Available Online: 16 June 2026 DOI: 10.54254/2977-3903/2026.34398

Research on remaining useful life prediction of rolling bearings based on adaptive variational mode decomposition and dual-branch temporal neural network

Wuchu Tang, Wenxin Dong*, Jiawei Yang, Guofu Wu

Dalian Jiaotong University, Dalian, China

*Corresponding Author. Email: 2415019618@qq.com

Abstract. To address the nonlinear and non-stationary characteristics of vibration signals during rolling bearing operation and the issue of insufficient degradation information representation, this paper designs a prediction framework that combines Adaptive Variational Mode Decomposition (AVMD) with a SETCN-BiGRU multi-head temporal attention mechanism for Remaining Useful Life (RUL) prediction. First, AVMD is used to decompose the bearing horizontal vibration signal into five Intrinsic Mode Functions (IMFs). Time-domain and frequency-domain statistics are extracted from each IMF and concatenated into a 115-dimensional degradation feature sequence. Subsequently, the model processes in parallel: a TCN-SENet branch extracts local temporal features and adaptively adjusts channel weights, while a BiGRU with multi-head temporal attention sub-network captures global bidirectional dependencies and critical degradation periods within the degradation sequence. Finally, the two types of features are fused, and the RUL prediction result is output. Experimental results demonstrate that the proposed model achieves an RMSE, MAE, and R^2 of 0.0582, 0.0477, and 0.9483 respectively on the IEEE PHM 2012 dataset, and average values of 0.0780, 0.0559, and 0.9133 on a self-built laboratory bearing dataset, indicating good prediction accuracy, robustness, and generalization ability.

Keywords: remaining useful life prediction, adaptive variational mode decomposition, temporal convolutional network, bidirectional gated recurrent unit, multi-head temporal attention mechanism

1. Introduction

Rolling bearings are indispensable components in various rotating machinery and are widely used in industrial manufacturing, rail transportation, aerospace equipment, and other fields. The health status of bearings directly determines the stable and safe operation of mechanical systems [1, 2]. Once a bearing fails, it not only causes equipment downtime and direct economic losses but may also induce safety accidents. Therefore, accurately predicting the Remaining Useful Life (RUL) of rolling bearings holds significant engineering importance for implementing predictive maintenance, reducing operational risks, and optimizing maintenance resources [3, 4].

Existing RUL prediction methods are mainly divided into model-based methods and data-driven methods. Due to the complexity and uncertainty of actual mechanical degradation mechanisms, accurately establishing a physical model is difficult. In contrast, data-driven methods learn degradation patterns directly from monitoring data and have become the mainstream direction [5]. Medjaher et al. [6] used a data-driven method to predict the RUL of key bearing components; Ben Ali et al. [7] combined the Weibull distribution with an artificial neural network for life prediction; Khelif et al. [8] designed a direct RUL estimation strategy based on support vector regression; Guo et al. [9] used a recurrent neural network to construct a health indicator for bearing RUL prediction; Li et al. [10] employed a deep convolutional neural network for degradation feature learning; Zheng et al. [11] utilized LSTM to extract implicit degradation patterns from multi-sensor sequences; Zhu et al. [12] proposed a DACN-ConvLSTM model to mine temporal information between adjacent signal samples; Cao et al. [13] combined multi-domain mixed features with TCN for rolling bearing RUL prediction. Existing research results indicate that combining multi-source feature construction with deep temporal modeling helps improve life prediction accuracy.

Although existing methods have achieved preliminary results, there are still shortcomings in degradation feature expression and temporal information utilization, especially in jointly modeling multi-scale degradation features, channel correlations, and key time steps, where there is significant room for improvement. To overcome these limitations, this paper constructs a parallel fusion model integrating TCN-SENet and BiGRU with multi-head temporal attention. The model first constructs a multi-domain degradation feature sequence through AVMD, then extracts local temporal features, channel attention features, and global temporal dependencies respectively, ultimately improving the accuracy and stability of RUL prediction. It is worth noting that feature robustness under complex operating conditions and model generalization ability remain core issues urgently needing resolution in current life prediction research, a viewpoint widely acknowledged in deep learning for machine health monitoring, cross-condition RUL transfer prediction, and review literature.

2. Remaining useful life prediction method

2.1. AVMD feature extraction

Rolling bearing vibration signals contain strong nonlinear and non-stationary components. The original waveform is often mixed with structural vibrations, environmental noise, and degradation-induced impacts. If the original signal is directly used for RUL prediction, the effect is compromised due to insufficient degradation information. Variational Mode Decomposition (VMD) can decompose non-stationary signals into several band-limited modes [14]. Adaptive decomposition ideas like EMD also provide an important foundation for non-stationary signal analysis [15]. To enhance feature expression capability, this paper adopts Adaptive VMD (AVMD) based on VMD to decompose the bearing vibration signal, then extracts time-domain and frequency-domain statistics from each mode component to form a degradation feature sequence.

Let the collected raw vibration signal be $x(t)$, and its VMD decomposition result can be expressed as in Equation (1):

$$x(t) = \sum_{k=1}^{K} u_k(t) + r(t) \tag{1}$$

where $u_k(t)$ represents the k-th intrinsic mode component, K is the number of modes, and $r(t)$ is the residual component. The mode number is fixed at $K = 5$, meaning each sampled signal is decomposed into 5 IMF components for subsequent degradation feature extraction. An adaptive parameter selection strategy is used to determine the VMD penalty parameter α. The set of candidate penalty parameters is shown in Equation (2):

$$\Omega = \{1000, 2000, 3000, 4000, 5000, 6000\} \tag{2}$$

For each bearing, the first m samples in chronological order are selected as parameter optimization samples, where $m = min(3, N_b)$, and N_b is the total number of samples for the b-th bearing. For each candidate parameter $\alpha \in \Omega$, VMD decomposition is performed with a fixed mode number $K = 5$, and the decomposition quality evaluation function $J(\alpha)$ is calculated. The optimal penalty parameter is defined as in Equation (3):

$$\alpha^* = \arg\min_{\alpha \in \Omega} J(\alpha) \tag{3}$$

where Ω is the set of candidate parameters, α^* is the optimal penalty parameter, and $J(\alpha)$ is the decomposition quality evaluation function. To balance the reconstruction capability after decomposition and the independence between modes, the evaluation function is defined as in Equation (4):

$$J(\alpha) = E_{rec}(\alpha) + \lambda E_{red}(\alpha) \tag{4}$$

where $E_{rec}(\alpha)$ represents the reconstruction error, $E_{red}(\alpha)$ represents the degree of redundancy between modes, and λ is a trade-off coefficient. This paper takes $\lambda = 0.05$ to balance the reconstruction error and the inter-mode correlation redundancy term. The reconstruction error is defined as in Equation (5):

$$E_{rec}(\alpha) = \frac{1}{m} \sum_{s=1}^{M} \frac{\| x_s(t) - \sum_{k=1}^{K} u_{k,s}(t) \|_2}{\| x_s(t) \|_2} \tag{5}$$

Where m is the total number of samples involved in parameter optimization, $x_s(t)$ represents the s-th optimization sample, and $u_{k,s}(t)$ is the k-th IMF component obtained by decomposing this sample. A smaller E_{rec} indicates a stronger ability of the decomposed modes to reconstruct the original signal. The inter-mode redundancy is defined as in Equation (6):

$$E_{red}(\alpha) = \frac{1}{m} \sum_{s=1}^{m} [\frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} |\rho(u_{i,s}, u_{j,s})|] \tag{6}$$

where $\rho(\cdot)$ represents the Pearson correlation coefficient. A smaller E_{red} indicates weaker correlation between different IMF components and lower modal redundancy.

