A Hybrid Semi-Supervised Framework for Full Life-Cycle Degradation Trajectory Learning and RUL Estimation

Xianpeng Qiao[1], Veronica Lestari Jauw[1], Tiyamike Banda[2], Chin Seong Lim[1,*]

[1]*Department of Mechanical, Materials & Manufacturing Engineering, University of Nottingham Malaysia, Semenyih, Selangor, 43500, Malaysia*

edxxq1@nottingham.edu.my
Veronica.Jauw@nottingham.edu.my
Corresponding author: ChinSeong.Lim@nottingham.edu.my

[2]School of Engineering, Malawi University of Business and Applied Sciences (MUBAS), Blantyre, Malawi
tbanda@mubas.ac.mw

ABSTRACT

Deep Learning (DL) has substantially expanded its role in Prognostics and Health Management (PHM), particularly by enabling automated feature extraction for Remaining Useful Life (RUL) prediction. Despite this progress, existing DL models such as Long Short-Term Memory (LSTM) networks still face challenges in accurately capturing complete life-cycle degradation trajectories. To address this limitation, this study introduces a hybrid semi-supervised model that integrates a Time-Transformer (TT) with a Denoising Autoencoder (DAE), termed TT-DAE. The DAE first extracts spatial features and suppresses noise through signal reconstruction from degraded inputs. These extracted features are then separated into source and target domains and normalized to a uniform sequence length using a padding strategy. Subsequently, the TT module leverages both source features and a Sliding Variable-Length Window (SVW) mechanism to learn full degradation trajectories. A comprehensive experimental evaluation conducted on the C-MAPSS dataset demonstrates the effectiveness of the proposed approach, achieving an average Pearson Correlation Coefficient (PCC) of 0.89 between the predicted and actual target signals.

Keywords: Time-Transformer, Denoise Auto-Encoder, Turbofan engine, Remaining Useful Life, Complete life cycle degradation trajectory learning.

1. INTRODUCTION

The cost-effectiveness and reliability of machinery maintenance can be significantly improved by Predictive Maintenance (PdM). PdM employs Machine Learning (ML) models to anticipate the Remaining Useful Life (RUL) for scheduling precise maintenance based on multi-sensor life cycle data. Comparing PdM to traditional methods, such as preventive maintenance, which performs maintenance based on expert knowledge, the PdM reduces maintenance expenses by 20–30% and machine downtime by 20–50%, as reported by carbon collective (Heng et al., 2009). These improvements in cost-effectiveness and reliability are primarily attributed to the accurate RUL prediction achieved through PdM.

In the realm of RUL prediction, there are two main approaches: direct and indirect mapping methods (Yu et al., 2019). Direct mapping utilizes ML techniques, such as Random Forest (RF) and Extreme Learning Machine (ELM), to establish an optimal mapping between signal data and corresponding RUL values (Thakkar & Chaoui, 2022; Xue et al., 2022). However, this method cannot capture the degradation features of machinery accurately, especially when the signal data are noisy, nonlinear, and non-stationary. Conversely, the indirect mapping method involves extracting abstract representations of features, referred to as Health Indicators (HIs), which are correlated with the machinery degradation trajectory (Guo et al., 2017). Once extracted, these features are then fed into a predictor to estimate the RUL. Notably, due to its superior prediction accuracy, the indirect mapping method has garnered more scholarly attention compared to the direct mapping method (De Beaulieu et al., 2022).

In the indirect mapping method, similarity, classification, and regression approaches can be utilized to categorize it effectively (Ahn et al., 2021; Qiao et al., 2025). The similarity approach matches a similar HI curve from the HI library (Xia et al., 2022). However, the prediction accuracy depends on the quantity of HI curves with various operating conditions

in the HI library. The classification approach attempts to establish a mapping function between representative features and RUL values (Ma et al., 2018). However, this approach is less interpretable and cannot foresee future degradation trajectories. In contrast, the regression approach aims to predict the representative features indicative of the time remaining until reaching the failure threshold (Jia et al., 2020). Figure. 1 demonstrates the difference between the regression approach and the classification approach. Where $\{X_t^d\}_{t=i}^{i+T}$ denotes the signals with the length of T, X_{i+T}^d is the signals at the timestamp $i+T$ with d dimensions, and RUL_{i+T-1} is the corresponding RUL value at the timestamp $i+T-1$. As a result of its inherent interpretability and intuition about the degradation trajectory, the regression approach has been progressively garnering substantial attention compared to the other two approaches in the field of RUL prediction.

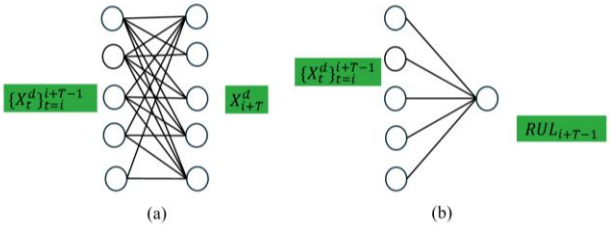

Figure. 1. The schematic of: (a) regression, (b) classification approach in the indirect mapping method.

