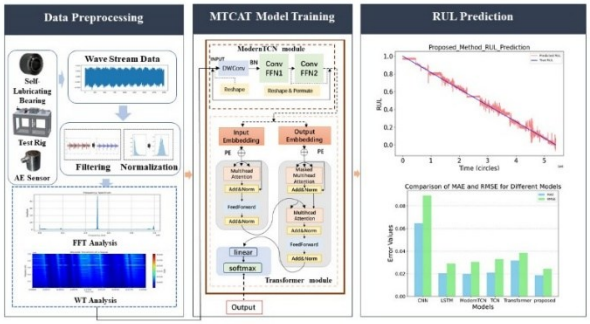
MTCAT: A Modern Temporal Convolution and Enhanced Attention Transformer Model for Remaining Useful Life Prediction of Aerospace Self-Lubricating Bearings

Yuheng Wang, Shuo Wang, Shichang Du, Member, IEEE, Xiaoxiao shen, Member, IEEE, Shanshan Li, Xianmin Chen

***Abstract*—As mission-critical components in aircraft flap actuation systems, aerospace self-lubricating bearings play a pivotal role in ensuring operational safety during critical flight phases. Conventional machine learning approaches for remaining useful life (RUL) prediction suffer from limited time-series and nonlinear modelling capabilities. Especially individual models, built on a single algorithmic paradigm, struggle to address the complex, multi-dimensional challenges of RUL prediction. In order to solve the above challenge, this study proposes an innovative hybrid deep learning method integrating modernized temporal convolutional Networks (ModernTCN) with attention-enhanced Transformer (MTCAT) for RUL prediction. First, ModernTCN employs optimized dilated convolutions to capture multi-scale local temporal patterns efficiently. Second, an attention-enhanced Transformer leverages multi-head self-attention to model long-range dependencies and global degradation trends. Finally, the MTCAT is validated on the custom test rig, replicating real-world conditions, where it achieves a 30-50% reduction in root mean square error (RMSE) compared to baseline models. By autonomously learning hierarchical features from raw data, the MTCAT model minimizes manual feature engineering, demonstrating robustness and scalability for RUL prediction.**



***Index Terms*—RUL prediction, Signal Processing, Transformer, ModernTCN, Self-lubricating bearing**

I. Introduction

As critical components in aircraft flap hinge mechanisms, self-lubricating bearings play a pivotal role in ensuring safe takeoff and landing operations, making the prediction of their remaining useful life particularly important [1]. Unlike conventional rolling bearings requiring periodic maintenance, self-lubricating films at friction interfaces, to achieve Automatic lubrication counterparts utilize integrated solid lubricants or self-adaptive effectively [2]. The operational lifespan of these bearings is governed by complex multi-physics interactions involving material degradation kinetics, tribological evolution, thermo-mechanical coupling effects, and stochastic load spectrum variations, creating substantial challenges for accurate RUL estimation.

RUL is a critical aspect of ensuring the effectiveness of prognostics and health management (PHM) [3]. There are two main types of approaches for predicting the life of self-lubricating bearings: the first is the statistical method based on test data, while the second involves developing the physical model of the bearing to derive a life prediction model [4-8]. The RUL prediction method based on statistical model estimates the remaining life of the machine by establishing a statistical model based on mathematical statistics. The prediction result is usually given in the form of conditional probability density function based on observation [9]. The statistical approaches use inspection data such as vibration signals and fit them to statistical models like artificial neural networks (ANN) and autoregressive models (ARMA) to predict the RUL of bearings. Common methods include autoregressive model, Markov model, principal component analysis (PCA), proportional hazard model, etc. [10-12]. This approach is beneficial for its adaptability and accuracy, especially when long-term data is available. The physical-based method mainly depends on comprehensive knowledge of physical mechanisms and environmental factors. Common models

This work was funded by National Natural Science Foundation of China (Grant No. 92467101 and 52275499) and Open Fund of National Key Laboratory of Strength and Structural Integrity (No. LSSIKFJJ202402003). (Corresponding author: Shichang Du)

Yuheng Wang, Shuo Wang Shichang Du, and Xiaoxiao Shen are with the Department of Industrial Engineering and Management, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China(e-mail: wangyuheng77@sjtu.edu.cn; SJD8013@sjtu.edu.cn; lovbin@sjtu.edu.cn; sjtusxx98@sjtu.edu.cn).

Shanshan Li and Xianmin Chen are with the National Key Laboratory of Strength and Aircraft Strength Research Institute of China (ASRI), Xi'an 310014, China (e-mail: shanshanmail2010@163.com; chenxm007@avic.com)

include the Paris and Forman models, which are primarily used to describe crack propagation [13], and Finite Element Analysis (FEA) is commonly used for predicting tool wear and die service life [14]. Physical model prediction methods, based on well-defined physical laws, offer high accuracy and interpretability, making them effective in cases with well-understood degradation mechanisms.

Traditional data-driven methods have effectively addressed RUL prediction by leveraging statistical and machine learning techniques. In recent years, deep learning approaches have significantly advanced this field, offering enhanced capabilities to model complex, nonlinear degradation patterns in bearings [15]. Compared to traditional machine learning methods, deep learning models possess strong feature learning capabilities, nonlinear modeling abilities, end-to-end learning capacities, enhanced model expressiveness, and superior generalization capabilities. These advantages enable deep learning models to effectively extract and leverage information from data, resulting in more accurate and reliable life predictions [16]. Common deep learning methods include Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), Generative Adversarial Networks (GAN), and Transformer models [17].

The emergence of large language models (LLMs) has fundamentally transformed time-series prognostics, establishing Transformer architectures as the dominant paradigm for industrial RUL prediction [18][19]. Recent innovations include Wang et al.'s [20] deep multi-scale window transformer (DMW-Trans), which implements multi-resolution time-frequency analysis through cascaded window transformer blocks and feature fusion modules, progressively expanding receptive fields to encapsulate cross-scale degradation characteristics. Hou et al. [21] developed a cross-domain adaptive framework employing multi-modal feature disentanglement and two-stage transformer architecture, effectively addressing operational condition variability through secondary feature engineering and domain-invariant representation learning. Cutting-edge solutions like Li et al.'s [22] knowledge-enhanced convolutional transformer (KE-CTEM) synergize physical knowledge embedding with data-driven approaches, introducing a feature extraction neural network (FENN) to construct physics-informed feature banks that enhance model generalizability under variable operating conditions. Parallel advancements in temporal convolution architectures include Qin et al.'s [23] adaptive multi-scale integration model, which innovatively embeds multi-scale evaluation indices into neural network dataflows through gated convolutional units (GCUs) and dilated TCN structures. The progression towards end-to-end learning is exemplified by Sun et al.'s [24] MDSCT framework, which combines multi-scale separable convolutions with optimized Transformer encoders (PPSformer) to directly process raw vibration signals while maintaining both local feature sensitivity and global dependency modelling.

Hybrid model approaches have also gained significant popularity. Liu et al. [25] introduced a Transformer-based model with spatio-temporal attention to predict bearing RUL from noisy vibration signals. The model captures both spatial dependencies (across multiple sensors) and temporal degradation trends, using a sliding window approach for time-series processing. Chen et al. [26] presented a domain-adaptive deep learning framework using a CNN-Transformer hybrid to predict RUL across different bearing types and machines. The model employs transfer learning to align source and target domain features, addressing variability in sensor data distributions. Wu et al. [27] propose a hybrid model combining multi-scale CNNs and gated recurrent units (GRUs) to extract and process degradation features from bearing vibration signals. The model uses a feature fusion strategy to integrate local and global patterns, improving RUL prediction under variable operating conditions.

Despite the development of various methods for predicting the RUL of self-lubricated bearings, conventional monolithic algorithms still encounter significant limitations in adapting to multidimensional data and modeling cross-domain degradation mechanisms in industrial settings. Moreover, these approaches often fail to incorporate time-series features. To overcome the persistent challenges in RUL prediction, an innovative hybrid method that synergistically integrates Modernized Temporal Convolutional Networks (ModernTCN) with attention-enhanced Transformer (MTCAT) is proposed, specifically engineered for aerospace self-lubricating bearing RUL prediction under real operational constraints. The contribution of this paper is as follows:

1) Dual-modal temporal fusion mechanism: A ModernTCN-Transformer co-modelling method is pioneered, it establishes cross-attention connections between ModernTCN's hierarchical dilated convolutions and Transformer's multi-head self-attention mechanisms. This architecture achieves complementary advantages: the ModernTCN excels in local temporal feature extraction through optimized causal dilation structures, while the Transformer captures global degradation trajectories via self-attention-based dependency modelling.

