinforcement learning and transfer learning have attracted increasing interest, with the aim of improving generalisation across varying operating conditions, different equipment types, and few-shot scenarios through domain adaptation or meta-learning strategies.

2.3.1.3. Hybrid methods

The development of hybrid approaches stems from the attempt to find an effective balance between physical modelling and data-driven learning. Early research mainly introduced physical constraints into the learning process, for example, using principles such as energy conservation or monotonic degradation limits for regularisation to guide neural networks toward the correct degradation pattern [48]. With the deepening of research, Physical Information Neural Networks (PINNs) have become an important research direction. These models directly embed physical partial differential equations into the loss function or network structure to ensure that the prediction results are consistent with the underlying physical laws [49]. This deeper integration improves the interpretability of the model and enhances its robustness, still be better than other models even with limited data or noise contamination. More recently, advances in simulation have brought renewed attention to digital-twin-based physical modelling. By using multi-physics simulation platforms to build high-fidelity virtual replicas of equipment [50], this approach allows for detailed simulation and prediction of system behaviour across its entire lifecycle.

2.3.2. Diversified expansion of prediction objectives

While Remaining Useful Life (RUL) prediction is a central indicator in predictive maintenance, a single time-point estimate is often not sufficient for complex industrial decision-making. As a result, practical prediction tasks have expanded to include several complementary objectives, such as assessing health states and estimating failure risk.

Health State Measurement: In the presence of heterogeneous data from multiple monitoring sources, establishing a direct relationship between high-dimensional signals and service life can be difficult. For this reason, constructing a virtual measure that reflects the overall degradation condition—commonly referred to as a Health Index—has become an important step linking raw data with life prediction models. Perry et al. [51] proposed a two-stage framework for guided-wave monitoring of aerospace composite structures, using an unsupervised Deep Time-Contrastive Variational Autoencoder (DTC-VAE) model to derive health indicators from time–frequency features, which led to clear performance gains.

Equipment Failure Risk: Reliability assessment aims to estimate the probability of failure occurring in a future time window. Such tasks often use Bayesian inference to generate reliability curves or confidence intervals [52]. In some fields such as aerospace, which focus on safety, using failure probabilities over a specific time period is more accurate than using a single estimate [53].

2.3.3. Challenges

Because RUL prediction tasks require time-series data to predict the future state of equipment, it faces more significant technical challenges than diagnostic tasks.

Lack of Complete Lifecycle Data: Accurate predictions rely on datasets covering all stages of failure. But in industrial scenarios, the frequency of actual failures is low due to regular maintenance, replacement, or repairs. Therefore, models have limited understanding of post-failure patterns, limiting the accuracy of their predictions.

High Uncertainty of Future Outcomes: RUL tasks require predicting future stochastic behaviour. Variations in operating conditions, noise, and model bias all affect the predictions. Models often fail to reflect actual risk levels. Most deep learning models also lack reliable methods for quantifying prediction uncertainty, making it difficult to provide credible confidence intervals, thus reducing the practical value of some strategies.

Recovery Phenomena During Degradation: Equipment in industrial scenarios does not always degrade. Certain maintenance measures or specific conditions may temporarily lower certain indicators, producing misleading recovery signals. These fluctuations affect model evaluations, leading to an overestimation of the remaining RUL and increasing safety risks.

3. Conclusion

This paper reviews the development of intelligent maintenance in three core tasks, their foundational technologies, and main challenges currently faced. Future research can further deepen this review, constructing a theoretical framework from a more systematic perspective to promote the comprehensive organisation and development of this field.

References

[1] Zhang, L., Lin, J., Liu, B., Zhang, Z., Yan, X., & Wei, M. (2019). A review on deep learning applications in prognostics and health management. *IEEE Access, 7*, 162415–162438. https: //doi.org/10.1109/ACCESS.2019.2950985

[2] Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing, 138*, 106587. https: //doi.org/10.1016/j.ymssp.2019.106587

[3] Yu, J., & Tang, D. (2022). Digital twin-driven prognostics and health management. In *Digital twin driven service* (pp. 205–250). Elsevier. https: //doi.org/10.1016/B978-0-323-91300-3.00005-X

[4] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41* (3), Article 15. https: //doi.org/10.1145/1541880.1541882

[5] Chalapathy, R., & Chawla, S. (2019). *Deep learning for anomaly detection: A survey* (arXiv: 1901.03407). arXiv. https: //doi.org/10.48550/arXiv.1901.03407

