A Two-Stage Framework for Small-Sample RUL Prediction on Structurally Complex Time-Series Data

Peng Gao[1], Fanyu Qi[1], Yizhang Zhu[1], Jianyu Zhang[1], Wenfei Li[1], and Jianshe Feng[1]

[1]*School of Advanced Manufacturing, Sun Yat-sen University, Shenzhen, 518107, China*
fengjsh7@mail.sysu.edu.cn

ABSTRACT

Remaining Useful Life prediction for high-value assets, such as aero-engines, presents a formidable challenge, compounded by sample scarcity, complex data structures and knowledge dilution. This paper proposes a two-stage framework designed to decouple representation construction from temporal pattern learning. Stage I mitigates data complexity by transforming multi-phase snapshot streams into standardized cycle-level sequences through hierarchical aggregation. Stage II addresses data scarcity and multi-target prediction using a Multi-task Shared Transformer. Furthermore, the model is optimized via a risk-aligned loss function that penalizes tardy predictions. The effectiveness of the proposed framework was validated by its strong generalization on PHM 2025 Data Challenge dataset, which ultimately secured a first-place result.

Keywords: remaining useful life, predictive maintenance, Transformer, prognostics and health management.

1. INTRODUCTION

Reliable Remaining Useful Life (RUL) estimation is a central tenet of predictive maintenance, fundamental to ensuring operational safety and optimizing economic efficiency. However, the development of generalizable RUL prediction models is impeded by three challenges: (i) Sample Scarcity, where small fleets of high-value assets render expressive models prone to overfitting; (ii) Deep Structural Complexity, where sensor data are organized in a deep hierarchy from engine level down through period, cycle, phase and snapshot in Figure 1; and (iii) Knowledge Dilution, arising from conventional per-device modeling approaches that fail to leverage shared physical degradation patterns across a fleet.

2. METHODOLOGY

To address the dual challenges of data complexity and scarcity, this study employs a two-stage framework designed for an effective bias-variance trade-off as shown in Figure 2. This approach firstly regularizes the problem by reducing data complexity in a structured manner (Stage I), thereby creating a simplified representation space where expressive temporal models can learn cross-unit patterns without overfitting (Stage II).

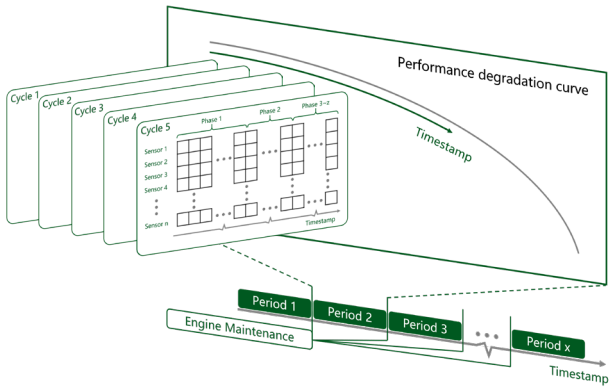

Figure 1. Hierarchical structure of aero-engine time-series data. This diagram illustrates deeply nested data where long-term degradation signals are obscured by short-term operational variations.

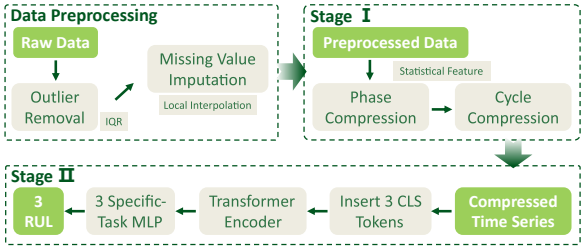

Figure 2. Overall architecture of two-stage framework. The framework first decouples data complexity via hierarchical aggregation (Stage I) and then learns temporal patterns using a Multi-task Shared Transformer (Stage II).

Prior to analysis, raw sensor signals were preprocessed to ensure data quality and integrity; this involved Interquartile Range (IQR) for outlier mitigation and local interpolation for handling missing values.

2.1. Stage I: Data Complexity Decoupling via Hierarchical Aggregation

Stage I transforms nested sensor data into a fixed-dimension time series. This transformation is achieved through a bottom-up aggregation process executed within each flight cycle as illustrated in Figure 3:

1. Phase Compression: For each operational phase, multi-sensor snapshots are averaged to yield a single phase-level vector.
2. Cycle Compression: The set of phase vectors is compressed into one cycle vector by computing statistics per channel, such as mean, variance, peak-to-peak, maximum and minimum.