2) Multi-scale representation learning: MTCAT implements three-tier temporal optimization through (a) ModernTCN's adaptive receptive field expansion with progressive dilation rates, (b) Transformer's dimensionally robust position encoding enhanced by degradation-aware attention weights, and (c) hybrid feature fusion gates that dynamically balance local-global temporal representations across varying operational phases.

The remainder of this paper is organized as follows: Section II elucidates the detailed innovations and mathematical formulations of the proposed method. Section III and section Ⅳ presents comprehensive experimental validation using proprietary aerospace bearing full-life testing rig. Section V discusses comparative performance benchmarks and provides concluding remarks with industrial implementation insights.

II. Proposed Method

In this section, MTCAT model is proposed, which aims to effectively deal with complex dependencies in temporal data, further improve the prediction accuracy and enhance the generalization ability. The ModernTCN module can expand the receptive field through a layer-by-layer inflated convolution, which can expand the sensory field and capture feature information of different time scales at multiple levels, the feature expression capability is strengthened by virtue of its powerful self-attention mechanism. The combination of the two enables the proposed method to fully utilize their respective strengths, forming an efficient solution for the prediction of complex time-series data. The entire design strategy of the model is

introduced first, covering the complete workflow from data preprocessing and model construction to lifespan prediction. Then, the model's dimensional enhancement and feature decoupling in time-series modelling is described. Finally, how optimization is performed in event sequence modelling is explained.

A. Model Design Strategy

The workflow diagram is shown in Fig. 1. First, raw bearing signals are preprocessed into frequency-domain features. Second, MTCAT adopts a serial design. Specifically, the ModernTCN module first processes the time domain data, frequency domain data, and time-frequency domain analysis data calculated from the original waveform stream data to capture local and global features. Then the Transformer module receives the feature representation output by the ModernTCN module, and further deepens the modelling capabilities of time-series data through a multiple attention mechanism to achieve dynamic modelling of global dependencies. Finally, the linear output layer maps the learned features into scalar prediction results, thus completing the entire modelling process. This design enables the model to integrate the short-term dynamic changes and long-term dependency patterns of time-series data, bringing significant prediction performance advantages.

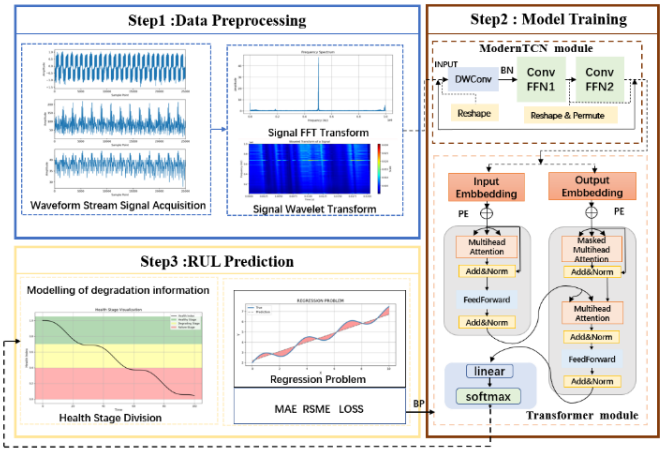


Fig. 1. MTCAT method workflow diagram

1) Step 1: Preprocessing raw bearing signals into frequency-domain features via Fast Fourier Transform (FFT) and wavelet transform (WT).

The workflow commences with the acquisition of raw bearing vibration signals as one-dimensional time-series data, typically sampled at 25.6 kHz for 1.28 seconds, yielding 32,768 data points per sample. These signals undergo FFT to convert them into the frequency domain, generating a spectrogram that captures frequency-domain features like dominant frequencies and their amplitudes, which are indicative of degradation patterns. The resulting data is structured into a two-dimensional format, such as [60, 1024], representing 60-time windows (1-minute intervals) and 1024 frequency bins, preparing it for model input.

2) Step 2: Training a hybrid ModernTCN-Transformer model for robust time-series modelling

The preprocessed features are fed into a hybrid ModernTCN-Transformer model, starting with the ModernTCN module, which uses depth-wise convolutions with dilated kernels to extract local temporal patterns (e.g., short-term vibration changes), enhanced by batch normalization and convolutional feedforward networks. The Transformer module then processes these features using multi-head self-attention to capture long-range dependencies (e.g., linking early degradation to failure trends), with positional encoding preserving temporal order and feedforward layers refining global representations. With the local sensitivity of ModernTCN with Transformer's global modelling, MTCAT ensuring robust time-series analysis across multi-scale degradation dynamics. Table I shows the training process of the proposed model.

TABLE I
TRAINING PROCESS OF THE MTCAT

Module	Description
RUL dataset	Custom dataset class used to generate bearing data based on time-series. It supports sliding feature windows and can add time-related features like relative position and differential features.
Positional encoding	Positional encoding module used to provide positional information within the sequence. It generates position encodings using sine and cosine functions and adds them to the input features.
Modernized conv1d	Modernized convolution module that performs convolution using multiple kernel sizes (e.g., 3, 5, 7) to extract multi-scale time-series features, and outputs the average of these features.
Improved ModernTCN	Improved Modern TCN that extracts time-series features through input projection, convolution modules, and residual connections, outputting the features after convolution.
Improved ModernTCN-Transformer	A hybrid model combining the improved Modern TCN and Transformer. It includes feature embedding, ModernTCN feature extraction, positional encoding, Transformer encoder, and global attention mechanism to extract key features in the time-series and predict RUL.
Monotonic regression loss	Monotonic regularization loss function that combines RMSE loss with regularization that encourages predictions to decrease over time, improving the accuracy of the predictions.
Training	The training process includes forward pass, loss computation, backpropagation, gradient updates, and early stopping strategy, using the AdamW optimizer and learning rate scheduler.
Evaluation	Evaluates the model's performance, calculates and outputs metrics like MAE and RMSE, and saves the prediction results to a file.

3) Step 3. Predicting RUL with regression, optimization, and degradation visualization

The model frames RUL prediction as a regression task, extracts feature to remaining life, and uses Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) losses through backpropagation. It also visualizes the bearing's degradation stages including healthy, mild, and severe stage, and a health index curve (1 to 0) over time is built [32]. This dual approach of regression-based prediction and stage visualization enhances practical applicability for bearing health management, enabling proactive maintenance strategies.

For the MTCAT building part, the input sequence is a one-dimensional time-series of preprocessed bearing vibration signals after FFT, reshaped into a 2D/3D tensor for the model input. It captures time-frequency features reflecting bearing degradation. ModernTCN extracts local patterns, Transformer models global trends, enabling accurate RUL prediction. The input sequence $X = \{x_1, x_2 \ldots x_T\}$ is first projected into the latent space

through a linear mapping:

$$z_t = W_{in} x_t + b_{in}, t = 1,...,T \quad (1)$$

where $W_{in} \in R^{d_{model} \times d_{input}}$ is the weight matrix and d_{model} is the hidden dimension of the model. Next, in the ModernTCN module, parallel convolution branches with multi-scale convolution kernels (for example, sizes of 3, 5, 7, and 11) [22] are used. Specifically, the convolution output of branch i is:

$$h_t^{(i)} = \mathrm{GELU}(\mathrm{BN}(\mathrm{Conv1d}^{(i)}(z_t))), i = 1,..., K \quad (2)$$

where K represents the number of convolution branches, $\sigma(\cdot)$ represents the activation function (GELU), and BN represents batch normalization. Finally, the fused feature representation is obtained by averaging or other aggregation models:

$$h_t = \frac{1}{K}\sum_{i=1}^{K} h_t^{(i)} \quad (3)$$

Then, it passes through the position encoding module, whose formula is the classic sine-cosine encoding:

$$PE_{(t,2i)} = \sin\left(\frac{t}{10000^{\frac{2i}{d_{model}}}}\right) \quad (4)$$

$$PE_{(t,2i+1)} = \cos\left(\frac{t}{10000^{\frac{2i}{d_{model}}}}\right) \quad (5)$$

After adding the positional encoding to the output of the convolutional module, it is input into the Transformer module for global dependency modelling. The output of the Transformer encoder is a sequence $\{t_1, t_2, \ldots, t_T\}$. In order to obtain a global sequence representation, global attention pooling is introduced:

$$\alpha_t = \frac{\exp(w_a^T t_T)}{\sum_{j=1}^{T}\exp(w_a^T t_j)}, g = \sum_{t=1}^{T}\alpha_t \quad (6)$$

where w_a are learnable parameters. The final prediction is mapped to a scalar output through a regression head:

$$\hat{y} = f_{reg}(g) = w_r^T\ \mathrm{GELU}(W_r g + b_r) + b_r \quad (7)$$

The ModernTCN module, which undertakes the key feature extraction task, is designed to consist of multiple Temporal Blocks, each of which uses convolution operations with different expansion rates to gradually expand the sensory field. Specifically, the first layer of convolution sets the convolution kernel size to 3 and the padding parameter to 2, which can efficiently capture short-term local features; the second layer further expands the receptive field by setting the convolution operation with an expansion rate of 2 to capture longer-term dependencies; and the expansion rates of the convolution operations in the third and the fourth layers are increased to 4 and 8, respectively, in order to enhance the ability of modelling long-term dependencies. This layer-by-layer expansion of the sensory field ensures that the model is able to comprehensively capture information at different time scales, significantly improving the learning of complex temporal patterns.