[6] Cattarius, J., & Inman, D. J. (1997). Time domain analysis for damage detection in smart structures. *Mechanical Systems and Signal Processing, 11*(3), 409–423. https: //doi.org/10.1006/mssp.1996.0086

[7] Wang, F., Lin, W., Liu, Z., Wu, S., & Qiu, X. (2017). Pipeline leak detection by using time-domain statistical features. *IEEE Sensors Journal, 17*(19), 6431–6442. https: //doi.org/10.1109/JSEN.2017.274022

[8] Chang, Y., Jiang, F., Zhu, Z., & Li, W. (2017). Fault diagnosis of rotating machinery based on time–frequency decomposition and envelope spectrum analysis. *Journal of Vibroengineering, 19*(2), 943–954. https: //doi.org/10.21595/jve.2017.17232

[9] Bendjama, H., Bouhouche, S., & Boucherit, M. S. (2012). Application of wavelet transform for fault diagnosis in rotating machinery. *International Journal of Machine Learning and Computing, 2*(1), 82–87. https: //doi.org/10.7763/IJMLC.2012.V2.93

[10] Fu, S., Wu, Y., Wang, R., & Mao, M. (2023). A bearing fault diagnosis method based on wavelet denoising and machine learning. *Applied Sciences, 13*(10), 5936. https: //doi.org/10.3390/app13105936

[11] Bao, H., & Gao, J. (2025). Network intrusion detection based on improved KNN algorithm. *Scientific Reports, 15*(1), Article 29842. https: //doi.org/10.1038/s41598-025-92860-4

[12] Lu, J., & Gweon, H. (2025). Random k conditional nearest neighbor for high-dimensional data. *PeerJ Computer Science, 11*, e2497. https: //doi.org/10.7717/peerj-cs.2497

[13] Warin, T., d'Anglade, P.-M., & De Marcellis-Warin, N. (2025). Optimising industrial efficiency: Integrating K-means clustering and data science for sustainable manufacturing and waste reduction. *International Journal of Sustainable Engineering, 18*(1), 2527300. https: //doi.org/10.1080/19397038.2025.2527300

[14] Wani, A. A. (2024). Comprehensive analysis of clustering algorithms: Exploring limitations and innovative solutions. *PeerJ Computer Science, 10*, e2286. https: //doi.org/10.7717/peerj-cs.2286

[15] Li, P., Xu, H., & Guo, J. (2010). Urban building damage detection from very high-resolution imagery using OCSVM and spatial features. *International Journal of Remote Sensing, 31*(13), 3393–3409. https: //doi.org/10.1080/01431161003727705

[16] Bountzis, P., Kavallieros, D., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2025). A deep one-class classifier for network anomaly detection using autoencoders and one-class support vector machines. *Frontiers in Computer Science, 7*, 1646679. https: //doi.org/10.3389/fcomp.2025.1646679

[17] Wang, W., & Wang, G. (2025). A study of improved isolation forest algorithm for data management of transmission line defects and hazards. *Energy Informatics, 8*(1), Article 130. https: //doi.org/10.1186/s42162-025-00585-7

[18] Xiang, H. (2024). Federated learning-based anomaly detection with isolation forest in the IoT-edge continuum. *ACM Transactions on Multimedia Computing, Communications, and Applications, 20*(5), Article 173. https: //doi.org/10.1145/3702995

[19] Monemizadeh, V., & Kiani, K. (2025). Detecting anomalies using rotated isolation forest. *Data Mining and Knowledge Discovery, 39*(3), 24. https: //doi.org/10.1007/s10618-025-01096-5

[20] Zhou, T., Zhang, G., & Cai, Y. (2025). Unsupervised autoencoders combined with multi-model machine learning fusion for improving the applicability of aircraft sensor and engine performance prediction. *Optimization and Applications of Machine Learning, 5*(1), 83–95. https: //doi.org/10.71070/oaml.v5i1.83

[21] Qiu, S. (2025). *Optimizing predictive maintenance in intelligent manufacturing: An integrated FNO-DAE-GNN-PPO MDP framework* (arXiv: 2511.05594). arXiv. https: //arxiv.org/abs/2511.05594

[22] Chen, Y., Li, H., Li, D., Yang, K., & Zhou, J. (2026). Diesel engine lubricating oil fault prognosis: A hybrid Bayesian LSTM and deep generative model architecture for multilayer anomaly detection. *Tribology International, 215*, 111434. https: //doi.org/10.1016/j.triboint.2025.111434

