Transformer-Based Architectures for Machinery Prognostics: A Review

Maxime Pierfederici[1], Mayank Shekhar Jha[2*], Chetan Kulkarni[3], and Didier Theilliol[2]

[1] *Polytech Nancy, University of Lorraine, France*

[2] *Université de Lorraine, CNRS, CRAN, Nancy, F-54000, France*
mayank-shekhar.jha@univ-lorraine.fr
didier.theilliol@univ-lorraine.fr

[3] *KBR Inc., NASA Ames Research Center, Mountain View, California, USA*
chetan.s.kulkarni@nasa.gov

ABSTRACT

Machinery prognostics requires robust modeling of multivariate degradation signals under noise, non-stationarity, variable operating conditions, and limited run-to-failure labels. Transformer-based deep learning architectures have recently attracted strong interest because self-attention can capture long-range temporal dependencies and inter-sensor interactions more directly than purely recurrent or convolutional models. This focused review presents Transformer-based approaches for machinery prognostics, with emphasis on remaining useful life (RUL) estimation and degradation representation learning. The literature is organized using a consistent taxonomy covering PHM task, Transformer backbone, hybridization strategy, and input representation. We also analyze preprocessing choices that strongly influence performance, including windowing, health-indicator construction, tokenization, embedding, and positional encoding. Across benchmark datasets, studied studies frequently show gains from Transformers and hybrid attention models, especially when long temporal context and multivariate dependencies are central. However, improvements are not universal and remain sensitive to evaluation protocol, signal representation, and model complexity. Key open challenges include data efficiency, computational cost, cross-condition generalization, interpretability, and uncertainty quantification. The review concludes by identifying methodological gaps in the current literature and outlining research directions for robust, efficient, and deployable Transformer-based prognostics.

1. INTRODUCTION

Prognostics and Health Management (PHM) has become a major research area in predictive maintenance, with the central objective of estimating degradation and predicting the Remaining Useful Life (RUL) of systems and components prior to functional failure (Sikorska, Hodkiewicz, & Ma, 2011). More broadly, PHM aims to support maintenance planning, operational reliability, and decision-making through continuous condition monitoring, early detection of fault precursors or incipient degradation, and assessment of the future health state of engineering assets. Within this framework, diagnostics and prognostics constitute two complementary but conceptually distinct functions that underpin effective PHM architectures (Jha, Bressel, Ould-Bouamama, & Dauphin-Tanguy, 2016). Diagnostic tasks are primarily concerned with fault detection, isolation, and identification, typically by comparing measured sensor signals with expected or nominal system behavior derived from process knowledge, reference models, or historical operating conditions (Jha, Dauphin-Tanguy, & Ould-Bouamama, 2016; Jardine, Lin, & Banjevic, 2006). Such analyses commonly involve variables such as temperature, pressure, vibration, or current, with the aim of identifying abnormal operating regimes at an early stage. Prognostics, in contrast, is specifically concerned with forecasting the future evolution of degradation and inferring the RUL, ideally together with a characterization of the associated uncertainty (Jha, Bressel, et al., 2016; Kanso, Jha, Galeotta, & Theilliol, 2022). Existing prognostic methodologies are generally classified into three broad categories: model-based approaches, which exploit physics-based degradation models; data-driven approaches, which infer degradation trends directly from measured data; and hybrid approaches, which combine partial physical knowledge, empirical degradation descriptions, and

Table 1. List of acronyms used in the manuscript.

Acronym	Full form / meaning
PHM and application terms	
PHM	Prognostics and Health Management
RUL	Remaining Useful Life
SOH	State of Health
HI	Health Indicator
OSA-CBM	Open System Architecture for Condition-Based Maintenance
PEM	Proton Exchange Membrane
Learning architectures and model families	
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
Bi-LSTM	Bidirectional Long Short-Term Memory
GRU	Gated Recurrent Unit
TCN	Temporal Convolutional Network
TCNN	Temporal Convolutional Neural Network
ConvLSTM	Convolutional Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-trained Transformer
ViT	Vision Transformer
DAST	Dual-Aspect Self-Attention Transformer
FTT	Fast Temporal Transformer
RMTF	Recurrence Multi-Information Time-Frequency Transformer
Normalization, signal processing, and evaluation	
LN	Layer Normalization
BN	Batch Normalization
RMSNorm	Root Mean Square Layer Normalization
EMD	Empirical Mode Decomposition
IMF	Intrinsic Mode Function
MSE	Mean Squared Error
RMSE	Root Mean Square Error
AUC	Area Under the Curve
F1	F1 score
DOI	Digital Object Identifier
Datasets, platforms, and benchmark names	
C-MAPSS	Commercial Modular Aero-Propulsion System Simulation
PRONOSTIA/PHM 2012	Bearing accelerated degradation test platform / dataset
XJTU-SY	Xi'an Jiaotong University–Sumyoung bearing dataset
SWaT	Secure Water Treatment
WADI	Water Distribution
MIT	Massachusetts Institute of Technology
HUST	Huazhong University of Science and Technology

estimation-theoretic tools to fuse *a priori* information with online observations (Sikorska et al., 2011; Chelouati, Jha, Galeotta, & Theilliol, 2021). In many hybrid strategies, the State of Health (SOH) is first estimated and then propagated over a prediction horizon to derive the corresponding RUL, an approach that generally presupposes the availability of sufficiently informative degradation models for reliable inference (Thuillier, Jha, Le Martelot, & Theilliol, 2024; Kanso et al., 2022). Over the past decade, deep learning (DL) has emerged as a particularly promising paradigm for prognostics owing to its strong representational capacity for nonlinear, high-dimensional, and non-stationary degradation processes encountered in complex engineering systems (Fink et al., 2020). In contrast to model-based and hybrid methods, DL-based prognostics typically aims to learn a direct mapping from raw or processed measurements to the target RUL, without explicitly requiring a first-principles or empirically derived degradation model (de Beaulieu, Jha, Garnier, & Cerbah, 2022, 2024; Jha, Theilliol, Belleoud, & Oriol, 2025). Representative early contributions include the convolutional neural network (CNN)-based regression framework introduced in (Babu, Zhao, & Li, 2016) for RUL estimation from multivariate sensor time series. Subsequently, (Zheng, Ristovski, Farahat, & Gupta, 2017) stated that Long Short-Term Memory (LSTM) networks could outperform CNN-based approaches, arguing that CNN models built on independent sliding windows do not fully capture the temporal dependencies and sequential degradation structure that are intrinsic to accurate RUL prediction.

Conventional deep-learning architectures, including recurrent models, face limitations when modeling dependencies over long horizons. These challenges have contributed to the emergence of Transformer architectures, which are generally better suited to parallel sequential processing. Transformers can model long-range temporal dependencies while enabling parallel training. Transformers build upon the attention prin-

ciple introduced in (Vaswani et al., 2017), which assigns varying importance to sequence elements, emphasizes salient information, and enables direct interactions between distant observations. By explicitly evaluating relationships between every pair of tokens, Transformers can identify the elements that are most informative for representation learning and prediction (Wen et al., 2022). Although initially developed for natural language processing, Transformers have been adapted to time-series learning and PHM through multiple structural choices (encoder-only models akin to BERT, decoder-only models akin to GPT, and encoder–decoder models), as well as hybrid variants combining CNNs and/or RNNs to exploit complementary inductive biases. Beyond architectural design, the performance of Transformer-based approaches also depends on methodological decisions, including signal representation, windowing strategies, attention-block parameterization, and normalization.