After obtaining the optimal penalty parameter α^* , it is used for VMD decomposition of all sampled signals of that bearing. Each sampled signal is decomposed into 5 IMF components. Subsequently, time-domain and frequency-domain statistical features are extracted from each IMF component. Time-domain features mainly describe the vibration amplitude distribution and waveform characteristics, while frequency-domain features mainly reflect the spectral energy distribution and center frequency variation patterns. The feature vector for a single IMF can be expressed as in Equation (7):

$$f_k = [T_{k1}, T_{k2}, \ldots, T_{k11}, F_{k1}, F_{k2}, \ldots, F_{k11}, F_{k12}] \tag{7}$$

where T_{ki} represents the i-th time-domain feature of the k-th IMF, and F_{kj} represents the j-th frequency-domain feature of the k-th IMF. Furthermore, the total feature vector corresponding to the t-th sampling instant can be written as Equation (8):

$$f_t = [f_{1,t}, f_{2,t}, \ldots, f_{5,t}] \tag{8}$$

Therefore, the total feature dimension is as in Equation (9):

$$D = K \times (11 + 12) = 5 \times 23 = 115 \tag{9}$$

That is, each sampled signal is ultimately represented as a 115-dimensional degradation feature vector. By repeating the above process for all sampling points of the same bearing, the feature sequence

$F = \{f_1, f_2, \ldots, f_N\}$ for the bearing throughout its life cycle can be constructed, where $f_t \in R^{115}$ is the number of sampling time steps.

2.2. Model architecture

2.2.1. SENet module

The Squeeze-and-Excitation Network (SENet) is a typical channel attention mechanism [16]. Its core idea is to automatically learn the dependencies between channels, then weight the channel features to enhance important degradation features and suppress irrelevant information. This paper embeds SENet into the TCN branch to perform channel recalibration on the local temporal features output by the TCN. Its structure is shown in Figure 1.

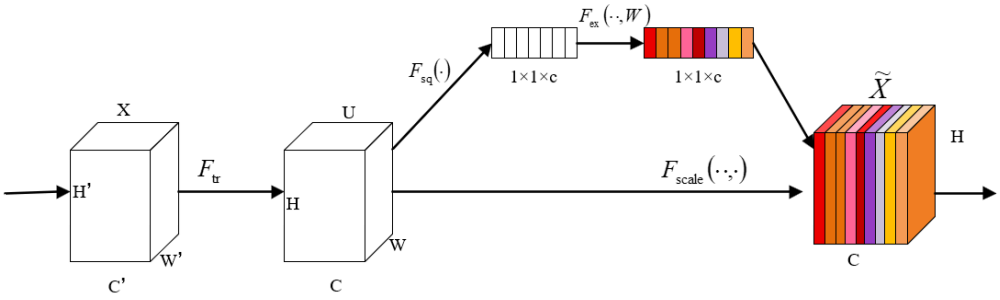

Figure 1. Schematic diagram of the SENet architecture

Let the temporal feature representation output by the TCN be as shown in Equation (10):

$$U = [u_1, u_2, \cdots, u_C] \in R^{T \times C} \tag{10}$$

where T denotes the time-step length, C denotes the number of channels, and u_c denotes the feature response of the c-th channel.

(1) Squeeze

Global average pooling is performed along the temporal dimension to compress the entire temporal response of each channel into a scalar, yielding a channel descriptor vector z, as shown in Equation (11):

$$z_c = F_{sq}(u_c) = \frac{1}{T} \sum_{t=1}^{T} u_c(t) \tag{11}$$

where z_c denotes the global response value of the c-th channel. This process compresses the temporal response of length T into scalar statistics, thereby aggregating global temporal information.

(2) Excitation

The compressed channel descriptor vector is then fed into two fully connected layers, combined with ReLU and Sigmoid activation functions, to generate channel weights, as shown in Equation (12):

$$s = F_{ex}(z, W) = \sigma(W_2 \delta(W_1 z)) \tag{12}$$

where denotes the ReLU activation function, denotes the Sigmoid activation function, and and denote the weight matrices of the two fully connected layers. Through this process, the model can automatically learn the importance of different channels for the RUL prediction task.

(3) Recalibration

Finally, the computed channel weights are multiplied back onto the original feature map to highlight important channels and suppress irrelevant channels, as expressed in Equation (13):

$$\widetilde{u}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c \tag{13}$$

where s_c denotes the weight coefficient corresponding to the c-th channel, and $\widetilde{u}_c$ denotes the recalibrated output feature. Through this process, SENet can highlight the key channel features related to the evolution of bearing degradation and improve the subsequent BiGRU's ability to model effective temporal information.

Through the above process, SENet can emphasize sensitive channels related to bearing degradation states and weaken the influence of ineffective or redundant features, thereby obtaining more discriminative local channel features F_{space} and providing effective input for subsequent feature fusion and RUL prediction.

2.2.2. TCN network model

Temporal Convolutional Network (TCN) is a deep architecture designed specifically for time-series modeling [17]. By integrating the temporal extension capability of one-dimensional convolution with the gradient-optimization mechanism of Residual Networks (ResNet) [18], it effectively addresses the difficulty of capturing long-term dependencies in traditional recurrent models. Compared with conventional CNNs, TCN has the following characteristics:

The convolution operation in TCN adopts a causal constraint, so the output at the current time step depends only on the current and previous time steps, thereby avoiding future information leakage and ensuring the validity of time-series prediction. Meanwhile, TCN expands the receptive field by introducing dilated convolution, enabling the model to capture dependencies over a longer time span without a significant increase in computational cost; with the residual connection structure, the network can be deepened further and the difficulty of training deep networks can be alleviated to a certain extent (see Figure 2).

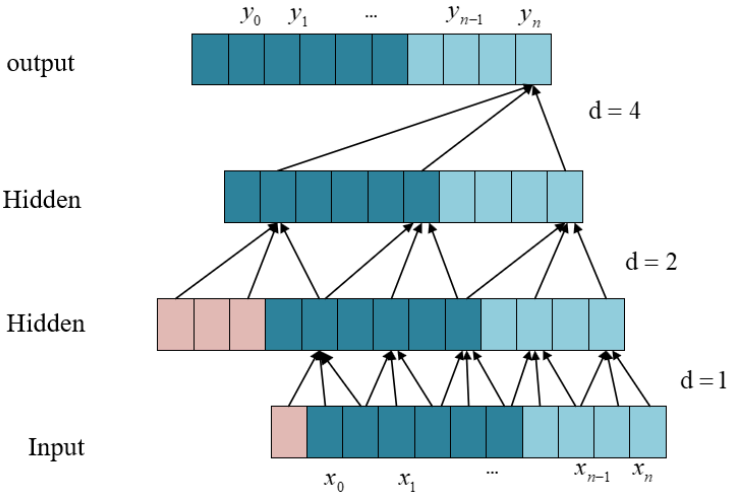

Figure 2. TCN architecture

In the TCN, the mapping between input and output is given in Equation (14):

$$Y = F_T(X) \tag{14}$$

where $F_T(\cdot)$ denotes the TCN model; $X = \{x_0, x_1, x_2, \ldots, x_{t-1}, x_t\}$ denotes the input mixed-domain feature sequence at time t; and $Y = \{y_0, y_1, y_2, \ldots, y_{t-1}, y_t\}$ denotes the corresponding output sequence.

