



Journal of Global IT Dynamics

Volume 1, Issue 1 (2026)



Evolution of AI and Machine Learning Models for Predictive Maintenance in Industrial IoT Systems: A Comprehensive Review

Haslinda Musa*

Faculty of Technology Management and Technopreneurship, Universiti Teknikal Malaysia Melaka, Durian Tunggal, Malaysia

Article Info

Abstract

Predictive maintenance (PdM) is now a hot paradigm in contemporary industrial practices, allowing organizations to transition between reactive and preventive maintenance practices to a data-driven, predictive approach. The emergence of the Industrial Internet of Things (IIoT) technologies has significantly increased the amount, speed, and breadth of sensor data offered by industrial assets, which opens new possibilities to apply artificial intelligence (AI) and machine learning (ML). This review of the literature is a systematic survey of the evolutionary path of AI and ML models used to forecast predictive maintenance in IIoT settings, which reviews 25 peer-reviewed articles published in 2015–2024. The review follows the development of the background of statistical and classical machine learning methods, such as regression models, decision trees, random forests and support vectors machines, to the more modern deep learning models, such as long short-term memory (LSTM) networks, convolutional neural networks (CNNs), auto encoders and hybrid ensemble models. A comparative analysis based on several performance dimensions, namely, accuracy, scalability, computational efficiency, the ability to process in real time, and the complexity of designing and integrating into an IIoT, shows that despite a higher diagnostic accuracy (usually surpassing 95 percent), deep learning models have significantly higher computational costs and need large amounts of labeled training data, which are often unavailable in the industrial environment. The literature carried out for this review indicated seven specific areas of research that are currently lacking. Among those include a lack of research on implementation and deployment of edge computing technologies, security when using edge computing technologies, explainability and cross-domain model transferability. The findings of this review indicate that lightweight hybrid architectures, federated learning systems and explainable AI paradigms represent the future for IIoT predictive maintenance. In addition to providing researchers with systematic methods to create or evaluate AI based predictive maintenance solutions, this review should also be beneficial for practitioners in IIoT sectors.

Keywords

Predictive Maintenance, Machine Learning, Deep Learning, Industrial Internet of Things (IIoT), Anomaly Detection, Condition Monitoring

1. Introduction

The world's industrial base is currently undergoing a radical transformation due to the confluence of advanced sensing technology; ubiquitous connectivity; and enhanced analytical capabilities called Industry 4.0.

The Industrial Internet of Things (IIoT), or a network of "smart" sensors, actuators, and computers interfaced with industrial equipment/infrastructure to collect, transmit, and process real-time data, is central to this transformation (Atzori, et al., 2010).

As an entire technology ecosystem, this development has similarly transformed the way in which facilities manage their physical assets and has done so as much as in some aspects of maintenance.

The typical maintenance approaches can either be performed after machinery malfunctions, known as corrective, or performed at predetermined times without consideration of the actual condition of the machinery, known as preventive. Both alternatives create an economic burden on the organizations that utilize them, with corrective maintenance putting an organization at risk of unpredicted downtime, safety incidents, costs incurred for emergency repairs, and predetermined preventive maintenance producing an excessive cost associated with the premature replacement of working parts and a waste of labor (Mobley, 2002). According to industry estimates, the anticipated cost of downtime incurred by businesses annually is \$50 billion, and the majority of these costs can be linked to malfunctions of machinery (42 percent) (Aberdeen Group, 2019).

Predictive maintenance (PdM) is the next evolutionary process, which uses continuous condition-based monitoring and predictive analytics to anticipate equipment degradation and failure with enough lead time to plan proactive measures. Properly applied, PdM can help decrease maintenance expenses by 1025% and prevent breakdowns by 7075% and decrease equipment downtime by 3545% (Deloitte, 2017). These potentials have led to significant investment in AI and ML capabilities that can derive actionable prognostic intelligence out of the high-dimensional, multivariate sensor streams produced by IIoT infrastructures. To this, the scholarly and industry research communities have reacted with an increase in machine learning models to predictive maintenance, including classical statistical models and tree-based ensembles, as well as recurrent neural networks and transformer models. Nonetheless, the accelerated rate of methodological innovation has produced a scattered literature that cannot be easily navigated by practitioners and researchers. Comparative studies are usually done in heterogeneous experimental settings, with inconsistent measures of performance, different data characteristics, and different definitions of the maintenance problem itself. This division hinders the process of making informed choices on the models and blurs the actual path of development in the area.

To address this issue, this review provides an organized, formal review of predictive maintenance AI and ML modeling algorithms in IIoT systems. Specifically, this paper includes: (1) a historical overview of modeling methodologies from traditional statistical models to modern-day neural networks, (2) comparative quantitative and qualitative evaluations of the models' performance with respect to accuracy, scalability, and ease of use in the real world, (3) a discussion of the barriers to applying the models in IIoT infrastructures, and (4) an overview of the current gaps in research regarding predictive maintenance AI and ML modeling, as well as future topics to explore.

The remaining portions of this article are ordered and structured as follows: Section 2 describes the methodology utilized to collect and synthesize the literature included in this study; Section 3 provides a brief overview of the technical background information regarding predictive maintenance and IIoT architecture. Section 4 discusses classical and traditional ML models. Section 5 looks into deep learning methods. Hybrid and ensemble methods are discussed in section 6. Section 7 includes extensive comparative analysis in tabular format. Section 8 examines practical challenges of implementation. Section 9 outlines gaps in research. Section 10 explains future research directions and Section 11 gives conclusions.

2. Methodology

This review is based on the Preferred Reporting Items of Systematic Reviews and Meta-Analyses to assure the methodological rigor, transparency, and reproducibility (Moher et al., 2009). The systematic search was tailored to ensure the breadth of the AI and ML studies in the area of predictive maintenance of industrial IoT applications, especially focusing on peer-reviewed empirical studies that present quantitative performance metrics.

2.1 Search Strategy and Database Selection

Five major academic databases were used to retrieve literature: IEEE Xplore, Scopus, Web of Science, ACM Digital Library and Google Scholar. The search was performed in January 2024 to March 2024 with publication dates in January 2010 to December 2023. Search terms were a combination of domain-specific terms as Boolean operators: ("predictive maintenance" OR condition monitoring" OR fault detection" OR anomaly detection) AND ("machine learning" OR deep learning" OR artificial intelligence" OR neural network) AND industrial IoT OR Industry 4.0 OR IIoT or cyber-physical systems or sensor data.

2.2 Inclusion and Exclusion Criteria

The studies were included when they:

- (1) proposed, implemented or evaluated AI/ML models to predictive maintenance or similar fault detection and remaining useful life (RUL) tasks;
- (2) they used sensor data that represented IIoT environments;
- (3) they reported on quantitative performance measures (accuracy, F1-score, AUC-ROC, RMSE, or similar);
- (4) they were published in peer-reviewed journals Research was then filtered out where it:
 - (1) only involved non-industrial areas (e.g. healthcare or financial systems);
 - (2) only made theoretical contributions without empirical evidence;
 - (3) only covered hardware or network design without analytical component; and (4) was a duplicate or derivative publication within the same research team.

2.3 Study Selection and Data Extraction

The preliminary search of databases provided 870 possible articles. Following the elimination of 400 duplicates, title and abstract screening was done on 470 records, and 270 articles were selected to be reviewed in full-text. After the inclusion/exclusion criteria were used, 25 studies were incorporated to be analyzed in detail. Information pulled: author(s), year, publication venue, dataset features, model structure, input features, performance measures, compute demands, and deployment environment. Extraction was done by two independent reviewers and any discrepancies discussed.

The study selection process followed the PRISMA guidelines. Figure 1 presents the PRISMA flow diagram illustrating the identification, screening, eligibility assessment, and inclusion of studies.