The Transformer module builds on the features extracted by ModernTCN to further enhance the modelling of global temporal dependencies through a multi-head attention mechanism. The multi-head attention mechanism achieves the modelling of dynamic dependencies between global time steps by calculating the relationship weights between different positions in the input sequence, thus making up for the shortcomings of the traditional temporal modelling models in global dependency capturing [33]. In addition, in order to fully exploit the expressive power of features, the Transformer module designs a set of feature dimension expansion and dimensionality reduction strategies, which firstly enhances the feature dimension from 16 to 2048 through a linear layer in order to learn more complex feature relationships in high-dimensional space, and then reduces the dimension to 16 through a second linear layer to adapt to the subsequent tasks. In order to prevent the overfitting problem, two layers of Dropout operation are introduced in the module, with the dropout probability of each layer set to 0.2, and the LayerNorm operation is added to each layer to ensure the stability of the input data, so as to accelerate the convergence speed of the model.

The linear output layer, as the final module of the model, further maps the high-dimensional features output from the Transformer into scalar values to adapt to the regression task. Through this design, the whole model realizes the organic integration of local features and global information in the process of combining ModernTCN and Transformer modules, which can efficiently extract useful information from time-series data and generate accurate prediction results.

B. Dimensionality Enhancement and Feature Decoupling

The decoupling design of features in the ModernTCN module is one of its innovations, focusing on independent modelling of time and variable dimensions. In traditional convolutional models, mixed modelling of time and variable dimensions often leads to confusion in information interaction between features, making it difficult to effectively capture dynamic relationships in a single dimension. ModernTCN achieves independent modelling in the time dimension by introducing deep separable convolution (DWConv), which is designed to capture the temporal dynamic characteristics of each variable separately and avoid interference from other variable information. In addition, DWConv significantly expands the effective receptive field (ERF) by applying larger convolution kernels, which can capture complex dependencies on long time scales, making the model modelling in the time dimension more accurate. Furthermore, in addition to the original features, position features and differential features are added to the input data.

In order to make full use of high-dimensional information, the model uses linear transformations between different modules to convert dimensions. Before the Transformer, feature expansion and dimensionality reduction are performed through a feedforward network. The process is as follows:

$$u = \mathrm{GELU}(W_1 z + b_1), z' = W_2 u + b_2 \quad (8)$$

ModernTCN enhances time-series modelling by integrating DWConv and Expansion Convolution, effectively capturing both short-term and long-term temporal dependencies. The architecture employs DWConv to model individual variable dynamics independently, while Expansion Convolution progressively ERFs through layered operations, enabling multi-scale temporal pattern extraction across different time horizons. Compared with traditional models, the considering of the long length of time-series data, multi-scale convolution is used in model design to adjust the receptive field. Assuming the

convolution kernel size is k and the expansion rate is d, the receptive field of each convolution operation is:

$$\text{ReceptiveField} = d \times (k-1) + 1 \tag{9}$$

C. Optimization for Time-Series Data

The design of ModernTCN-Transformer takes into account the unique characteristics of time-series data, and provides proprietary optimizations in multiple dimensions to enhance the model's modelling capability and performance. Time-series data often have complex time-dependent patterns, multivariate interactions, and dynamically changing feature distributions, which pose challenges for model design. To address these characteristics, MTCAT achieves exclusive optimization through innovative design.

For the multivariate interaction characteristics in time-series data, ModernTCN introduces an exclusive optimization for cross variable dependency modelling. In the ConvFFN module [28], univariate features (ConvFFN1) and inter-variate dependencies (ConvFFN2) are captured separately through the decomposition mechanism, ensuring that the model can effectively model complex interactions in multivariate time-series. For example, in the multivariate missing value filling task, the complete data of certain variables can be used to infer the values of the missing variables through cross-variate dependencies, thus improving the filling accuracy. The optimization is particularly applicable to the analysis of multivariate time-series data and can significantly enhance the model's adaptability in complex task scenarios.

In order to further adapt to the dynamically changing feature distributions in time-series data, the model introduces feature normalization and regularization strategies. LayerNorm is used to normalize the features of each time step, and the calculation formula is:

$$\text{LayerNorm}(z) = \gamma \odot \frac{z-\mu}{\sqrt{\sigma^2+\varepsilon}} \tag{10}$$

where μ and σ^2 are the feature mean and variance respectively, and γ, β are learnable parameters.

The model can dynamically adjust the feature distribution to cope with the numerical fluctuations of different time periods or variables, thus improving the stability of the model. To prevent overfitting and stabilize the training process, regularization measures are introduced in each module. Common regularization formulas include Dropout, which sets the output of some neurons to zero with probability p in the forward propagation:

$$\text{Dropout}(z) = z \odot \text{m}, \text{m} \sim \text{Bernoulli}(1-p) \tag{11}$$

Meanwhile, the Dropout operation (with a dropout probability of 0.2) is introduced into the Transformer module. These measures make the model more robust in dealing with non-stationarity and noise in time-series data.

D. Loss Function Design

To enhance the transparency and reproducibility of the work, the explicit form of the composite loss function have now been included and provided. A detailed explanation for incorporating monotonicity into the loss design.

The formulation is as follows:

$$L_{mono} = \sum_{i=1}^{n} \left| \frac{d\hat{y}_i}{dx_i} - sign\left(\frac{dy_i}{dx_i}\right) \right| \tag{12}$$

The monotonic regularization loss is defined as: where $\hat{y}_i$ is the predicted value, y_i is the actual value, x_i represents the input feature, and sign denotes the direction of the gradient. The purpose of this regularization term is to ensure that the predicted changes align with the actual data's trend, especially in cases where monotonicity needs to be preserved, helping the model better reflect physical laws or domain-specific constraints.

To combine monotonicity with the RMSE loss, the authors integrate both components into a composite loss function as follows:

$$L_{total} = \lambda L_{RMSE} + (1-\lambda) L_{mono} \tag{13}$$

III. Case Study Ⅰ: Self-Lubricating Bearing

A. Self-Lubricating Bearing Experiment Rig

The experimental object is a self-lubricating bearing with precise geometric specifications, as depicted in the technical drawing. Fig. 2 shows the self-lubricating bearing of our experiment. It features an outer diameter of 25 mm, an inner diameter of 15 mm, and a width of 19.5 mm, with a spherical outer raceway curvature of 0.254 mm. The bearing's design, including a 6° angular misalignment capability, ensures reliable performance under varying loads and oscillatory conditions, making it suitable for testing in simulated aircraft operating environments.

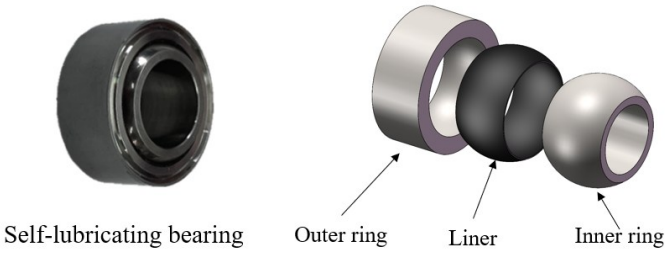


Fig. 2. Self-lubricating bearing of the experiment

The bearing is first mounted to the bearing housing to be measured in the axial direction, and the tension and pressure transfer is realized by connecting the loading plate via the loading electric cylinder, while the loading plate is connected to the position of the bearing housing to be measured via the long screw to realize the tension and pressure loading. Measurement of the axial loading force is realized by means of a force transducer. In the radial direction, the loading module is connected to the servo-electric cylinder through the loading actuator, the cylinder is mounted on the bracket through the front flange, and the bracket of the servo-electric cylinder is fastened to the cast iron base through the connecting rod and bolt. The test rig is shown in Fig. 3.

