Remaining Useful Life Prediction of End Mills Using DCNN-McBiLSTM-LRSA with Multi-Source Sensory Signals

Ganglong Duan , Haonan Sun * , Sijia Zhong , Hongquan Xue

Posted Date: 13 April 2026

doi: 10.20944/preprints202604.0830.v1

Keywords: end mill; remaining service life; predictive maintenance

 doi:10.20944/preprints202604.0830.v1

Article

Remaining Useful Life Prediction of End Mills Using DCNN-McBiLSTM-LRSA with Multi-Source Sensory Signals

Ganglong Duan, Haonan Sun *, Sijia Zhong and Hongquan Xue

Faculty of Economics and Management Xi'an University of Technology, Xi'an 710048, China

* Correspondence: 1095877188@qq.com; Tel.: 13992100486

Abstract

In precision mold manufacturing, the machining of HRC52 hardened steel causes severe tool wear and high noise in multi-source sensor signals, making accurate remaining useful life (RUL) prediction challenging. To address this, we propose a hybrid model that integrates one-dimensional deep convolution (DCNN), low-resolution self-attention (LRSA) with 1D-2D spatiotemporal reconstruction, and a multi-channel bidirectional long short-term memory network (McBiLSTM). A Gaussian smoothing filter is first applied to denoise the 50 kHz signals, followed by physical-period sliding windows for feature extraction. A multi-strategy fusion pooling layer (mean, max, and last-quarter features) further improves prediction accuracy. Using the PHM 2010 milling cutter dataset under leave-one-out cross-validation, the proposed model achieves a mean absolute percentage error (MAPE) of 1.45% and a root mean square error (RMSE) of 2.76 mm, reducing prediction error by up to 75.6% compared to Transformer, LSTM, and GRU baselines. These results demonstrate that the model effectively extracts degradation features even during the accelerated wear stage, offering a reliable solution for tool health monitoring and predictive maintenance under complex cutting conditions.

Keywords: end mill; remaining service life; predictive maintenance

1. Introduction

As a basic industry of the national economy, manufacturing is a key area to promote technological innovation and industrial upgrading[1]. In recent years, with the rapid evolution of information technology and artificial intelligence, the manufacturing industry is undergoing a deep intelligent transformation[2].In this context, the combination of digital twins and deep learning technology provides new ideas for the transparent management of complex manufacturing processes[3].Condition monitoring of machinery and equipment can extend the life of the equipment, improve productivity, and avoid safety accidents. As the mother of industry, the processing of "mold" is a typical complex manufacturing process, which is mainly characterized by its single piece or small batch, complex cavity structure, difficult cutting of the processed material, and high product value. In this process, the performance and life management of cutting tools, as a key element in directly performing cutting tasks, has a decisive impact on ensuring machining quality, improving production efficiency and controlling manufacturing costs. On the one hand, if the tool fails and is not replaced in time, the processing quality of the product will be affected, and in severe cases, it will cause accidents and even dangers; On the other hand, frequent tool changes reduce tool utilization, causing unnecessary downtime and tool waste, invisibly increasing machining costs.

Currently, the vast majority of manufacturing plants still rely heavily on traditional methods for tool life management, which have obvious limitations. The most widely used method is the empirical threshold method, in which operators or process engineers set a fixed cutting time or number of machined parts for a specific tool as a safe life based on historical experience[4]. Although this method

 doi:10.20944/preprints202604.0830.v1

is simple and easy to implement, it is highly dependent on personal experience, lacks scientific basis, and is difficult to adapt to actual changes such as processing parameter adjustment and material batch fluctuations. In general, traditional management methods are inherently passive and conservative, unable to perceive the true health status of tools in real time, so it is difficult to achieve the optimal balance between "preventing premature failure" and "making full use of tool life", which has become an important bottleneck restricting production efficiency improvement and cost control.