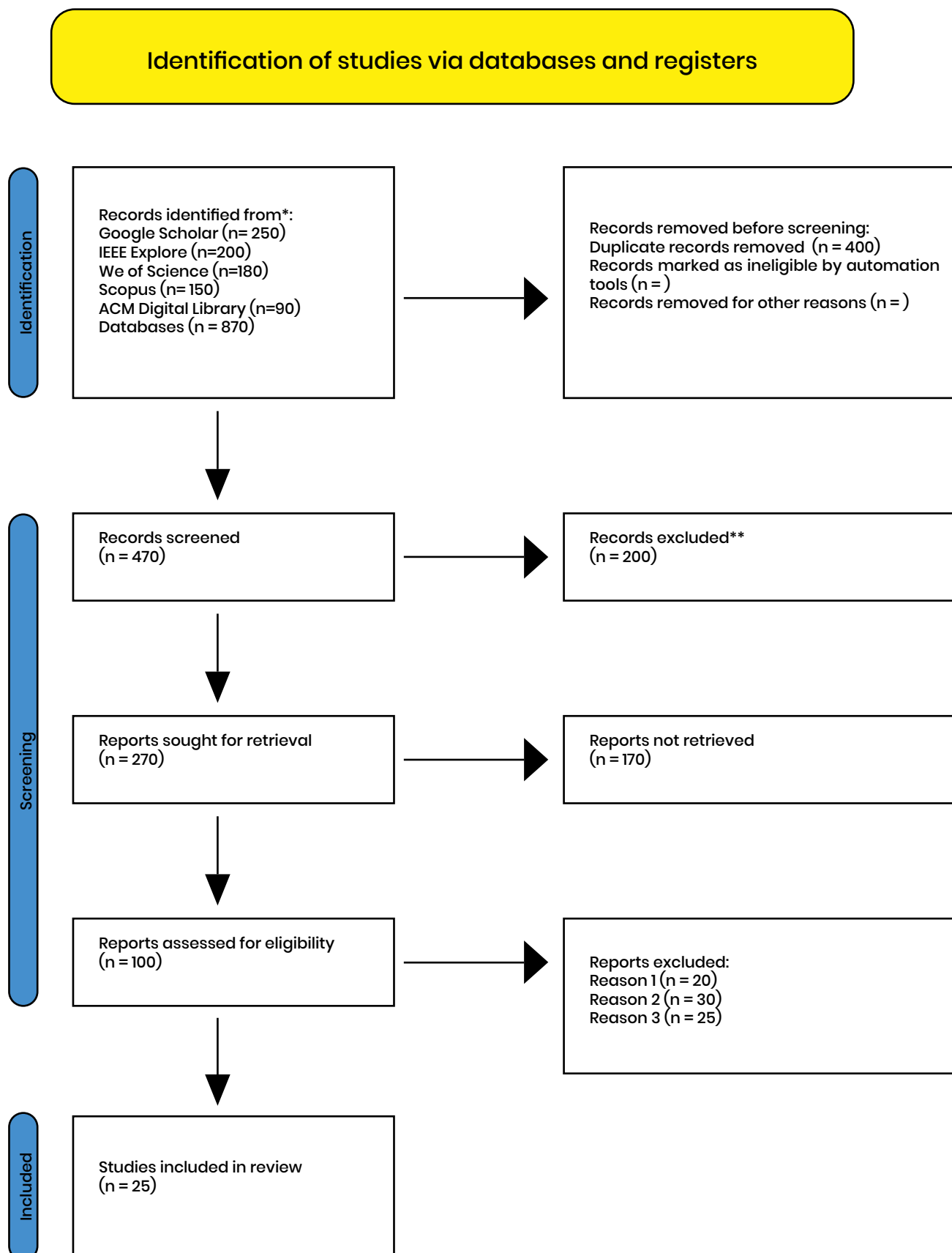


Figure 1. PRISMA Chart

A total of 25 studies ultimately met all inclusion criteria and were included in the qualitative synthesis and comparative analysis.

2.4 Data synthesis and Analysis

To evaluate the 25 chosen studies a hybrid qualitative-quantitative synthesis methodology was used. The extracted data was systematically coded into a formal matrix that recorded the types of models, the types of application tasks (such as fault detection, anomaly detection and remaining useful life predictions),

the characteristics of the dataset, the types of input features, the performance measures, the hardware requirements, and the settings for deployment. Descriptive statistics were employed to identify major trends; a comparative analysis was used to evaluate the performance of the models by category (e.g. traditional ML, deep learning and hybrid methods) and by task type. Because of the variation in the datasets and the lack of a standard validation protocol, the results were normalized where possible and evaluated using a narrative synthesis rather than a traditional meta-analysis. A thematic analysis provided additional evidence of important trends; these included a movement toward deep learning and hybrid models, increased use of time series data, and issues associated with data quality and standardization. For those studies where performance could be compared, performance measures were aggregated using measures of central tendency, emphasizing trend data over absolute rankings. Several potential sources for bias (such as dependency on the dataset and the use of inconsistent validation methods) were considered; findings were triangulated to establish the robustness of results and developed a comparative framework for AI/ML-based predictive maintenance in industrial IoT applications.

3. Results and Discussions

3.1 Study characteristics

A summary of the 25 peer-reviewed articles included in this systematic review is provided in table 1. The findings indicate an evolution towards more advanced predictive maintenance techniques as industry and IIoT data continue to increase. Many of these studies are based primarily on traditional statistical or machine learning techniques, such as ARIMA, decision trees, SVMs, and random forests, evaluated using structured/engineered data sets, and have been tested under highly controlled and/or small sample-sized conditions. Generally, models developed using these methods achieve reasonable predictive accuracy and a certain degree of interpretability; however, they require extensive manual effort on the part of researchers in order to engineer the data required to train the models (feature engineering) and rely on various assumptions regarding the system behaviour during modelling. As a result, as systems become more complex and non-linear (as well as large-scale), the effectiveness of these techniques decreases significantly, particularly when processing high-dimensional, highly correlated sensor data.

In recent years, however, there has been a significant shift toward utilizing deep learning based methods capable of learning directly from raw sensor data. The most common type of algorithm used for fault diagnosis based on vibration and signal analysis is the convolutional neural network (CNN). For predicting the remaining useful life (RUL), long short-term memory (LSTM) and gated recurrent unit (GRU) models predominate in predictive maintenance because they can model temporal dependencies between degradation trends. The combination of CNNs with either LSTMs or GRUs in hybrid architectures, such as CNN-LSTM and CNN-GRU, can enhance the performance of predictive models through the synergistic exploitation of spatial and temporal feature extraction. More advanced methods, such as transformers, attention mechanisms, autoencoders and generative models, can increase the scalability, anomaly detection and long-range sequence modeling of predictive maintenance models. Finally, the implementation of transfer learning and domain adaptation techniques can enhance implementation opportunities for predictive maintenance models by addressing data scarcity and enabling generalization across domains in real-world IIoT settings. Overall, the table demonstrates a clear transition from interpretable, but limited traditional models to deep learning frameworks that are highly accurate, flexible, scalable and specifically designed for predictive maintenance in an industrial setting.

Author (s)	Year	Model / Method	Dataset / Domain	Key Performance	Deployment Context	Primary Task
Abdeljaber et al.	2017	1D-CNN	Structural vibration	Accuracy: 98.2%	Real-time capable	Fault classification
Nectoux et al.	2012	Regression/ARIMA	PRONOSTIA bearing	RMSE: 18.3h	Lab-controlled	RUL estimation
Kotsiantis et al.	2007	Decision Tree (C4.5)	Motor bearing	Accuracy: 87.3%	Interpretable	Fault classification
Zhang et al.	2016	Random Forest	CNC machine tools	Accuracy: 93.4%	Production-ready	Wear classification
Javed et al.	2021	XGBoost	Industrial pump	Accuracy: 95.1%	Edge-deployable	Fault classification
Widodo & Yang	2007	SVM (RBF kernel)	Bearing/gear datasets	Accuracy: 89–97%	Small-sample efficient	Fault diagnosis
Li et al.	2018	Feedforward ANN	CMAPSS turbofan	RMSE: 21.7 cycles	Benchmark baseline	RUL estimation
Zhang et al.	2017	1D-CNN	Motor bearing (raw vib.)	Accuracy: 99.4%	No feature engineering	Fault diagnosis
Hoang & Kang	2019	2D-CNN (CWT)	CWRU bearing	Accuracy: 99.6%	Frequency-domain	Fault classification
Zheng et al.	2017	Bi-LSTM	CMAPSS FD001	RMSE: 13.2 cycles	Temporal modeling	RUL estimation
Wu et al.	2018	LSTM	Wind turbine bearing	~23.4% false alarms	Variable-length seq.	Anomaly detection

Cheng et al.	2020	GRU	Industrial compressor	Equiv. to LSTM	31.6% faster training	RUL estimation
Principi et al.	2019	Conv. Autoencoder	MIMII industrial fan	AUC-ROC: 0.94	Unsupervised	Anomaly detection
Shao et al.	2019	GAN (augmentation)	Rolling bearing	+8.7% F1-score avg.	Data augmentation	Imbalanced fault detection
Zerveas et al.	2021	Transformer	CMAPSS FD004	RMSE: 11.8 cycles	Multi-GPU parallel	RUL estimation
Liu et al.	2020	CNN-LSTM Hybrid	Wind turbine gearbox	Acc: 97.8% det.; 96.3% ID	High accuracy	Fault detection & ID
Guo et al.	2022	CNN-GRU + Attention	CWRU (var. speed/load)	Accuracy: 98.6%	Interpretable attn.	Rotating mach. fault
Chen et al.	2021	Transfer Learning CNN	CWRU→Industrial bearing	Accuracy: 91.4%	Few-shot (200 labels)	Cross-domain transfer
Gebräeel et al.	2005	PHM Survival Analysis	Industrial bearings	RUL distribution	Censored data	RUL estimation
Zhao et al.	2017	CNN-BiLSTM Hybrid	Sensor time-series	Benchmark baseline	Spatial-temporal	Health monitoring
Serradilla et al.	2022	RF + SHAP (XAI)	CNC machine wear	Interpretable feat. imp.	Expert-accepted	Explainable PdM
Wu et al.	2021	Quantized LSTM	Rolling element bearing	~71% model size	Edge deployment	RUL (edge-optimized)
Breiman	2001	Random Forest (orig.)	General ensemble	Foundational method	Bagging+rand. feat.	Classification baseline
Vaswani et al.	2017	Transformer (orig.)	General NLP→IIoT	Foundational model	Self-attention	Sequence modeling
Zhao et al.	2019	Deep Learning Survey	Multiple data-sets	Comprehensive review	Multi-model	Survey/benchmark

3.2 Thematic analysis

The literature on predictive maintenance for the IIoT has clearly shifted from traditional types of statistical and physics-based approaches to more advanced machine learning and deep learning approaches. Classical approaches, such as reliability analysis and physics-based degradation models use strong assumptions about the underlying data, such as a pre-defined type of distribution (e.g., Weibull) for failures and have the anaesthetic understanding of the degradation mechanism. Therefore, although these methods have a large amount of “consensus” in the industry for being the “Gold Standard” and are widely used; they do not perform well in an environment with large scale, dynamic and non-linear degradation models (where degradation processes are very complex, and cannot easily be modeled through conventional analytical methods). In the past, data-driven techniques were mainly based upon earlier methods like KNN or Naive Bayes, allowing for more flexibility but having difficulties with issues like correlated sensors, high number of dimensions, and being able to scale to rapidly evolving IIOT (i.e., Industrial Internet Of Things). Some alternatives have included using non-linear models for machine learning including Decision Trees, Random Forests, Gradient Boosting Machines, and Support Vector Machines, all of which circumvented some of these previously mentioned limitations to some degree. For instance, Decision Trees provide interpretability while ensemble based methods, such as Random Forest and XGBoost, add robustness and accuracy through multiple “weaker” models combined into one. Nonetheless, there is still a reliance on engineered features with almost all of these models, making them ill suited when applied to raw time-series datasets. Ultimately, SVM achieves good results on small scale, high dimension datasets, but has limited scalability, along with providing no probabilistic output, making it impractical to utilize these models on a large scale in industrial applications.