The test rig for evaluating the lifespan of self-lubricating bearings comprises four key systems: the transmission system, control system, loading system, and measurement system. The

transmission system, driven by a servo motor (Delta ECMA-J11010RS) and a 35:1 reduction gearbox. The control system, utilizing an industrial computer and servo drivers, ensures accurate motion and load parameter regulation, while the loading system employs servo-electric cylinders and load cells to apply radial and axial forces, mimicking real-world stresses. The measurement system records essential parameters such as torque, force, and displacement for comprehensive performance assessment.

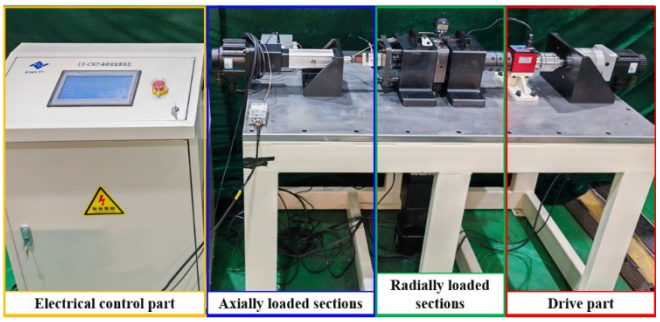


Fig. 3. Overall structure of test rig for self-lubricating bearings

Fig. 4 shows the detailed rig, the signal acquisition process starts with the bearing generating mechanical responses under controlled loading and motion, which are captured by Acoustic emission (AE), torque, axial, and radial force sensors. These raw signals are amplified by signal amplifiers, digitized, and processed to extract features for bearing health evaluation. The workflow ensures comprehensive data collection, enabling accurate analysis of the bearing's performance and degradation behavior in simulated operational environments.

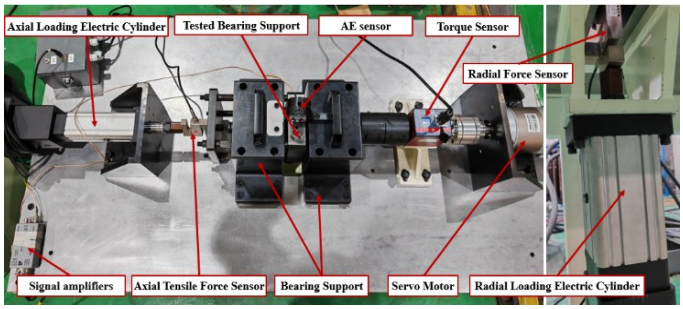


Fig. 4. Detailed part of test rig for self-lubricating bearings

B. Feature Analysis

The dataset contains data from the operation of self-lubricating bearings under simulated real operating conditions. These data were collected by a variety of sensors, including signals from torque sensors, temperature sensors, and AE sensors. AE signals and torque signals were focused. The data come from long-term experimental studies and field tests, covering a wide range of failure modes and operating conditions, which can effectively reflect the loss and wear characteristics of flap hinges under different operating conditions. Table II-III demonstrates the difference of different data processing methods. The features are in Table III. The feature selection is shown in Fig. 5, the correlation heatmap of top 10 features of RUL is showed and Top 5 feature trends in Fig. 6-7.

Feature importance by scoring features based on their prognosability, monotonicity, and trendability is accessed to evaluate their predictive power, relationship consistency, and trend capture ability. Prognosability measures a feature's ability to predict changes in the target variable. Monotonicity evaluates whether the relationship between the feature and the target variable consistently increases or decreases. Trendability assesses the feature's effectiveness in capturing trends over time or across other variables. Through the scoring method, the features with the most impact on the model can be identified.

TABLE II
DATA PREPROCESSING METHOD

Feature	FFT	WT
Time-frequency resolution	Fixed, with clear frequency-domain information but blurred time-domain information.	Dynamic, providing both time and frequency local information with multi-scale analysis capability.
Applicable scenario	Global frequency analysis, suitable for periodic signals or stable frequency components.	Local feature analysis, suitable for non-stationary signals, transient signals, edge detection, etc.
Computational complexity	Depends on the signal length and FFT algorithm, generally less complex.	Depends on the specific wavelet basis and decomposition level, usually more complex than FFT.
Frequency resolution	Improves with signal length, suitable for analyzing periodic signals.	Analyzes signals at different scales, suitable for analyzing complex and non-stationary signals.

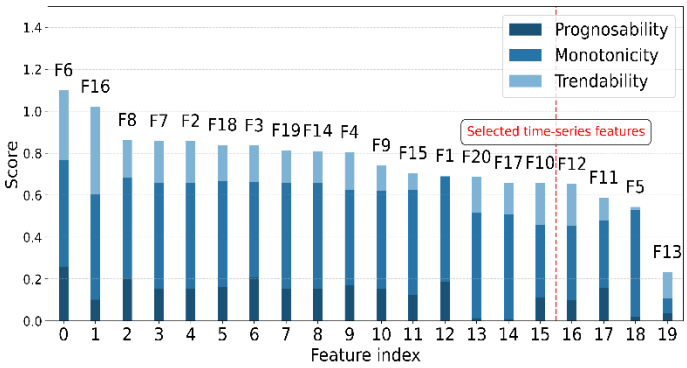


Fig. 5. Selected features based on the rank

Fig. 5 presents a bar chart of selected time-series features for RUL prediction of self-lubricating bearings, derived from acoustic emission and torque signals collected under simulated operating conditions. It displays 20 features (F0 to F19) with their scores (0 to 1.4) based on prognosability (dark blue) and monotonicity and trendability (light blue), with a red dashed line separating the top 10 features (e.g., F6, F16, F8) for their high predictive relevance. The dataset, sourced from long-term tests capturing diverse failure modes, supports the feature selection process outlined in reference [2], with further analysis in subsequent correlation heatmaps and trend visualizations.

TABLE III
CALCULATED FEATURES

Feature name	Explanation
mean	The average value of the signal, reflecting the overall trend of the signal.
std	The standard deviation of the signal, indicating the fluctuation and stability of the signal.
max	The maximum amplitude in the signal, used to detect extreme abnormal conditions.
min	The minimum amplitude in the signal, used to detect extreme abnormal conditions.
skew	The asymmetry of the signal; large skewness may indicate nonlinear faults.
kurtosis	The sharpness of the signal; high kurtosis is often associated with sudden changes during faults.
rms	Root Mean Square value of the signal, typically used to represent the vibration or power of the bearing.
Peak to peak	The range of signal fluctuations, helpful for analyzing the operating amplitude of the bearing.
Max slope	The rate of change of the signal, helping to capture rapid changes in the bearing's status.
q1	The first quartile of the data, representing the lower distribution of the signal.
q2	The second quartile (median) of the data, reflecting the middle trend of the signal.
q3	The third quartile of the data, representing the upper distribution of the signal.
Peak frequency	The primary frequency component of the signal, reflecting the vibration characteristics of the bearing.
Total energy	The total energy of the signal, measuring the work intensity of the bearing.
Spectral entropy	The complexity or disorder of the signal; high entropy indicates high complexity, often related to faults.
bandwidth	The frequency range of the signal, helping to analyze the breadth of the bearing's vibration spectrum.
RMSfrequency	The RMS frequency in the frequency domain, representing the frequency characteristics of the signal.
Mean amplitude	The average amplitude of the signal, evaluating the energy level of the bearing's operation.
Amplitude variance	The variance of the amplitude fluctuations, reflecting the operational stability of the bearing.
Frequency center of mass	The overall position of the energy distribution in the frequency spectrum, helping to understand the concentration of the spectrum.

The correlation heatmap is a 10×10 grid where each cell displays the Pearson correlation coefficient between feature pairs, ranging from -1.00 (strong negative correlation, deep blue) to 1.00 (strong positive correlation, deep red). The color gradient intuitively indicates the strength of the correlation. Fig. 6 shows the correlation heatmap of top 10 features of RUL. The diagonal cells have an implicit value of 1.00, representing perfect correlation of features with themselves, while off-diagonal cells display the correlations between different features. For example, the correlation coefficient between q1 and q3 is as high as 0.99 (deep red), indicating a very strong positive correlation, while the correlation coefficient between min and q1 is -0.99 (deep blue), showing a very strong negative correlation. Additionally, the correlation between scaled RUL and other features is generally weak, with bandwidth (0.42) and kurtosis (0.33) showing the highest positive correlations.