The prediction accuracy of the regression approach in the indirect mapping method heavily relies on performance of the features extractor (Qiao et al., 2024). Most feature extractors segment the complete life cycle degradation trajectory into several windows of length T during the data sampling process, referred to as the Sliding Window (SW) technique. However, the extracted features utilizing the SW technique do not provide an overview information of the complete life cycle degradation trajectory. Therefore, this paper does not segment the complete life cycle data but inputs the complete life cycle data into the features extractor, termed as the Sliding Variable-length Window (SVW) technique. The extracted features using the SVW technique provide a global perspective of the degradation trajectory. The principles of proposed SW and SVW are explained below in formula. The objective of feature extractor using the SW technique is to construct a mapping function between the past signal with the length of T and next time's signal, represented as follows:

$$X_{T+i}^d = f\left(X_i^d, X_{i+1}^d, \dots, X_{T+i-1}^d\right), with\ l < i < S - T, \quad (1)$$

where S is the working life of the machinery, T is the length of the window, and l is the length of the existing data referred to as source features. In contrast, the objective of the feature extractor using the SVW technique is to construct a mapping function between all past signals and next time's signal. It is described as:

$$X_{l+i}^d = f\left(X_l^d, \quad X_{l+1}^d, \dots, X_{l+i-1}^d\right), with\ 0 < i < S - l. \quad (2)$$

In order to further enhance the complete life cycle degradation trajectory, the source features are fed into the model. The objective of the feature extractor is represented as follows, as used in this paper:

$$X_{l+i}^d = f\left(X_0^d, X_1^d, \dots, X_{l+i-1}^d\right), \quad with\ 0 < i < S - l. \quad (3)$$

However, the complete life cycle is often lengthy for machinery. Regarding the turbofan engine, the life cycle ranges from 182 to 362. This presents a challenge for the feature extractor to avoid losing previous information when extracting temporal features. Statistical and Deep Learning (DL) models are commonly used to extract features for regression approach. While statistical model like the nonlinear wiener process proposed by Wang Zezhou et al. (Wang et al., 2020) is to approximate the degradation trajectory of machinery. However, these statistical models fail to capture the intricate and varied degradation trajectory. In response to these constraints, DL models have emerged as a solution. Guo et al. (Guo et al., 2017) introduced the Recurrent Neural Network (RNN) to discern degradation trajectory for predicting HI up to the failure threshold. However, processing long time series data with RNN may lead to the loss of long-term dependencies when gradient explosion and vanishing occur. To address these issues, a variant of RNN, the Long Short-Term Memory (LSTM) network, utilizes a forget gate to control which information should be discarded from the memory cells, thus mitigating gradient explosion and vanishing, as introduced by Zijian Ye et al. (Ye et al., 2022). Despite the LSTM demonstrating the capacity on alleviating gradient explosion and vanishing than RNN in RUL prediction (Wu et al., 2020), it still struggles with the loss of previously provided information in processing complete life cycle data (Zhou et al., 2021). Inspired by the extreme success of the transformer model in Natural Language Process (NLP) and time series prediction, this paper employs the transformer model in RUL prediction. However, the transformer model exhibits its limitation in extracting the spatial features, due to the self-attention mechanism to weigh the attention score of data in a sequence. In addition, the existing transformer model used in RUL prediction (J. Zhang et al., 2023) direct estimate RUL instead of predict next time's signals.

In order to address these limitations, this paper proposes a hybrid model of the TT-DAE to effectively extract both long-term temporal and spatial features utilizing the regression approach. Specifically, the representative spatial features are first extracted by reconstructing the original signals from the noisy signals, which are subjected to white noise. Subsequently, the extracted spatial features are fed into the TT model for temporal feature extraction. In the TT model,

the transformer encoder module is integrated with the transformer decoder module as a regression approach in the indirect mapping method to learn the complete life cycle degradation trajectory. The transformer encoder processes source features for prediction and sends them to the Transformer decoder via the encoder-decoder attention. Subsequently, the decoder generates successive features (referred to as target features) step by step based on the output from the transformer encoder and previously predicted features. Finally, the trained TT-DAE is used to predict target signals step by step and calculate the RUL. This study makes the following primary contributions:

- The TT-DAE model is first proposed to learn the complete life cycle degradation trajectory by utilizing the SVW technique, rather than a portion of it.
- The TT-DAE model is employed to calculate the RUL by predicting the degradation trajectory instead of directly predicting RUL.
- The efficiency is analyzed between the LSTM model and the TT model in handling the complete life cycle data.

The overall organization is described below. Section 2 provides a detailed introduction to the proposed TT-DAE model. Section 3 describes the case study using the Commercial Modular Aero-propulsion System Simulation (C-MAPSS) dataset and comparison with other state-of-the-art models. Finally, the conclusion and the future research direction are discussed in Section 4.

2. METHODOLOGY

In this section, the hybrid TT-DAE model is briefly introduced, followed by a detailed description of the DAE and TT model.

2.1. Time-Transformer and denoise autoencoder (TT-DAE) model architecture

This section provides an in-depth representation of the proposed TT-DAE. The framework involves data pre-processing, complete life cycle degradation trajectory learning, and RUL calculation. Among these steps, the data pre-processing technique aims to eliminate irrelevant sensors that remain constant over time, normalize the data from 0 to 1 for faster convergence, and split the data into 90% for training and 10% for testing. Subsequently, the complete life cycle degradation trajectory learning includes the extraction of spatial features by the self-supervised model DAE and the degradation trajectory learning by TT using training data. The final step is to test the performance of the trained model in RUL prediction. This step involves predicting target features and converting these features into target signals via the decoder in the trained DAE. Following this, RUL calculation is performed, as illustrated in Figure. 2. Detailed explanations of the DAE and the innovative complete life cycle degradation trajectory recognition methods, TT, are provided in the remainder of this section.

2.2. Denoise Auto-Encoder (DAE)

In Figure. 4, the schematic of the proposed DAE is depicted. The DAE is a modification of the auto-encoder, wherein white noise is introduced to the input signal to generate a noisy signal. Subsequently, the DAE is trained to reconstruct the clean output from the noisy signal, ensuring the output signal closely matches the input signal. Following this reconstruction, latent spatial features are extracted for aiding the subsequent learning of complete life cycle degradation trajectories.