In the TCN model, causal dilated convolution can expand the receptive field of the network and enable the model to better capture dependencies in long time series. However, as the number of layers increases, model performance may deteriorate. On the one hand, a deeper architecture increases the complexity of parameter updates and backpropagation, making convergence more difficult during training; on the other hand, after the network becomes deeper, it is harder for the model to learn an identity mapping directly, which in turn affects optimization performance. To alleviate these issues, residual connections are usually introduced in TCN.

Compared with directly stacking multiple convolutional layers to fit the mapping between input and output, residual connections transform the learning objective into a residual function, making the model easier to optimize and helping improve training convergence speed and feature transfer efficiency, as defined in Equation (15):

$$x_{n+1} = \phi(x_n + F(x_n, \zeta_n)) \tag{15}$$

where x_n denotes the input sequence at the n-th level; x_{n+1} denotes the output of the residual block; $F(x_n, \zeta_n)$ denotes the residual mapping, which usually consists of 2 to 3 causal dilated convolution operations; and $\phi(\cdot)$ denotes the activation function.

2.2.3. BiGRU network model

Gated Recurrent Unit (GRU) is an improved structure of recurrent neural networks [19], mainly composed of an update gate and a reset gate. Compared with LSTM, GRU has a simpler structure and fewer parameters, while also mitigating the vanishing-gradient problem in long-sequence modeling to a certain extent [20]. Its basic computation process can be expressed by Equation (16) to (19):

$$z_t = \sigma(W_z \cdot |h_{t-1}, x_t| + b_z) \tag{16}$$

$$r_t = \sigma(W_r \cdot |h_{t-1}, x_t| + b_r) \tag{17}$$

$$\widetilde{h}_t = tanh(W \cdot |r_t h_{t-1}, x_t| + b_h) \tag{18}$$

$$h_t = (1 - z_t) \otimes h_t - 1 + z_t \widetilde{h}_t \tag{19}$$

where x denotes the input data; W_Z , W_r and W denote the corresponding weights; b_z , b_r and b_h denote the corresponding bias terms; z_t denotes the update gate; r_t denotes the reset gate; $\widetilde{h}_t$ denotes the candidate activation state; h_t denotes the activation state; σ denotes the sigmoid activation function; $[\cdot]$ denotes the concatenation of the two vectors; x_t denotes the input to the GRU at time t; $tanh(\cdot)$ denotes the hyperbolic tangent activation function; and $\otimes$ denotes element-wise matrix multiplication.

A unidirectional GRU can only propagate forward in time, making it difficult to fully exploit the forward and backward contextual information in the degradation sequence. Therefore, this paper uses a Bidirectional Gated Recurrent Unit (BiGRU) to model the bearing degradation sequence, as shown in Figure 3. BiGRU consists of a forward GRU and a backward GRU, which extract temporal dependencies in the forward and reverse directions, respectively, and concatenate the hidden states from both directions to obtain the comprehensive temporal representation at the current time step, as shown in Equation (20) to (22):

$$\vec{h}_t = F_{GRU}(x_t, \vec{h}_{t-1}) \tag{20}$$

$$\overleftarrow{h}_t = F_{GRU}(x_t, \overleftarrow{h}_{t-1}) \tag{21}$$

$$h_t = W_t \vec{h}_t + v_t \overleftarrow{h}_t + b_t \tag{22}$$

where $\vec{h}_t$, $\overleftarrow{h}_t$ and h_t denote the hidden states of the forward propagation, backward propagation, and final output, respectively; F_{GRU} denotes the function used to map the input vector to the GRU hidden state; b_t denotes the bias vector; W_t and V_t denote the corresponding weight matrices.

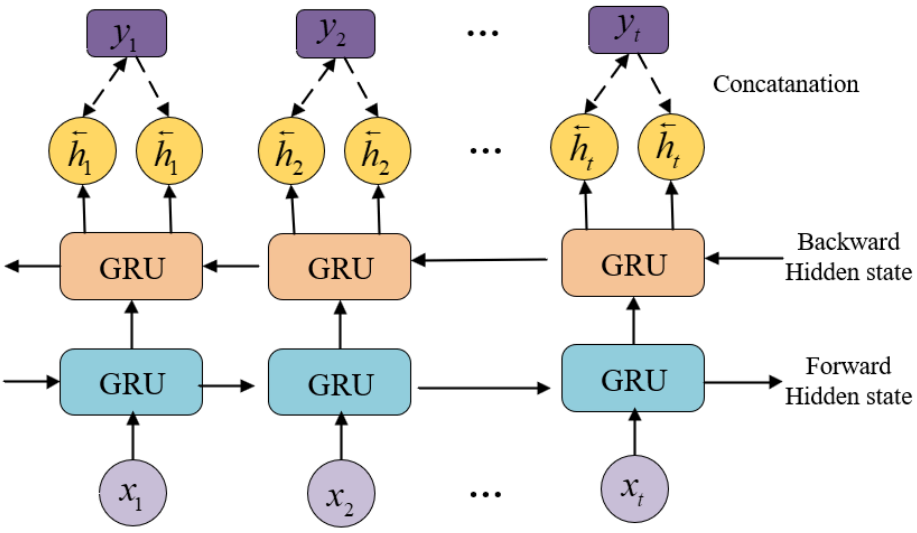

Figure 3. BiGRU architecture

2.2.4. Multi-head temporal attention mechanism

The multi-head temporal attention mechanism [21] is used to assign weights to key time steps in the temporal features output by BiGRU. Let the input sequence after feature mapping in BiGRU be denoted as $X \in R^{T\times d}$, where T is the time-step length and d is the feature dimension. First, query matrix Q, key matrix K, and value matrix V are obtained through linear mapping, as shown in Equation (23):

$$Q = XW_Q, K = XW_K, V = XW_V \tag{23}$$

Then, scaled dot-product attention is used to calculate the correlation between different time steps, as shown in Equation (24):

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{24}$$

To enhance the model's ability to represent different degradation stages and temporal dependencies, multi-head attention is adopted to model the temporal features in parallel, as shown in Equation (25) and (26):

$$head_i = Attention(Q_i, K_i, V_i) \tag{25}$$

$$H = Contact(head_1, head_2, \ldots, head_h)W_O \tag{26}$$

where W_Q , W_K , W_V and W_O are learnable parameter matrices, d_k denotes the key-vector dimension, and h denotes the number of attention heads. Through the multi-head temporal attention mechanism, the model can capture key time-step information in the degradation sequence from multiple subspaces and highlight the degradation segments that contribute more to RUL prediction. Finally, residual connection and normalization are applied to the attention output and the original mapped features to obtain the temporal-branch feature representation F_{time} .