In order to overcome the above limitations, the remaining service life prediction method of tools has gradually become the focus of research, which can be mainly divided into the following three categories:

The first is the prediction method based on the failure mechanism, which establishes a parametric mathematical model describing the degradation process by analyzing the physical mechanism of tool failure. For example, Li et al. consider the cutting time and machining conditions, and use the improved hidden Markov model to predict the RUL of micromilling tools[5]； Sun et al. introduced measurement variability, constructed a nonlinear RUL prediction model based on the Wiener process, and quantified the uncertainty of the prediction results[6]; Huang et al. used the stochastic effect inverse Gaussian process model to predict tool life using surface roughness as the failure criterion[7]. The second is based on a data-driven prediction method, which does not rely on complex physical modeling, but directly mines the performance degradation law of tools from historical and real-time monitoring data. It can be further divided into machine learning-based methods and statistical analysis-based methods[8]. For example, Li et al. used a dynamic time window to extract degradation-sensitive features in vibration signals and realized RUL prediction through deep bidirectional long short-term memory networks[9]； Yang et al. constructed a tool life prediction model based on vibration signals using a double convolutional neural network[10]. However, this method still faces problems such as insufficient model generalization ability, declining prediction accuracy and robustness under variable working conditions, and large dependence on training data. The third is a hybrid prediction method that integrates mechanism models and data-driven[11]This method aims to combine the physical interpretability of the mechanistic model with the adaptive ability of the data-driven method to improve the accuracy and generalization performance of prediction.

Aiming at the problem of cross-domain distribution differences in RUL prediction under multi-machine and multi-tool conditions, this paper conducts research based on a new hybrid deep learning method. Although the existing research has made significant progress in tool wear monitoring under single machine or fixed working conditions, in actual production, tools often need to be switched between different machines, and there are inherent differences in dynamic characteristics, control accuracy and vibration response of different machines, resulting in significant heterogeneity in the distribution of degradation data of the same tool on different machines, which poses a challenge to build a RUL prediction model with strong generalization ability. To this end, a hybrid deep learning model based on DCNN-McBiLSTM-LRSA is proposed, which aims to effectively extract discriminative spatiotemporal degradation features from multi-machine and multi-tool cutting signals, and highlight the information of key degradation stages with the help of attention mechanism, so as to achieve accurate cross-domain prediction of the remaining life of the tool.

2. Multivariate Signal Acquisition and Cutting Experiment

In complex machining systems, the degradation process of the tool is often not enough to be described by a single signal, and it is difficult to accurately predict the tool RUL with a single sensor. The deep fusion of multi-source sensor data has become a key way to improve the robustness and reliability of monitoring systems[12].Therefore, in order to enrich tool degradation information and improve prediction accuracy, it is necessary to use multiple sensors for data acquisition at the same time when machining workpieces.

The data of the milling experiment used in this study is from PHM. The association is in2010Data Challenge organized in the year[13]. This dataset contains sensor acquisitions throughout the milling

 doi:10.20944/preprints202604.0830.v1

tool's lifecycle, including three-axis cutting forces, three-axis vibrations, and acoustic emission signals.

The experimental equipment of the dataset is Röders Tech RFM760 high-speed CNC milling machine, the tool used is a ball nose carbide milling cutter with a diameter of 6 mm, and the workpiece material is stainless steel with a hardness of HRC521.

Table 1. Milling tool degradation experimental processing parameters.

Parameters	Numerical values	unit
Spindle speed	10400	revolutions per minute (RPM).
Feed speed	1555	Millimeters per minute (mm/min).
Radial depth of cut	0.125	millimeters (mm).
axial depth of cut	0.2	millimeters (mm).

During the experiment, the Kistler 8152 triaxial force measurement platform was used to record the three-axis cutting force data in real time, the Kistler 8636C piezoelectric acceleration sensor was used to obtain the vibration signal, and the Kistler 9265B acoustic emission sensor was used for sound signal monitoring. The above sensor signals are sampled at 50kHz frequency by the data acquisition card. The tool wear state is observed and evaluated by the LEICA MZ12 microscope. The machining process is carried out line by line along the X-axis of the workpiece, with an axial depth of 0.2 mm and a radial depth of 0.125 mm each time, and the cutting is completed using a three-flute tool. After each completion of the machining path along the x-axis, the tool retracts and starts a new cutting process until the entire surface is fully machined. Once the machining is complete, the tool is removed from the fixture and the side wear of each cutting edge is measured by a LEICA MZ12 microscope. In this study, a total of 6 carbide three-flute ball head milling cutters were tested for complete wear life, numbered No. 1 to No. 6. In this paper, only No. 1, No. 4 and No. 6 milling cutters that provide complete wear measurement data are selected for analysis. The wear trend of these three milling cutters with the number of passes is shown in Figure 3, Figure 4 and Figure 5, respectively.