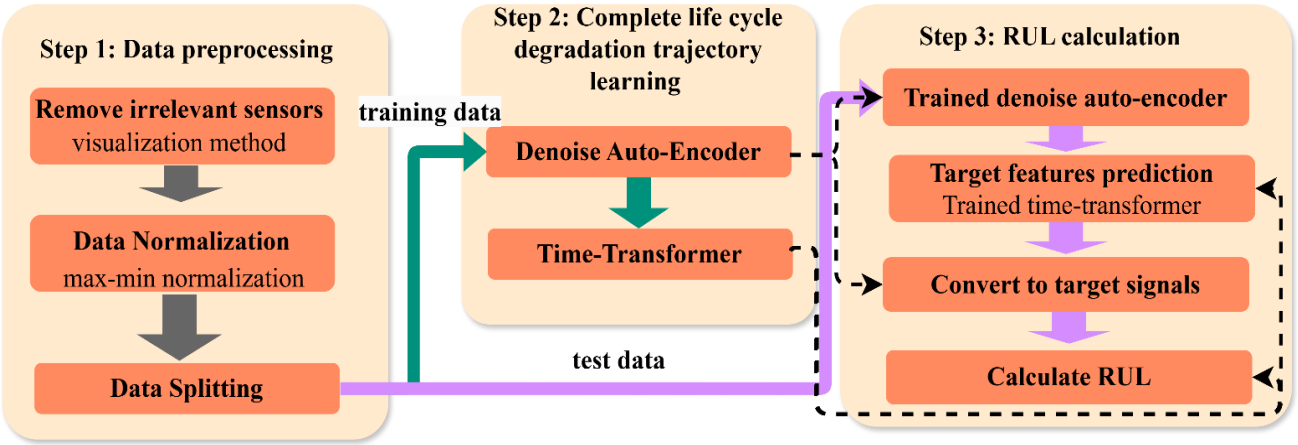

Figure. 2. The overall framework of the proposed TT-DAE model for RUL prediction.

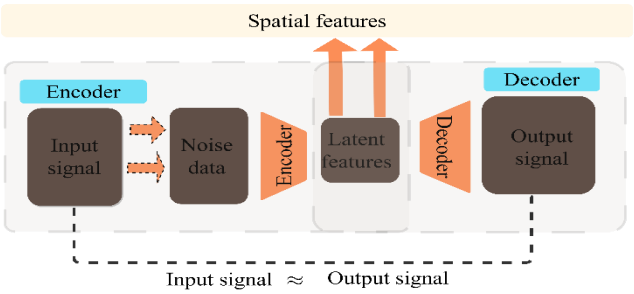

Figure. 4. The architecture of the DAE.

2.3. Time-Transformer (TT)

The detailed structure and procedures of the TT model for complete life cycle degradation trajectory recognition, is shown in Figure. 3.

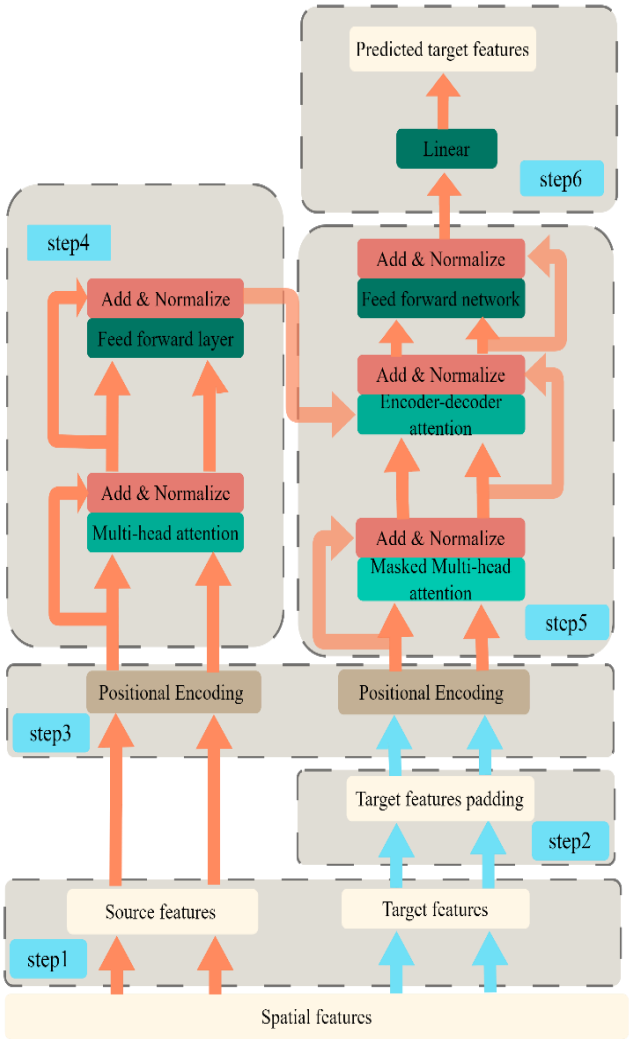

Figure. 3. Schematic of the TT model in the entire degradation trajectory recognition.

2.3.1. Separation of source and target features

The purpose of the TT model in the RUL prediction aims to establish the mapping function f between the source and target features, as expressed in Eq. (3). Therefore, the first step in utilizing the TT model is the randomized partition of entire multi-sensor run-to-failure data into the source and target features, as illustrated in the Figure 5. Herein, the source features mean the existing features, and the target features imply the predicted features.

2.3.2. Target features padding

Figure 5 indicates that the target features possess variable lengths as a result of random segmentation operation. To overcome this issue, a data padding technique is employed to standardize the length of target features. In this study, the value representing the end of the cycle is chosen as the padding value, which is appended to the end of the target features. During training, these padding values are masked in the self-attention score calculations because they are irrelevant to the degradation trajectories.

2.3.3. Position encoding

Position is the order of the multi-sensor run-to-failure data. Unlike the RNN which inherently takes the order of multi-sensor run-to-failure data into account, the TT model needs to insert the positional encoding block to make sure the source and target features include the temporal information before feeding them to the self-attention mechanism. Below is the detailed method for the positional encoding.