3. Model description

As shown in Figure 4, according to the nonlinear, non-stationary, and long-sequence-dependence characteristics of rolling bearing degradation signals, this paper designs a parallel fusion prediction model based on AVMD feature extraction, TCN-SENet, and BiGRU multi-head temporal attention. Traditional single-path prediction networks usually focus on only one aspect of local features or temporal features, making it difficult to simultaneously account for local impact changes, channel correlations, and global temporal dependencies in the bearing degradation process. Therefore, this paper fuses TCN-SENet and BiGRU with multi-head temporal attention: TCN extracts local temporal degradation features through causal and dilated convolution while preserving temporal causality and expanding the receptive field; SENet recalibrates the features output by TCN through a channel-attention mechanism to enhance degradation-sensitive channels and suppress redundant features; BiGRU models the degradation sequence in both forward and backward directions to capture the global temporal evolution during bearing operation; and the multi-head temporal attention mechanism further computes correlations among different time steps to highlight key degradation segments that contribute more to RUL prediction. With this structure, the model can simultaneously extract local temporal features, channel-sensitive features, and global temporal dependencies, thereby improving the accuracy and stability of rolling bearing RUL prediction [22-25].

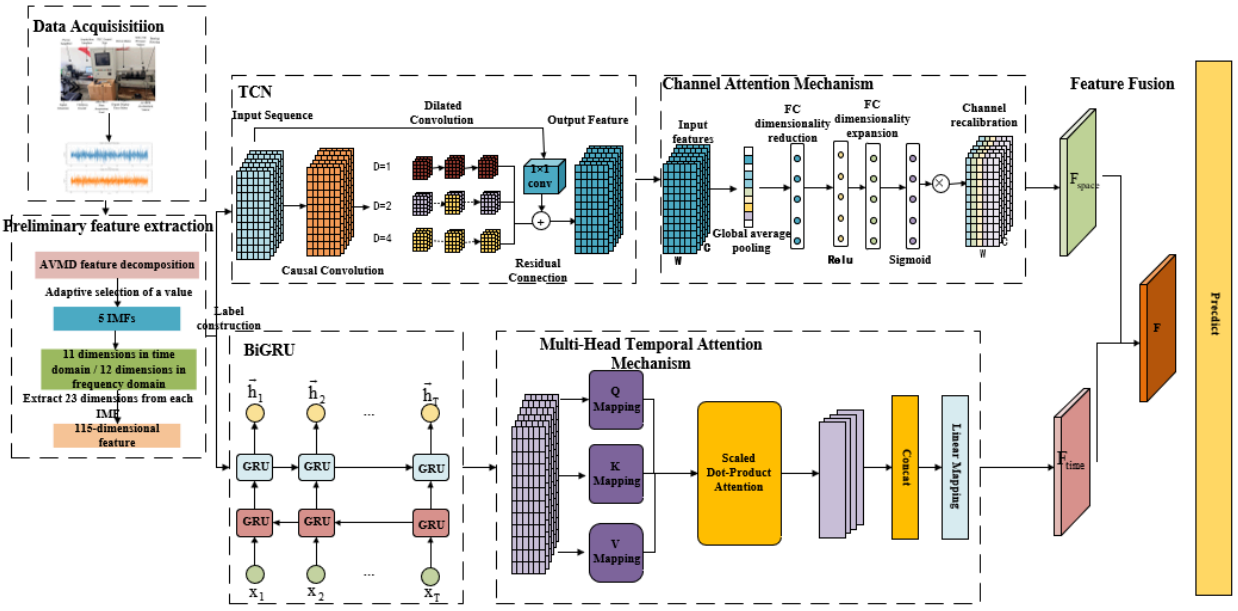

Figure 4. Framework of the SETCN-BiGRU multi-head temporal attention model

First, in the data acquisition and feature construction stage, the bearing vibration signal is initially feature-extracted and then adaptively decomposed by AVMD into five IMF components. Subsequently, 11 time-domain features and 12 frequency-domain features are extracted from each IMF component, i.e., 23 features per IMF, ultimately forming a 115-dimensional degradation feature sequence as the model input. In the feature modeling stage, the model uses parallel branches for feature extraction. The upper branch is the TCN-SENet branch: the input features first pass through the TCN module to extract local temporal degradation features and then through the SENet channel-attention module to obtain spatial feature representations F_{space} ; the lower branch is the BiGRU-multi-head temporal attention branch: after being modeled by BiGRU, the input features are further mapped to Q, K, and V and processed by scaled dot-product attention to compute time-step

correlations, thereby obtaining temporal feature representations F_{time} . Finally, the feature outputs obtained from the two branches are fused to form the comprehensive degradation representation F, which is then fed into the prediction layer to output the rolling bearing RUL prediction result. Through multi-branch feature complementarity and attention-weighting mechanisms, the model achieves a comprehensive representation of the bearing degradation state, which helps improve the generalization ability and prediction accuracy of RUL prediction under complex operating conditions.

4. Experiments

To systematically verify the effectiveness and generalization capability of the proposed SETCN-BiGRU multi-head temporal attention model for rolling bearing remaining useful life prediction, experiments are conducted on the IEEE PHM 2012 bearing dataset and a self-built laboratory bearing dataset under Condition 1 and Condition 2, respectively. The IEEE PHM 2012 rolling bearing dataset includes horizontal vibration signals, vertical vibration signals, and temperature signals, while the self-built laboratory dataset contains horizontal vibration signals and vertical vibration signals. To extract complete data features and better train the model to learn the fault-state characteristics of the bearing, the horizontal vibration signals are preprocessed using AVMD. The data are decomposed into mode components with central frequencies and finite bandwidths, after which 11 time-domain statistical features and 12 frequency-domain statistical features are computed from the decomposed components. Time-domain statistical features include the mean, standard deviation, root mean square, skewness, peak-to-peak value, waveform factor, and so on; frequency-domain features include the mean of spectral amplitude, variance of spectral amplitude, skewness of spectral amplitude, spectral frequency centroid, coefficient of spectral variation, and so on. During training, for bearings under the same operating condition, one bearing is used as the test set and the remaining bearings are used as the training set. The optimizer and hyperparameter settings are listed in Table 1. The model uses the AdamW optimizer for parameter updates and introduces weight decay to suppress overfitting [26, 27].

Table 1. Optimizer and hyperparameter settings of the SETCN-BiGRU multi-head temporal attention model

Iteration count	170
Batch size	128
Learning rate	0.00007
Patience	22
Sliding window length	128
Optimizer	AdamW

The evaluation metrics for the prediction results are Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). The formulas for RMSE, MAE, and R^2 are given in Equation (27) to (29), where R^2 measures how well the model fits the true life trend.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{27}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}| \tag{28}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{29}$$

where n is the number of data samples; y_i is the actual life value; $\hat{y}$ is the predicted life value; and $\overline{y}$ is the mean of all actual life values.