By analyzing the heatmap results of the features, significant multicollinearity between features is found. For instance, the correlation coefficient between rms and std is 1.00, indicating potential redundancy. Moreover, the correlation analysis between scaled RUL and other features suggests that bandwidth and kurtosis could be important variables for predicting RUL. Highly correlated feature pairs (such as q1 and q3) may only need to retain one to reduce model complexity and improve prediction accuracy. Fig. 7 shows the Top 5 feature trends. The results provide important references for subsequent feature engineering and machine learning model development, aiming to enhance the robustness and accuracy of scaled RUL prediction.

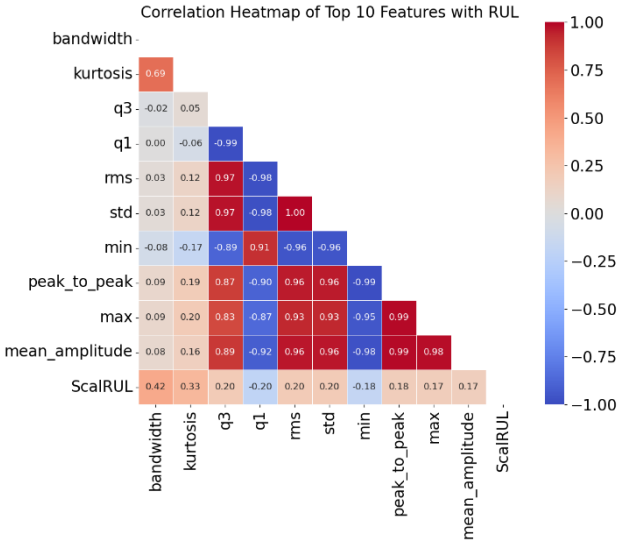


Fig. 6. Correlation heatmap of top 10 features of RUL

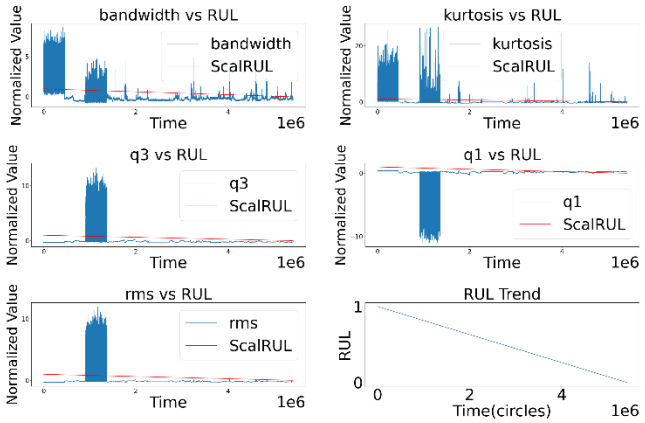


Fig. 7. Top 5 feature trends of RUL

C. Results and Performance Analysis

Because the service life of self-lubricating bearings can reach more than 100,000, it is very difficult to collect experimental data for the entire life cycle. Segmented division for one tested self-lubricating bearing is used, 0.7 for training set, 0.15 for validation set and 0.15 for test set. The experimental results are shown in Fig 8. For each model, the x-axis represents time in seconds, and the y-axis represents the RUL. The predictions of the models are shown as red lines, while the true RUL values are represented by blue lines.

Fig. 8 demonstrates the prediction accuracy of each model and their ability to track the true RUL values over time, with variations observed across different models. Fig. 8 actually

presents a comparative analysis of RUL predictions over a 50000 circles period for six models: (a) CNN, (b) LSTM, (c) ModernTCN, (d) TCN, (e) Transformer, and (f) the proposed model. Each subplot displays the predicted RUL (red line) against the true RUL (blue line), with the x-axis representing time in hours and the y-axis showing RUL values ranging from 0.0 to 1.0.

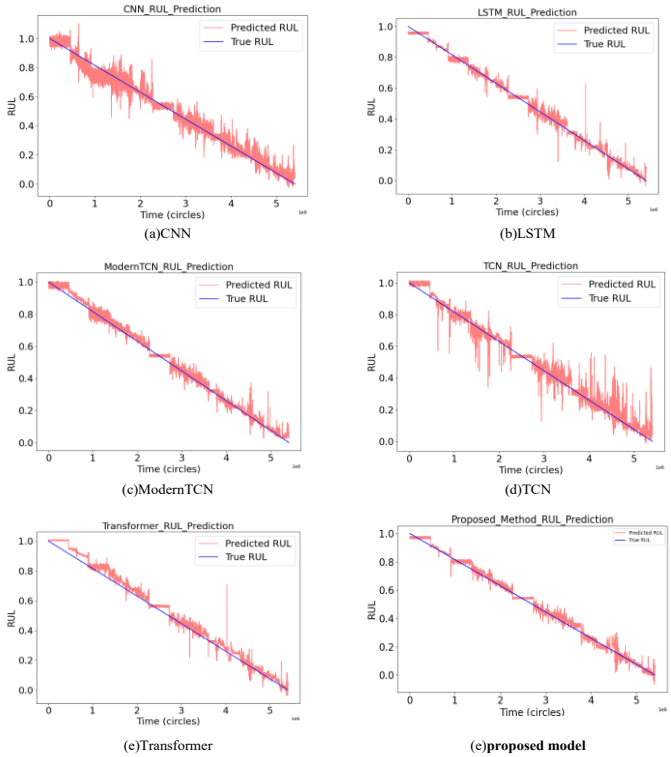


Fig. 8. RUL prediction result of the tested self-lubricating bearing

The subfigures (a)-(e) correspond to: (a)CNN: Displays the RUL prediction using CNN. (b)LSTM: Shows the RUL prediction using a long short-term memory (LSTM) network. (c)ModernTCN: Represents the RUL prediction using a modern Temporal Convolutional Network. (d)TCN: Shows the prediction using a Temporal Convolutional Network. (e)Transformer: Demonstrates the RUL prediction using a Transformer model. (f)Proposed model: Represents MTCAT model for RUL prediction.

All models capture the overall downward trend of RUL over time, but they exhibit varying degrees of accuracy. The CNN, LSTM, ModernTCN, TCN, and Transformer models display noticeable deviations between predicted and true RUL values, with frequent over- or under-predictions, particularly in regions of rapid RUL changes. In contrast, the proposed model demonstrates the closest alignment with the true RUL values, maintaining consistency across the time series and effectively capturing both the trend and fluctuations. The superior performance highlights the proposed model's enhanced capability in RUL prediction compared to the baseline models. The proposed model effectively reduces noise compared to other models, and the prediction accuracy is significantly higher than other models. Table IV and Fig. 9 show the MAE and RSME of each model.

MAE measures the average error of the prediction and is insensitive to outliers, while RMSE emphasizes large errors by squared errors and can more sensitively reflect the impact of outliers.

TABLE IV
PREDICTION ERROR COMPARISONS OF THE MODELS

Model	MAE	RSME
CNN	0.0647	0.0892
LSTM	0.0205	0.0290
ModernTCN	0.0200	0.0304
TCN	0.021	0.033
Transformer	0.0317	0.0385
Proposed	**0.0188**	**0.0244**

The bar chart illustrates a comparative analysis of MAE and RMSE across six different models: CNN, LSTM, ModernTCN, TCN, Transformer, and the proposed model. The x-axis represents the models, while the y-axis denotes the error values, ranging from 0.00 to 0.09. The CNN model exhibits the highest error, with an MAE of approximately 0.07 and an RMSE of 0.09, indicating the poorest performance. In contrast, the proposed model demonstrates the lowest error values, with an MAE and RMSE both around 0.02, suggesting superior predictive accuracy. The remaining models like LSTM, ModernTCN, TCN, and Transformer show moderate performance, with MAE values around 0.02–0.03 and RMSE values ranging from 0.03 to 0.04. These results highlight the proposed model's effectiveness in minimizing prediction errors compared to established models.