Historically, CNNs were among the first deep architectures considered for prognostics, due to their ability to capture nonlinear degradation patterns under varying operating conditions (Suh, Jang, Won, Jha, & Lee, 2020; Mittal et al., 2023; Suh et al., 2024). For instance, (Liu, Hsaio, & Tu, 2019) proposed a multivariate convolutional neural network tailored to multivariate sequential time series, integrating both multivariate features and lag characteristics. However, standard CNNs primarily focus on local patterns, and may neglect long-range temporal dependencies unless substantially deepened. To address temporal modeling more explicitly, recurrent neural networks (RNNs) and their gated variants, such as long short-term memory (LSTM) and gated recurrent unit (GRU) networks, have been extensively adopted. Hybrid combinations have also been proposed, e.g., CNN–LSTM–GRU fusion for aircraft-engine RUL prediction (Patra, Sethi, & Behera, 2025), Bi-LSTM with attention and transfer learning for rolling-bearing RUL prediction (Dong et al., 2023), and other CNN/LSTM variants that aim to enrich temporal context. Nevertheless, these models can exhibit structural complexity and high computational cost, and they remain challenged by long-range dependency modeling, in particular due to optimization issues such as vanishing gradients in recurrent training (H. Zhang et al., 2022). These limitations motivate attention-based architectures capable of processing long sequences while improving memory of distant yet relevant historical information.

This paper provides a structured and critical review of Transformer architectures applied to PHM. We analyze and compare major model families studied in the literature, and we contrast them with conventional convolutional and recurrent methodologies on widely used benchmarks. The goal is to clarify the effectiveness of Transformers for prognostics, their practical contributions and limitations, and promising research directions, particularly those related to data preprocessing and model-learning strategies.

The novelty of this review lies in its PHM-specific synthesis of Transformer-based prognostics from the joint perspective of PHM task, Transformer backbone, hybridization strategy, and signal representation. Unlike general surveys of Transformers for time-series learning, this paper focuses on how Transformer design choices interact with PHM-specific requirements, including run-to-failure data scarcity, variable operating conditions, degradation representation, benchmark-protocol differences, and deployment constraints. The main contributions are: (i) a clarified taxonomy of Transformer-based PHM studies according to task, backbone, hybridization, and input representation; (ii) a synthesis of preprocessing and tokenization choices that directly affect prognostic performance; (iii) a comparison of studied benchmark results on C-MAPSS, PHM2012/PRONOSTIA, and XJTU-SY; and (iv) a discussion of practical limitations and research directions for robust and deployable Transformer-based PHM.

Organization of the paper. The remainder of this paper is organized as follows. Section 2 defines the review scope, the core corpus, and the inclusion criteria used to study the Transformer-based PHM literature. It also relates the proposed taxonomy to a PHM/OSA-CBM-style pipeline and provides a quantitative overview of the reviewed studies. Section 3 presents the Transformer-based PHM pipeline, progressing from benchmark datasets and PHM tasks to data preprocessing, tokenization, embedding, positional encoding, attention mechanisms, and Transformer architecture families. Section 4 summarizes comparative evidence from widely used PHM benchmarks, including C-MAPSS, PHM2012/PRONOSTIA, and XJTU-SY, while emphasizing the limitations of direct numerical comparison across studies with different preprocessing and evaluation protocols. Section 5 then synthesizes the reviewed architectures at a higher level by comparing their typical PHM uses, strengths, limitations, and practical implications. Finally, the conclusion summarizes the main methodological lessons and outlines research directions for robust, efficient, interpretable, and deployable Transformer-based prognostics.

2. REVIEW SCOPE, CORPUS, AND TAXONOMY

2.1. Corpus definition and inclusion criteria

The quantitative part of this review is based on a core corpus of 38 Transformer-PHM studies selected. A study is included or elaborated in this core corpus when it explicitly proposes, adapts, or evaluates a Transformer or attention-based neural architecture for a PHM-related task, including RUL estimation, degradation forecasting, fault diagnosis, anomaly detection, or predictive-maintenance decision support. Additional references are cited to provide background on PHM, deep learning, benchmark datasets, and general time-series Transformers; however, those background references are not

included in the quantitative body counts studied in Table 2.

2.2. Relation to the PHM/OSA-CBM pipeline

The taxonomy used here is not intended to replace established PHM process models such as OSA-CBM. Instead, it provides a coding framework for organizing Transformer-based studies. In relation to a PHM pipeline, data acquisition is followed by data manipulation, preprocessing, and feature representation; Transformer representation learning then supports state detection, health assessment, diagnosis, degradation forecasting, or RUL regression; finally, decision-layer studies connect prognostic outputs to maintenance advisories or operational planning.

2.3. Taxonomy used for literature coding

The reviewed studies are organized using four consistent and explicitly separated dimensions. The first dimension is the PHM task: prognostics/RUL estimation, fault diagnosis, anomaly detection, or predictive-maintenance decision support. The second dimension is the Transformer backbone: encoder-only, encoder–decoder, efficient/long-sequence, ViT-style, graph-style, or related variants. The third dimension is the hybridization strategy, because many PHM models combine a Transformer with a CNN, TCN, RNN, LSTM, GRU, decomposition block, or frequency-domain module. The fourth dimension is the input representation, including raw or windowed multivariate sensor sequences, patch-based tokens, health indicators, time-frequency representations, and multimodal or topology-aware representations. Operational issues such as domain shift, scarce run-to-failure data, long-sequence computational cost, interpretability, and uncertainty quantification are treated as cross-cutting PHM challenges rather than as taxonomy axes.

2.4. Quantitative overview of the core corpus

Table 2 summarizes the distribution of the 38-study works that have been studied in this paper. The counts are descriptive and non-exclusive in that a single study may contribute to more than one PHM task or hybridization category.

3. TRANSFORMER-BASED PHM PIPELINE

3.1. Benchmark datasets and PHM tasks

In this study, we focus on works using widely adopted PHM benchmarks, C-MAPSS (Saxena, Goebel, Simon, & Eklund, 2008a), PHM2012/PRONOSTIA (Nectoux et al., 2012), and XJTU-SY (Lei et al., 2019), due to their popularity and shared characteristics, namely multivariate time series and run-to-failure trajectories. C-MAPSS is based on simulated degradation of aircraft turbofan engines and is organized into four main subsets commonly referred to as FD001, FD002, FD003, and FD004; each subset corresponds to a specific operating and degradation scenario (Saxena, Goebel, Simon, & Eklund, 2008b). For PHM2012, released in the context of the IEEE PHM Data Challenge, we rely on studies using the PRONOSTIA dataset (Nectoux et al., 2012), which focuses on accelerated degradation of ball bearings. The XJTU-SY dataset similarly addresses rolling-element bearing degradation and RUL prognostics (Lei et al., 2019). We also consider architectures evaluated on less standardized datasets, as studied in (H. Zhang et al., 2022; Biggio, Bendinelli, Kulkarni, & Fink, 2022; Zerveas, Jayaraman, Patel, Bhamidipaty, & Eickhoff, 2020; Jiao, Pan, Fan, Xu, & Chen, 2022).