4.1. Experiment 1

4.1.1. Data description

To verify the effectiveness of the proposed model, the IEEE PHM 2012 rolling bearing dataset is selected for experimental analysis. The operating conditions and the number of bearings of this dataset are shown in Table 2. The data were collected from the PRONOSTIA test rig, which can be used for accelerated bearing degradation experiments and validation of life prediction methods, as shown in Figure 5. The test platform is equipped with two acceleration sensors to collect vibration signals in the horizontal and vertical directions of the bearing. The vibration signal sampling frequency is 25.6 kHz, the temperature signal sampling frequency is 10 Hz, and data are acquired for 0.1 s every 10 s. To ensure consistency of the experimental conditions, all model training and testing are conducted under the same operating condition. For example, if the B1-1 bearing data under Condition 1 are selected as the test set, then the data from B1-2, B1-3, B1-4, B1-5, B1-6, and B1-7 under the same condition are selected as the training set.

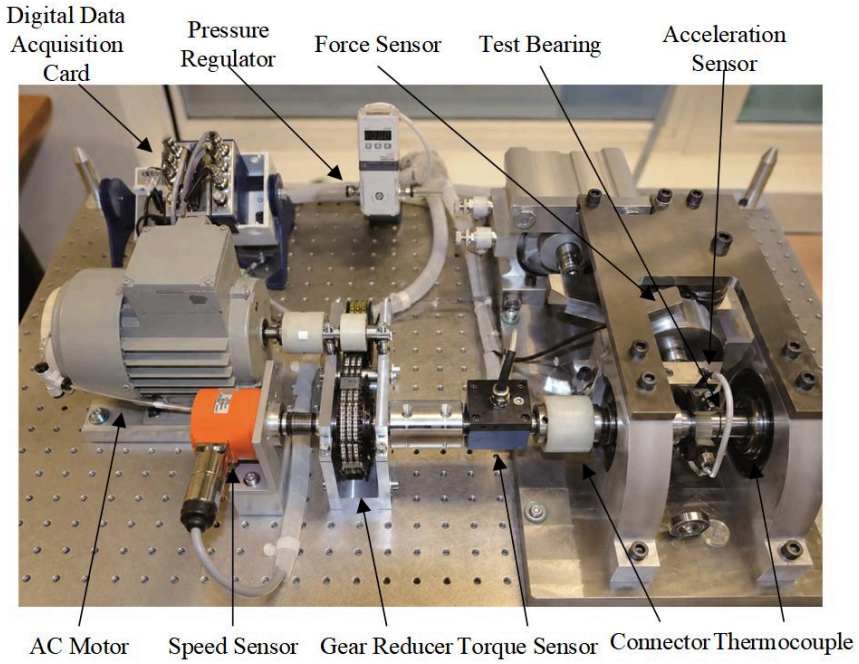

Figure 5. Schematic diagram of the PRONOSTIA test rig

Table 2. Information of the IEEE PHM 2012 dataset

	Condition 1	Condition 2
Speed-load	1,800 r/minand 4,000 N	1,650 r/minand 4,200 N
Bearing	B1_1	B2_1
	B1_2	B2_2
	B1_3	B2_3
	B1_4	B2_4
	B1_5	B2_5
	B1_6	B2_6
	B1_7	B2_7

4.1.2. Experimental results and analysis

Table 3 presents a comparison of the prediction performance between the SETCN-BiGRU model and other models on the IEEE PHM 2012 dataset, from which it can be clearly seen that the proposed model achieves the best results. The calculated overall mean RMSE of the SETCN-BiGRU model is 0.0582, while that of the TCN-BiLSTM model is 0.0782. As shown in Figure 6 and 7, the predicted curves of bearings B1-1 and B2-6 closely fit the actual value curves. Especially at the end of life, the model is still able to maintain stable predictions, indicating that the SETCN-BiGRU combination effectively mitigates the error accumulation problem caused by abrupt feature changes in the later stage of degradation that plague traditional methods.

Table 3. Performance evaluation of different models on the IEEE PHM 2012 dataset

Metric	CNN-BiGRU	GRU	TCN-Transformer	TCN-BiLSTM	Proposed model
RMSE	0.0794	0.0745	0.0845	0.0782	0.0582
MAE	0.0657	0.0596	0.0687	0.0625	0.0477
R^2	0.8998	0.9133	0.8905	0.9005	0.9483

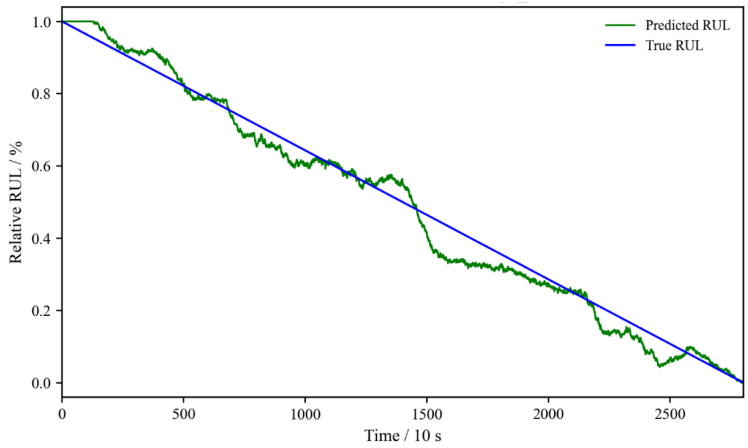

Figure 6. Prediction result of B1-1

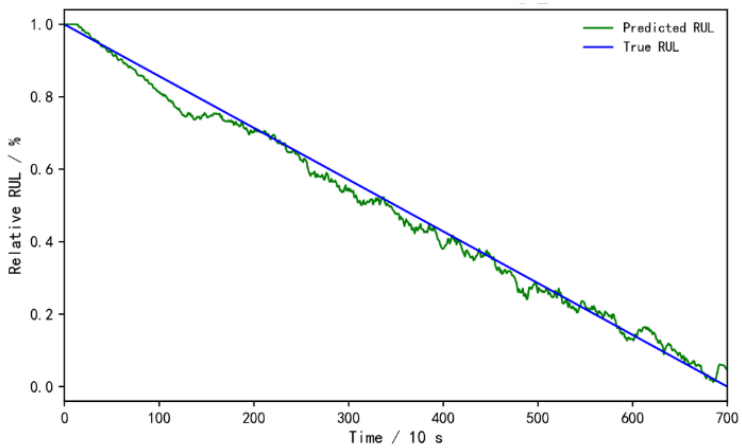

Figure 7. Prediction result of B2-6

By comparing and visualizing the RUL prediction results of different models for B1-1, B1-3, and B1-6, the prediction accuracy and robustness of each model can be understood more clearly. Figure 8 to 10 show the prediction results of these three bearings. The orange line represents the proposed model, while the green and red lines represent the prediction results of the CNN-BiGRU model and the TCN-BiLSTM model, respectively. As can be seen from Figure 8 to 10, the fitting result of the orange line is the best, with lower overall error and higher prediction accuracy. Across all bearing datasets, the proposed model can accurately capture the variation trend of RUL. In particular, in Figure 9, the prediction performance at the terminal stage, that is, close to the failure time, is significantly better than that of the other two conventional models. The prediction curves of the CNN-BiGRU and TCN-BiLSTM models show obvious deviations at the end, resulting in larger errors and even failing to reflect the RUL variation trend near failure effectively. In contrast, the orange line can track RUL variations more accurately near the failure time and predict equipment failure in advance. This phenomenon reflects the superiority of the proposed model in complex environments, especially when different operating conditions and different bearing states are considered. The proposed method can more precisely capture the subtle changes in equipment health status and thus maintain high prediction accuracy in complex scenarios.

