International Journal of Energy
ISSN: 2957-9473 | Vol. 9, No. 3, 2026

A Transformer-based Domain Generalization Method for Equipment Remaining Useful Life Prediction

Shaoshuai Qiu [1,*], Zheng Han [2], Jianguang Liu [1], Hang Liu [1], Zi'an Chen [3]

[1] China Nuclear Power Engineering Co., Ltd. Shenzhen Guangdong, 518000, China
[2] Yangjiang Nuclear Power Co., Ltd. Yangjiang Guangdong, 529500, China
[3] Zhejiang University, Hangzhou Zhejiang, 310000, China

[*] **Corresponding author**: Shaoshuai Qiu (Email: 155039800318@qq.com)

Abstract: Remaining Useful Life (RUL) prediction is crucial for the maintenance decision-making and operational safety of nuclear power systems. However, neural-network-based RUL prediction methods that emerged in recent years face the risk of insufficient generalization when confronted with operating-condition drift caused by power generation peak-shaving, individual differences among components, and so on, which is inconsistent with the high-safety and high-reliability requirements of nuclear power. To address this problem, and also to break through the limitation of conventional domain adaptation methods that rely on target-domain data, this paper proposes a Transformer-based domain generalization method for RUL prediction of nuclear power equipment. The core ideas of the proposed model are "paired learning" and "paired querying." During training, the model uses a cross-attention mechanism to pair-wise learn the positional correspondence between local monitoring signals and the global degradation history to which they belong. During inference, the model pairs the online-monitored local signal one by one with every historical global signal in the training set and queries them, and computes similarity weights based on the Dynamic Time Warping (DTW) algorithm to obtain the final weighted prediction. Cross-domain generalization experiments on the C-MAPSS dataset show that the proposed method, without requiring any target-domain data, significantly outperforms multiple domain adaptation methods that do require target-domain data. This demonstrates that the extracted features, which embed positional information, possess stronger generalization capability and are well suited to nuclear-power scenarios with stringent safety and reliability requirements.

Keywords: Remaining Useful Life; Fault Prognostics; Domain Generalization.

1. Introduction

Remaining Useful Life (RUL) prediction is an important branch of the fault prognostics field. By linking an asset's current health status to its eventual functional failure, RUL prediction delivers three core benefits to nuclear power systems: it supports maintenance decision-making, optimizes upstream and downstream spare-parts planning, and extends the service life of the system [1]. At present, neural-network-based RUL prediction methods gradually become mainstream due to their advantages of requiring less expert experience, being compatible with high-dimensional multi-source heterogeneous data, and mining nonlinear degradation features in an end-to-end fashion [2]. However, nuclear power scenarios are generally characterized by diversity and heterogeneity—for example, operating-condition drift caused by peak-shaving power generation, and individual equipment differences caused by different component models, different reactor types, and different management modes. These factors result in a substantial risk of failure when neural-network-based RUL prediction methods are applied in practice in the nuclear power field. Therefore, RUL prediction methods for the nuclear power field have higher requirements in terms of transferability and generalizability.

In recent years, transfer learning for RUL prediction of nuclear power equipment has become a research hotspot, giving rise to three main directions: instance transfer, feature transfer, and model transfer. Among these, feature transfer, i.e., Domain Adaptation, has received extensive attention because it is compatible with a variety of model architectures. Domain adaptation can be divided into two approaches: one is to design statistical metrics, such as MMD [3] and the Wasserstein distance [4], to minimize the overall or local distribution discrepancy between the source and target domains; the other is to use adversarial methods [5], employing a gradient reversal layer (DANN) so that the feature extractor fools the domain discriminator and thus learns domain-invariant degradation representations. However, domain adaptation relies on target-domain observational data obtained in advance, whereas in many practical situations the target domain is entirely unknown. For example, when transferring an RUL prediction model to a completely new piece of nuclear power equipment or to a new operating condition, this prerequisite cannot be satisfied. Against this background, Domain Generalization (DG) has gradually become a research frontier spanning fields such as fault detection and fault diagnosis. Its goal is to use data from multiple source domains only to train an RUL prediction model that remains robust for unknown operating conditions and unknown equipment types, without touching any observational data from the target domain.