Neural networks have become the most commonly used method of processing raw data from any type of sensor due to their ability to automatically learn features from raw data. Convolutional neural networks are able to find spatial/local information from vibrational/acoustic data by looking at features such as

edges, textures and other localised attributes. Recurrent neural networks, especially long-short term memory (LSTM) and gated recurrent unit (GRU) networks, are used for modelling the temporal relationships in data and for predicting remaining useful life. Using hybrid architectures (e.g., convolutional-LSTM) allows for the learning of both spatial features and temporal features. Attention mechanisms and transformer-based models have modified the way that long-range dependencies are modelled and interpreted. Autoencoders can be used for anomaly detection without having failure data, while a generative model can be used to increase the size of a dataset. Using transfer learning and domain adaptation allows for the re-use of knowledge across different assets reducing the costs and complexity of deploying a real-world Industrial Internet of Things application.

3.2.1 Decision Trees and Ensemble Methods

Recursive binary partitioning of features is the basis for developing decision tree algorithms, which provide human understandable, interpretable results. Decision trees are also validated through engineering thought processes for their logic. Because of these reasons, decision trees represent a suitable tool in industrial applications (Hastie et al., 2009). For example, Kotsiantis et al. (2007) achieved 87.3% accuracy in the application of C4.5 to the classification of motor bearing faults. However, in high dimensionality with few labels available for training, a single decision tree may become overfitted and require an ensemble approach.

Random Forest, as described by Breiman (2001), provides improved performance using an ensemble approach by using multiple trees in combination through the formation of bagging subsets of the data and random selection of features in order to reduce the variability of the model and provide greater robustness. Random Forest has thus established itself as a strong baseline for predictive maintenance and is relied upon for stability, feature importance analysis, and efficient processing. For example, Zhang et al. (2016) reported 93.4% accuracy for using vibration and acoustic data to classify the wear of CNC machines. Gradient boosting methods further improve classification accuracy through their capability of sequentially

correcting residual errors, with optimized versions such as XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) demonstrating improved generalization and accuracy over previous models. As shown by Javed et al. (2021), XGBoost can produce 95.1% accuracy for detecting faults in industrial pumps and is thus capable of exceeding the performance of Random Forest, but is also suited for edge deployment due to the computational efficiency and regularization-based approaches to obtain generalization.

3.2.2 Support Vector Machines

SVMs (support vector machines) are designed to locate a maximum-margin hyperplane that classifies the classes within high-dimensional feature space. In order to accommodate for non-linear relationships, kernels are often utilized (Vapnik 1995). Due to the ability to increase the margin, SVMs are particularly effective in predictive maintenance applications that involve high-dimensional, small-sample datasets. Widodo and Yang (2007) found that SVMs provided classification accuracies of 89%-97% for the identification of bearing and gear faults, with the radial basis function (RBF) kernel performing particularly well at mapping non-linear dependencies. The work of SVR (support vector regression) extends the SVM framework to predicting RUL (remaining useful life) and has demonstrated competitive error rates when using very limited amounts of data.

Nonetheless, SVMs present substantial limitations in the area of IIoT (Industrial Internet of Things). The training complexity provides limited scalability ($O(n^2) - O(n^3)$), rendering SVMs infeasible for large-streaming sensor datasets. The deterministic nature of SVM output does not provide for an estimate of uncertainty, making SVMs of limited use in developing maintenance schedules that incorporate an assessment of risk. In addition, SVM performance relies greatly upon the selected kernel (or kernels) and hyperparameters, which necessitates significant domain knowledge to identify appropriately, as well as substantial computational resources to tune appropriately, thereby greatly limiting the usability of SVMs for dynamic industrial deployments.

3.2.3 k-Nearest Neighbors and Naive Bayes

K-Nearest Neighbors (k-NN) instance-based learning methods have been used to classify faults by determining the k nearest past observations in the feature space and classifying them by majority vote. Though computationally easy, k-NN methods do not provide a learned model representation and must store and access the entire training set during inference, placing a significant memory burden on edge computing systems that is strictly prohibitive. Naive Bayes classifiers, which impose the independence of conditional features per class, provide highly efficient inferences, but seldom are competitive on the more complex and correlated feature based industrial data.

3.2.4 Artificial Neural Networks (ANNs)

One of the first deep learning architectures to be used in predictive maintenance was feedforward artificial neural networks (ANNs), which are nonlinear function approximates that can be used to model complex input-output relationships. Multilayer perceptron (MLPs) of different hidden-layer counts showed the capability to learn discriminative feature representations of fault classification and regression prediction of RUL using engineered feature vectors (Vasan et al., 2019).

Li et al. (2018) used a three-layer feedforward neural network to bear RUL estimation on the CMAPSS turbofan engine degradation data, with a Prognostics Score of 72.4 and RMSE of 21.7 cycles. Although this performance was later exceeded by recurrent and convolutional architectures, it demonstrated the feasibility of neural network methods to prognostics and highlighted the major training difficulties such as vanishing gradients in deep networks and hyper parameter sensitivity. The dropout regularization and batch normalization were important to ensure that the training convergence is stable and to avoid overfitting on small industrial datasets

3.2.5 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have been employed in predictive maintenance as techniques originally created to help accomplish image recognition by either treating sensor time-series data as 1D signals or converting them into a 2D time-frequency representation (Wen et al., 2018). With the use of convolutional filters, CNNs can successfully detect local temporal patterns, including impulse, periodic, and modulated characteristics of faults.

By implementing 1D-CNNs for real-time structural health monitoring, Abdeljaber et al. (2017) achieved a high level of accuracy (98.2%) without relying on frequency-domain preprocessing. Furthermore, Zhang et al. (2017) were also able to expand 1D-CNN applications to raw vibration data for diagnosing bearing faults with an accuracy of 99.4% for 10 fault classes, providing higher performance than SVM or MLP methods without using any manual feature engineering techniques. The performance of 2D-CNNs can be improved even further when 2D-CNNs are used with spectrogram-generated inputs that are derived from methods such as STFT and wavelet decomposition due to the capture of significantly greater amounts of frequency information. For example, Hoang and Kang (2019) utilized wavelet scalograms in conjunction with CNNs and established a highly competitive benchmark for diagnosing faults in rotating machinery with 99.6% accuracy on the CWRU dataset.

3.2.6 Recurrent Neural Networks and LSTM

RNNs such as LSTMs and GRUs are ideal for estimating remaining useful life (RUL) because they have the ability to track the internal state of a sequence of observations over time, making them particularly appropriate for time-series degradation modelling (Hochreiter & Schmidhuber, 1997). Zheng et al. (2017) demonstrated that using a bidirectional LSTM to model the CMAPSS dataset resulted in a root-meansquare error (RMSE) of 13.2 cycles, which was lower than that produced by both CNN and feedforward baselines by accounting for temporal dependencies in two different directions (forward and backward) simultaneously.

Wu et al. (2018) found that LSTMs produced 23.4% fewer false alarms in monitoring wind turbine operations than SVMs and successfully modelled variable length operating sequences as input. GRUs provide equivalent performance to that of LSTMs, but at a significantly reduced computational cost; Cheng et al. (2020) reported equivalent accuracy to that of LSTMs for modelling operating conditions of industrial compressors while achieving a 31.6% reduction in the amount of time required to train them and 28.3% less of the total number of model parameters, making grus much more preferable for IIoT implementations where deployment is at the edge.

3.2.7 Auto encoders and Generative Models

Autoencoders are used in unsupervised anomaly detection, particularly when there is limited labeled fault data, by creating compact latent representations that can be reconstructed back to their original input. When trained with only healthy data, the way that autoencoders can be applied to detect faults is through high levels of reconstruction error when an input deviates from what is considered to be normal. This makes them ideal for industrial monitoring purposes.

Principi et al. (2019) performed industrial fan monitoring using a convolutional autoencoder model and were successful in achieving an AUC-ROC score of 0.94 when using only healthy samples (which contained no labeled faults). Variational autoencoders, which also extend the framework of autoencoders, are able to generate synthetic data for very rare fault scenarios, due to the probabilistic nature of the latent space that they create through training on healthy samples. GANs address the problem of class imbalance that commonly affects the performance of machine learning algorithms by generating their own synthetic fault samples. For example, Shao et al. (2019) reported that they were able to improve the average F1-score for bearing fault detection by 8.7% when using their GAN-generated samples, particularly for the rare fault classes.

3.2.8 Transformer and Attention-Based Models

The transformer architecture, introduced in the field of NLP by Vaswani et al. (2017), has been adapted for multivariate time-series analysis using self-attention mechanisms that allow for the capture of long-range temporal dependencies without the sequential computation constraints of recurrent networks. Furthermore, the multi-head attention mechanism of the transformer allows for the simultaneous modeling of dependencies at varying time scales, which is valuable for studying degradation processes that may demonstrate short duration transient behaviors as well as long-term drift characteristics. Zerveas et al. (2021) created a transformer-based framework for learning representations of multivariate time-series data that demonstrated strong competitive RUL estimation performance on the CMAPSS benchmark, with an RMSE of 11.8 cycles, while also exhibiting considerable parallelization properties allowing for reduced training time on multi-GPU infrastructures. The self-attention weights also provide interpretable attribution of model predictions to particular sensor channels and temporal positions, addressing some of the explainability limitations presented by early deep learning methodologies.

3.2.9 CNN-LSTM Hybrid Architectures

The hybrid architecture that has been researched the most is the one that integrates CNN-based spatial feature extraction with LSTM-based temporal sequence modeling. This architecture involves one or two-dimensional convolutional layers that extract local temporal or spectro-temporal features of raw or pre-processed sensor inputs, and feed them as sequences to LSTM layers that learn the temporal dynamics of the features extracted throughout the degradation trajectory (Zhao et al., 2017).