[23] Liu, X., Gong, Z., & Zhang, X. (2025). Research on anomaly detection in wastewater treatment systems based on a VAE-LSTM fusion model. *Water, 17*(19), 2842. https: //doi.org/10.3390/w17192842

[24] Chang, J., Zhen, P., Yan, X., Yang, Y., Gao, Z., & Chen, H. (2025). MemATr: An efficient and lightweight memory-augmented transformer for video anomaly detection. *ACM Transactions on Embedded Computing Systems, 24*(3), 1–26. https: //doi.org/10.1145/3719203

[25] Liu, X., Li, X., Li, Y., Tang, F., & Zhao, M. (2025). RTdetector: Deep transformer networks for time series anomaly detection based on reconstruction trend. *Journal of Systems Engineering and Electronics, 36*(4), 1201–1215. https: //doi.org/10.23919/JSEE.2025.000112

[26] Zhang, W., & Luo, C. (2025). Decomposition-based multi-scale transformer framework for time series anomaly detection. *Neural Networks, 187*, 107399. https: //doi.org/10.1016/j.neunet.2025.107399

[27] Riaz, R., Han, G., Shaukat, K., Khan, N. U., Zhu, H., & Wang, L. (2025). A novel ensemble Wasserstein GAN framework for effective anomaly detection in industrial Internet of Things environments. *Scientific Reports, 15*(1), Article 26786. https: //doi.org/10.1038/s41598-025-07533-1

[28] Beizaee, F., Lodygensky, G. A., Desrosiers, C., & Dolz, J. (2025). Correcting deviations from normality: A reformulated diffusion model for multi-class unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1578–1587). IEEE. https: //doi.org/10.1109/CVPR52734.2025.01778

[29] Li, W. (2022). A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges. *Mechanical Systems and Signal Processing, 167*, 108487. https: //doi.org/10.1016/j.ymssp.2021.108487

[30] Liu, Y., Du, Z., Zheng, H., Zhang, Q., Chen, C., & Wu, N. (2025). A novel temporal classification prototype network for few-shot bearing fault detection. *Scientific Reports, 15*(1), Article 14321. https: //doi.org/10.1038/s41598-025-98963-4

[31] Ciabattoni, L., Ferracuti, F., Freddi, A., & Monteriu, A. (2018). Statistical spectral analysis for fault diagnosis of rotating machines. *IEEE Transactions on Industrial Electronics, 65*(5), 4301–4310. https: //doi.org/10.1109/TIE.2017.2762623

[32] Kang, M., Kim, J., Wills, L. M., & Kim, J.-M. (2015). Time-varying and multiresolution envelope analysis and discriminative feature analysis for bearing fault diagnosis. *IEEE Transactions on Industrial Electronics, 62*(12), 7749–7761. https: //doi.org/10.1109/TIE.2015.2460242

[33] Tun, W., Wong, J. K.-W., & Ling, S.-H. (2021). Hybrid random forest and support vector machine modeling for HVAC fault detection and diagnosis. *Sensors, 21*(24), 8163. https: //doi.org/10.3390/s21248163

[34] D, V., Palit, S., Mehta, A., Saranya, G., Joseph, D., & Pathak, A. (2025). Machine fault diagnosis using random forest with recursive feature elimination and cross validation. *Journal of Machine Computing, 5*(1), 1700–1711. https: //doi.org/10.53759/7669/jmc202505134

[35] Zhang, S., Zhang, S., Wang, B., & Habetler, T. G. (2020). Deep learning algorithms for bearing fault diagnostics—A comprehensive review. *IEEE Access, 8*, 29857–29881. https: //doi.org/10.1109/ACCESS.2020.2972859

[36] Saravanan, B., D, P. K. M., & Vengateson, A. (2025). *Benchmarking traditional machine learning and deep learning models for fault detection in power transformers* (arXiv: 2505.06295). arXiv. https: //doi.org/10.48550/arXiv.2505.06295

[37] Wang, Y., Li, D., Li, L., Sun, R., & Wang, S. (2024). A novel deep learning framework for rolling bearing fault diagnosis enhancement using VAE-augmented CNN model. *Heliyon, 10*(15), e35407. https: //doi.org/10.1016/j.heliyon.2024.e35407

[38] Iqbal, M., Lee, C. K. M., Keung, K. L., & Zhao, Z. (2024). Intelligent fault diagnosis across varying working conditions using triplex transfer LSTM for enhanced generalization. *Mathematics, 12*(23), 3698. https: //doi.org/10.3390/math12233698