At present, there is still relatively little work on domain generalization for RUL prediction of nuclear power equipment, and the existing works typically set a finite-length time window that slides over complete degradation cases of nuclear equipment (also called global signals) to obtain local signals within a number of windows, and then extract degradation features from the local signals inside each window. It is worth noting that if taken from different degradation cases, similar local features may correspond to drastically different RULs [6], which is an important reason for the poor generalization of existing methods. To address

this problem, this paper proposes a Transformer-based degradation-feature extraction method that primarily exploits the idea of paired learning and paired querying between global and local features. Because the extracted features embody positional information, they exhibit strong generalization and high robustness. In the paired-learning (training) stage, offline window samples and historical cases are randomly paired, and a cross-attention mechanism is used to learn the correspondence between local and global features. In the paired-querying (inference) stage, an online window sample is paired one-by-one with each historical case, and a weighted-averaging mechanism based on Dynamic Time Warping (DTW) is designed to obtain the final RUL prediction. Experiments on the C-MAPSS dataset verify the effectiveness of the proposed method.

2. A Transformer-based Domain Generalization Method for RUL Prediction

This paper proposes a Transformer-based RUL prediction method that aims to improve the model's generalization ability across different operating conditions by means of cross-sequence feature comparison. This section is discussed from two perspectives—model training and online inference. Section 1.1 introduces the overall structure of the model in the training stage, and Section 1.2 elaborates on the inference strategy in the testing stage.

2.1. Paired Learning Method Based on the Cross-Attention Mechanism

In the proposed method, paired learning of local and global features is realized through the encoder–decoder architecture of the Transformer. For equipment degradation cases $\{U_i\}_{i=1}^{C}$ of lengths c_i, these complete degradation cases are called global signals. A sliding time window of length l and step size s is used to extract segments; for example, for the i-th degradation case U_i, $\lceil \frac{c_i-l+1}{s} \rceil$ local signals can be obtained, where $\lceil \cdot \rceil$ denotes rounding up. In this paper, to better preserve the relative positional information of local information within global information, native pairing is adopted between the N global signals and the $\lceil \frac{c_i-l+1}{s} \rceil$ local signals—that is, any local signal is paired only with the global signal from which it was taken—yielding a total of $\sum_{i=1}^{C} \lceil \frac{c_i-l+1}{s} \rceil$ sample pairs.

The local information in each sample pair serves as the input to the encoder of the Transformer, while the global information serves as the input to the decoder, as shown in Figure 1. In the encoder, the local information passes through N embedding layers, self-attention layers, and feed-forward layers, and finally yields the local feature X_{en}^{N}. The above process in the i-th layer of the encoder can be expressed as

$$Z_{en}^{i} = \mathrm{SelfAttention}\left(X_{en}^{i-1}\right) + X_{en}^{i-1}. \quad (1)$$

$$X_{en}^{i} = \mathrm{FeedForward}\left(Z_{en}^{i}\right) + Z_{en}^{i}. \quad (2)$$

where X_{en}^{i-1} is the input of the i-th layer, Z_{en}^{i} is the intermediate state, and the output X_{en}^{i} of the i-th layer serves as the input of the next layer, until the last layer of the encoder is reached. The calculation flow of the self-attention and feed-forward layers in the decoder is largely similar, with the difference that the decoder uses a cross-attention layer to compare and learn the local feature $X_{en}^{N,1}$ delivered by the encoder against the global feature extracted by the decoder itself. At the i-th layer of the decoder, the above process can be expressed as

$$Z_{de}^{i,1} = \mathrm{SelfAttention}\left(X_{de}^{i-1}\right) + X_{de}^{i-1} \quad (3)$$

$$Z_{de}^{i,2} = \mathrm{CrossAttention}\left(Z_{en}^{i,1}\right) + Z_{de}^{i,1} \quad (4)$$

$$X_{de}^{i} = \mathrm{FeedForward}\left(Z_{de}^{i,2}\right) + Z_{de}^{i,2} \quad (5)$$