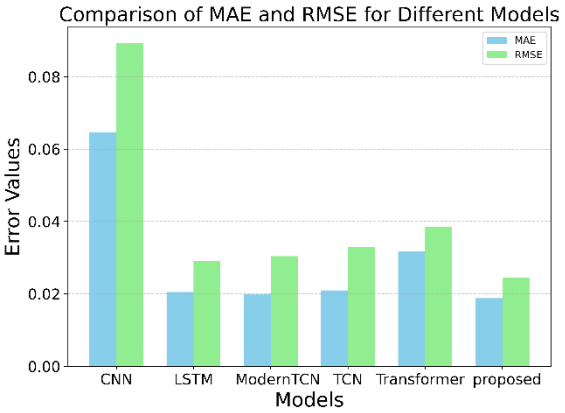


Fig. 9. Model error comparison of the tested bearing

The interpretability of various types of models for prediction accuracy is analyzed as follows:

1) CNN model: convolutional neural network is widely used in image and time-series data analysis, which effectively extracts local features through convolutional layers. Although CNN has a great advantage in capturing local features, it does not perform as well as other models with more global modelling capabilities when dealing with long sequences because it is unable to effectively deal with global dependencies in long time-series.

2) LSTM model: the LSTM network is good at capturing long-time dependencies and is particularly suitable for time-series prediction. However, LSTM may encounter the problem of vanishing or exploding gradients during training and high

computational complexity when dealing with high-dimensional features.

3) TCN and ModernTCN model: the models overcome the shortcomings of using CNN or LSTM alone to a certain extent, but still suffer from higher computational efficiency and model complexity.

4) Original Transformer model: the Transformer model effectively models global dependencies in sequences through the self-attention mechanism, which is particularly suitable for long time sequences. However, its learning ability for local features is weak, and the computational complexity is also higher when long sequences are input.

Compared to the above traditional models, the proposed model exhibits superior performance by combining ModernTCN to enhance the ability to capture local features, while leveraging on Transformer's global dependency modelling capability. Specifically, the RMSE and MAE of the improved model in the RUL prediction task are lower than those of the other compared models, indicating a significant improvement in accuracy and generalization ability.

The computational time for each model was recorded across all experimental runs, and the average time per training epoch and total training time are now provided. Additionally, the hardware and software configurations used during the experiments are clarified in Table V.

TABLE V
COMPUTATIONAL TIME AND EXPERIMENTAL SETTINGS

Model	Platform Specifications	Training Time (h)	Inference Time (s)
CNN	NVIDIA RTX 3090, Ubuntu 20.04, PyTorch 2.0	2.5	10.2
LSTM		4.8	15.6
ModernTCN		3.2	12.4
TCN		3.0	11.8
Transformer		5.5	18.3
Proposed Model		**2.8**	**11.0**

CASE STUDY II : XJTU-SY BEARING EXPERIMENT

D. XJTU-SY Bearing Dataset

The XJTU-SY bearing dataset is widely used for bearing RUL prediction and health management, containing vibration and sensor data under various loads and speeds [29]. It records bearing status and operating time, supporting the evaluation of machine learning and deep learning models for time-series analysis. In order to validate the effectiveness and robustness of the model, a different kind of bearing for experiment validation is used. Table V describes the XJTU-SY dataset and shows the detailed information of the RUL task [30].

In order to verify the generalization and robustness of the model, multiple sets of experiments using bearings under different working conditions as training sets, validation sets, test sets, and validation sets are designed. It can effectively prove the generalization of MTCAT and solve possible overfitting problems. Table VI lists the different tasks.

TABLE VI
XJTU-SY DATASET DESCRIPTION

Bearing Dataset	Operating condition	Bearing number	FPT (min)	RUL (min)
XJTU-SY	C1	B1_1	78	45
	2100rpm	B1_2	36	125
	12kN	B1_3	58	100
		B1_4	29	23
	C2	B2_1	452	39
	2250m	B2_2	46	115
	11kN	B2_3	302	231
		B2_4	121	218
	C3	B3_1	2346	192
	2400rpm	B3_2	340	30
	10kN	B3_3	1417	98
		B3_4	10	104

TABLE VII
RUL DATASET LISTING

Task	Training	Validation	Testing
A1	B1_1 ~ B1_4	B2_1	B2_2
A2	B1_1 ~ B1_4	B2_2	B2_3
A3	B2_1 ~ B2_4	B3_1	B3_3
A4	B2_1 ~ B2_4	B3_1	B3_4
A5	B3_1 ~ B3_4	B1_1	B1_2
A6	B3_1 ~ B3_4	B1_2	B1_4

E. Feature Analysis

Signal analysis is performed on the original waveform stream file and obtained a total of 32 features in the horizontal and vertical directions, as shown in the Table VII.

TABLE VIII
CALCULATED FEATURES

Label	Feature (Vertical and Horizontal signal)	Explanation
F1, F2	maximum	Maximum value of the vertical and horizontal vibration signals.
F3, F4	minimum	Minimum value of the vertical and horizontal vibration signals.
F5, F6	mean	Mean value of the vertical and horizontal vibration signals.
F7, F8	peak_to_peak	Peak-to-peak value of the vertical and horizontal vibration signals (dynamic range).
F9, F10	variance	Variance of the vertical and horizontal vibration signals, indicating the spread of data points.
F11, F12	std_dev	Standard deviation of the vertical and horizontal vibration signals, representing deviation from mean.
F13, F14	rms	Root Mean Square value of the vertical and horizontal vibration signals, quantifying the signal power.
F15, F16	skewness	Skewness of the vertical and horizontal vibration signals, indicating asymmetry of the signal's distribution.
F17, F18	kurtosis	Kurtosis of the vertical and horizontal vibration signals, indicating the "tailedness" of the signal.
F19, F20	mean_absolute	Mean absolute value of the vertical and horizontal vibration signals, measuring signal magnitude.
F21, F22	mean_absolute_square	Mean square of the absolute value of the vertical and horizontal vibration signals.
F23, F24	impulse_factor	Impulse factor of the vertical and horizontal vibration signals, identifying sharp spikes in signals.
F25, F26	crest_factor	Crest factor of the vertical and horizontal vibration signals, identifying peaks relative to noise.
F27, F28	margin_factor	Margin factor of the vertical and horizontal vibration signals, indicating abnormal signal behavior.

For features analysis, trendiness is the consistency of the trend of the feature value changing over time and the RUL change trend. Monotonicity is the consistency of the direction of the feature value change, that is, whether it always increases or always decreases. Predictability is the degree of correlation between the feature value and the RUL. Fig. 10. shows the score of the features of task A1, derived from acoustic emission and torque signals collected under simulated operating conditions. It displays 19 features (F0 to F31) with their scores (0 to 1.4) based on prognosability (dark blue) and monotonicity & trendability (light blue), with a red dashed line separating the top 16 features (e.g., F29 F32, F19) for their high predictive relevance.

F. Results and Performance Analysis

The RMSE and MAE are used to measure the quantitative prediction performance of the proposed model. Two sets of task results are selected for display and used the following models as control groups: CNN, LSTM, TCN, ModernTCN, Transformer Fig. 11 and Fig. 12 show the RUL prediction results of A1 and A2 respectively. In the plots, the red predicted RUL values are compared to the true RUL values (black line), and the prediction error (blue shaded area) shows the range within which the model's predictions are expected to vary. The plot helps visualize the accuracy and uncertainty of the model's RUL predictions.

Fig. 11 and Fig. 12 present the RUL prediction results with error bounds for Task A1 and Task A2, respectively, across six models: (a) CNN, (b) LSTM, (c) ModernTCN, (d) TCN, (e) Transformer, and (f) the proposed model. Each subplot illustrates the true RUL (black line), predicted RUL (red line), 90% confidence interval (CI) of the predicted RUL (light blue shaded area), and prediction error (blue shaded area) over time in cycles, with Task A1 spanning 70 cycles and Task A2 extending to 175 cycles.

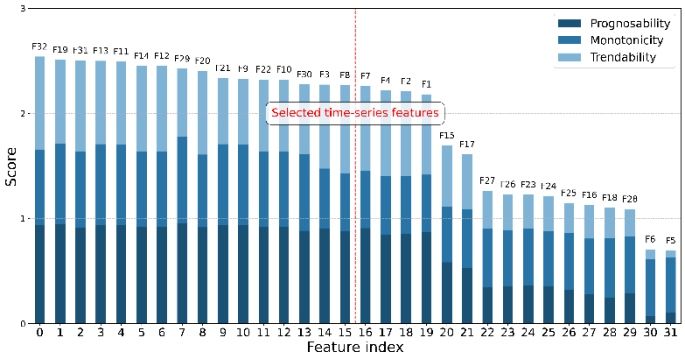


Fig. 10. Selected features based on the score

The y-axis represents RUL values from 1.0 — 0.0. In both tasks, the CNN, LSTM, ModernTCN, TCN, and Transformer models exhibit significant deviations between predicted and true RUL, with wider 95% CI bounds and larger prediction errors, particularly in regions of rapid RUL decline. Notably, the ModernTCN and TCN models show the most pronounced fluctuations and overestimations in both tasks. Some model predicts wrong RUL values from 0.8 — 0.0.