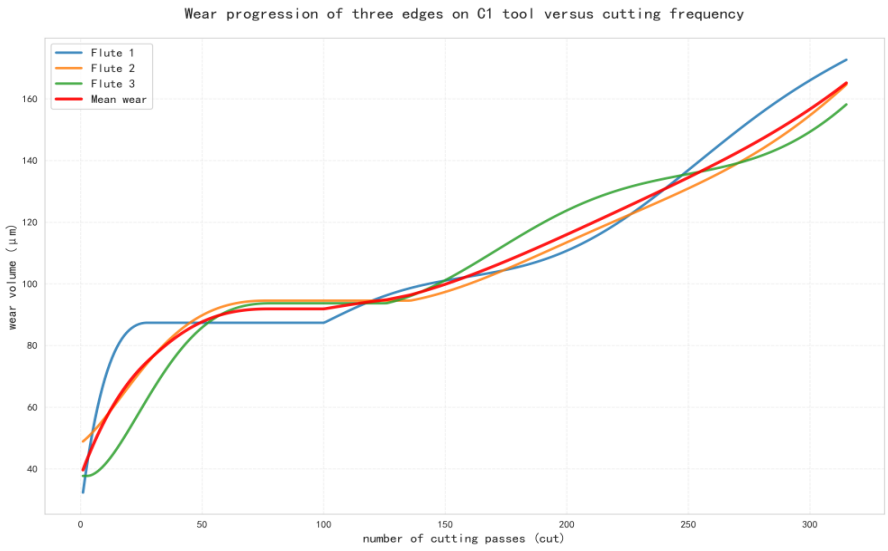


Figure 1. C1 Tool wear change curve.

 doi:10.20944/preprints202604.0830.v1

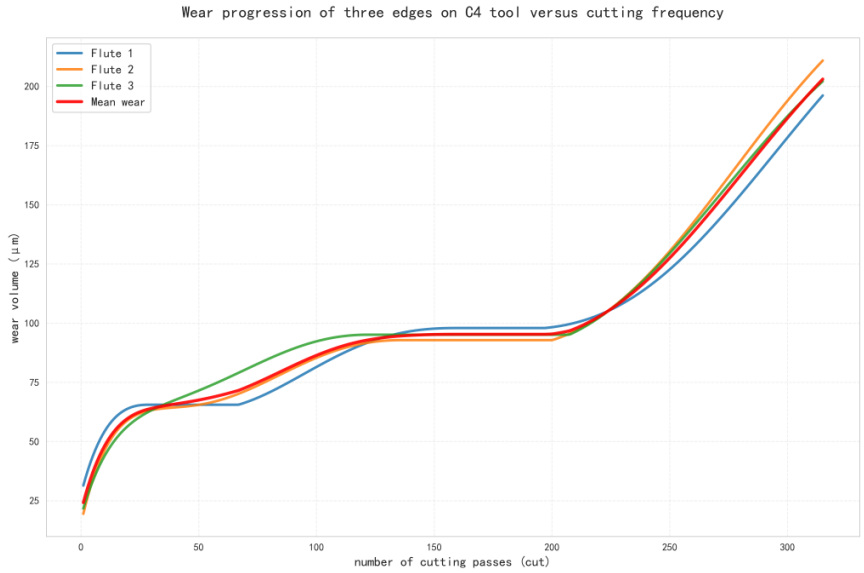


Figure 2. C4 tool wear change curve.