This comparative learning is analogous to the manual procedure of judging RUL, i.e., one observes information such as the amplitude and trend of the current monitoring signal (corresponding to the encoder), compares it with the amplitude and trend of historical cases (corresponding to the self-attention layer of the decoder), and judges the RUL based on its positional information (corresponding to the cross-attention layer of the decoder). The positional information obtained via paired learning is passed through a linear layer to obtain the final RUL prediction:

$$\tilde{y} = \mathrm{Linear}\left(X_{de}^{N}\right) \quad (6)$$

The above model is trained using the MSE loss function, i.e., for the input pair

$$\mathcal{X}_{i,j} = \{u_i, p_{i,j}\} \quad (7)$$

$$\mathcal{L}\left(y_{i,j}; \mathcal{X}_{i,j}\right) = \mathrm{MSE}\left(y_{i,j}, \hat{y}_{i,j}\right) = \mathrm{MSE}\left(y_{i,j}, f_\theta(\mathcal{X}_{i,j})\right) \quad (8)$$

where $p_{i,j}$ denotes the k-th local sample taken from the i-th degradation case, $\hat{y}_{i,j}$ denotes the model output, f_θ denotes the proposed model, and θ denotes the learnable model parameters.

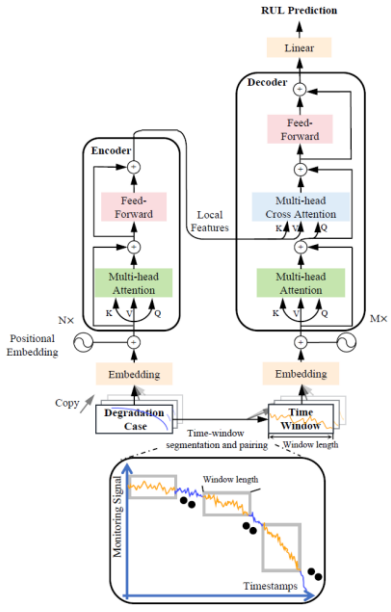

Figure 1. Overall structure of the model during the training phase

2.2. Paired Query Method Based on DTW Weighting

As mentioned above, if taken from different degradation cases, similar monitoring data may represent very different RULs. Therefore, in the online inference stage, to achieve accurate RUL prediction it is still necessary to pair local features with global features. Since the equipment has not yet fully failed at this stage, all the data obtained are right-censored samples; the global feature natively paired with the current local feature cannot be obtained (because it has not

yet happened). Therefore, this paper proposes to approximate native pairing by placing the current local feature into a historical global-feature library (i.e., the training set) for paired querying. Among such schemes, the simplest method is to pair with each historical global feature one by one and take the arithmetic mean—i.e., for the monitoring signal $x \in \mathbb{R}^{l\times d}$ within the current window, C pairs $\{U_i, x\}_{i=1,\cdots,C}$ are obtained, and the RUL prediction is

$$y_{pred} = \frac{1}{C}\sum_{i=1}^{C} f_\theta(\{U_i, x\}) \quad (9)$$

However, this approach ignores the differences among different degradation cases. Intuitively, historical cases whose degradation pattern is similar to that of the current censored data should be assigned larger weights, which requires measuring the similarity between the current censored data and historical global information. Given the significant difference in time lengths between the two, traditional methods such as the L2 distance cannot be used to measure similarity; this paper therefore adopts segmental Dynamic Time Warping (sDTW) to compute the weighting coefficients. Standard DTW requires the two sequences to be fully aligned head-to-head and tail-to-tail, whereas a censored segment is only a small early part of a degradation case. In nonlinear degradation, the former has a gentler trend while the latter has a steeper trend, so forcing full head-and-tail alignment would forcibly stretch the tail of the censored data to the last point of the paired degradation case, causing loss of trend information in the data. sDTW permits local alignment in which the query segment can start and end at arbitrary positions in the long sequence, which fits the requirement that the censored data are the early part of a degradation case yet still need to be compared against the similarity of the degradation case. Specifically, for

$$y_{pred} = \mathrm{Softmax}(\alpha)^{T} f_\theta(\{U_i, x\}) \quad (10)$$

where the weight vector $\alpha \in \mathbb{R}^{C\times 1}$ and the element α_i is given by sDTW [7], that is, for $i = 1,2,\ldots,C$