In contrast, the proposed model consistently demonstrates the closest alignment with the true RUL, with narrower 90% CI bounds and minimal prediction errors across both tasks, indicating superior predictive accuracy and reliability in RUL estimation. The proposed model outperforms the CNN, LSTM, Modern TCN, TCN, and Transformer models in terms of both prediction accuracy and error minimization. It shows a better fit to the true RUL values, with reduced prediction errors and more stable performance across the entire dataset. The other models, while performing relatively well, especially Transformer and Modern TCN, still exhibit greater fluctuation in the predictions and higher prediction errors.

Compared with the above traditional models, the improved Transformer-ModernTCN model enhances the ability to capture local features by combining conventional network, and at the same time, with the help of Transformer's global dependency modelling capabilities, it shows better generalization and robustness.

IV. Conclusion

For the prediction of RUL of aerospace self-lubricating bearings, this study proposes a RUL prediction model based on the MTCAT model. MTCAT by combining ModernTCN and enhance local feature capture capability and using Transformer's global dependency modelling capability. MTCAT effectively improves the model's performance in long time-series prediction tasks. Through the constructed full life cycle wear test rig of self-lubricated bearings, the AE signal features are extracted to get the full life cycle dataset for validation, and compared with other single deep learning models, the following conclusions can be obtained:

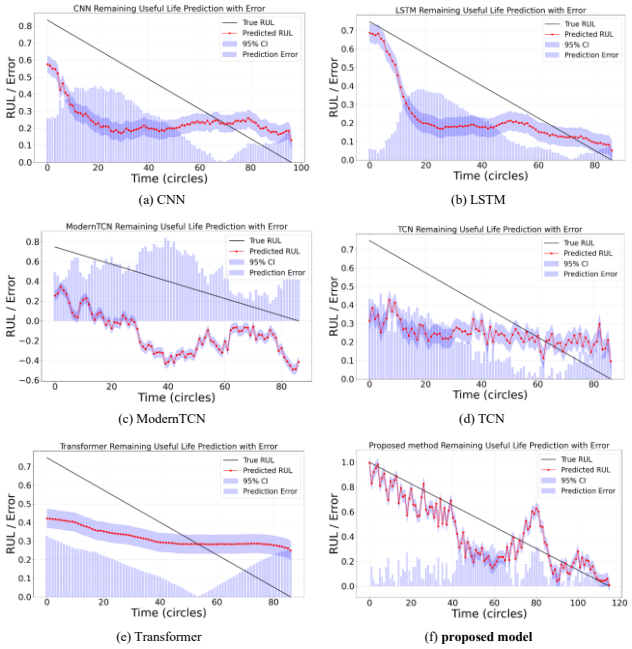


Fig. 11. Task A1 RUL prediction result

Based on multi-channel signals, multi-scale data based on experts' experience is constructed, and the features of acoustic emission signals and torque signals in the tiny scale changes of self-lubricated bearings are extracted by FFT and WT, and the

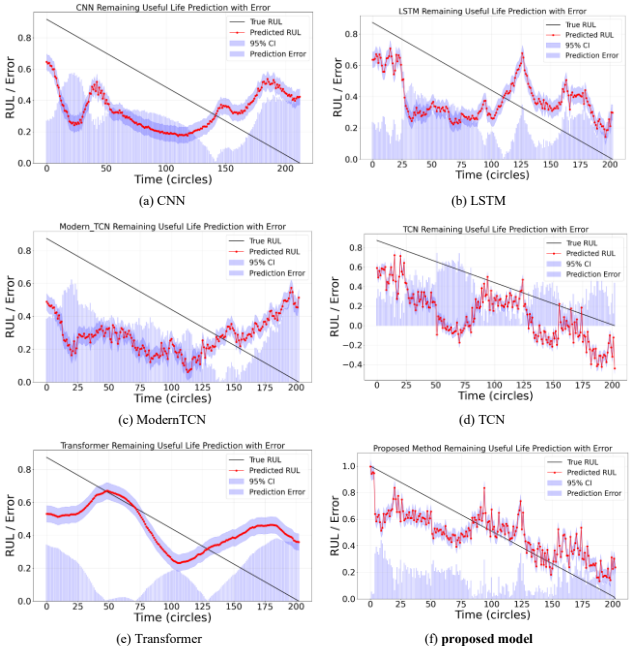


Fig. 12. Task A2 RUL prediction result

tiny vibration changes and longtime stress accumulation are analyzed, and compared with the commonly used statistical features, the dataset construction model proposed has a non-redundancy and better trend.

In the RUL prediction stage, MTCAT model is proposed, ModernTCN can handle dependencies on different time scales through its extended convolutional kernel sensing field, and the attention mechanism can dynamically adjust the multi-scale information. Transformer makes full use of the complex data features, and the global modelling capability can in turn helps to capture the long-time dependencies, the model does not rely on empirical formulas or physical models, and digs deeper into the depth features of the wear process of self-lubricating bearings, which is more accurate and has a smaller error margin through MTCAT in comparison with other single algorithmic models.

In the future research, the method of combining physical constraints will be considered to construct a model by introducing a combination of physical and data-driven models [31], utilizing the attention mechanism and capturing local features in the data, while ensuring that the model's outputs follow the physical a priori knowledge, to enhance the innovation and practical application value of the model, thus improving the accuracy of the remaining life prediction results.

REFERENCES

[1] *P. Revill, A. Clarke, R. Pullin, and G. Dennis, "Acoustic emission monitoring of wear in aerospace self-lubricating bearing liner materials," Wear, vol. 486–487, pp. 204102, 2021.*

[2] *M. F. Wani and O. P. Gandhi, "Maintainability design and evaluation of mechanical systems based on tribology," Reliab. Eng. Syst. Safety, vol. 77, no. 2, pp. 181–188, 2002.*

[3] *Y. Deng, D. Shichang, J. Shiyao, Z. Chen, and X. Zhiyuan, "Prognostic study of ball screws by ensemble data-driven particle filters," Journal of Manufacturing Systems, vol. 56, pp. 359–372, Jul. 2020, doi: 10.1016/j.jmsy.2020.06.009.*

[4] *Y. Deng, S. Du, D. Wang, Y. Shao, and D. Huang, "A Calibration-Based Hybrid Transfer Learning Framework for RUL Prediction of Rolling Bearing Across Different Machines," IEEE Trans. Instrum. Meas., vol. 72, pp. 1–15, 2023, doi: 10.1109/TIM.2023.3260283.*

[5] *S. Jia, Y. Deng, J. Lv, S. Du, and Z. Xie, "Joint distribution adaptation with diverse feature aggregation: A new transfer learning framework for bearing diagnosis across different machines," Measurement, vol. 187, p. 110332, Jan. 2022, doi: 10.1016/j.measurement.2021.110332.*

[6] *Y. Deng, J. Lv, D. Huang, and S. Du, "Combining the theoretical bound and deep adversarial network for machinery open-set diagnosis transfer," Neurocomputing, vol. 548, p. 126391, 2023.*

[7] *Y. Deng, D. Huang, S. Du, G. Li, C. Zhao, and J. Lv, "A double-layer attention based adversarial network for partial transfer learning in machinery fault diagnosis," Computers in Industry, vol. 127, p. 103399, May 2021, doi: 10.1016/j.compind.2021.103399.*

[8] *J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications," Mech. Syst. Signal Process., vol. 42, no. 1–2, pp. 314–334, 2014.*

[9] *J. Yu, "Bearing performance degradation assessment using locality preserving projections and Gaussian mixture models," Mech. Syst. Signal Process., vol. 25, no. 7, pp. 2573–2588, 2011.*

[10] *S. Schwendemann, Z. Amjad, and A. Sikora, "A survey of machine-learning techniques for condition monitoring and predictive maintenance of bearings in grinding machines," Comput. Ind., vol. 125, p. 103380, 2021.*

[11] *S. Sankararaman, "Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction," Mech. Syst. Signal Process., vol. 52–53, pp. 228–247, 2015.*

[12] *J. Wang, Y. Li, Z. Huang, and Q. Qiao, "Digital twin-driven fault diagnosis service of rotating machinery," in Digital Twin Driven Service, F. Tao, Q. Qi, and A. Y. C. Nee, Eds., pp. 119–138, Academic Press, 2022.*