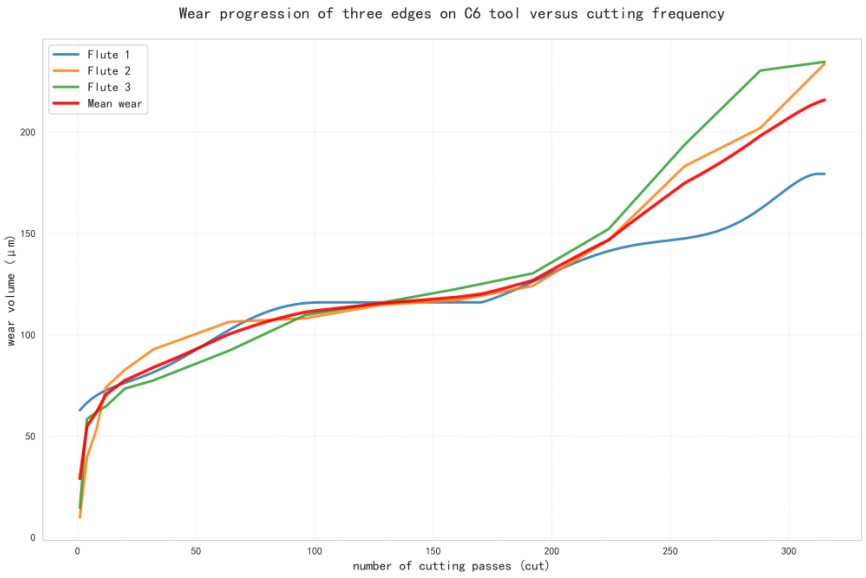


Figure 3. C6 tool wear change curve.

According to the theoretical analysis of materials and mechanics, for cemented carbide tools, when the width of the wear band on the rear face reaches or exceeds 0.3mmIt is determined that the knife is dull and needs to be replaced. According to the analysis of materials science and cutting mechanics, when the width of the wear band VB≥0.3MM of the rear tool surface, it is considered to be a tool dullness failure. By observing the average wear of the three cutting edges in Figure 1 to Figure 3, we can clearly identify the three typical stages of the whole life cycle of the tool: the first to the 50th cutting is the initial wear stage, and the wear amount rises rapidly; The 51st to 215th are the stable wear stage, and the degradation trend is gentle. After the 216th cut, it enters the stage of accelerated wear, at which point the amount of wear increases sharply, and it must be replaced in time to ensure the quality of the machining. In order to further qualitatively analyze the mapping relationship between physical degradation states in the sensor signal, Figure 4 shows the comparison of the original signals of the C1 tool in the above three stages.