$$\alpha_i = \mathrm{sDTW}(U_i, x) \quad (11)$$

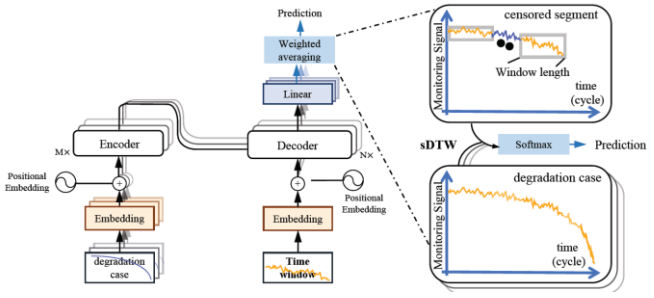

Figure 2. Overall structure of the model during the testing phase

3. Case Study

C-MAPSS is an RUL prediction benchmark dataset generated by a simulation model developed by NASA. It simulates the operation and degradation process of large engines under different operating conditions and provides a reproducible and comparable experimental platform for data-driven research on Prognostics and Health Management (PHM). The C-MAPSS dataset contains multiple subsets (FD001–FD004), each of which simulates the full-life-cycle degradation of several equipment units under different operating conditions and fault modes, until complete failure. The number of units, operating conditions, and fault modes in each subset are listed in Table 1. Because the operating conditions and fault modes differ across subsets, the data distributions among them differ significantly, and a model trained on one subset cannot simply be reused on another, otherwise the model would face the risk of failure.

Table 1. A Brief Overview of the C-MAPSS Dataset

Subset	Train / Test Units	Operating Conditions	Fault Modes
FD001	100/100	1	HPC
FD002	260/259	6	HPC
FD003	100/100	1	HPC+Fan
FD004	248/249	6	HPC+Fan

This paper selects several subsets from the C-MAPSS dataset and performs transfer tasks among them, as listed in the first column of Table 2. For example, 1→3 (FD001→FD003) means training on subset FD001 (called the source domain) and performing inference on subset FD003 (called the target domain). As noted earlier, research on domain generalization in the RUL prediction field is relatively scarce, so this paper settles for the next best option and lists several typical domain adaptation methods for performance comparison, including the Deep Adversarial Neural Network (DANN), the CORAL-distance-based cross-domain alignment method, and Adversarial Discriminative Domain Adaptation (ADDA). Note that the proposed method belongs to domain generalization and only requires source-domain data during training, whereas all listed baselines belong to domain adaptation and thus additionally require unlabeled target-domain data during training. The results in Table 2 show that, with less data required, the proposed method nevertheless outperforms the comparison methods that rely on more information. This fully indicates that the features extracted by the proposed method embody positional information and therefore exhibit strong generalization and high robustness. The last column of Table 2 shows the results of a baseline trained only on the source domain and directly inferred on the target domain, which also demonstrates that simply copying a model trained on one subset to another subset leads to a substantial degradation in prediction performance and poses a very high risk of failure.

Table 2. Experimental Results of Transfer Tasks on the C-MAPSS Dataset (Lower is Better)

Task	Proposed	DANN	CORAL	ADDA	Source Only
Metric	**RMSE**	**RMSE**	**RMSE**	**RMSE**	**RMSE**
1→3	21.6	39.8	41.2	39.7	50.2
2→3	32.0	44.5	42.3	32.6	142.2
2→4	32.0	43.6	55.3	34.4	47.4
3→1	14.3	29.2	46.9	20.0	43.6
4→1	21.3	41.7	70.6	37.9	191.3
4→2	19.7	40.0	46.9	28.7	36.4
4→3	22.6	40.6	48.5	14.1	147.8
Average	**23.2**	**39.9**	**50.2**	**29.6**	**94.13**

Furthermore, this paper uses all degradation cases of FD001 as the training data, and the degradation cases of engines No. 1, 40, 70, and 90 in FD003 as the test data. The results are shown in Figure 3, where the orange line denotes the label (the true RUL value) and the blue line denotes the predicted RUL. As can be seen from Figure 3, even though the training and test data come from different operating

conditions, because the degradation features extracted by the proposed method possess broad generality, the method can transfer and generalize under different operating conditions, exhibiting relatively robust RUL prediction capability.