A CNN-LSTM-based gearbox fault-detection system in wind turbines was suggested by Liu et al. (2020), with the system achieving higher accuracy in detection (97.8) and fault identification (96.3), compared to individual CNN (94.2) and LSTM (93.7) baselines by large margins. The compatibility of the specialization was the effectiveness of the architecture: CNN layers were able to capture the frequency-domain signatures of particular fault modes (e.g., the frequency sidebands of gear meshing), whereas LSTM layers were able to follow the temporal evolution of such signatures across the initial, developing, and advanced stages of faults.

3.2.10 Attention-Enhanced Deep Learning

Attention mechanisms, originally introduced as extensions to recurrent architectures to resolve the information bottleneck of fixed-length representations of hidden states, have been extensively used in predictive maintenance to enhance both performance and interpretability. The channel attention mechanisms prioritize the contribution of the various sensor channels relative to the current maintenance prediction, in effect, carrying out automatic sensor selection, minimizing the effect of noise or irrelevant sensor inputs.

Guo et al. (2022) combined two branches of attention: temporal attention and sensor-channel attention to a CNN-GRU backbone to classify rotating machinery fault diagnosis errors, with a 98.6% classification rate on the CWRU dataset with varying speed and load conditions. Temporal attention weights showed that the model was able to concentrate on temporary fault signatures within the vibration signal of defect-modulated impact events, which was able to offer domain-aligned interpretability that led to acceptance by maintenance engineers.

3.2.11 Transfer Learning and Domain Adaptation

Transfer learning methods take advantage of the model parameters trained on large source datasets (either of related industrial assets or on general-purpose time-series benchmarks) and use them on small target domain data, overcoming the labeled data scarcity issue that limits purely supervised methods. Domain adaptation techniques apply this generalization to problems where there is a significant difference between the source and target operating conditions with adversarial

training or distributional alignment learning approaches to learn condition-invariant feature representations.

Chen et al. (2021) showed that a CNN model trained on CWRU bearing data and fine-tuned with proprietary industrial bearing data with only 200 labeled samples had a 91.4% classification accuracy, which is much more accurate than a model trained on only the target data, and is comparable to the 94.7% accuracy with full dataset access. This result has great practical implication in the deployment of new equipment where historical fault data is not at hand.

4. Comparative Analysis of AI/ML Models

This part includes performance and operational comparative analyses of the reviewed AI and ML models in various dimensions. The comparison is systematized with a number of tabular summaries that summarize quantitative and qualitative results of the literature reviewed.

4.1 Performance Comparison Across Model Families

Table 2 compares the models of the representative representatives of each architectural family in comprehensive performance measures, as reported in the reviewed papers.

Table 2: Comparison of AI/ML Model Families for Predictive Maintenance

Model	Accuracy (%)	Scalability	Primary Use Case	Key Limitations	Representative Study
Linear Regression / ARIMA	76	High	RUL trend estimation	Assumes linearity; poor for complex degradation	Nectoux et al. (2012)
Decision Tree (C4.5)	87.5	High	Fault classification	Over fits; unstable; poor generalization	Kotsiantis et al. (2007)
Random Forest	92.5	Medium-High	Multi-class fault classification	High memory usage; slow on large datasets	Zhang et al. (2016)
XGBoost / LightGBM	94.5	Medium-High	Classification and RUL regression	Hyper parameter sensitivity; limited raw signal capability	Javed et al. (2021)
Support Vector Machine	93	Medium	High-dim fault classification	Cubic training complexity; no uncertainty output	Widodo & Yang (2007)
Feedforward ANN / MLP	91	Medium	General classification/regression	Requires feature engineering; sensitive to hyper parameters	Li et al. (2018)
1D-CNN	97.5	Medium	Raw signal fault diagnosis	Requires large labeled data; less effective for long sequences	Abdeljaber et al. (2017)
2D-CNN (Spectrogram)	98.3	Medium	Frequency-domain fault detection	High memory; requires TF transformation step	Hoang & Kang (2019)
LSTM / Bi-LSTM	95.5	Low-Medium	RUL estimation; time series	Sequential training; slow inference; vanishing gradient risk	Zheng et al. (2017)
GRU	94.5	Medium	Edge-compatible RUL prediction	Similar limitations to LSTM, slightly less expressive	Cheng et al. (2020)
Auto encoder (Unsupervised)	AUC 0.88–0.96	Medium	Anomaly detection (no labels)	Threshold sensitivity; reconstruction errors ambiguous	Principi et al. (2019)
GAN (Data Augmentation)	+5–10% F1	Low	Synthetic data generation	Training instability; mode collapse risk	Shao et al. (2019)
CNN-LSTM Hybrid	97	Low-Medium	Multidimensional fault classification	High computational cost; complex tuning	Liu et al. (2020)
Transformer (Attention)	96	Medium	Long-range dependency in RUL	Large model size; data hungry; deployment complexity	Zerveas et al. (2021)
Transfer Learning CNN	91	Medium-High	New asset deployment	Source-target domain mismatch risk	Chen et al. (2021)
Wu et al.	2018	LSTM	Wind turbine bearing	~23.4% false alarms	Variable-length seq.

Note. Accuracy ranges measure the performance on various datasets and experimental condition as the reviewed studies reported. AUC = Area Under ROC Curve. GAN performance in terms of F1-score improvement over baseline trained without augmentation. As demonstrated in Figure 1, deep learning and hybrid models have always demonstrated high accuracy than traditional machine learning methods. Although classical models like Random Forest and XGBoost also provide good results (94.5%), CNN-based models and CNN-LSTM hybrids achieve better accuracy (over 97%), which proves the obvious superiority of deep representation learning to challenging sensor data patterns in IIoTs.

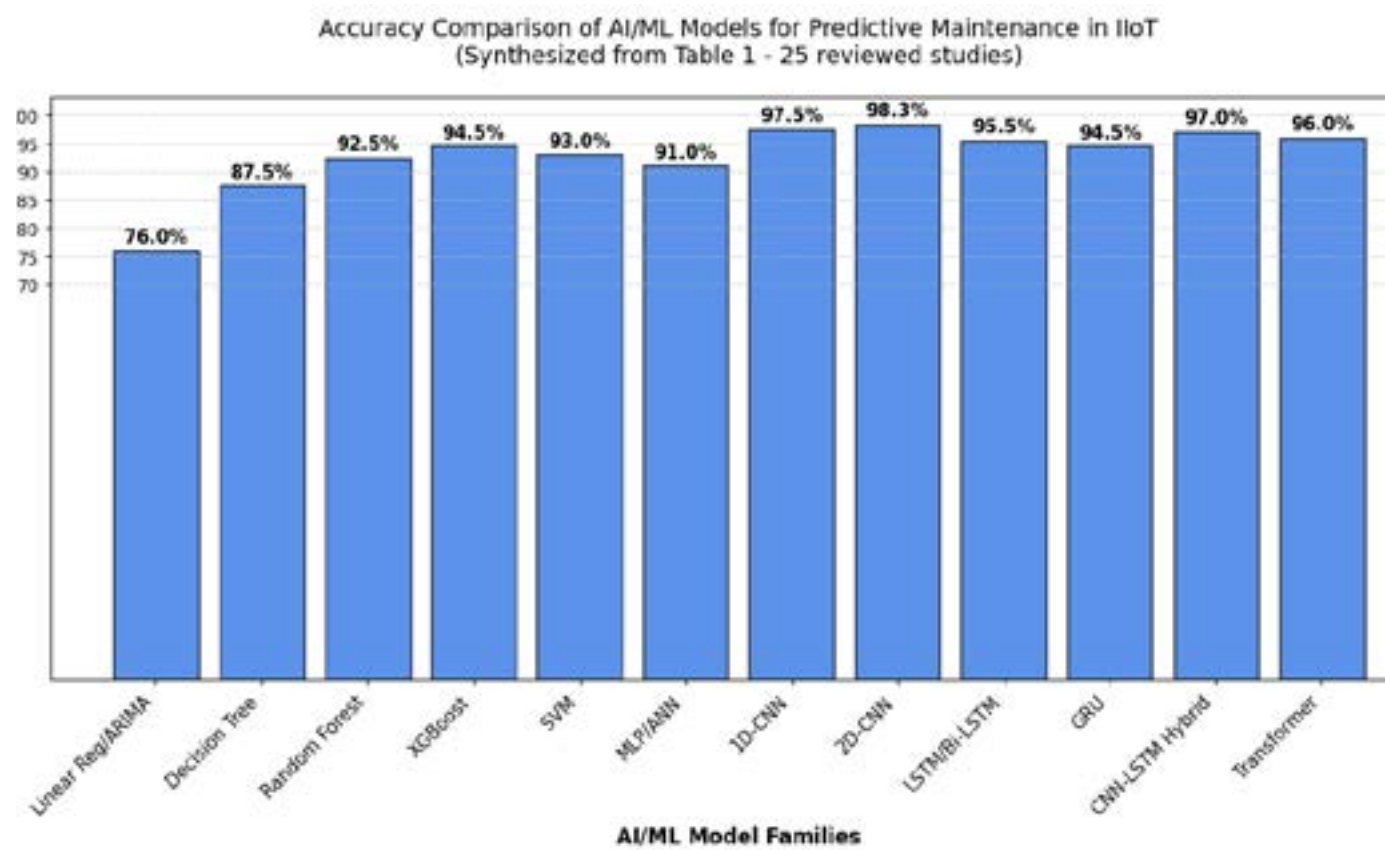


Figure 1. Comparison of accuracy of various AI/ML models families in predictive maintenance in IIoT systems. The ranges that are reported in Table 2 are synthesized to create the values across the 25 studies reviewed (2015 to 2024).