Let $\overrightarrow{p_t} \in \mathbb{R}^d$ is the corresponding positional encoding. Where the t is the desired position within an input sequence of length $\mathbb{N}$, and d is the encoding dimension (also the dimension of spatial features). Then $f: \mathbb{N} \to \mathbb{R}^d$ is the function that produces the positional encoding vector $\overrightarrow{p_t}$ and it is defined as follows:

$$\overrightarrow{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_j . t), & if\ i = 2j, \\ \cos(\omega_j . t), & if\ i = 2j + 1. \end{cases} \quad (4)$$

where,

$$\omega_j = \frac{1}{10000^{2j/d}},$$

and i, j are the position of the input sequence, $0 \le i, j < \mathbb{N}/2$.

2.3.4. Time-Transformer (TT) encoder

The TT encoder is the vital part of the proposed TT to extract the temporal features based on the self-attention mechanism and output the temporal features to the transformer decoder for further processing. The TT encoder includes the N_x encoder layer to extract dive temporal features. Each encoder layer primarily comprises multi-head attention and feed-

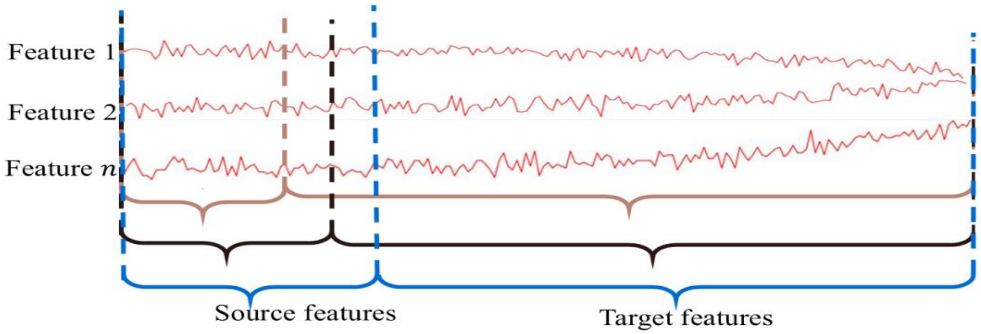

Figure 5. Schematic of the separation of the source and target features from complete life cycle data.

.

forward layer, which are elaborated individually in the subsequent sections

The multi-head attention layer includes the multi-head attention and a residual layer to mitigate the vanishing gradient problem. The multi-head attention concatenates several scaled dot-product attentions running in parallel, as shown in Figure. 6.

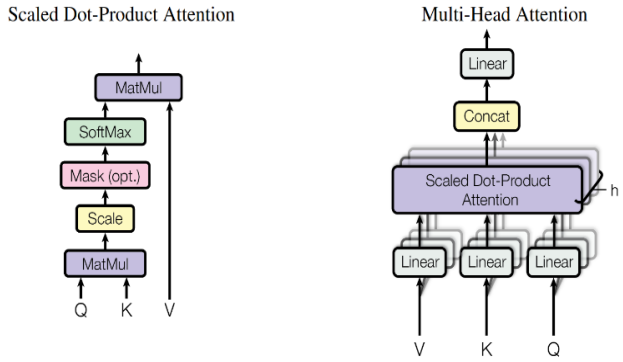

Figure. 6. (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.(Vaswani et al., 2017)

The scaled dot-product attention takes its inputs X_t^d in the form of three represented vectors query Q, key K, and value V, which are obtained by the product of learnable weight W_q, W_k, and W_v and its inputs X_t^d. Its calculation process is as follows:

$$\begin{aligned} Q &= X_t^d W_q, \\ K &= X_t^d W_k, \\ V &= X_t^d W_v. \end{aligned} \tag{5}$$

Then, the dot-product attention matrix is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{6}$$

where the d_k is the dimension of the key vectors. The multi-head attention projects the query, key, and value h times and concatenates the h attention matrix. The purpose of multi-head attention is to learn attention mechanisms for different operating scenarios.

Subsequently, the result of the multi-attention layer passes to the feed-forward layer, sending its output (referred to as memory) to the subsequent encoder or decoder layers. The feed-forward layer operates with the residual connection as its primary components, transforming linear temporal features into nonlinear temporal features.

2.3.5. Time-Transformer (TT) decoder

The TT decoder tries to predict the target features based on the memory fed from the TT encoder. Therefore, each TT decoder layer includes the masked multi-head attention layer, the encoder-decoder attention layer, and the feed-forward layer, as shown in Figure. **3**.

The masked multi-head attention layer consists of the masked multi-head attention and residual network. The mask mechanism is utilized in multi-head attention to prevent attached future features when predicting subsequent time steps. Following this, the output of the masked multi-head attention layer is passed to the encoder-decoder attention layer. As the vital component in the encoder-decoder attention layer, encoder-decoder attention is a variant of multi-head attention. The query Q in target features queries all the key K and value V in memory instead of querying the key K and value V in target features again, resulting in the output attention value taking into account the information of the source features. Finally, the output from the encoder-decoder attention layer is passed to the feed-forward layer, which is the same layer in the TT encoder to convert the linear temporal features to the non-linear temporal features.

2.3.6. Target features prediction

As the final step in the proposed TT model, the output of the TT decoder is fed into one-layer linear neural network to

predict the target features. To ensure the predicted target features align closely with the actual target features via minimizing the Mean Absolute Error (MAE) between the predicted and actual target features during training. It is described as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{X_t^d} - X_t^d\right|, with\ t = (T+1, T+2, \ldots, S_i). \tag{7}$$

3. CASE STUDY

To validate the prediction accuracy of RUL and the degradation trajectory, the proposed TT-DAE model employed the turbofan engine dataset, named C-MAPSS, which was published by NASA (Saxena et al., 2008). Firstly, the dataset is described in detail. Following this, detailed preliminary processes were conducted, including data preprocessing and features extraction. In the features extraction, a hybrid model of TT-DAE is proposed. The DAE model extracts the representative spatial features by reconstructing the original data, then the TT model was applied for recognizing the complete life cycle degradation trajectory. Finally, the superiority of our proposed TT model was justified through a comprehensive comparison with existing transformer-based RUL prediction models.