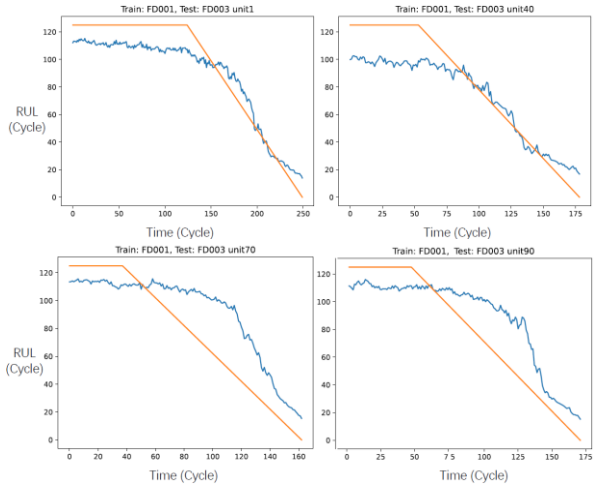

Figure 3. RUL prediction results of the proposed method on the specified engine units

4. Conclusion

To address the deficiency that data-driven RUL prediction models suffer significant performance degradation when facing unknown operating conditions and unknown equipment, this paper proposes and verifies a Transformer-based domain generalization method for RUL prediction of nuclear power equipment. The study identifies a key deficiency of existing methods: for monitoring signals of nuclear power equipment, similar local degradation features located at different global degradation stages may correspond to drastically different RULs. To solve this problem, the core contribution of this research is to propose the innovative framework of "paired learning" and "paired querying." By exploiting the cross-attention mechanism within the Transformer architecture, the model can effectively encode the relative positional information of a local signal within its complete life cycle. In the online prediction stage, by adopting a strategy of weighted paired querying between real-time data and the historical database, the model draws on the historical experience most similar to the current degradation pattern, thereby providing more accurate predictions. Experimental results on multiple transfer tasks on the C-MAPSS benchmark dataset verify the effectiveness of the proposed method. Compared with various mainstream domain adaptation methods, the proposed method achieves lower prediction errors (average RMSE of 23.2) without any access to target-domain data. This fully demonstrates that by explicitly learning the positional information of a local signal within the global signal, one can extract equipment degradation representations with stronger generalization and higher robustness, effectively avoiding the risk of substantial performance degradation when a model is transferred to a new domain, and better meeting the high-safety and high-reliability requirements of nuclear power.

References

[1] WANG Xiaopeng, WANG Lei, HAN Xiaowei, et al. K-Means clustering-based particle swarm optimized CNN-BiGRU-HAM method for engine remaining useful life prediction [J]. Machine Tool & Hydraulics, 2024, 52(20).

[2] JIA Xiaolin, LIU Jia. Prediction of the remaining useful life of bearings [J]. Advances in Applied Mathematics, 2025, 14: 140.

[3] Cui J, Zhang Y, Miao Q. Remaining Useful Life Prediction for Electro-Mechanical Actuator with Scale-Aware Domain Adaptive Deep Transfer Learning [J]. IEEE Transactions on Instrumentation and Measurement, 2025.

[4] Shi H, Huang C, Zhang X, et al. Wasserstein distance based multi-scale adversarial domain adaptation method for remaining useful life prediction [J]. Applied Intelligence, 2023, 53(3): 3622–3637.

[5] CHEN Renxiang, ZHANG Yanfeng, XU Xiangyang, et al. Remaining life prediction of rolling bearings of different models based on a subspace domain adversarial discriminative network [J]. Chinese Journal of Scientific Instrument, 2024, 45(3): 119–127.

[6] Chen Z, Jin X, Kong Z, et al. Global and local information integrated network for remaining useful life prediction [J]. Engineering Applications of Artificial Intelligence, 2023, 126: 106956.

[7] Tsai T J. Segmental DTW: A parallelizable alternative to dynamic time warping [C] // ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 106–110.