Most bearing-oriented Transformer prognostics are validated on public run-to-failure datasets (notably PHM2012 or PRONOSTIA and XJTU-SY), often alongside smaller industrial datasets to test domain realism (C. Jin et al., 2025; M. Zhang, He, Huang, & Yang, 2024). Battery-oriented Transformer studies typically rely on multi-battery cycling datasets and highlight knee-point behavior and regeneration effects; recent work explicitly uses MIT/HUST-type datasets for multimodal attention modeling (Suh et al., 2024). Beyond prognostics, Transformer-based anomaly detection has become a competitive baseline for multivariate industrial monitoring and is frequently evaluated on process-control benchmarks.

3.2. Data preprocessing and representation for Transformer-based PHM

In a PHM pipeline, preprocessing is not merely an implementation step but a key modeling decision that determines what degradation information is available to the Transformer. Before attention is applied, raw sensor streams are typically normalized, segmented into windows, aligned with RUL labels or health states, and converted into token sequences. These operations correspond to the data manipulation and feature representation stages of a PHM pipeline and strongly influence whether the model learns degradation-relevant temporal dependencies or spurious operating-condition correlations.

3.2.1. Tokenization

Transformers operate on sequences of discrete elements, referred to as *tokens*. Unlike natural language processing, where tokens correspond to linguistic units, PHM data typically consist of continuous multivariate time series, where each variable represents a physical measurement associated with system health (e.g., vibration, temperature, current, or pressure). A representation step is therefore required to convert raw signals into sequences compatible with Transformer processing.

3.2.2. Patch-based tokenization

Most tokenization methods for time series build input sequences by segmenting signals either at each time step or

Table 2. Revised quantitative overview of the core Transformer-based PHM works.

Coding dimension	Category	Count	Interpretation / counting note
Corpus size	Core Transformer-PHM studies	38	Total corpus used for descriptive coding.
PHM task	Prognostics / RUL	26	Non-exclusive; some studies address more than one task.
PHM task	Fault diagnosis	8	Includes classification-oriented PHM studies.
PHM task	Anomaly detection	4	Includes screening/monitoring studies relevant to PHM.
PHM task	Predictive-maintenance decision layer	1	Study explicitly linking model outputs to maintenance decisions.
Transformer backbone	Encoder-only Transformer	18	Largest single backbone group; not an absolute majority of the 38 papers.
Transformer backbone	Encoder–decoder Transformer	3	Primarily used for sequence-to-sequence forecasting or reconstruction.
Transformer backbone	Efficient / long-sequence variant	4	Attention modified to reduce long-sequence cost.
Transformer backbone	ViT-style Transformer	1	Image-like or patch-based representation.
Transformer backbone	Graph Transformer	1	Topology-aware or graph-structured representation.
Hybridization strategy	CNN/TCN–Transformer hybrid	6	Adds local temporal or vibration-pattern inductive bias.
Hybridization strategy	RNN/LSTM/GRU–Transformer hybrid	6	Adds sequential memory or short-range temporal modeling.
Input representation	Raw/windowed multivariate time series; health indicators; time-frequency/decomposition features; multimodal or topology-aware tokens	–	Discussed qualitatively in Section 3.2; mutually exclusive counts should be added only after separate recoding of all 38 studies.

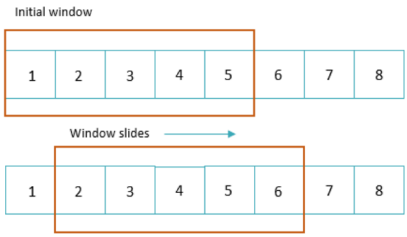

Figure 1. Sliding-window tokenization process, adapted from (Ogunfowora & Najjaran, 2023).

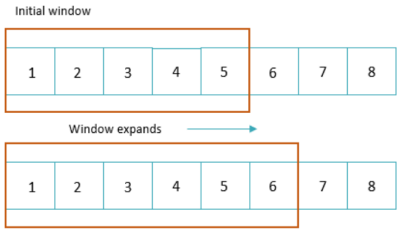

Figure 2. Expanding-window tokenization process, adapted from (Ogunfowora & Najjaran, 2023).

using temporal windows. Fixed-length *sliding windows* are widely used (M. Zhang et al., 2024; Wang, Cheng, & Song, 2021; Chirukiri, Cheerala, Kanta, Karim, & Damacharla, 2025; Hu, Zhao, & Ren, 2023; Biggio et al., 2022; Zerveas et al., 2020), as illustrated in Fig. 1. Variable-length *expanding windows* have also been considered (Ogunfowora & Najjaran, 2023; Zerveas et al., 2020; Pour, Karimi, & Mazloumi, 2025), as shown in Fig. 2. Expanding windows facilitate batching sequences of varying lengths by padding shorter sequences (e.g., with zeros) and applying padding masks to prevent the self-attention mechanism from attending to padded positions.

When applied directly to raw input sequences, windowing can yield very long and highly redundant token sequences, which increases computational burden due to the quadratic cost of standard self-attention. Furthermore, naive patching may fail to explicitly capture nonlinear temporal motifs, non-stationary regimes, or recurring events that are informative for degradation modeling.

3.2.3. Health-indicator tokenization

To alleviate the limitations of patch-based tokenization, multiple studies construct a health indicator (HI) (H. Zhang et al., 2022; X. Jin et al., 2025; C. Jin et al., 2025; Deng et al., 2022; Peng, Jiang, Mao, & Liu, 2023; Sun et al., 2024). The objective is to extract informative features from raw signals in the time, frequency, and time-frequency domains, and to retain those that effectively capture the global degradation state.

The resulting HI is then segmented into temporal windows that represent successive stages of the degradation process; these segments serve as input tokens to the Transformer.

3.2.4. Advanced tokenization strategies

Beyond windowing, explicit tokenization methods aim to reduce sequence length while preserving salient patterns. The goal is to convert continuous time series into compact token sequences representing meaningful motifs (e.g., oscillations, peaks, plateaus). In (Talukder, Yue, & Gkioxari, 2025), segments are mapped to discrete tokens via a supervised codebook. Approaches inspired by byte-pair encoding (Götz, Kollovieh, Günnemann, & Schwinn, 2025) build a variable-length motif vocabulary by iteratively merging frequently adjacent symbols. Other works exploit specialized tokenizers to produce patch sequences that are more informative along the temporal direction (Chang, Li, Chen, Liu, & Li, 2022; Che, Lu, Bao, Zhang, & Liu, 2023; Jiao et al., 2022). Once tokens are defined, they must be encoded into numerical vectors suitable for Transformer processing. For continuous signals, this step typically includes normalization and the construction of embedding vectors that feed the Transformer embedding layer.