[39] Fu, G., Wei, Q., & Yang, Y. (2024). Bearing fault diagnosis with parallel CNN and LSTM. *Mathematical Biosciences and Engineering, 21*(2), 2385–2406. https: //doi.org/10.3934/mbe.2024105

[40] Siddique, M. F., Saleem, F., Umar, M., Kim, C. H., & Kim, J.-M. (2025). A hybrid deep learning approach for bearing fault diagnosis using continuous wavelet transform and attention-enhanced spatiotemporal feature extraction. *Sensors, 25*(9), 2712. https: //doi.org/10.3390/s25092712

[41] Wang, Y., Sun, K., Li, Y., & Liang, H. (2025). Adaptive hybrid attention mechanism deep residual threshold networks for bearing fault diagnosis under noisy environments. *Electronic Research Archive, 33*(9), 5301–5322. https: //doi.org/10.3934/era.2025237

[42] Zhang, Q., Xu, C., Li, J., Sun, Y., Bao, J., & Zhang, D. (2025). LLM-TSFD: An industrial time series human-in-the-loop fault diagnosis method based on a large language model. *Expert Systems with Applications, 264*, 125861. https: //doi.org/10.1016/j.eswa.2024.125861

[43] Liu, R. (2025). Knowledge enhanced industrial question-answering using large language models. *Engineering, 36*, 126–138. https: //doi.org/10.1016/j.eng.2025.07.035

[44] Ma, Y., Zheng, S., Yang, Z., Pan, H., & Hong, J. (2025). A knowledge-graph enhanced large language model-based fault diagnostic reasoning and maintenance decision support pipeline towards Industry 5.0. *International Journal of Production Research*. Advance online publication. https: //doi.org/10.1080/00207543.2025.2472298

[45] Jia, M., Wu, Z., Jiang, X., Yu, R. C., Zhang, X., & Wang, Y. (2024). Modified Paris law for mode I fatigue fracture of concrete based on crack propagation resistance. *Theoretical and Applied Fracture Mechanics, 131*, 104383. https: //doi.org/10.1016/j.tafmec.2024.104383

[46] Xu, K., Guo, Y., & Zhou, Q. (2025). Research on the remaining useful life prediction algorithm for aero-engines based on Transformer–KAN–BiLSTM. *Aerospace, 12*(11), 998. https: //doi.org/10.3390/aerospace12110998

[47] Zhao, Q., Zhang, X., Xie, J., Liang, S., & Mbeka, E. (2026). KFC-Former: An interpretable kernel function convolutional former with phase space reconstruction for remaining useful life prediction of mechanical equipment. *Expert Systems with Applications, 302*, 130472. https: //doi.org/10.1016/j.eswa.2025.130472

[48] Lu, W., Wang, Y., Zhang, M., & Gu, J. (2024). Physics-guided neural network: Remaining useful life prediction of rolling bearings using long short-term memory network through dynamic weighting of degradation process. *Engineering Applications of Artificial Intelligence, 127*, 107350. https: //doi.org/10.1016/j.engappai.2023.107350

[49] Kapoor, T. (2025). Neural differential equation-based two-stage approach for generalization of beam dynamics. *IEEE Transactions on Industrial Informatics, 21*(3), 2481–2490. https: //doi.org/10.1109/TII.2024.3507213

[50] Yang, H., Feng, C., Jiang, G., & Mei, X. (2024). Digital twin-enabled health prognostics for smart manufacturing systems under uncertain operating conditions. *IEEE Transactions on Industrial Informatics, 20*(12), 14072–14082. https: //doi.org/10.1109/TII.2024.3441633

[51] Perry, J. J., Ortiz, P. G.-C., Konstantinou, G., Vergouwen, C., Kumaran, E. S., & Moradi, M. (2025). *Semi-supervised and unsupervised learning for health indicator extraction from guided waves in aerospace composite structures* (arXiv: 2510.24614). arXiv. https: //doi.org/10.48550/arXiv.2510.24614

[52] Xie, S. (2024). Incremental contrast hybrid model for online remaining useful life prediction with uncertainty quantification in machines. *IEEE Transactions on Industrial Informatics, 20*(12), 14308–14320. https: //doi.org/10.1109/TII.2024.3450003

[53] Wang, C., Ning, G., Deng, Q., Liu, R., & Luo, Q. (2025). Joint optimisation of cooperative maintenance and inventory control for multiple k-out-of-n: F systems considering component interchange and shared spare parts. *Reliability Engineering & System Safety, 262*, 111181. https: //doi.org/10.1016/j.ress.2025.111181