3.1. Dataset description

The C-MAPSS dataset consists of simulated multi-sensor run-to-failure data from turbofan jet engines generated using the C-MAPSS dynamic model (Frederick et al., 2007). The C-MAPSS dataset is divided into four subsets, FD001, FD002, FD003, and FD004, according to the operating conditions and fault modes, as described in Figure **7**. Moreover, each subset encompasses three operating settings and 21 monitored sensor signals, including temperature, pressure, speed signals, et al. to monitor the degradation trajectory, as depicted in Figure **7**.

	FD001	FD002	FD003	FD004
Number of engines in the training set	100	260	100	249
Number of engines in the test set	100	259	100	248
Training samples	17,731	48,819	21,820	57,522
Testing samples	100	259	100	248
Operating conditions	1	6	1	6
Fault modes	1	1	2	2

Table 1. Description of the C-MAPSS dataset.

3.2. Data pre-processing

Data pre-processing techniques are crucial for selecting dependable attributes that convey degradation information. It is apparent that certain monitored signals are discounted as they fail to carry any insight into degradation due to their constancy throughout the entire life cycle (C. Zhang et al., 2017). Based on the observation, the sensor signals such as 'Total temperature at fan inlet', 'Pressure at fan inlet', 'Engine pressure ratio (P50/P2)', 'Burner fuel-air ratio', 'Demanded fan speed', 'Demanded corrected fan speed' were removed, due to their consistent values throughout the entire life cycle, as illustrated in Figure **7**. Finally, the data are normalized to a range of 0 to 1 using the min–max normalization method, as described below:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}, \tag{8}$$

where X is the data before normalization, X' denotes the data after normalization, X_{min} is the minimum value of the X, and X_{max} is the maximum value of the X.

3.3. Features extraction

3.3.1. Representative spatial features extraction

The selected signals remain contaminated with white noise and exhibit high dimensionality, posing challenges for accurate prediction (Saxena et al., 2008). This approach also addresses the limitations of Transformer-based models in spatial feature extraction. Consequently, we propose the DAE extract representative spatial features while eliminating noise.

As a self-supervised model, the DAE introduces white noise to the original signals, creating noisy inputs, and is trained to reconstruct the clean signals. The features extracted through this process demonstrate strong representativeness, particularly when the evaluation metric—the Mean Absolute Error (MAE) loss between the reconstructed and original signals—is minimized. The representative spatial features extracted by the optimized DAE, as illustrated in Figure 8, achieve a dimensionality reduction from 14 signals to 8 features.

3.3.2. Complete life cycle degradation trajectory learning

The TT model recognizes complete life-cycle degradation trajectories by constructing the mapping function outlined in Eq. (3). In this study, the length of the source features, denoted as l, is set to 30. To meet its operational requirements, data separation and padding techniques were implemented prior to training. Furthermore, to enhance the TT model's performance in degradation trajectory recognition, the Hyperband algorithm—a hyperparameter optimization method identical to that employed in the DAE—was utilized.