3.2.5. Embedding and positional encoding

Following tokenization, each token is mapped to a dense vector of size d. In PHM, where tokens typically consist of real-valued vectors derived from multivariate signals, this projection is most often implemented using a linear mapping (Ogunfowora & Najjaran, 2023; X. Jin et al., 2025; C. Jin et al., 2025; Zerveas et al., 2020; Chirukiri et al., 2025; Pour et al., 2025; Biggio et al., 2022; Che et al., 2023), which is frequently considered sufficient.

However, patch embedding based solely on linear projection (or a shallow convolution) may not capture temporal correlations effectively (M. Zhang et al., 2024). These correlations may evolve with the degradation level, thereby affecting RUL estimation. Several studies therefore introduce adaptive weighting or enriched embeddings to better inform the attention mechanism (M. Zhang et al., 2024; Götz et al., 2025; C. Jin et al., 2025).

Manually designed positional encoding. Unlike recurrent or convolutional architectures, Transformers do not intrinsically encode token order. In prognostics, temporal ordering is critical because degradation evolution is inherently sequential. A positional encoding is therefore added to token embeddings to inject temporal information. Following (Vaswani et al., 2017), sinusoidal positional encodings are commonly used in time-series Transformers.

Learned positional encoding. Fully learnable positional encodings are also possible (Zerveas et al., 2020). This choice is motivated by the hypothesis that temporal order may be implicitly represented through the continuity and temporal correlations of multivariate signals, allowing the model to infer temporal structure primarily from input data.

Positional information embedded in tokens. Alternatively, temporal ordering can be partially encoded during tokenization by applying operations such as 3D convolutions (M. Zhang et al., 2024) or multi-layer convolutions (Deng et al., 2022). Such preprocessing can enrich tokens with local temporal features prior to attention, potentially improving the identification of relevant temporal correlations.

3.3. Transformer building blocks

3.3.1. Canonical Transformer architecture

A canonical Transformer consists of two main components: an encoder and a decoder. The encoder maps an input sequence to an intermediate representation that captures the overall context while emphasizing relevant information. This transformation is performed progressively through stacked layers, enabling increasingly abstract feature extraction, a property that is particularly valuable in PHM, where sensor signals are noisy, complex, and evolve over time. Conversely, the decoder leverages the encoder representation together with its own previous outputs to generate the target sequence in an auto-regressive manner, refining predictions at each stage to yield temporally coherent forecasts.

The original Transformer architecture (Vaswani et al., 2017) comprises an encoder and a decoder built from repeated identical layers. Each layer includes a multi-head attention block, a position-wise feed-forward network, residual connections, and layer normalization. The attention block captures dependencies across sequence elements, residual connections facilitate information flow, the feed-forward network transforms representations, and normalization stabilizes activations (Fig. 3).

In PHM-oriented Transformers, layer normalization (LN) remains the most commonly used normalization technique. However, LN may be sensitive to outliers and can incur non-negligible computational overhead. Moreover, when applied across many stacked layers, normalization can interact with attention statistics and degrade training stability. Alternative normalization schemes have therefore been explored. Root mean square layer normalization (RMSNorm) (B. Zhang & Sennrich, 2019) reduces the normalization cost by relying on the root mean square of activations rather than the full mean-variance normalization. Batch normalization (BN) has also been considered as a potential mitigation for outlier sensitivity; nevertheless, its effectiveness strongly depends on

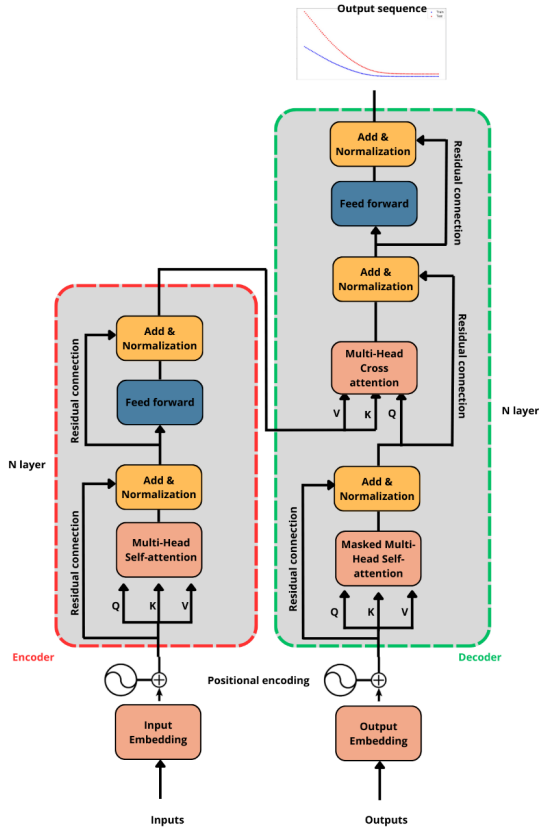

Figure 3. Canonical Transformer architecture with encoder and decoder blocks.

batch statistics and data characteristics, which can limit its applicability for non-stationary PHM time series. More recently, UnitNorm (Huang, Kümmerle, & Zhang, 2024) was proposed as a time-series-oriented normalization strategy for Transformers, aiming at improving stability and information propagation in long-range dependency modeling.

In PHM time series, purely global self-attention may underrepresent sharp local transients (e.g., impulsive vibration signatures). Multiple PHM studies therefore augment self-attention with convolutional or temporal-convolution modules to capture local structures while preserving long-range dependencies (Ding & Jia, 2022; X. Jin et al., 2025).

3.3.2. Scaled dot-product attention

In deep learning, attention mechanisms enable a model to dynamically weight the elements of a sequence and to extract task-relevant information. In Transformers, attention is commonly implemented as scaled dot-product attention (Vaswani et al., 2017):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \qquad (1)$$

where Q, K, and V denote the query, key, and value matrices, respectively, obtained by learned linear projections of the input representations. Queries encode the information sought by the model; keys act as descriptors to assess the relevance of available information; values contain the content that is aggregated to produce context-aware representations.

The dot product $QK^\top$ measures similarity between each query and all keys, producing attention scores that are normalized by $\sqrt{d_k}$ (with d_k the key dimension) to stabilize learning and prevent saturation of the softmax function. The resulting attention weights are used to compute a weighted combination of values, yielding a contextual representation that integrates relevant dependencies across the sequence.

3.3.3. Self-attention and multi-head attention

Self-attention is the core mechanism enabling Transformers to model long-range dependencies efficiently. Unlike recurrent architectures that process data sequentially, self-attention considers all tokens in parallel and establishes direct interactions between any pair of positions, regardless of temporal distance. In self-attention, queries, keys, and values are derived from the same sequence, so that each position attends to other positions in the sequence and receives a representation enriched by global context.

To enhance representational capacity, Transformers employ *multi-head attention*, which performs multiple self-attention operations in parallel across different representation subspaces. In PHM, multi-head attention is particularly suitable for multivariate time series, as it can jointly model temporal dependencies and inter-sensor correlations. Attention maps may also support a degree of interpretability by highlighting time indices and variables that contribute most to a prediction.

3.3.4. Masked self-attention and cross-attention

Masked self-attention prevents certain positions from attending to future tokens, thereby enforcing causal dependence. In PHM, masked attention is primarily used in decoder components (Hu et al., 2023; Biggio et al., 2022; Deng et al., 2022; C. Jin et al., 2025) to ensure that predictions at time t are based only on information available up to t (Fig. 4).