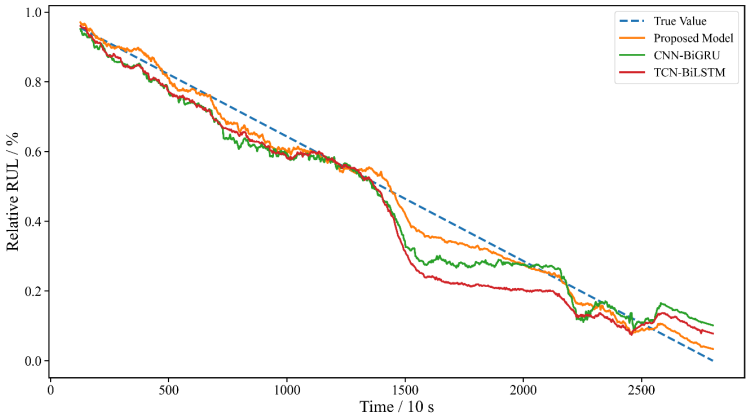

Figure 8. Comparison of all models on B1-1

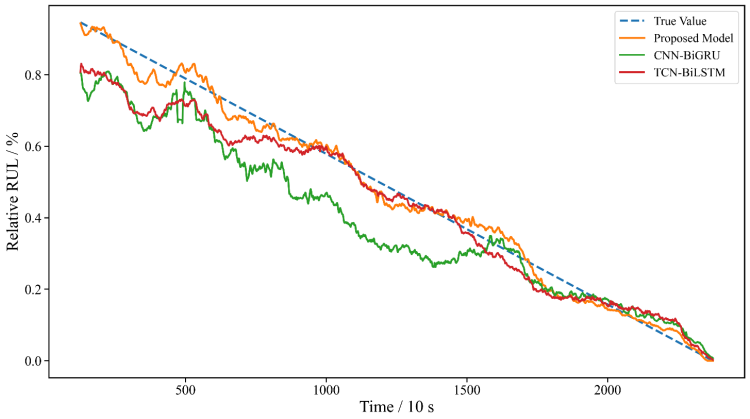

Figure 9. Comparison of all models on B1-3

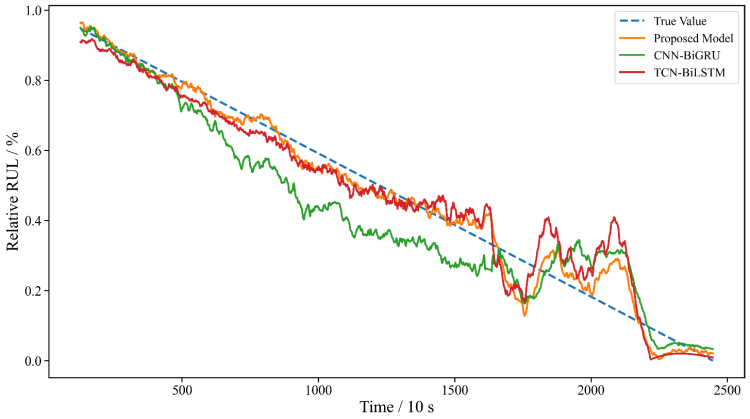

Figure 10. Comparison of all models on B1-6

Overall, the visualization results in Figure 8 to 10 show that the proposed model exhibits clear advantages in prediction accuracy, robustness, and adaptability to complex scenarios, and has strong practical application potential.

4.1.3. Ablation experiment

To further verify the effectiveness and rationality of each module in the proposed method, ablation experiments are conducted on the IEEE PHM 2012 bearing dataset. The experiments use the same data splitting strategy, feature extraction procedure, and hyperparameter settings to ensure comparability among the ablated models. The specific ablation settings are as follows:

1) Ablation method 1: Remove the SENet channel-attention mechanism, reducing the TCN-SENet branch to a standard TCN branch.

2) Ablation method 2: Remove the TCN-SENet branch and retain only the BiGRU-multi-head temporal attention branch for temporal feature modeling.

3) Ablation method 3: Remove the BiGRU module and rely mainly on the TCN-SENet branch to extract local temporal features and output the prediction result.

According to Table 4, after removing SENet, the TCN-SENet branch, or the BiGRU module, both RMSE and MAE increase to varying degrees, while R^2 decreases, indicating that each module can effectively improve prediction performance. Among them, removing BiGRU leads to the most obvious performance degradation, showing that BiGRU plays an important role in modeling temporal dependencies during degradation; removing the TCN-SENet branch also increases the error significantly, indicating that this branch can effectively extract local degradation features. In summary, the prediction performance of the complete model is superior to that of each ablated model, which verifies the rationality and effectiveness of the proposed architecture.

Table 4. Ablation experiment results on the IEEE PHM 2012 dataset

Model	RMSE	MAE	R^2
Remove SENet	0.1018	0.0834	0.8535
Remove TCN-SENet branch	0.1060	0.0869	0.8366
Remove BiGRU	0.1158	0.0930	0.8074
Complete model	0.0582	0.0477	0.9483

4.2. Experiment 2

4.2.1. Data description

To further verify the applicability and effectiveness of the proposed model in a real experimental scenario, Experiment 2 collects vibration signals based on a laboratory self-built bearing fault simulation platform, constructs a self-built bearing fault dataset, and uses it for further testing and analysis of the model. As shown in Figure 11, the test platform mainly consists of a PLC control unit, drive motor, transmission shaft system, test bearing seat, loading device, vibration signal acquisition terminal, and vibration excitation and calibration unit, among other components, and can realize bearing operation tests under different speed-load combinations. Among them, the signal generator is used to generate controllable standard excitation signals; the power amplifier is used to amplify the excitation signal to meet subsequent excitation requirements; and the vibration exciter is used to output stable and controllable mechanical vibration excitation, thereby completing sensor calibration, acquisition-chain debugging, and verification of the operating state of the test platform. Compared with public datasets, this self-built dataset is closer to the actual acquisition environment and can be used to examine the model's feature representation capability and generalization performance under multiple fault types and multiple operating conditions.