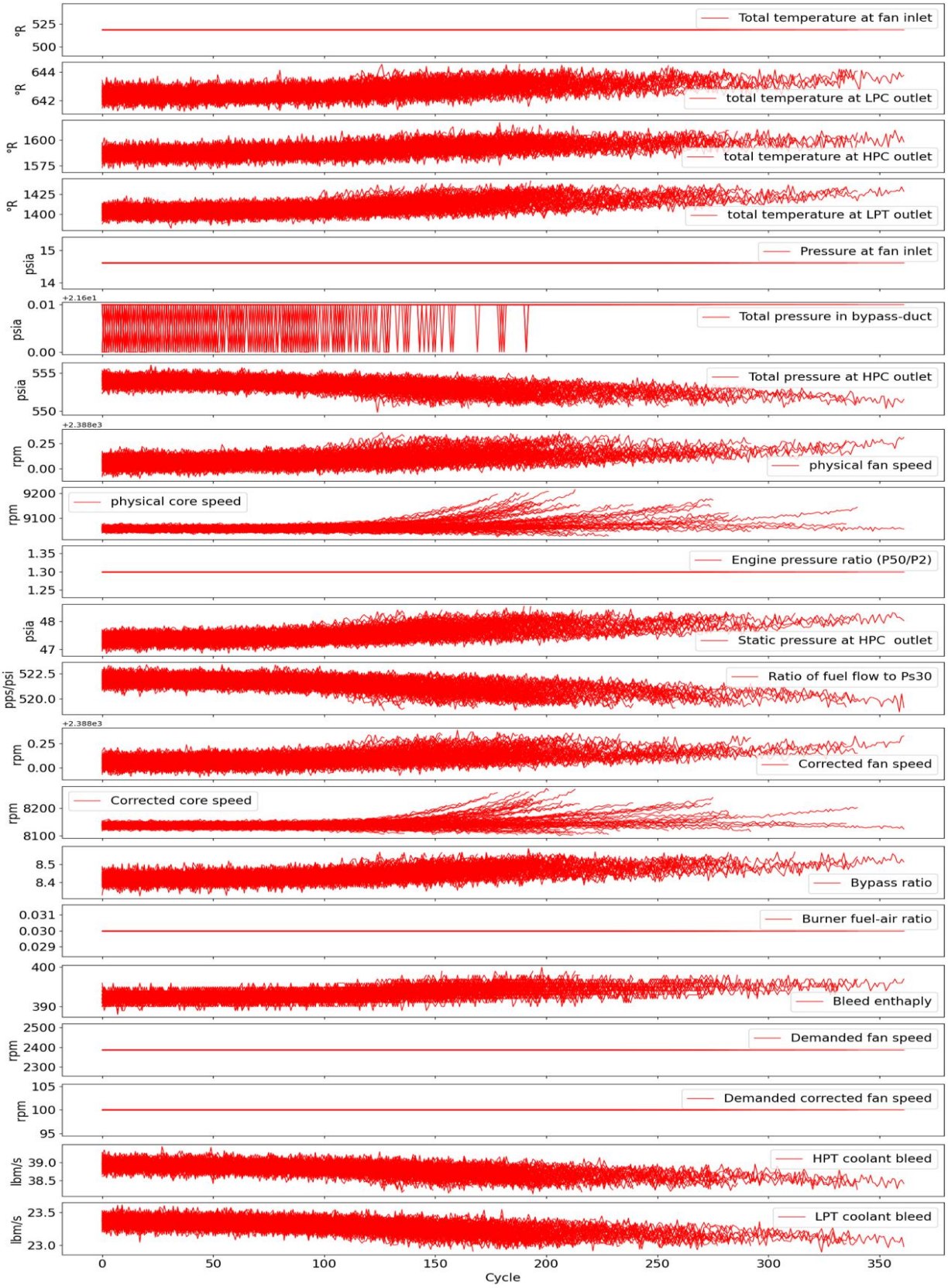

Figure 7. Visualization of signals within the FD001 dataset.

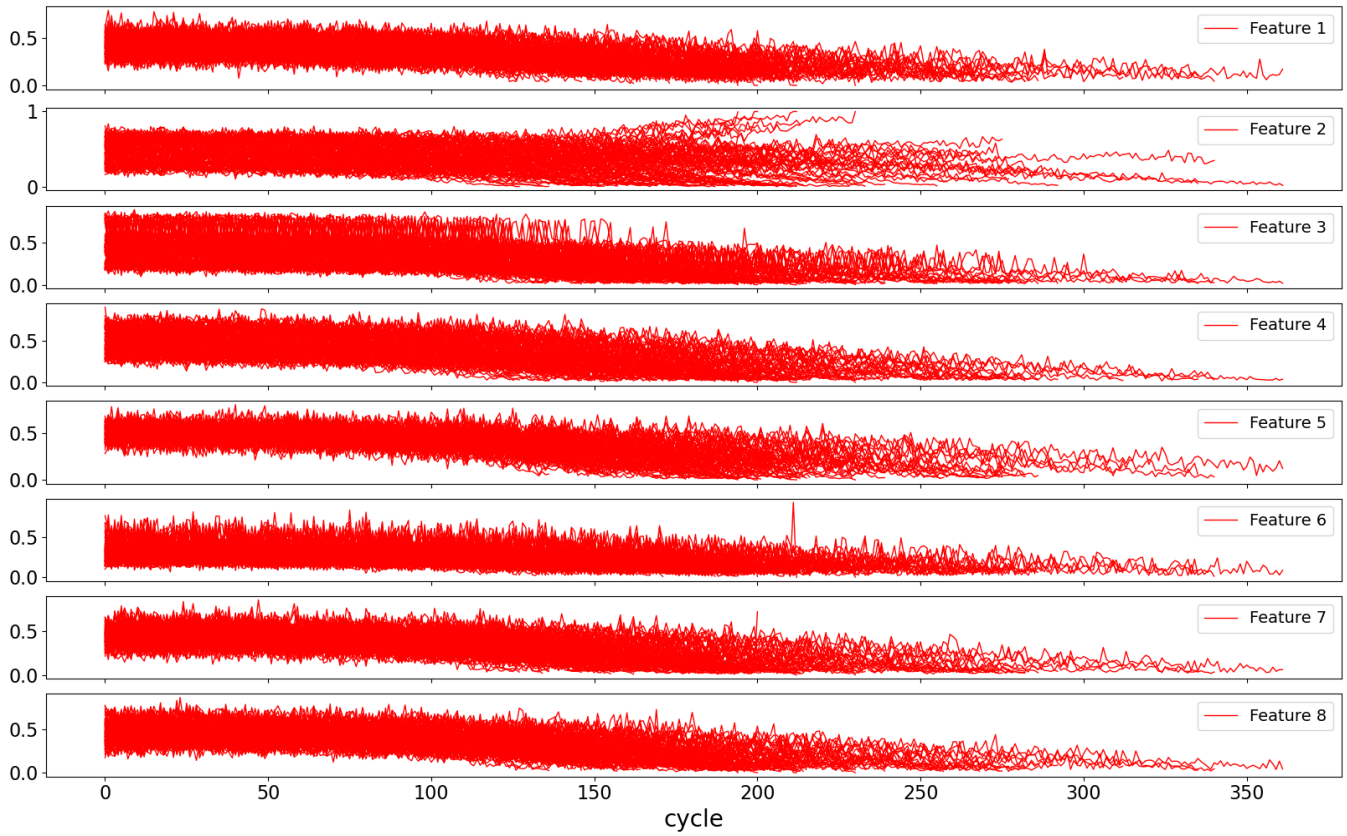

Figure 8. Visualization of extracted features in the FD001 dataset.

3.4. Assessment and comparison of the hybrid Time-Transformer and Denoise Auto-Encoder (TT-DAE)

Performing a testing stage on unknown testing turbofan engine data is essential to comprehensively assess the performance of trained TT model in recognizing complete life cycle degradation trajectories. In this section, the evaluation of the predicted target signals was conducted using the Pearson Correlation Coefficient (PCC) as shown in step 3 in Figure. 3.