Cross-attention enables a model to align and integrate information from different sequences. In encoder–decoder Transformers, cross-attention occurs in the decoder: decoder queries interact with encoder-produced keys and values. This mechanism allows the decoder to condition its predictions on the encoded input sequence. Several PHM architectures exploit cross-attention (Hu et al., 2023; Biggio et al., 2022; Deng et al., 2022; C. Jin et al., 2025; Che et al., 2023; Zerveas et al., 2020).

Table 3. RMSE comparison of selected architectures on C-MAPSS. studied values are extracted from the cited papers and are indicative rather than strictly head-to-head, because preprocessing, RUL labeling, train-test protocols, and evaluation settings may differ. Best values in each column are shown in bold. N.R. = not studied.

Work	Model	FD001	FD002	FD003	FD004	Avg.
(Ogunfowora & Najjaran, 2023)	Transformer (encoder-only)	14.21	12.75	15.57	**12.09**	13.66
(Chirukiri et al., 2025)	Transformer + GRU	30.76	N.R.	N.R.	N.R.	30.76
(Wang et al., 2021)	Transformer encoder + TCNN	12.31	15.35	12.32	18.35	14.58
(Hu et al., 2023)	Transformer + CNN + LSTM	8.31	16.42	7.10	17.06	12.22
(Hu et al., 2023)	Transformer + CNN	12.24	19.53	7.40	17.51	14.17
(Hu et al., 2023)	Transformer + LSTM	**3.52**	13.29	**4.44**	13.79	8.76
(Pour et al., 2025)	TFT + TCN	11.53	13.65	9.31	14.24	12.18
(Z. Zhang et al., 2022)	DAST	11.43	15.25	11.32	18.36	14.09
(Lv et al., n.d.)	RMTF Transformer	11.17	14.02	11.48	17.07	13.44
(Nguyen et al., 2025)	CNN + Transformer + Fourier transform	6.97	**8.03**	6.52	12.24	**8.44**
Baseline mean from selected studies	Transformer	6.18	17.34	6.82	19.30	12.41
Baseline mean from selected studies	CNN	21.04	30.43	20.35	28.21	25.01
Baseline mean from selected studies	LSTM	14.61	22.77	14.74	24.75	19.21

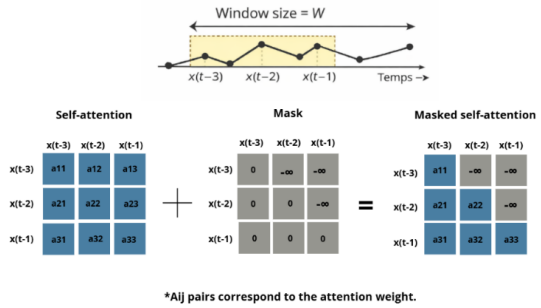

Figure 4. Illustration of masked self-attention for causal modeling.

3.3.5. Enhanced attention variants

Standard attention relies on linear projections and global dot products. While effective for capturing long-range dependencies, it may be less efficient at representing detailed local structures. Multiple works thus propose enhanced attention formulations that jointly capture local structured dependencies and global context. For example, convolution-enhanced Transformers have been proposed to learn long-term degradation features and local contextual associations directly from input sequences (Peng et al., 2023; Sun et al., 2024). ProbSparse or probabilistic attention has also been introduced to improve attention efficiency on long sequences (Chang et al., 2022). In addition, frequency-domain linearization strategies have been explored to reduce complexity while preserving global context, as in the fast temporal Transformer (FTT) (Chirukiri et al., 2025).

3.4. Transformer architecture families in PHM

The adoption of Transformers in PHM has produced multiple architecture families. These families differ not only in whether the model uses an encoder, decoder, or encoder–decoder backbone, but also in how the Transformer is coupled to PHM-specific preprocessing, local feature extraction, and degradation-representation modules.

3.4.1. Encoder-only and encoder–decoder architectures

Several studies employ Transformers directly for time-series analysis, aiming to learn both short- and long-term temporal dependencies in univariate or multivariate signals. In such settings, self-attention captures global dependencies, while local inductive biases may still benefit from convolutional or recurrent front-ends. Transformers may be used as encoder-only models (Ogunfowora & Najjaran, 2023; Zerveas et al., 2020), combined with regression heads for RUL estimation, or as encoder–decoder models when forecasting an entire multivariate sequence (Biggio et al., 2022). More elaborate designs include dual-encoder architectures that extract complementary features before decoding (Z. Zhang et al., 2022).

3.4.2. Hybrid and efficient architectures

Depending on the benchmark and signal characteristics, self-attention can capture global dependencies effectively but may be less effective at extracting certain local structures. Consequently, many works combine Transformers with recurrent (C. Jin et al., 2025; Chirukiri et al., 2025; Pour et al., 2025) and/or convolutional networks (X. Jin et al., 2025; Wang et al., 2021; Sun et al., 2024; Jiang, Zhang, Lei, Zhuang, & Li, 2023), or with both (Hu et al., 2023; Deng et al., 2022), in order to incorporate local inductive biases. Other approaches introduce additional mechanisms to improve attention learning in the time, frequency, or time-frequency domains (M. Zhang et al., 2024; H. Zhang et al., 2022; Sun et al., 2024; Chang et al., 2022; Nguyen et al., 2025; Lv et al., n.d.; Che et al., 2023; Jiao et al., 2022).

Table 4. RMSE comparison on PHM2012/PRONOSTIA and XJTU-SY. Values are studied in the scales used in the original studies and should only be compared within the same dataset. Best values within each dataset block are shown in bold.

Work	Model	Avg.
PHM2012 / PRONOSTIA		
(X. Jin et al., 2025)	Transformer + TCN	0.0514
(X. Jin et al., 2025)	Transformer	0.0852
(X. Jin et al., 2025)	RNN	0.1093
(X. Jin et al., 2025)	LSTM	0.0969
(X. Jin et al., 2025)	GRU	0.0991
(C. Jin et al., 2025)	Transformer + Bi-LSTM	0.0578
(C. Jin et al., 2025)	TCN	0.0586
(C. Jin et al., 2025)	GRU	0.0767
(C. Jin et al., 2025)	Bi-LSTM	0.0721
(Deng et al., 2022)	Transformer + ConvLSTM	**0.0230**
(Sun et al., 2024)	MDSCT	0.1240
XJTU-SY		
(C. Jin et al., 2025)	Transformer + Bi-LSTM	**0.0531**
(C. Jin et al., 2025)	TCN	0.1203
(C. Jin et al., 2025)	GRU	0.1485
(C. Jin et al., 2025)	Bi-LSTM	0.0698
(Sun et al., 2024)	MDSCT	0.1600

4. COMPARATIVE EVIDENCE FROM PHM BENCHMARKS

This section compares predictive performance studied for the selected Transformer architectures on the C-MAPSS as well as PHM2012 or PRONOSTIA, and XJTU-SY benchmarks. The comparison relies on the root mean square error (RMSE), which is consistent with the mean squared error (MSE) loss typically used during training:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{2}$$

where y_i and $\hat{y}_i$ denote the ground-truth and predicted targets (e.g., RUL), and N is the number of evaluated samples. RMSE is commonly used in PHM because it penalizes large prediction errors more severely.