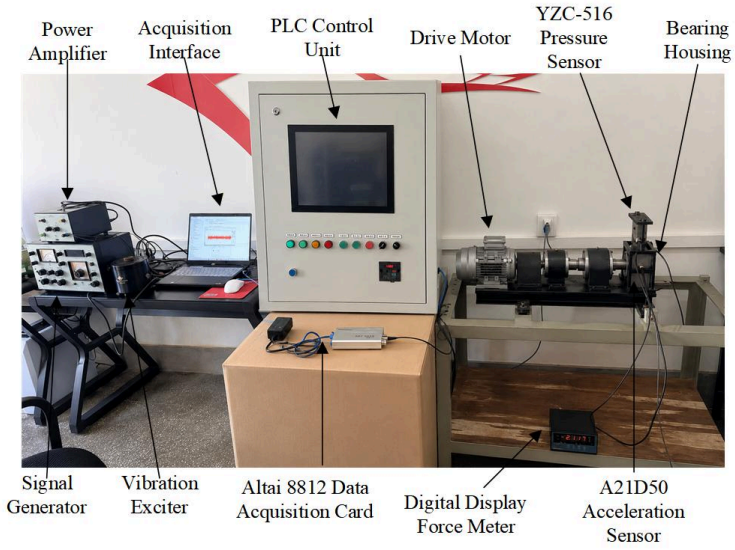

Figure 11. Laboratory small-sample self-built bearing fault simulation platform

SKF 31306 tapered roller bearings are selected as the test objects in this study. Due to laboratory constraints, two speed-load combinations are set: 1,000 r/min / 2,500 N (Condition 1) and 1,500 r/min / 2,000 N (Condition 2). Under each condition, full-life vibration data are collected separately for inner-race, outer-race, and rolling-element faulty bearings. The information of the self-built dataset is listed in Table 5. During signal acquisition, the vibration signals are collected by an A21D50 acceleration sensor mounted at the test bearing position and converted into digital signals by an Alientek 8812 data acquisition card. The sampling frequency is set to 25.6 kHz to ensure effective capture of impact components and high-frequency features of bearing faults. To ensure the rationality of the training and testing process, independent bearings are used to split the training and test sets under the same operating condition. For example, when B2-1 is selected as the test set, the data from B2-2 and B2-3 are used to form the training set.

Table 5. Information of the self-built small-sample bearing dataset

Condition ID	Dataset	Total samples	Acquisition time (min)
Condition 1 (1,000 r/min / 2,500 N)	B1-1	65	40.14
	B1-2	89	55.32
	B1-3	81	50.02
Condition 2 (1,500 r/min / 2,000 N)	B2-1	128	87.85
	B2-2	103	66.42
	B2-3	92	60.38

4.2.2. Experimental results and analysis

Table 6 presents the prediction results on the self-built dataset. Under Condition 1, the average RMSE of the three test bearings is 0.0809, the average MAE is 0.0625, and the average R^2 is 0.9205. Among them, the R^2 values of B1-1 and B1-3 are both higher than 0.94, indicating that the model can well fit the life-degradation trends of inner-race fault and rolling-element fault samples; the error of B1-2 is relatively higher, reflecting that outer-race fault samples under small-sample conditions may exhibit unstable degradation processes and more pronounced local impacts. Nevertheless, the average R^2 under this condition still exceeds 0.92, indicating that the model has certain applicability to laboratory-collected data. From the results under Condition 2, the model achieves an average RMSE of 0.0751, an average MAE of 0.0492, and an average R^2 of 0.9061. The prediction performance of B2-1 is the best, with RMSE, MAE, and R^2 values of 0.0179, 0.0148, and 0.9962, respectively, and the prediction curve closely follows the true life variation. The error of B2-3 is larger, indicating that the degradation process in this sample may involve stronger nonlinear changes or a sudden change in the later stage, making prediction more difficult.

Table 6. Main experimental results on the self-built small-sample dataset

Test bearing	RMSE	MAE	R^2
B1-1	0.0704	0.0521	0.9423
B1-2	0.1030	0.0799	0.8756
B1-3	0.0694	0.0556	0.9436
B2-1	0.0179	0.0148	0.9962
B2-2	0.0712	0.0604	0.9403
B2-3	0.1363	0.0724	0.7818

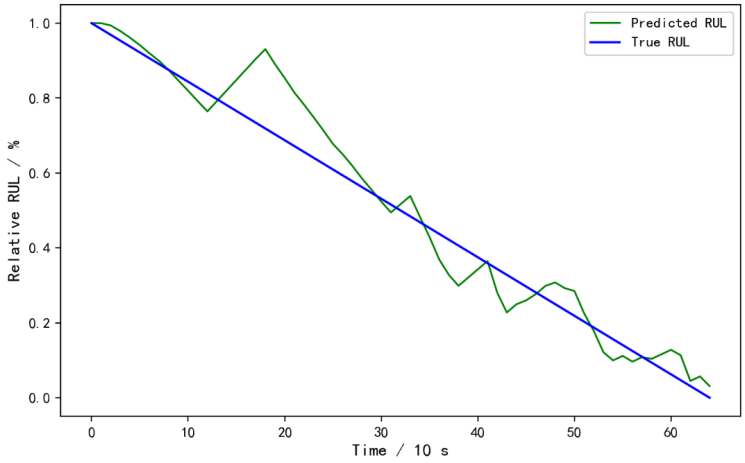

Figure 12. Prediction results for bearing B1-1 under Condition 1

As can be seen from Figure 12, the true RUL of bearing B1-1 under Condition 1 decreases in an approximately linear manner, and the prediction curve generally follows the decay trend of the true curve. At the early stage of operation, the predicted and actual values are relatively close, indicating that the multi-domain features extracted by AVMD can well characterize the early health state of the bearing; in the mid-degradation stage, the prediction curve exhibits certain fluctuations in local intervals, with short-term underestimation and overestimation near approximately 12 to 20 sampling points, which is related to the gradually intensified impact features of the inner-race fault and the relatively large fluctuation of local degradation features. In the later life stage, the prediction curve gradually returns close to the true RUL and can better reflect the rapid decay trend of the remaining useful life. Combined with Table 6, where the RMSE, MAE, and R^2 of B1-1 are 0.0704, 0.0521, and 0.9423, respectively, it can be concluded that the proposed model has good trend-tracking capability and prediction stability for inner-race fault samples under low-speed and heavy-load conditions.

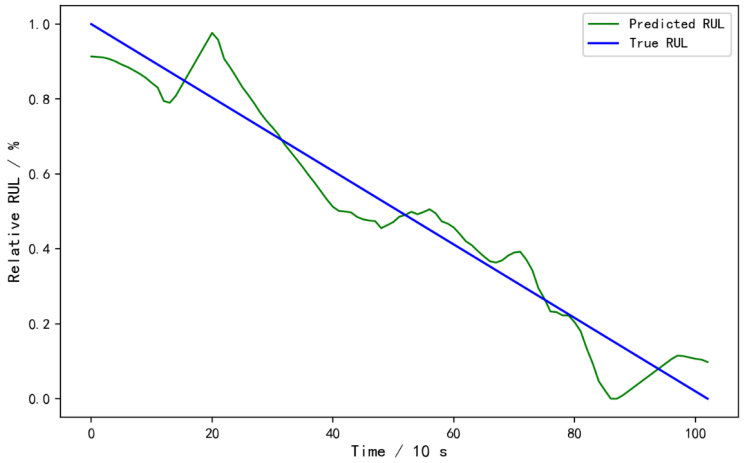

Figure 13. Prediction results for bearing B2-2 under Condition 2

As can be seen from Figure 13, the true RUL of bearing B2-2 under Condition 2 also decreases gradually over time, and the overall trend of the prediction curve is consistent with the true curve. At the early stage of life, the predicted result is slightly lower than the true value, indicating that the model is sensitive to the initial

degradation signs of the outer-race fault; near approximately 20 sampling points, the prediction curve shows a brief overestimation, suggesting that the outer-race fault impact and acquisition noise under medium-speed and medium-load conditions can affect local prediction to a certain extent. As the degradation process progresses, the model can track the downward trend of RUL well and shows strong responsiveness in the rapid-decay stage in the middle and later periods, although slight fluctuations still exist near the end of life. Combined with Table 6, where the RMSE, MAE, and R^2 of B2-2 are 0.0712, 0.0604, and 0.9403, respectively, it can be concluded that the proposed model also maintains high fitting accuracy and prediction stability on medium-speed and medium-load outer-race fault samples.