Hyperparameter	Value
Number of attention heads in encoder	4
Neurons of feedforward layers in encoder	230
Number of encoder layers	4
Dropout	0.1
Number of decoder layers	4
Neurons of feedforward layers in decoder	330
Learning rate	1e-3
Batch size	512
Neurons of linear layer	230

Table 2. Optimized hyperparameter configuration of the TT model.

3.4.1. Assessment of target signals prediction

The predicted target signals generated by the TT-DAE model, as depicted in Figure 9, demonstrate that the predicted signals (solid red line) closely align with the actual target signals (solid blue line), achieving an average Pearson Correlation Coefficient (PCC) of 0.89. In comparison, the Transformer Encoder-Denoising Autoencoder (TE-DAE) hybrid model yields an average PCC of 0.80, while the Long Short-Term Memory-Denoising Autoencoder (LSTM-DAE) hybrid model attains a score of 0.61. Both TE-DAE and LSTM-DAE employed the Sliding Window (SW) technique for data sampling, in contrast to the Sliding Variable-Length Window (SVW) technique utilized by TT-DAE.

Both TT-DAE and TE-DAE exhibit superior prediction accuracy over LSTM-DAE, highlighting the effectiveness of self-attention mechanisms compared to LSTM for the regression approach. Notably, TT-DAE outperforms TE-DAE (represented by the solid cyan line). Initially, the predicted degradation trajectory from TE-DAE follows the actual trajectory, but it eventually deviates because TE-DAE learns only a portion of the complete life-cycle degradation trajectory. By contrast, the TT-DAE's predicted trajectory remains consistently close to the actual one. This demonstrates that learning the complete life-cycle degradation trajectory is more effective than focusing on a

single portion, resulting in greater accuracy for predicting target signals.

In summary, the proposed TT-DAE model demonstrates superiority in complete life cycle degradation trajectory learning, evident in its heightened prediction accuracy in both target signals and RUL values.

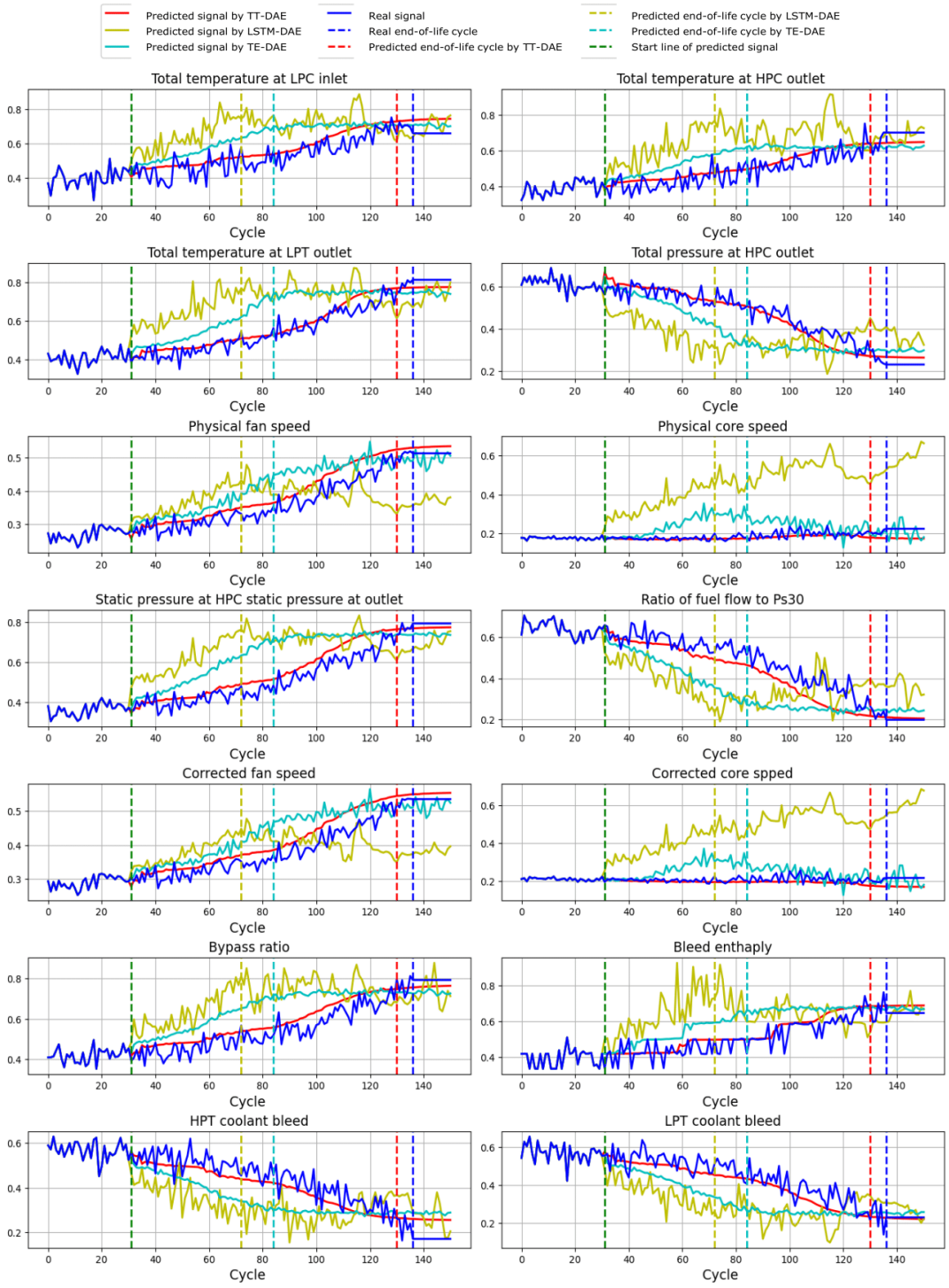

Figure 9. Comparison of target signals predictions: TT-DAE Model, TE-DAE, and LSTM-DAE versus actual target signal in the FD001 dataset.