In addition to RMSE, C-MAPSS-oriented studies often report domain-specific scoring functions that impose asymmetric penalties on early versus late predictions, reflecting maintenance-driven risk asymmetry. Many Transformer RUL studies explicitly report both RMSE and such scores, and comparisons should ideally consider both criteria (Wang et al., 2021). For diagnosis and anomaly detection tasks, evaluation commonly relies on accuracy, F1-score, and AUC-type measures, which better reflect class imbalance and detection trade-offs.

Since this work focuses on Transformers for prognostics, we use classical CNN and RNN models as baseline references. For C-MAPSS, baseline RMSE values stated in different studies vary due to differences in learning protocols. To enable a consistent comparison, Table 3 reports mean baseline RMSE values for CNN and LSTM models extracted from the selected studies.

5. ARCHITECTURE-LEVEL SYNTHESIS, ADVANTAGES, AND LIMITATIONS

The studied results suggest that Transformer-based and hybrid attention architectures can achieve strong performance on standard PHM benchmarks. However, these comparisons should be interpreted cautiously because preprocessing, RUL labeling, train-test protocols, and baseline implementations differ across studies. The strongest studied results are often obtained by hybrid models that combine global attention with local temporal, convolutional, recurrent, decomposition, or frequency-domain modules.

The best mean RMSE values on C-MAPSS are studied by the Transformer–LSTM fusion model in (Hu et al., 2023) and by the hybrid CNN–Transformer model with a Fourier-transform component in (Nguyen et al., 2025). In (Hu et al., 2023), an LSTM layer is placed prior to positional encoding and the Transformer encoder, effectively providing an LSTM-based feature extractor; a second LSTM layer after the Transformer decoder further enriches temporal decoding. With tuned hyperparameters, this hybridization appears to improve degradation-feature learning and RUL regression.

In (Nguyen et al., 2025), a 1D convolution layer precedes positional encoding to capture local context. The CNN features are fused with positional information to facilitate learning of correlations within the input sequence. The encoder output is then processed with a Fourier transform to reduce training time and model complexity. This hybrid architecture exploits CNN inductive biases for local modeling, while leveraging frequency-domain processing to improve computational efficiency.

Table 4 shows that (Deng et al., 2022) achieves the best RMSE on PHM2012/PRONOSTIA (0.0230). The proposed ConvLSTM module extracts and filters spatio-temporal fea-

Table 5. Comparative synthesis of Transformer architecture families in PHM. This table converts the cited studies into a compare-and-contrast review summary.

Architecture group	Typical PHM use	Strengths	Limitations	Representative evidence
Encoder-only Transformer	Window-to-RUL regression, diagnosis, anomaly screening	Simple parallel backbone; captures global temporal and inter-sensor dependencies.	Quadratic attention cost; weak local inductive bias for sharp transients.	(Ogunfowora & Najjaran, 2023; Zerveas et al., 2020)
Encoder–decoder Transformer	Sequence forecasting, reconstruction, degradation trajectory prediction	Suitable for sequence-to-sequence prediction and causal decoding.	More complex training; performance depends strongly on target construction and decoding protocol.	(Biggio et al., 2022; Deng et al., 2022)
CNN/TCN–Transformer hybrid	Bearing vibration, turbofan RUL, local transient modeling	Combines local pattern extraction with global attention.	More hyperparameters and higher architectural complexity.	(Wang et al., 2021; X. Jin et al., 2025; Nguyen et al., 2025)
RNN/LSTM/GRU–Transformer hybrid	Ordered degradation sequences with short- and long-range dynamics	Adds sequential memory and local temporal smoothing before or after attention.	Less parallel than pure Transformer models; recurrent modules may be redundant for some datasets.	(Hu et al., 2023; C. Jin et al., 2025; Chirukiri et al., 2025)
Efficient / frequency-domain Transformer	Long histories, high-frequency signals, non-stationary bearing or engine data	Reduces sequence cost or emphasizes spectral degradation signatures.	May sacrifice fine temporal detail; performance depends on preprocessing and transformation choices.	(Chang et al., 2022; Che et al., 2023; Lv et al., n.d.)
Graph / ViT-style Transformer	Sensor topology, image-like or patch-like signal representations	Can encode spatial, structural, or image-like dependencies.	Limited benchmark evidence in machinery prognostics compared with encoder and hybrid models.	Representative graph/ViT-style studies in the coded corpus

tures from the input sequence prior to a Transformer composed of six stacked encoder–decoder layers. On the XJTU-SY benchmark, (C. Jin et al., 2025) reports the best RMSE among the considered works. That approach combines empirical mode decomposition (EMD) with a Bi-LSTM–Transformer model: EMD decomposes non-stationary vibration signals into intrinsic mode functions, from which multi-scale statistical indicators are extracted. These features are then modeled by a Bi-LSTM to capture local temporal dependencies, and by a Transformer to learn global relationships and degradation trends.

Transformer architectures represent a significant opportunity for PHM, notably due to their ability to capture global context over very long sequences via self-attention. This property is particularly attractive for degradation tracking, where long-term dependencies can be decisive. Nonetheless, in their standard form, Transformers may struggle to capture fine-grained local correlations in industrial signals and may exhibit limited robustness to non-stationarity (Deng et al., 2022; Peng et al., 2023; Sun et al., 2024). In addition, learning attention over long sequences entails high computational cost, which can hinder deployment under resource constraints (Ogunfowora & Najjaran, 2023). Consequently, many architectures combine Transformers with CNNs or RNNs to introduce local inductive biases (X. Jin et al., 2025; C. Jin et al., 2025; Chirukiri et al., 2025; Wang et al., 2021; Hu et al., 2023; Pour et al., 2025; Deng et al., 2022). However, such hybridization often increases architectural complexity and computational overhead. As an alternative, several works explore frequency-domain modeling (M. Zhang et al., 2024; H. Zhang et al., 2022; Sun et al., 2024; Chang et al., 2022; Nguyen et al., 2025; Lv et al., n.d.; Che et al., 2023; Jiao et al., 2022) to reduce training time, sometimes at the cost of a modest decrease in predictive accuracy (e.g., RMSE).

From the perspective of review synthesis, the current evidence suggests that the main advantage of Transformers is not simply replacement of CNNs or RNNs, but the ability to combine global attention with domain-aware representations. Hybrid models are attractive because PHM signals frequently contain both local fault signatures and long-range degradation context. However, the same hybridization also makes fair comparison difficult because performance can depend on window length, feature engineering, RUL labeling, normalization, optimizer choices, and benchmark split. Therefore, future studies should report preprocessing, label construction, hyperparameter ranges, and baseline implementations with enough detail to support replication.