[13] *X. Wan, D. Wang, P. W. Tse, G. Xu, and Q. Zhang, "A critical study of different dimensionality reduction methods for gear crack degradation assessment under different operating conditions," Measurement, vol. 78, pp. 138–150, 2016, doi: 10.1016/j.measurement.2015.09.032..*

[14] *P. Shakya, M. S. Kulkarni, and A. K. Darpe, "A novel methodology for online detection of bearing health status for naturally progressing defect," J. Sound Vib., vol. 333, no. 21, pp. 5614–5629, 2014.*

[15] *F. Ahmadzadeh and J. Lundberg, "Remaining useful life prediction of grinding mill liners using an artificial neural network," Miner. Eng., vol. 53, pp. 1–8, 2013.*

[16] *Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 6999–7019, Dec. 2022.*

[17] *Y. Mo, Q. Wu, X. Li, and B. Huang, "Remaining useful life estimation via transformer encoder enhanced by a gated convolutional unit," J. Intell. Manuf., vol. 32, no. 7, pp. 1997–2006, Oct. 2021.*

[18] *L. Tao, H. Liu, G. Ning, W. Cao, B. Huang, and C. Lu, "LLM-based framework for bearing fault diagnosis," arXiv preprint arXiv:2411.02718, 2024.*

[19] *Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Adv. Neural Inf. Process. Syst., vol. 30, pp. 5998–6008, 2017.*

[20] *C. Wang, T. Wang, M. Ma, and M. Wang, "Remaining useful life prediction of bearings using multi-scale window-based hybrid-domain transformer," IEEE Trans. Instrum. Meas., vol. 72, pp. 1–12, 2023.*

[21] *D. Hou, J. H. Chen, R. Cheng, X. Hu, and P. Shi, "A bearing remaining life prediction method under variable operating conditions based on cross-transformer fusioning segmented data cleaning," Reliab. Eng. Syst. Saf., vol. 245, Art. no. 110021, 2024.*

[22] *Y. Li, J. Li, H. Wang, C. Liu, and J. Tan, "Knowledge enhanced ensemble method for remaining useful life prediction under variable working conditions," Reliab. Eng. Syst. Safety, vol. 242, p. 109748, 2024.*

[23] *Y. Qin, F. Gan, B. Xia, D. Mi, and L. Zhang, "Remaining useful life estimation of bearing via temporal convolutional networks enhanced by a gated convolutional unit," Eng. Appl. Artif. Intell., vol. 133, Art. no. 108308, 2024.*

[24] *N. Sun, J. Tang, X. Ye, C. Zhang, S. Zhu, S. Wang, and Y. Sun, "Remaining useful life prognostics of bearings based on convolution attention networks and enhanced transformer," Heliyon, vol. 10, no. 19, p. e38901, 2024.*

[25] *Y. Liu, H. Wang, and J. Chen, "Spatio-temporal transformer for bearing remaining useful life prediction under noisy conditions," IEEE Sensors J., vol. 24, no. 10, pp. 15678–15689, May 2024.*

[26] *X. Chen, Y. Zhang, and H. Zhang, "Domain-adaptive deep learning for cross-machine bearing RUL prediction," IEEE Sensors J., vol. 24, no. 15, pp. 23456–23467, Aug. 2024.*

[27] *J. Wu, Z. Zhao, and F. Li, "Multi-scale feature fusion with gated recurrent units for bearing prognostics," IEEE Sensors J., vol. 23, no. 21, pp. 24567–24578, Nov. 2023.*

[28] *D. Luo and X. Wang, "ModernTCN: A modern pure convolution structure for general time series analysis," Proc. 12th Int. Conf. Learn. Represent., OpenReview, 2024.*

[29] *B. Wang, Y. G. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," IEEE Trans. Reliab., vol. 69, pp. 401–412, 2020.*

[30] *Z. Wang, L. Zhang, and X. Li, "Attention-enhanced multi-scale convolutional neural network for bearing RUL prediction," IEEE Sensors J., vol. 23, no. 22, pp. 26789–26800, Nov. 2023.*

[31] *S. Wang, L. Yan, S. Du, S. Li, and X. Chen, "Bearing prognostics and health management based on hybrid physical mechanism and data models: A systematic review," Meas. Sci. Technol., vol. 36, no. 5, p. 052002, May 2025.*

[32] *C. Zhao, E. Zio, and W. Shen, "Uncertainty-weighted domain generalization for remaining useful life prediction of rolling bearings under unseen conditions," IEEE Sensors J., vol. 24, no. 8, pp. 9341–9352, Apr. 2024, doi: 10.1109/JSEN.2024.3364327.*

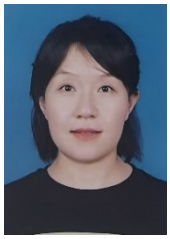
[33] *L. Chen, M. Li, X. Wang, and J. Zhang, "A customized dual-Transformer framework for remaining useful life prediction of mechanical systems," Mech. Syst. Signal Process., vol. 205, p. 110563, May 2025, doi: 10.1016/j.ymssp.2025.110563.*

Yuheng Wang received the B.S. degree in vehicle engineering from the Southwest Jiaotong University, Chengdu, China, in 2023. He is currently pursuing the M.S. degree in mechanical engineering with Shanghai Jiao Tong University, Shanghai, China. His research interests include hybrid deep learning model for prognostics and health management, and time-series large language model (LLM) for prognostics and health management

Shuo Wang received the B.S. degree in industrial engineering from Xinjiang University, Urumqi, China, in 2024. He is currently pursuing the Ph.D. degree in mechanical engineering with Shanghai Jiao Tong University, Shanghai, China. His research interests include hybrid physical model and data-driven model for prognostics and health management, condition monitoring, and intelligent fault diagnosis.

Xiaoxiao Shen (Member, IEEE) received the B.S. degree from the College of Information Management, Nanjing Agricultural University, Nanjing, China in 2019. She is currently pursuing the Ph.D. degree in Department of Industrial Engineering and Management, Shanghai Jiao Tong University, and dual Ph.D. degree in Department of Mechanical Engineering, Politecnico di Milano. Her current research interests include analysis, modeling, and decision-making optimization for operations management problems in manufacturing systems and service systems. Her technical specialties include mathematical modeling, stochastic optimization, exact optimization algorithm, intelligent optimization algorithm, statistical modeling analysis, and simulation optimization. She is a reviewer for international high-level journals such as IEEE Transactions on Automation Science and Engineering, Computers & Operations Research, and Computers & Industrial Engineering

Shichang Du (Member, IEEE) received the B.S. and M.S.E. degrees in mechanical engineering from the Hefei University of Technology, Hefei, China, in 2000 and 2003, respectively, and the Ph.D. degree in industrial engineering and management from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2008. He was with the University of Michigan, Ann Arbor, MI, USA, from 2006 to 2007, as a Visiting Scholar. He is currently a Professor with the Department of Industrial Engineering and Management, Shanghai Jiao Tong University. He has authored/coauthored 100 articles in journals, including IEEE TIM, IEEE TASE, ASME Transaction and IISE Transaction, etc. His current research interests focus on quality control, PHM and smart manufacturing methods. His research has been supported by the National Natural Science Foundation of China, the Ministry of Science and Technology of China, and the Shanghai Association for Science and Technology. He receives various awards for his research and teaching, including IEEE T-ASE Googol Best New Application Paper Award, QR2MSE Best Paper Award, and Most Popular Professor Award of ME School in SJTU. He serves for several journals an Area Editor including Computers and Industrial Engineering, Flexible Services and Manufacturing Journal, Precision Engineering, Measurement, Journal of Intelligent Manufacturing, and Computers in Industry.



Shanshan Li received the B.S. degree in Engineering Mechanics from Shandong University, Jinan, China, in 2008, and the M.S. degree in Engineering Mechanics from China Institute of Aeronautics and Astronautics, Beijing, China, in 2011. She is currently working in the National Key Laboratory of Strength and Structural Integrity & Aircraft Strength Research Institute of China (ASRI). Her research interests include fatigue reliability analysis and verification techniques for aircraft structures.

Xianmin Chen received his Ph.D. degree in engineering from Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently working in the National Key Laboratory of Strength and Structural Integrity & Aircraft Strength Research Institute of China (ASRI) as a professor. His research interests include aircraft design, fatigue reliability analysis and verification techniques for aircraft structures.