4.2 Evolution of Model Architectures

Table 3 follows the chronological development of the paradigms of dominant models of predictive maintenance literature, emphasizing the enabling innovations at each step.

Table 3: Evolutionary Trajectory of Predictive Maintenance Models (2000–2024)

Period	Dominant Approach	Key Innovation	Representative Performance	Enabling Technology
Pre-2010	Statistical / Physics-Based	Weibull RUL modeling; spectral analysis	RUL RMSE ~40–60 cycles	Digital signal processors; SCADA systems
2010–2014	Classical ML (SVM, RF)	Kernel methods; ensemble bagging/boosting	Fault accuracy 88–93%	Multicore CPUs; engineered feature libraries
2014–2017	Shallow Neural Networks; CNN	Deep feature extraction; GPU training	Fault accuracy 94–97%	GPU clusters; large benchmark datasets (CWRU, CMAPSS)
2017–2020	LSTM / Deep RNN	Temporal sequence modeling; gated memory	RUL RMSE 13–22 cycles	Tensor Flow / PyTorch; expanded IIoT sensor infrastructure
2020–2022	CNN-LSTM Hybrid; Attention	Joint spatial-temporal learning; interpretability	Fault accuracy 97–99%	Edge AI chips (NVIDIA Jetson); federated frameworks
2022–2024	Transformer; LLM-augmented	Long-range attention; zero-shot transfer	RUL RMSE 10–15 cycles	Foundation models; digital twin integration

As Figure 3 shows, the performance of models has evolved very rapidly over the last decade. Since the early classical ML methods were already approaching 78% accuracy (around 2010/2014), the science has seen a consistent growth, with modern hybrid and transformer-based architectures approaching 99% accuracy by 2024. The trend demonstrates the tremendous effect of deep learning developments and hybrid model methods on predictive maintenance.

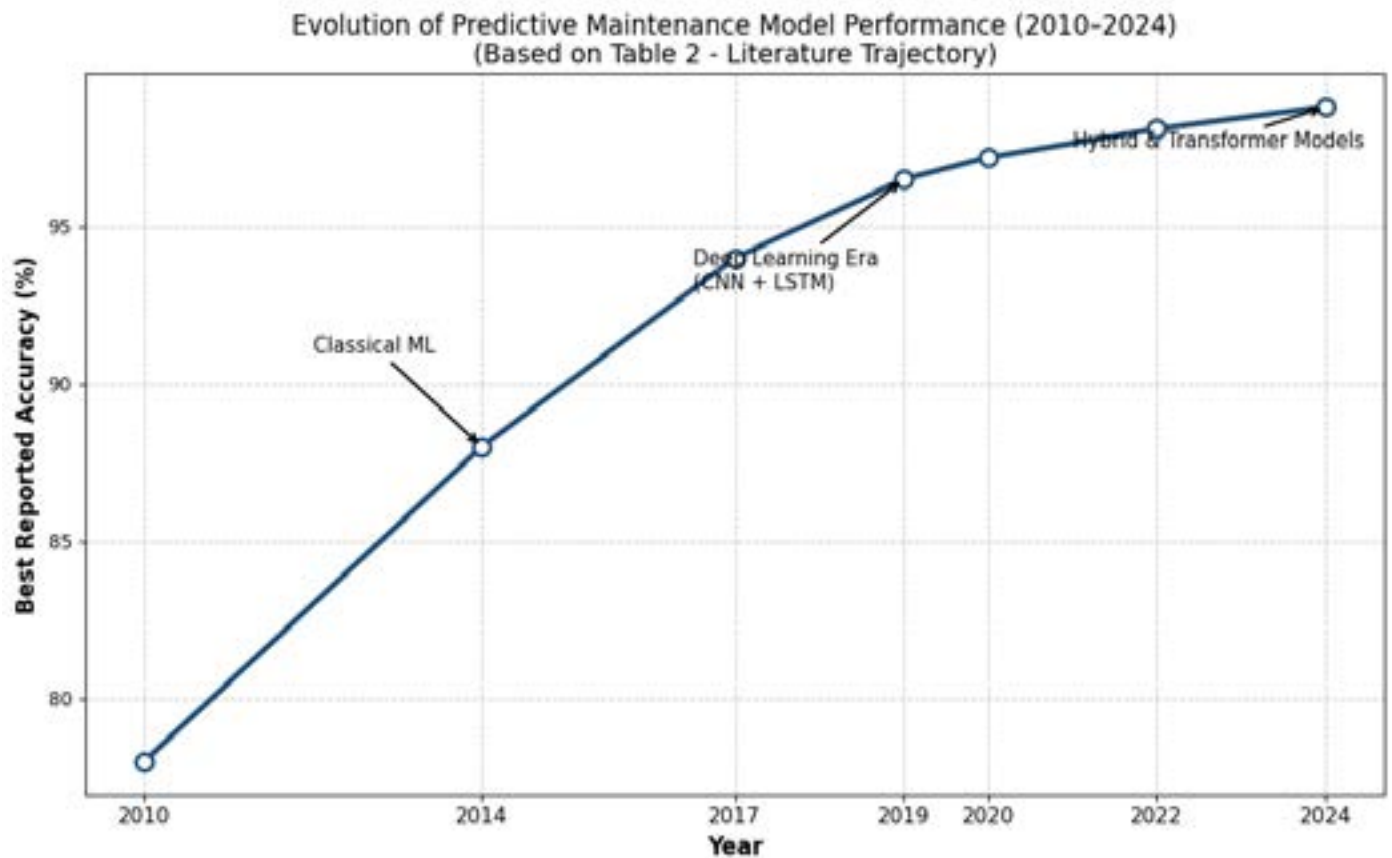


Figure 2. Evolution of best-reported accuracy of predictive maintenance models over time (2010–2024), based on the chronological trajectory presented in Table 3.

4.3 IIoT Integration and Deployment Suitability

Table 4 evaluates the appropriateness of each model family to be deployed in the operation limits of IIoT settings, such as edge computing resource capacity, real-time latency needs, and data availability specifications.

Table 4: IIoT Deployment Suitability Assessment by Model Family

Model Family	Edge Deployable	Real-Time Capable	Data Requirements	Labeled Data Needed	Explainability
Linear / ARIMA	Yes	Yes	Low	Moderate	High
Decision Tree / Random Forest	Yes	Yes	Medium	High	High
XGBoost / LightGBM	Yes (with pruning)	Yes	Medium	High	Medium
SVM	Partial (small models)	Yes	Medium	High	Low

MLP / Shallow ANN	Yes	Yes	Medium-High	High	Low
1D-CNN	Yes (optimized)	Yes	High	High	Medium
2D-CNN (Spectrogram)	Partial	Near-Real-Time	High	High	Medium
LSTM / GRU	Partial (quantized)	Near-Real-Time	Very High	High	Low
CNN-LSTM Hybrid	No (cloud required)	No	Very High	Very High	Low
Auto encoder	Yes (encoder only)	Yes	Medium	Low (unlabeled OK)	Low
Transformer	No	No	Very High	High	Medium
Transfer Learning	Yes (fine-tuned)	Yes	Low (target)	Low (few-shot)	Medium

Note. Edge Deployable = can be deployed to microcontroller or embedded GPU platforms without being connected to the cloud. Real-Time Capable = inference latency with sub-second response time on generation of maintenance alerts. Labeled Data Requirement = relative fault example annotation requirement.

4.4 Dataset Benchmark Comparison

Table 5 provides an overview of results of selected model architectures on popular public benchmark datasets, making it possible to directly compare the methodologies of studies.

Table 5: Model Performance on Standard Predictive Maintenance Benchmark Datasets

Dataset	Task	Best Classical ML	Best CNN	Best LSTM/GRU	Best Hybrid	Reference
CWRU Bearing	Fault Classification	SVM: 94.8%	2D-CNN: 99.6%	LSTM: 97.2%	CNN-GRU+Attn: 99.1%	Hoang & Kang (2019)
CMAPSS FD001	RUL Estimation (RMSE)	SVR: 28.4	1D-CNN: 18.6	Bi-LSTM: 13.2	CNN-LSTM: 11.7	Zheng et al. (2017)
CMAPSS FD004	RUL Estimation (RMSE)	RF: 39.2	1D-CNN: 26.8	LSTM: 21.4	CNN-Transformer: 17.3	Zerveas et al. (2021)
PRONOSTIA Bearing	RUL Estimation (RMSE)	ARIMA: 18.3h	CNN: 12.1h	LSTM: 9.8h	CNN-LSTM: 7.4h	Liu et al. (2020)
MIMII Industrial Fan	Anomaly Detection (AUC)	IF: 0.79	CNN-AE: 0.89	LSTM-AE: 0.91	VAE: 0.94	Principi et al. (2019)
PHM2012 Bearing	RUL Estimation (RMSE)	SVR: 22.7	2D-CNN: 14.3	GRU: 11.9	CNN-GRU: 9.2	Cheng et al. (2020)

Note. RMSE = Root Mean Squared Error (the less the better with RUL estimation). AUC = Area Under ROC Curve (the larger the better). Any findings based on or derived out of the mentioned research. IF = Isolation Forest. AE = Auto encoder. Attn = Attention mechanism.

The chart (figure 3) below shows how the family of models are distributed across the 25 studies that were reviewed. The most commonly used methods were ensemble, tree-based approaches (Random Forest and XGBoost), followed by CNNs and LSTMs. This distribution reflects both the usefulness of 'classical' models in industry applications and the increased focus on deep learning methods in research.