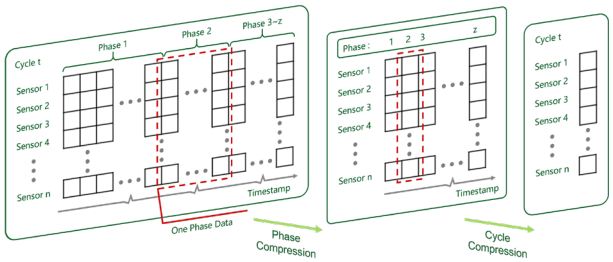

Figure 3. Schematic of Hierarchical Aggregation in Stage I, which transforms snapshot data into standardized, cycle-level sequences.

2.2. Stage II: Multi-task Shared Transformer with Risk-Aligned Optimization

This stage performs temporal pattern learning using a Multi-task Shared Transformer as illustrated in Figure 4. A Transformer encoder is shared across all devices to learn universal degradation features from cycle-level sequences. Task specialization is achieved by prepending three learnable [CLS] tokens to each input sequence, corresponding to three maintenance targets. The final hidden state of each [CLS] token serves as the input to an independent MLP head.

The training is guided by a risk-aligned, asymmetric time-weighted loss function designed to penalize tardy predictions more heavily:

$$L(y,\hat{y}) = w(y,\hat{y}) \cdot (\hat{y}-y)^2 \cdot \beta_k \qquad (1)$$

where the weight $w(y,\hat{y}) = (1 \; or \; 2)/\,(1+\alpha y)$ applies a higher penalty for late predictions ($\hat{y} > y$) and incorporates a time-decay factor α, and β_k is a task-specific factor to balance loss magnitudes across different RUL scales.

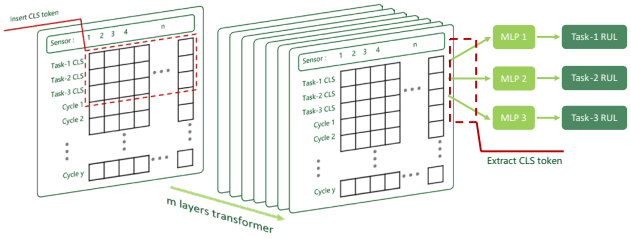

Figure 4. Architecture of Multi-task Shared Transformer in Stage II. A shared encoder learns universal features, while task-specific [CLS] tokens and MLP heads enable specialized predictions for distinct maintenance targets.

3. EXPERIMENTS AND RESULTS

The proposed framework was evaluated on the PHM 2025 Data Challenge dataset, which comprises complete operational histories from four commercial aero-engines. A leave-one-out cross-validation strategy was employed, where in each fold, the model was trained on data from three engines and tested on the remaining unseen engine. The model architecture consists of a 7-layer, 8-head Transformer encoder with an input sequence length of 151 cycles.

3.1. Ablation Study on Loss Function

To validate the effectiveness of the proposed loss function, its performance was compared against standard Mean Squared Error (MSE). As summarized in Table 1, the risk-aligned loss consistently yields substantial performance gains across all held-out engines, confirming its superiority for this prediction task.

Table 1. Comparison of custom loss vs. MSE on local cross-validation scores (lower is better).

Engine (ESN)	101	102	103	104
Scores with custom loss	**45.20**	**52.02**	**47.47**	**51.49**
Scores with MSE loss	107.08	157.85	121.29	194.61

3.2. Performance

As illustrated in Figure 5, highly consistent performance was observed between the validation and test datasets, with the horizontal axis ordered according to the final validation leaderboard rankings. This consistency indicates that the overfitting effect was minimized in the proposed framework. The small performance gap between the two datasets can be attributed to the regularization introduced by Stage I and the cross-unit learning strategy in Stage II.

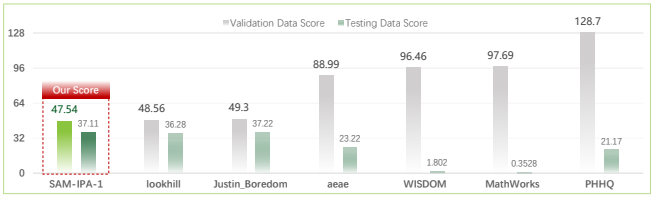

Figure 5. Final leaderboard scores of the PHM 2025 Challenge. Our framework (highlighted) demonstrates superior performance and generalization compared to other top-ranking solutions.

4. CONCLUSION

A two-stage framework is proposed in this paper to address the dual challenges of data complexity and sample scarcity in RUL prediction. In Stage I, structural complexity is reduced through hierarchical aggregation, creating a regularized representation space. In Stage II, a multi-task shared Transformer is employed to learn cross-unit degradation patterns, enabling effective knowledge transfer across limited samples. Strong generalization capability was demonstrated on the PHM 2025 Data Challenge, validating the effectiveness of the proposed framework for predictive maintenance in high-value industrial assets.