4. CONCLUSION

This paper primarily addresses the problem of recognizing complete life-cycle degradation trajectories using an indirect mapping method with a regression approach. In this study, we introduce the TT-DAE model for predicting target signals and estimating RUL. Specifically, the DAE first extracts spatial features and mitigates signal noise by reconstructing signals from noisy inputs. These features are then partitioned into source and target domains and standardized to uniform lengths via padding. Subsequently, the TT module reconstructs the mapping function between source and target features to recognize degradation trajectories. To validate its effectiveness, the TT-DAE model was applied to the C-MAPSS dataset, yielding an average Pearson Correlation Coefficient (PCC) of 0.89 between predicted and actual target signals. This performance surpasses that of the TE-DAE, LSTM-DAE, and Transformer-DAE models, highlighting the TT-DAE's superior ability to capture entire degradation trajectories. Looking ahead, the TT-DAE framework could form the basis for diverse multivariate time-series prediction applications, such as traffic flow forecasting. Future research will prioritize evaluating the TT model's performance across varied datasets for RUL prediction.

Acknowledgments The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the industrial collaboration grant from Syarikat Letrik Chen Guan Sdn. Bhd. (NVLD0001).

Author contributions Xianpeng Qiao: Writing – review & editing, Writing – original draft, Methodology. Veronica Lestari Jauw: Writing – review & editing, Supervision, Methodology. Tiyamike Banda: Writing – review & editing. Chin Seong Lim: Supervision, Writing – review & editing.

Data availability The datasets used in this article are public datasets and do not involve any additional data privacy and protocols.

Code availability Code will be made available on request.

Declarations

Competing Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical and informed consent for data used The authors are aware of ethical and informed consent for data used.

REFERENCES

Ahn, G., Yun, H., Hur, S., & Lim, S. (2021). A Time-Series Data Generation Method to Predict Remaining Useful Life. *Processes*, *9*(7), 1115. https://doi.org/10.3390/pr9071115

Frederick, D. K., DeCastro, J. A., & Litt, J. S. (2007). *User's Guide for the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS)* (p. NASA/TM-2007-215026) [Technical Memorandum (TM)]. NASA. https://ntrs.nasa.gov/citations/20070034949

Guo, L., Li, N., Jia, F., Lei, Y., & Lin, J. (2017). A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing*, *240*, 98–109. https://doi.org/10.1016/j.neucom.2017.02.045

Heng, A., Zhang, S., Tan, A. C. C., & Mathew, J. (2009). Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing*, *23*(3), 724–739. https://doi.org/10.1016/j.ymssp.2008.06.009

Jia, X., Li, W., Wang, W., Li, X., & Lee, J. (2020). Development of Multivariate Failure Threshold with Quantifiable Operation Risks in Machine Prognostics. *Annual Conference of the PHM Society*, *12*(1), 9. https://doi.org/10.36001/phmconf.2020.v12i1.1288

Ma, J., Su, H., Zhao, W., & Liu, B. (2018). Predicting the Remaining Useful Life of an Aircraft Engine Using a Stacked Sparse Autoencoder with Multilayer Self-Learning. *Complexity*, *2018*, 1–13. https://doi.org/10.1155/2018/3813029

Qiao, X., Jauw, V. L., Chin Seong, L., & Banda, T. (2024). Advances and limitations in machine learning approaches applied to remaining useful life predictions: A critical review. *The International Journal of Advanced Manufacturing Technology*. https://doi.org/10.1007/s00170-024-14000-0

Qiao, X., Liow, H. Y., Jauw, V. L., & Lim, C. S. (2025). Comparative Study of Deep Learning Model Based Equipment Fault Diagnosis and Prognosis. *International Journal of Prognostics and Health Management*, *16*(1). https://doi.org/10.36001/ijphm.2025.v16i1.4254

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. *2008 International Conference on Prognostics and Health Management*, 1–9. https://doi.org/10.1109/PHM.2008.4711414

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. 652–658. https://doi.org/10.48550/arXiv.1706.03762

Wang, Z., Chen, Y., Cai, Z., Gao, Y., & Wang, L. (2020). Methods for predicting the remaining useful life of equipment in consideration of the random failure threshold. *Journal of Systems Engineering and Electronics*, *31*(2), 415–431. https://doi.org/10.23919/JSEE.2020.000018

Wu, J., Hu, K., Cheng, Y., Zhu, H., Shao, X., & Wang, Y. (2020). Data-driven remaining useful life prediction via multiple sensor signals and deep long short-term

memory neural network. *ISA Transactions*, *97*, 241–250. https://doi.org/10.1016/j.isatra.2019.07.004

Xia, T., Shu, J., Xu, Y., Zheng, Y., & Wang, D. (2022). Multiscale similarity ensemble framework for remaining useful life prediction. *Measurement*, *188*, 110565. https://doi.org/10.1016/j.measurement.2021.110565

Ye, Z., Zhang, Q., Shao, S., Niu, T., & Zhao, Y. (2022). Rolling Bearing Health Indicator Extraction and RUL Prediction Based on Multi-Scale Convolutional Autoencoder. *Applied Sciences*, *12*(11), 5747. https://doi.org/10.3390/app12115747

Zhang, C., Lim, P., Qin, A. K., & Tan, K. C. (2017). Multiobjective Deep Belief Networks Ensemble for Remaining Useful Life Estimation in Prognostics. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2306–2318. https://doi.org/10.1109/TNNLS.2016.2582798

Zhang, J., Li, X., Tian, J., Luo, H., & Yin, S. (2023). An integrated multi-head dual sparse self-attention network for remaining useful life prediction. *Reliab. Eng. Syst. Saf.*, *233*, 109096. https://doi.org/10.1016/j.ress.2023.109096

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(12), 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325