Several priorities emerge for future research. First, the field needs more reproducible and standardized benchmarking protocols to support fair comparison across studies. Second, data-efficient learning remains essential because run-to-failure labels are scarce in many realistic industrial settings. Third, robust deployment requires stronger handling of domain shift, noise, missing data, and variable operating regimes. Fourth, efficient long-sequence attention mechanisms, better uncertainty quantification, and more interpretable model behavior are necessary for trustworthy use in safety-critical environments. Finally, PHM research would benefit from moving beyond isolated point prediction toward end-to-end frameworks that connect prognostic outputs to maintenance actions, operational risk, and decision-making.

6. CONCLUSION

This review examined the growing use of Transformer-based architectures in Prognostics and Health Management (PHM),

with primary emphasis on machinery prognostics and remaining useful life (RUL) estimation, while also considering adjacent tasks such as fault diagnosis, anomaly screening, and decision-oriented maintenance support. By organizing the literature according to PHM task, Transformer backbone, hybridization strategy, and input representation, and by reviewing preprocessing choices such as windowing, health-indicator construction, embedding, and positional encoding, the survey shows that predictive performance depends not only on the attention mechanism itself, but also on how degradation information is represented before it reaches the model.

Across widely used benchmarks, including C-MAPSS, PHM 2012, PRONOSTIA, and XJTU-SY, the studied literature indicates that Transformer-based models can capture long-range temporal dependencies and cross-sensor interactions effectively, especially when degradation evolves over long horizons and under variable operating conditions. At the same time, the strongest results are frequently achieved by hybrid architectures that combine Transformers with convolutional, recurrent, decomposition-based, or frequency-domain modules. This suggests that, in the current state of the field, local inductive biases remain highly valuable for PHM and often complement global self-attention rather than being replaced by it. An important outcome of this review is therefore that their effectiveness is highly contingent on methodological choices.

Overall, Transformer-based methods represent a powerful and promising direction for next-generation PHM. Their long-term industrial value, however, will depend less on architectural novelty alone than on rigorous evaluation, domain-aware representation design, transparent reporting of preprocessing and benchmark protocols, and dependable deployment under realistic operating constraints.

REFERENCES

Babu, G. S., Zhao, P., & Li, X.-L. (2016). Deep convolutional neural network based regression approach for estimation of remaining useful life. In *Database systems for advanced applications* (p. 214-228). doi: 10.1007/978-3-319-32025-0_14

Biggio, L., Bendinelli, T., Kulkarni, C., & Fink, O. (2022). Dynaformer: A deep learning model for ageing-aware battery discharge prediction. *arXiv preprint arXiv:2206.02555*. Retrieved from https://arxiv.org/abs/2206.02555

Chang, Y., Li, F., Chen, J., Liu, Y., & Li, Z. (2022). Efficient temporal flow transformer accompanied with multi-head probsparse self-attention mechanism for remaining useful life prognostics. *Reliability Engineering & System Safety*, *226*, Article number 108701. doi: https://doi.org/10.1016/j.ress.2022.108701

Che, S., Lu, J., Bao, C., Zhang, C., & Liu, Y. (2023). Multiscale time-frequency sparse transformer based on partly interpretable method for bearing fault diagnosis. *Shock and Vibration*, *2023*(1), Article number 1639287. doi: 10.1155/2023/1639287

Chelouati, M., Jha, M. S., Galeotta, M., & Theilliol, D. (2021). Remaining useful life prediction for liquid propulsion rocket engine combustion chamber. In *2021 5th international conference on control and fault-tolerant systems (systol)* (pp. 225–230). Saint-Raphaël, France. doi: 10.1109/SysTol52990.2021.9595286

Chirukiri, V. T., Cheerala, U. B., Kanta, S., Karim, A., & Damacharla, P. (2025). FTT-GRU: A hybrid fast temporal transformer with GRU for remaining useful life prediction. *arXiv preprint arXiv:2511.00564*. Retrieved from https://arxiv.org/abs/2511.00564

de Beaulieu, M. H., Jha, M. S., Garnier, H., & Cerbah, F. (2022). Unsupervised prognostics based on deep virtual health index prediction. In *Phm society european conference* (Vol. 7, pp. 193–199). Turin, Italy.

de Beaulieu, M. H., Jha, M. S., Garnier, H., & Cerbah, F. (2024). Remaining useful life prediction based on physics-informed data augmentation. *Reliability Engineering & System Safety*, *252*, Article number 110451.

Deng, F., Chen, Z., Liu, Y., Yang, S., Hao, R., & Lyu, L. (2022). A novel combination neural network based on ConvLSTM-transformer for bearing remaining useful life prediction. *Machines*, *10*(12), Article number 1226. doi: 10.3390/machines10121226

Ding, Y., & Jia, M. (2022). Convolutional transformer: An enhanced attention mechanism architecture for remaining useful life estimation of bearings. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1-10. doi: 10.1109/TIM.2022.3181933

Dong, S., Xiao, J., Hu, X., Fang, N., Liu, L., & Yao, J. (2023). Deep transfer learning based on Bi-LSTM and attention for remaining useful life prediction of rolling bearing. *Reliability Engineering & System Safety*, *230*, Article number 108914. doi: https://doi.org/10.1016/j.ress.2022.108914

Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.-J., & Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, *92*, Article number 103678. doi: 10.1016/j.engappai.2020.103678

Götz, L., Kollovieh, M., Günnemann, S., & Schwinn, L. (2025). Byte pair encoding for efficient time series forecasting. *arXiv preprint arXiv:2505.14411*.

Hu, Q., Zhao, Y., & Ren, L. (2023). Novel transformer-based fusion models for aero-engine remaining useful life estimation. *IEEE Access*, *11*, 52668–52685. doi: 10.1109/ACCESS.2023.3277730

Huang, N., Kümmerle, C., & Zhang, X. (2024). *Unitnorm: Rethinking normalization for transformers in time series.* Retrieved from https://arxiv.org/abs/2405.15903

Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, *20*(7), 1483-1510. doi: 10.1016/j.ymssp.2005.09.012

Jha, M. S., Bressel, M., Ould-Bouamama, B., & Dauphin-Tanguy, G. (2016). Particle filter based hybrid prognostics of proton exchange membrane fuel cell in bond graph framework. *Computers Chemical Engineering*, *95*, 216-230. doi: https://doi.org/10.1016/j.compchemeng.2016.08.018

Jha, M. S., Dauphin-Tanguy, G., & Ould-Bouamama, B. (2016). Particle filter based hybrid prognostics for health monitoring of uncertain systems in bond graph framework. *Mechanical Systems and Signal Processing*, *75*, 301-329. doi: https://doi.org/10.1016/j.ymssp.2016.01.010

Jha, M. S., Theilliol, D., Belleoud, P., & Oriol, S. (2025). Deep learning based prognostics of nonlinear systems under degradation in closed-loop. In *2025 6th international conference on control and fault-tolerant systems (systol)* (p. 172-179). doi: 10.1109/SysTol66549.2025.11267335

Jiang, L., Zhang, T., Lei, W., Zhuang, K., & Li, Y. (2023). A new convolutional dual-channel transformer network with time window concatenation for remaining useful life prediction of rolling bearings. *Advanced Engineering Informatics*, *56*, Article number 101966. doi: https://doi.org/10.1016/j.aei.2023.101966

Jiao, Z., Pan, L., Fan, W., Xu, Z., & Chen, C. (2022). Partly interpretable transformer through binary arborescent filter for intelligent bearing fault diagnosis. *Measurement*, *203*, Article number 111950. doi: https://doi.org/10.1016/j.measurement.2022.111950

Jin, C., Li, B., Yang, Y., Yuan, X., Tu, R., Qiu, L., & Chen, X. (2025). Remaining useful life prediction of rolling bearings based on empirical mode decomposition and transformer Bi-LSTM network. *Applied Sciences*, *15*(17), Article number 9529.