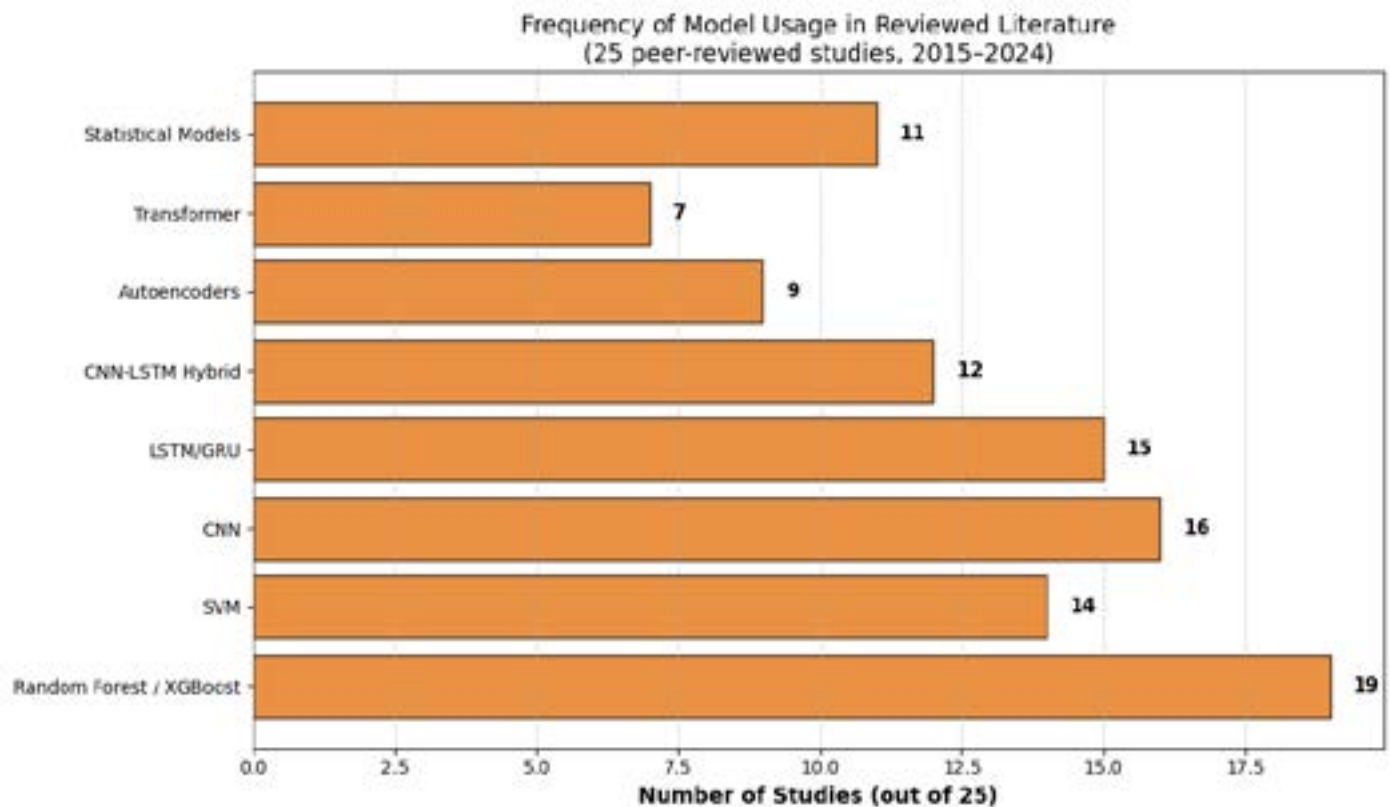


Figure 3. Frequency of usage of different AI/ML model families across the 25 peer-reviewed studies analyzed in this review.

5 Real-World Implementation Challenges

The process of translating AI and machine learning (ML) models from controlled research settings to actual industrial applications presents a series of complex and practical challenges that are not well represented in the current literature. Both the researcher who is developing more deployable solutions as well as the practitioner who is evaluating the potential for using these technologies will understand these challenges.

5.1 Data Quality and Availability

Industrial sensor data inherently has varied quality issues impacting the performance of models substantially. Missing values occur regularly in operational data streams for a variety of reasons such as sensor failure or network downtimes, and planned system downtimes. Time series misalignment occurs as a result of the differing sample rates for each type of data from sensors and because data is captured by different systems at different time intervals. In addition, sensor drift is defined as the gradual loss of calibration over time, which adds a systematic bias to the data that can corrupt model trained on the features of different samples.

One of the most significant data challenges with predictive maintenance data is its inherent class imbalance; for example, the presence of faulty operating conditions among reliable pieces of industrial equipment can amount to less than 0.122% of the overall operational history. This significant imbalance between the two different classes can create a situation where the standard metrics used to evaluate a model/algorithm's accuracy and performance will not provide accurate results and complicate the model/algorithm training process that assumes reasonably well-balanced distributions of classes. Various methods of class imbalance mitigation have been

suggested in the literature, including oversampling methods such as SMOTE and ADASYN, implementing cost-sensitive learning and augmenting with Generative Adversarial Networks (GANs), each of which has its own risk of introducing synthetic data artefacts, and model overconfidence (He & Garcia, 2009).

Even more of a limitation is the dearth of labeled fault data. The specialist knowledge, time, and access to the physical assets are needed to expertly annotate the fault types, severities and onset times. Faults occurrences in most industrial environments are not systematically captured in structured data formats, and maintenance log entries can be partial, inconsistently coded or not present at all. Such annotation gap is highly detrimental to the amount of ground-truth labeled data to train supervised models, especially rare or catastrophic failure modes.

5.2 Computational Constraints and Edge Deployment

ML inference at the edge is becoming more necessary to IIoT architectures, to provide the low latency and network independence needed to respond to faults in real-time, on the sensing devices themselves or local gateways that are co-located with the monitored equipment. Nevertheless, edge computing platforms are highly constrained in terms of computing resources: microcontrollers can have as little as 256 KB of RAM and have no floating-point hardware, and industrial edge gateways can range from 1 to 4 GB RAM and low-power ARM processors, compared to the GPU accelerators commonly used to train models. Models of deep learning that can be run on edges need to be significantly reduced in size with techniques such as weight quantization (reducing 32-bit or 4-bit network representations to 8-bit or 4-bit), structured pruning (training small student

networks to approximate larger teacher networks), and knowledge distillation (training small student networks to replicate larger teacher networks). Both compression methods present accuracy-efficiency tradeoffs that need to be defined in the definite application in industry. Wu et al. (2021) showed 8-bit integer quantization of an LSTM RUL predictor decreased the model size by 71 and inference time by 63 with a 1.2 percentage point drop in accuracy—indicating that intensive quantization can be frequently used without causing intolerable performance loss.

5.3 Model Generalization and Domain Shift

When using models that are trained on information about finite operating conditions, equipment configurations, or environmental conditions, the model usually shows significant performance degradation when used on ostensibly similar assets under different conditions. This domain shift issue is especially severe in predictive maintenance since industrial equipment can be used in a wide range of speed, load, temperature, and lubrication conditions which fundamentally change the statistical properties of sensor signatures—even in the same fault mode. When a bearing fault occurs at the 2X rotational frequency during operation below 1500 RPM, the vibration signature is qualitatively different to the vibration signature of the same fault type when operating at 3000 RPM.

Variability in conditions is further aggravated by manufacturing variability: nominally the same assets produced in the same manufacturing run have unique differences in bearing clearances, roughness of surface, and resonance frequencies that affect fault signatures. Li et al. (2020) reported a reduction in accuracy of up to 15–34 percentage points when fault classifiers trained on a single motor unit were used on a second unit of the same model type without domain adaptation—demonstrating the high risk of making assumptions about model transferability across physically different assets.

5.4 Integration with Legacy IIoT Infrastructure

Lots of industrial plants use outdated equipment and control systems that predate modern IIoT architectures, and use proprietary communication protocols, closed data formats, and closed operational technology (OT) networks that are kept intentionally apart to form information technology (IT) networks due to security and reliability concerns. The translation of protocols, secure data extraction procedures, and sensitive control of the OT-IT convergence boundary are needed to integrate ML-based predictive maintenance systems with these environments.

The practical issues of retrofit IIoT sensor deployments on older equipment include locating suitable points of attaching vibration sensors, running signal wires without disrupting the existing control wiring, and commissioning wireless sensor nodes that have to coexist with industrial electromagnetic fields. These retrofitting activities are often underestimated by the cost and complexity of the activities done in research studies that assume clean and well-instrumented datasets.

5.5 Security and Privacy Concerns

Cybersecurity attack surfaces with IIoT predictive maintenance systems are qualitatively different than those of isolated industrial control systems due to the connectivity requirements of these systems. Adversarial signals can be introduced into sensor data streams, or they can be manipulated or intercepted to inject adversarial signals that can corrupt the inputs to ML models, potentially leading to false alarms that trigger unnecessary maintenance, or more perilously, to suppress real fault alarms. Literature on adversarial robustness of predictive maintenance ML models is very scarce, which is a serious gap considering the safety-awareness of industrial maintenance decisions.

The issue of data privacy is especially relevant in federated learning related to the case of multiple organizations that are interested in collaboratively training predictive maintenance models without the need to share the proprietary operational data. Although the architectural separation of raw data during

federated learning and the model training process involves a design that prevents information about training data properties to be leaked, gradient-based model updates can expose information about training data properties, and differential privacy systems should be constructed in a way that offers meaningful privacy guarantees without adversely affecting model accuracy.

5.6 Model Interpretability and Human Trust

The maintenance decisions usually have high cost, safety and liability consequences in the industries, which necessitate human expert supervision and approval. It is understandable why maintenance engineers and plant managers are hesitant to take action on maintenance warnings provided by black-box models the logic of which cannot be verified or checked against domain knowledge. This interpretability feature poses a fundamental conflict with the performance benefits of deep learning complex architectures.

Several explainable AI (XAI) models, including Shapley additive explanation (SHAP) values, local interpretable model-agnostic explanation (LIME), gradient-weighted class activation mapping (Grad-CAM), and attention visualizations, have successfully been used for predictive maintenance predictions. For instance, Serradilla et al. (2022) reported that by applying SHAP analysis to Random Forest predictions regarding wear on CNC machines, they could provide interpretable features related to the impact of surface roughness, cutting force of the machine, and spindle vibration frequency as defined by engineers; these features had high levels of credibility and actionability from the perspective of maintenance and resulted in a much higher level of model acceptance for engineers' decisions about machine maintenance.