Considering the results of both conditions comprehensively, the average RMSE of the six test bearings is 0.0780, the average MAE is 0.0559, and the average R^2 is 0.9133. Compared with the PHM 2012 dataset, the self-built dataset contains fewer samples, and the noise in the acquisition environment, assembly errors, and individual fault differences are more pronounced, making prediction more difficult. The experimental results show that the proposed model can still maintain good trend-tracking capability under small-sample, multi-fault-type, and different speed-load combinations, indicating that it is not only suitable for public benchmark datasets but also has certain potential for practical engineering applications [28, 29].

5. Conclusion

This paper proposes a rolling bearing RUL prediction model that combines AVMD with the SETCN-BiGRU multi-head temporal attention mechanism. The model uses AVMD to adaptively decompose bearing vibration signals and extract multi-domain degradation features; the TCN-SENet branch enhances the representation of local temporal features and key channel features, while BiGRU with multi-head temporal attention captures long-term temporal dependencies and key degradation segments during the degradation process. According to the evaluation metrics RMSE and MAE, on the IEEE PHM 2012 dataset, compared with the CNN-BiGRU, GRU, TCN-Transformer, and TCN-BiLSTM models, the proposed model reduces the prediction error by approximately 27.1%, 20.9%, 30.8%, and 24.6%, respectively; on the self-built small-sample bearing dataset, the average RMSE, MAE, and R^2 reach 0.0780, 0.0559, and 0.9133, respectively. Ablation experiments show that SENet, the TCN-SENet branch, and the BiGRU module can all effectively improve model performance, with BiGRU exerting the most significant influence. Overall, the proposed method demonstrates good performance in prediction accuracy, robustness, and generalization across different operating conditions, and can provide an effective reference for rolling bearing condition monitoring and remaining useful life prediction.

References

[1] Zhang, J., Zou, T., & Wang, M. (2023). A review of remaining useful life prediction for rolling bearings. *Mechanical Science and Technology for Aerospace Engineering, 42*(1), 1–23.

[2] Yang, X., Tian, L., Zhang, Y., & Zhao, Y. (2025). Rolling bearing remaining useful life prediction method based on deep learning. *Journal of Xi'an University of Technology, 41*(3), 370–380.

[3] Guo, Y., Mao, J., & Zhao, M. (2023). Bearing remaining useful life prediction method based on CNN and attention BiLSTM. *Journal of Shanghai University of Engineering Science, 37*(1), 96–104.

[4] Zhou, Z., Liu, L., Song, X., & Chen, K. (2023). Remaining useful life prediction method of rolling bearing based on Transformer model. *Journal of Beijing University of Aeronautics and Astronautics, 49*(2), 430–443.

[5] Dong, Z., & Dong, J. (2024). Rolling bearing remaining life prediction method based on S-MCLSTM and DANN. *Application Research of Computers, 41*(9), 2787–2793.

[6] Medjaher, K., Tobon-Mejia, D. A., & Zerhouni, N. (2012). Remaining useful life estimation of critical components with application to bearings. *IEEE Transactions on Reliability, 61*(2), 292–302.

[7] Ben Ali, J., Chebel-Morello, B., Saidi, L., Malinowski, S., & Fnaiech, F. (2015). Accurate bearing remaining useful life prediction based on Weibull distribution and artificial neural network. *Mechanical Systems and Signal Processing, 56–57*, 150–172.

[8] Khelif, R., Chebel-Morello, B., Malinowski, S., Laajili, E., Fnaiech, F., & Zerhouni, N. (2017). Direct remaining useful life estimation based on support vector regression. *IEEE Transactions on Industrial Electronics, 64*(3), 2276–2285.

[9] Guo, L., Li, N., Jia, F., Lei, Y., & Lin, J. (2017). A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing, 240*, 98–109.

[10] Li, X., Ding, Q., & Sun, J. Q. (2018). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety, 172*, 1–11.

[11] Zheng, S., Ristovski, K., Farahat, A., & Gupta, C. (2017). Long short-term memory network for remaining useful life estimation. In *2017 IEEE International Conference on Prognostics and Health Management (ICPHM)* (pp. 88–95). IEEE. https: //doi.org/10.1109/ICPHM.2017.7998311

[12] Zhu, G., Zhu, Z., Xiang, L., Hu, A., & Xu, Y. (2023). Prediction of bearing remaining useful life based on DACN-ConvLSTM model. *Measurement, 211*, Article 112600.

[13] Cao, X., Zhang, F., Zhao, J., Duan, Y., & Guo, X. (2024). Remaining useful life prediction of rolling bearing based on multi-domain mixed features and temporal convolutional networks. *Applied Sciences, 14*(6), Article 2354. https: //doi.org/10.3390/app14062354

[14] Dragomiretskiy, K., & Zosso, D. (2014). Variational mode decomposition. *IEEE Transactions on Signal Processing, 62*(3), 531–544.

[15] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A, 454*(1971), 903–995.

[16] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132–7141). IEEE.

[17] Bai, S., Kolter, J. Z., & Koltun, V. (2018). *An empirical evaluation of generic convolutional and recurrent networks for sequence modeling*. arXiv. https: //doi.org/10.48550/arXiv.1803.01271

[18] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. https: //doi.org/10.1109/CVPR.2016.90

[19] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1724–1734). Association for Computational Linguistics.

[20] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates.

[22] Wang, Y., Liu, Q., & Peng, Y. (2023). Remaining useful life prediction of rolling bearings under non-uniform monitoring conditions. *Journal of Mechanical Engineering, 59*(23), 96–104.

[23] Chen, J., Mao, W., Liu, J., & Wang, G. (2023). Online remaining useful life assessment of rolling bearings under unknown working conditions based on time series transfer recurrent prediction. *Control and Decision, 38*(1), 112–122.

[24] Nie, L., Zhang, L., Xu, S., Cai, W., & Yang, H. (2023). Remaining useful life prediction of rolling bearings based on similarity feature fusion and CNN. *Noise and Vibration Control, 43*(5), 115–121.

[25] Song, L., Jin, Y., Guo, X., & Wang, H. (2024). Remaining useful life prediction method based on adaptive weight temporal convolutional network. *Journal of Beijing University of Chemical Technology (Natural Science Edition), 51*(3), 76–87.

[26] Chen, B., Chen, Z., Chen, X., & Guo, K. (2021). Remaining useful life prediction method for rolling bearings based on attention TCN. *Electronic Measurement Technology, 44*(24), 153–160.

[27] Yang, J., Li, B., & Liu, X. (2025). Cross-domain remaining useful life prediction of rolling bearings based on transfer learning. *Machinery, 52*(7), 17–24.

[28] Liu, K., Zhang, Y., Mao, W., & Wang, N. (2025). Online remaining useful life prediction method based on unsupervised deep domain adversarial adaptation. *Journal of Zhengzhou University (Natural Science Edition), 57*(1), 81–87.

[29] Zhang, J., Wang, L., Xiao, Y., & Ma, Y. (2025). Degradation feature information fusion and remaining useful life prediction of rolling bearings. *China Mechanical Engineering, 36*(7), 1553–1561.