Jin, X., Ji, Y., Li, S., Lv, K., Xu, J., Jiang, H., & Fu, S. (2025). Remaining useful life prediction for rolling bearings based on TCN–transformer networks using vibration signals. *Sensors*, *25*(11), Article number 3571. doi: 10.3390/s25113571

Kanso, S., Jha, M. S., Galeotta, M., & Theilliol, D. (2022). Remaining useful life prediction with uncertainty quantification of liquid propulsion rocket engine combustion chamber. *IFAC-PapersOnLine*, *55*(6), 96-101. (11th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes SAFEPROCESS 2022) doi: https://doi.org/10.1016/j.ifacol.2022.07.112

Lei, Y., Han, T., Wang, B., Li, N., Yan, T., & Yang, J. (2019). XJTU-SY rolling element bearing accelerated life test datasets: A tutorial. *Journal of Mechanical Engineering*, *55*(16), 1–6. doi: 10.3901/JME.2019.16.001

Liu, C.-L., Hsaio, W.-H., & Tu, Y.-C. (2019). Time series classification with multivariate convolutional neural network. *IEEE Transactions on Industrial Electronics*, *66*(6), 4788-4797. doi: 10.1109/TIE.2018.2864702

Lv, S., Liu, S., & Li, H. (n.d.). New method for remaining useful life prediction based on recurrence multi-information time-frequency transformer networks. *Quality and Reliability Engineering International*, *41*(5), 1643-1663. doi: 10.1002/qre.3740

Mittal, D., Bello, H., Zhou, B., Jha, M. S., Suh, S., & Lukowicz, P. (2023). *Two-stage early prediction framework of remaining useful life for lithium-ion batteries.* Retrieved from https://arxiv.org/abs/2308.03664

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. (2012, 06). Pronostia: An experimental platform for bearings accelerated degradation tests. In *Conference on prognostics and health management.* (p. 1-8). Denver, CO, USA.

Nguyen, H.-D., Nguyen, X. H., Dao, B. T., Do, C. N., Truong, H., & Tran, K. P. (2025). *Remaining useful lifetime prediction of turbofan engines based on hybrid transformer deep architecture.* SSRN. Retrieved from https://ssrn.com/abstract=5358720 (Available at SSRN: #5358720) doi: 10.2139/ssrn.5358720

Ogunfowora, O., & Najjaran, H. (2023). *A transformer-based framework for multi-variate time series: A remaining useful life prediction use case.* Retrieved from https://arxiv.org/abs/2308.09884

Patra, K. C., Sethi, R., & Behera, D. K. (2025). Estimation of the remaining useful life of aircraft engines using a CNN-LSTM-GRU hybrid model. *International Journal of System Assurance Engineering and Management*, *16*(12), 3968–3982. doi: 10.1007/s13198-025-02911-4

Peng, H., Jiang, B., Mao, Z., & Liu, S. (2023). Local enhancing transformer with temporal convolutional attention mechanism for bearings remaining useful life prediction. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1-12. doi: 10.1109/TIM.2023.3291787

Pour, M. A., Karimi, M. S., & Mazloumi, A. H. (2025). Temporal convolutional and fusional transformer model with Bi-LSTM encoder–decoder for multi-time-window remaining useful life prediction. *IEEE Access*, *13*, 203705–203722. doi: 10.1109/ACCESS.2025.3634285

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008a). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management* (pp. 1–9). doi: 10.1109/PHM.2008.4711414

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008b). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management* (pp. 1–9). Denver, CO, USA.

Sikorska, J. Z., Hodkiewicz, M., & Ma, L. (2011). Prognostic modelling options for remaining useful life estimation by industry. *Mechanical Systems and Signal Processing*, *25*(5), 1803-1836.

Suh, S., Jang, J., Won, S., Jha, M. S., & Lee, Y. O. (2020). Supervised health stage prediction using convolutional neural networks for bearing wear. *Sensors*, *20*(20), Article number 5846. doi: 10.3390/s20205846

Suh, S., Mittal, D. A., Bello, H., Zhou, B., Jha, M. S., & Lukowicz, P. (2024). Remaining useful life prediction of lithium-ion batteries using spatio-temporal multimodal attention networks. *Heliyon*, *10*(16), Article number e36236. doi: 10.1016/j.heliyon.2024.e36236

Sun, N., Tang, J., Ye, X., Zhang, C., Zhu, S., Wang, S., & Sun, Y. (2024). Remaining useful life prognostics of bearings based on convolution attention networks and enhanced transformer. *Heliyon*, *10*(19), Article number e38317. doi: 10.1016/j.heliyon.2024.e38317

Talukder, S., Yue, Y., & Gkioxari, G. (2025). *Totem: Tokenized time series embeddings for general time series analysis.* Retrieved from https://arxiv.org/abs/2402.16412

Thuillier, J., Jha, M. S., Le Martelot, S., & Theilliol, D. (2024). Prognostics aware control design for extended remaining useful life: Application to liquid propellant reusable rocket engine. *International Journal of Prognostics and Health Management*, *15*(1). doi: 10.36001/ijphm.2024.v15i1.3460

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). *Attention is all you need* (Vol. 30).

Wang, H.-K., Cheng, Y., & Song, K. (2021). Remaining useful life estimation of aircraft engines using a joint deep learning model based on TCNN and transformer. *Computational Intelligence and Neuroscience*, *2021*(1), Article number 5185938. doi: 10.1155/2021/5185938

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2022). Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2020). *A transformer-based framework for multivariate time series representation learning.* Retrieved from https://arxiv.org/abs/2010.02803

Zhang, B., & Sennrich, R. (2019). *Root mean square layer normalization.* Retrieved from https://arxiv.org/abs/1910.07467

Zhang, H., Zhang, S., Qiu, L., Zhang, Y., Wang, Y., Wang, Z., & Yang, G. (2022). A remaining useful life prediction method based on PSR-former. *Scientific Reports*, *12*(1), Article number 17887. doi: 10.1038/s41598-022-22941-3

Zhang, M., He, C., Huang, C., & Yang, J. (2024). A weighted time embedding transformer network for remaining useful life prediction of rolling bearing. *Reliability Engineering & System Safety*, *251*, Article number 110399. doi: 10.1016/j.ress.2024.110399

Zhang, Z., Song, W., & Li, Q. (2022). Dual-aspect self-attention based on transformer for remaining useful life prediction. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–11. doi: 10.1109/TIM.2022.3160561

Zheng, S., Ristovski, K., Farahat, A., & Gupta, C. (2017). Long short-term memory network for remaining useful life estimation. In *Ieee int. conf. prognostics and health management (icphm)* (pp. 88–95). Dallas, TX, USA.