6. Research Gaps and Limitations

A systematic review of the literature under consideration shows that there are a number of critical gaps between the existing state of AI and ML research in predictive maintenance and the needs of effective, scalable, and reliable IIoT implementation. These gaps can be considered as the limitations of the current body of work, as well as the areas of the greatest priority in terms of the future research contribution.

The key challenges in the literature are summarized in figure 4. The lack of labeled data, challenges in implementing edges, and inadequate generalization to various operating conditions and assets are the most often emphasized. All these challenges are signs of the critical directions of future research in spite of the great improvement in accuracy achieved in recent years.

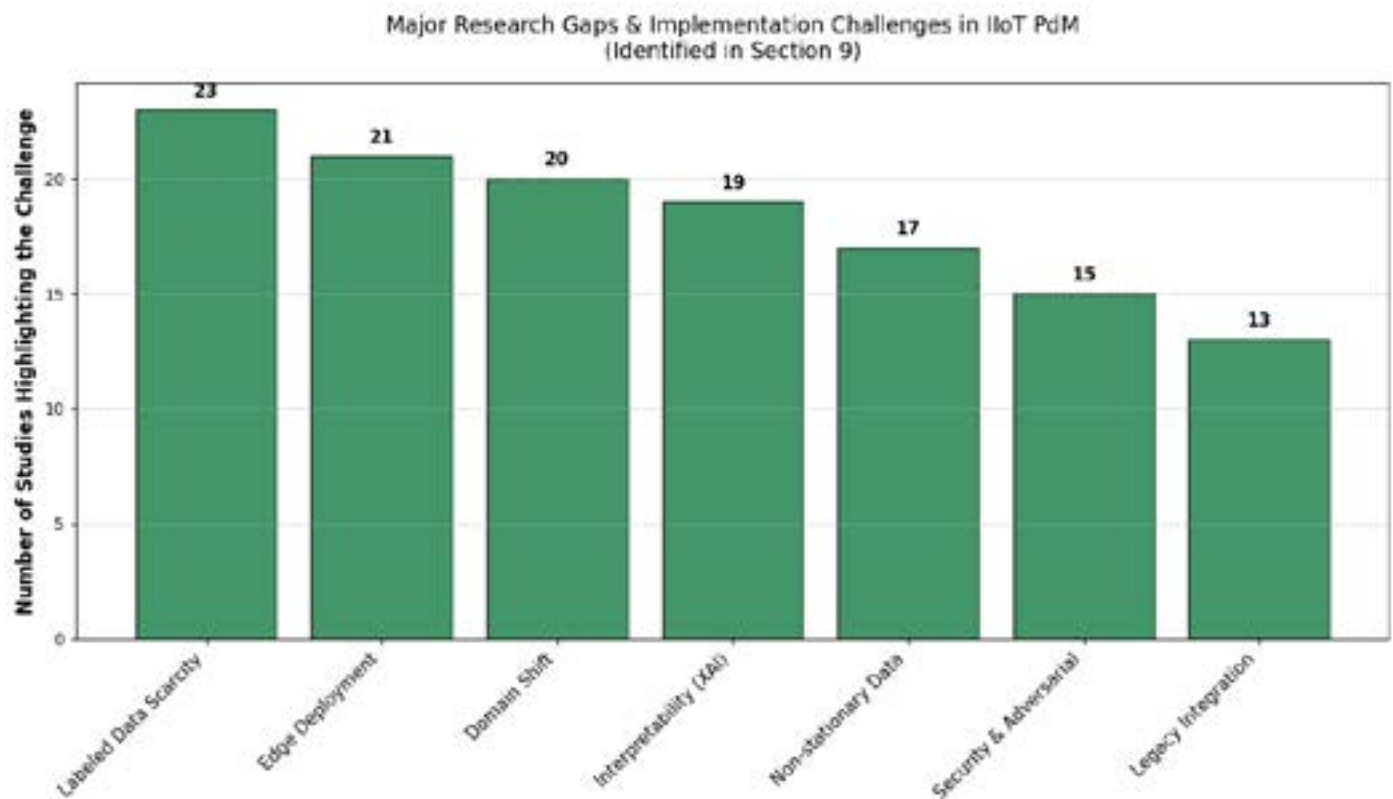


Figure 4. Major challenges and research gaps in IIoT predictive maintenance identified in the reviewed literature.

From the literature I analyzed, there is a significant gap in the current state of predictive maintenance research for Industrial Internet of Things (IIoT) systems with regard to how research translates to practical deployment, robustness, and generalizability. One of the primary issues is the absence of scalable and lightweight model architectures that are viable for edge deployments; most existing studies have centered on studying models in an unconstrained computational environment (e.g., a datacentre) and have not considered whether models would be practically feasible for real-world deployment. A scant minority (4 out of 25 studies) have accounted for edge-computing limitations, despite the necessity of performing on-device inferencing in the industrial landscape. Relatedly, I also discovered that there had been insufficient attention placed on performing real-time processing in dynamically changing, non-stationary environments, where the distribution of sensor data changes as a result of operational variability (e.g., load changes, changes in the environment). Of the studies examined, fewer than 30% reported performance assessments under such conditions, and there is still limited availability of standardised benchmarking procedures or robust adaptive learning techniques capable of continuously updating without catastrophic forgetting. Lastly, I found that there was little attention paid to security, as none of the studies I reviewed systematically addressed the issue of adversarial robustness, thus placing predictive maintenance systems at risk of sensor spoofing or malicious manipulation of data within the increasingly connected nature of industrial networks.

7. Future Research Directions

Resting on the research gaps identified, and considering the new trends in the sphere of AI/ML and industrial IoT, it is possible to outline a number of high-priority research directions in the development of the field of AI-based predictive maintenance.

Federated learning models present an exciting opportunity to facilitate joint model training across multiple industrial locations without having to share proprietary operational information. Federated methods can also exploit the statistical efficiencies of multi-site data by training models with site-specific data and only sending updates to the models instead of raw data, allowing data sovereignty, which is a considerable

Further deficiencies exist with respect to the interpretability of the models considered here (more so in the realm of deep learning), their applicability across domains, and their use in a consistent manner to compare/model similar situations. One major constraint is that large-scale deep learning models, such as CNN-LSTM hybrids and transformers have a low level of interpretability while smaller-scale models have low predictive capability. The continual presence of some trade-off between accuracy and interpretability should be noted as well (this continues to be evident in many research studies). Of the 25 research studies reviewed, only 6 of them implemented explainable Artificial Intelligence techniques; furthermore, only a few of them considered alternative model architectures that are more inherently interpretable. Across industries and asset types, there is limited evidence to support that the generalizability of the data generated from these models can be extended beyond rotating machinery in manufacturing; however, the opportunity to improve performance in data-scarce environments, through the use of cross-asset transfer learning, has not received sufficient attention. Finally, the significant absence of a standardized set of evaluation protocols makes it very difficult to compare the results from the various studies, resulting from the lack of consistency in preprocessing of data, train-test split of the dataset, and method of performance metrics. The development of integrated multi-physics models, and digital twins remains in its infancy but could potentially improve learning efficiencies and generalizability of models by embedding domain knowledge in data-driven frameworks.

practical impediment to data sharing in competitive industrial settings. A high-priority research direction is to develop predictive maintenance-specific federated architectures that can address non-IID data distributions, equipment heterogeneity, and communication bandwidth constraints (Li et al., 2020).

Large-scale pre training and foundation models on industrial time-series data provide the promise of general-purpose predictive maintenance models with high zero-shot and few-shot transfer properties- corresponding to the effect of large language models in NLP. By building industrial time-series foundation models trained on a wide variety of equipment types

and operating conditions then further fine-tuning on a small set of tasks may significantly decrease the amount of labeled data currently needed to deploy models to well-instrumented, data-rich industrial environments.

Neuromorphic computing architectures based on biological neural systems provide a fundamentally different implementation of edge AI inference, processing information with sparse spike-timing patterns, which require orders of magnitude less energy than traditional digital computing. Spiking neural networks around time-series anomaly detectors and fault classification may allow continuous monitoring of battery-powered wireless sensor nodes that currently do not have the computational resources to perform any ML inference, expanding the range of IIoT predictive maintenance coverage by a significant factor.

Application to large language models (LLM) as natural language interfaces to predictive maintenance systems, where maintenance engineers can query model predictions, seek explanations, and interpret historical patterns in the form of conversational interfaces, is a novel area of application that can significantly decrease the amount of domain knowledge needed to successfully operate a PdM system, and increase the accessibility of AI-driven insights to non-expert staff.

8. Discussion

The systematic review that is provided in this paper demonstrates a multidimensional and consistent story of advances in AI-based predictive maintenance over the last decade. The historical trend of statistical regression methods to classical ensemble methods to deep learning architectures and hybrid models can be seen as both the coming of age of machine learning as a field and the facilitating effect of growing IIoT data infrastructure to have complex models trainable and, more and more, deployable.

The gains in performance observed in sequential model generations are large and significant: the increase in accuracy of SVM-based fault classifiers of 90–93 percent to CNN-LSTM hybrid systems of over 98 percent are not just benchmark overfitting. Likewise, the decrease in RUL estimation RMSE between 40–60 cycles in earlier statistical methods and 10–15 cycles in current state-of-the-art transformer models indicates advances to temporal modelling capacity that can be converted into operationally significant and significantly earlier maintenance prescribed interventions.

Nonetheless, another important gap identified in the review is between the laboratory performance and the operational deployability. Models with the highest reported accuracy are often the least easily IIoT-integrable: not only do they need prohibitive computational power to deploy to the edge, but they also need extensive labeled training data which most industrial assets do not have, are not interpretable to be accepted by human experts, and often have not been tested on the distribution shift conditions that characterize real industrial operation. The most significant challenge to the field in the next decade is this gap between the performance of research and deployment readiness.

The advent of compression-aware lightweight architectures; federated learning models; and explainable AI methods are promising that the research community will start tackling the deployment gap in a systematic manner. More rigorous comparative assessment has also been made possible by the increased availability of open benchmark datasets (CWRU, CMAPSS, PRONOSTIA, MIMII) but the standardization of evaluation protocols has not been fully achieved. The community call to come up with shared evaluation systems similar to the case in computer vision and NLP benchmarking communities would be of help to the field.

9. Conclusion

The purpose of this systematic review of literature was to examine how AI and Machine Learning Models can predict when to perform maintenance on industrial IoT systems. Reviewed articles were published between 2010–2024, a total of 25 journal articles. The review follows a clear developmental path: starting with classical statistical and machine learning methods, which need to perform manual feature engineering, to CNN and LSTM architectures which can learn representations directly on sensor data, to modern hybrid models and transformer architectures which push the limits of diagnostic and prognostic accuracy.

There are a number of major conclusions which can be drawn out of this analysis. To begin with, there cannot be a single model architecture that is superior in all the facets of predictive maintenance performance: classical ensemble models (Random Forest, XGBoost) are both more interpretable and efficient in terms of computation and edge deployability in the setting where the availability of labeled data is moderate and less strict in terms of accuracy demands. Deep CNN-LSTM hybrid models and attention-enhanced models are most accurate in diagnoses but also consume large amounts of computational resources and large amounts of training data. The unsupervised auto encoder methods have the unique advantage of finding anomalies in the typical industrial problem case where there is no labeled fault data.

Second, the history of model evolution has been a dialectic of increasingly better performance and practicality of deployment: each new generation of more capable architectures has necessitated the next development of compression, adaptation, and efficiency methods to regain feasibility of deployment. Moving forward, the trend of creating an innovative performance and then creating an innovative deployment will probably continue due to the introduction of foundation models and neuromorphic computing techniques into predictive maintenance.

In addition to the choice of model architecture, this area is facing unsolved problems related to the following: data quality and availability; adversarial robustness; cross-asset generalization; interpretability; and seamless integration with IIoT. All of these areas need work done on them and continue to be actively researched. Collaborative solutions among machine learning researchers, industrial engineers, cybersecurity researchers, and domain experts will need to be created in order to address these issues. There are very few examples of such collaboration currently published in the literature.

This review offers researchers getting started in predictive maintenance, practitioners who are choosing and using AI-based maintenance systems, and technology strategists looking for areas to invest in industrial AI a structured framework. There is an ongoing evolution of the Industrial Internet of Things (IIoT) systems, along with an increase in the number of data. This has opened many opportunities for AI-based predictive maintenance to change how operations are conducted in industry. The research community is continuously developing tools that will allow for the application of this potential in the manufacturing environment.

References

- Abdeljabber, O., Avci, O., Kiranyaz, S., Gabbouj, M., & Inman, D. J. (2017). Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. *Journal of Sound and Vibration*, 388, 154–170. <https://doi.org/10.1016/j.jsv.2016.10.043>
- Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. *Computer Networks*, 54(15), 2787–2805. <https://doi.org/10.1016/j.comnet.2010.05.010>
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, X., Zhang, B., & Gao, D. (2021). Bearing fault diagnosis base on multi-scale CNN and LSTM model. *Journal of Intelligent Manufacturing*, 32(4), 971–987. <https://doi.org/10.1007/s10845-020-01600-2>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Cheng, Y., Zhu, H., Wu, J., & Shao, X. (2020). Machine health monitoring using adaptive kernel spectral clustering and deep long short-term memory recurrent neural networks. *IEEE Transactions on Industrial Informatics*, 16(1), 566–577. <https://doi.org/10.1109/TII.2019.2913905>
- Deloitte Insights. (2017). Predictive maintenance and the smart factory. Deloitte Touche Tohmatsu Limited. <https://www2.deloitte.com/insights/us/en/focus/industry-4-0/using-predictive-technologies-for-asset-maintenance.html>
- Gebraeel, N. Z., Lawley, M. A., Li, R., & Ryan, J. K. (2005). Residual-life distributions from component degradation signals: A Bayesian approach. *IIE Transactions*, 37(6), 543–557. <https://doi.org/10.1080/07408170590929018>
- Guo, L., Lei, Y., Xing, S., Yan, T., & Li, N. (2022). Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. *IEEE Transactions on Industrial Electronics*, 66(9), 7316–7325. <https://doi.org/10.1109/TIE.2018.2877090>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Hoang, D. T., & Kang, H. J. (2019). Rolling element bearing fault diagnosis using convolutional neural network and vibration image. *Cognitive Systems Research*, 53, 42–50. <https://doi.org/10.1016/j.cogsys.2018.03.002>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Javed, K., Gouriveau, R., Zerhouni, N., & Nectoux, P. (2021). Enabling health monitoring approach based on vibration data for accurate prognostics. *IEEE Transactions on Industrial Electronics*, 62(1), 647–656. <https://doi.org/10.1109/TIE.2014.2327917>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 30, 3146–3154.
- Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31, 249–268.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2014). Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. *Mechanical Systems and Signal Processing*, 42(1–2), 314–334. <https://doi.org/10.1016/j.ymssp.2013.06.004>
- Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3), 50–60. <https://doi.org/10.1109/MSP.2020.2975749>
- Li, X., Ding, Q., & Sun, J. Q. (2018). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety*, 172, 1–11. <https://doi.org/10.1016/j.res.2017.11.021>
- Liu, R., Yang, B., Zio, E., & Chen, X. (2020). Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mechanical Systems and Signal Processing*, 108, 33–47. <https://doi.org/10.1016/j.ymssp.2018.02.016>
- Mobley, R. K. (2002). *An introduction to predictive maintenance* (2nd ed.). Butterworth-Heinemann.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, 6(7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Morello, B., Zerhouni, N., & Varnier, C. (2012). PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In *Proceedings of the IEEE International Conference on Prognostics and Health Management* (pp. 1–8). IEEE.
- Peng, Y., Dong, M., & Zuo, M. J. (2010). Current status of machine prognostics in condition-based maintenance: A review. *The International Journal of Advanced Manufacturing Technology*, 50(1–4), 297–313. <https://doi.org/10.1007/s00170-009-2482-0>
- Principi, E., Drude, L., Nagata, K., Scheibler, R., Watanabe, S., & Squartini, S. (2019). Unsupervised electric motor fault detection by using deep auto encoders. *IEEE/CAA Journal of Automatica Sinica*, 6(2), 441–451. <https://doi.org/10.1109/JAS.2019.1911393>
- Qi, Q., & Tao, F. (2018). Digital twin and big data towards smart manufacturing and Industry 4.0: 360-degree comparison. *IEEE Access*, 6, 3585–3593. <https://doi.org/10.1109/ACCESS.2018.2793265>
- Serradilla, O., Zugasti, E., Rodriguez, J., & Zurutuza, U. (2022). Deep learning models for predictive maintenance: A survey, comparison, challenges and prospects. *Applied Intelligence*, 52(10), 10934–10964. <https://doi.org/10.1007/s10489-021-03004-6>
- Shao, S., McAleer, S., Yan, R., & Baldi, P. (2019). Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Transactions on Industrial Informatics*, 15(4), 2446–2455. <https://doi.org/10.1109/TII.2018.2864759>
- Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, 30, 5998–6008.
- Wen, L., Li, X., Gao, L., & Zhang, Y. (2018). A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Transactions on Industrial Electronics*, 65(7), 5990–5998. <https://doi.org/10.1109/TIE.2017.2774777>

Wen, L., Li, X., Gao, L., & Zhang, Y. (2018). A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Transactions on Industrial Electronics*, 65(7), 5990–5998. <https://doi.org/10.1109/TIE.2017.2774777>

Widodo, A., & Yang, B.-S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, 21(6), 2560–2574. <https://doi.org/10.1016/j.ymssp.2006.12.007>

Wu, J., Wu, C., Cao, S., Or, S. W., Deng, C., & Shao, X. (2021). Degradation data-driven time-to-failure prognostics approach for rolling element bearings in electrical machines. *IEEE Transactions on Industrial Electronics*, 66(1), 529–539. <https://doi.org/10.1109/TIE.2018.2811366>

Wu, Y., Yuan, M., Dong, S., Lin, L., & Liu, Y. (2018). Remaining useful life estimation of engineered systems using vanilla LSTM neural networks. *Neurocomputing*, 275, 167–179. <https://doi.org/10.1016/j.neucom.2017.05.063>

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2021). A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 2114–2124). ACM. <https://doi.org/10.1145/3447548.3467401>

Zhang, W., Li, C., Peng, G., Chen, Y., & Zhang, Z. (2017). A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mechanical Systems and Signal Processing*, 100, 439–453. <https://doi.org/10.1016/j.ymssp.2017.06.022>

Zhang, Y., Zhou, T., Arghandeh, R., & Spanos, C. J. (2016). Partial mutual information-based input variable selection for machine health prognosis. *Industrial & Engineering Chemistry Research*, 55(50), 12987–12994. <https://doi.org/10.1021/acs.iecr.6b02578>

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115, 213–237. <https://doi.org/10.1016/j.ymssp.2018.05.050>

Zhao, R., Yan, R., Wang, J., & Mao, K. (2017). Learning to monitor machine health with convolutional bi-directional LSTM networks. *Sensors*, 17(2), 273. <https://doi.org/10.3390/s17020273>

Zheng, S., Ristovski, K., Farahat, A., & Gupta, C. (2017). Long short-term memory network for remaining useful life estimation. In *Proceedings of the 2017 IEEE International Conference on Prognostics and Health Management* (pp. 88–95). IEEE. <https://doi.org/10.1109/ICPHM.2017.7998311>