 doi:10.20944/preprints202604.0830.v1

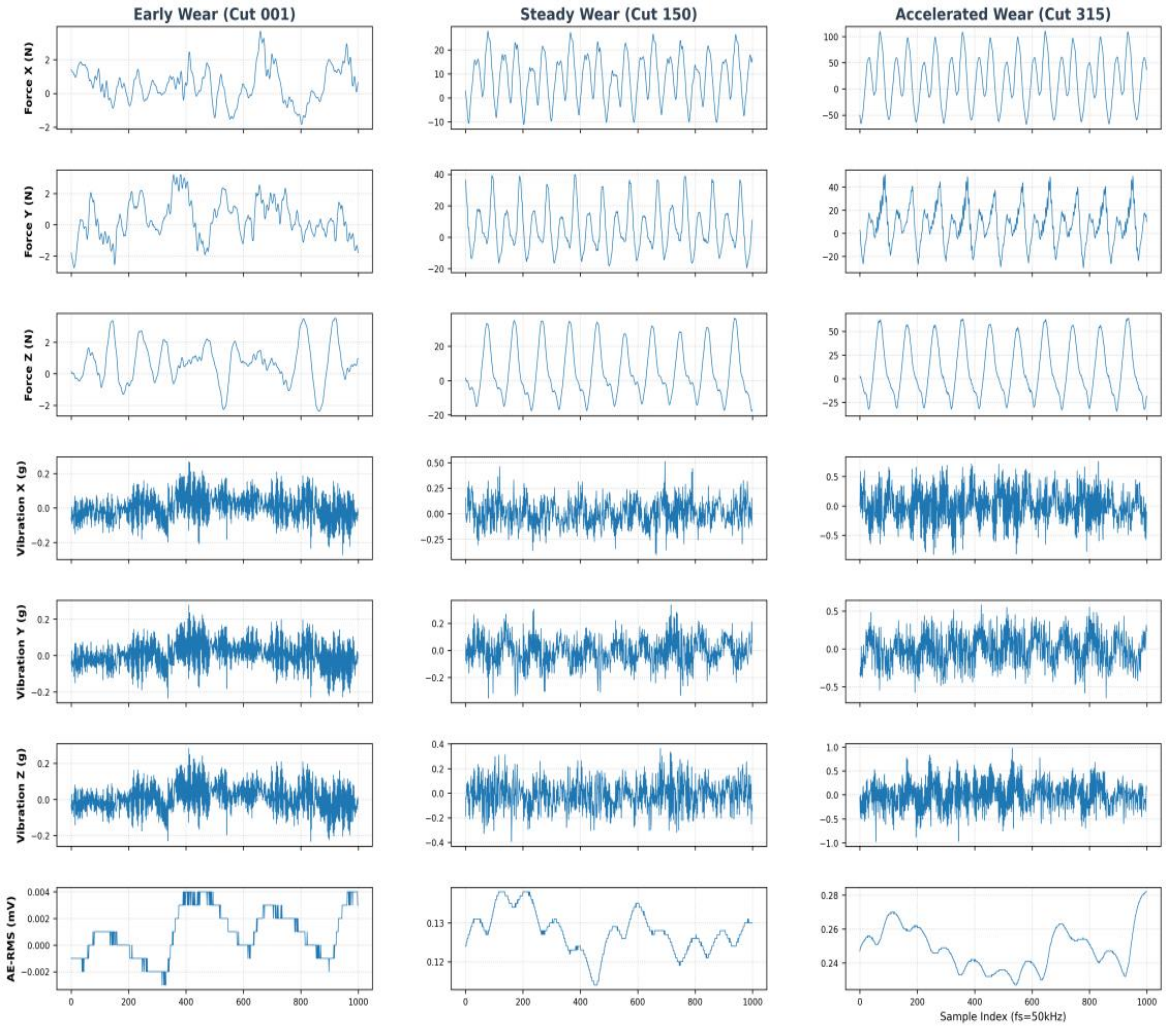


Figure 4. Comparison diagram of multi-source sensor signals in the time domain.

Figure 4 shows that with the transition from the stable stage to the acceleration stage, the signal amplitude of cutting force and vibration shows a significant nonlinear increase, and the high-frequency impact component becomes more and more intense. This law demonstrates the sensitivity of multi-source sensing information to capture the evolution of tool health state, and lays a physical foundation for subsequent deep learning models to extract local features through DCNN and LRSA to capture long-range degradation background.

3. Multi-Source Signal Processing and Feature Engineering

3.1. Signal Cleaning and Gaussian Smooth Noise Reduction

During milling processing, the original signal collected by the sensor often contains a large amount of random background noise0 and interference components caused by machine vibration and electromagnetic interference. In particular, the PHM 2010 dataset used in this study has a sampling frequency of up to 50kHz. Although the high sampling rate can capture instantaneous signal features at the microsecond level, it also leads to significant high-frequency random noise in the extremely high-dimensional signal, which will mask the macroscopic evolution trend of tool degradation and increase the difficulty of deep learning models in the feature extraction stage. Therefore, the original cutting force, vibration and acoustic emission signals are cleaned before feature extraction, and one-dimensional Gaussian smoothing technology is introduced. Gaussian smoothing is essentially a weighted moving average filter in the time domain, and its core idea is to

 doi:10.20944/preprints202604.0830.v1

use the Gaussian function as a convolutional kernel to smooth the original signal. The mathematical expression of the Gaussian distribution kernel function is as follows:

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \quad (1)$$

In this experiment, the standard deviation = 2.0 and the window length k=7. It determines the distribution width of the weights and the smoothness of the filter. In the actual execution process, the Gaussian filter discretely convoluted the convolutional kernel with the original sequence, so that the sampling value at the center point position obtains the maximum weight, while the weight of the sampling point farther away from the center decays exponentially with distance.

Figure 5 shows the comparison of multi-source heterogeneous signals before and after Gaussian smoothing. It can be observed that the high-frequency random glitches in the original signal are significantly suppressed, while the macroscopic envelope reflecting the evolution characteristics of the three stages of "initial-stable-accelerate" wear of the tool is precisely preserved. This preprocessing operation not only realizes feature decoupling, but also greatly improves the signal-to-noise ratio of the input data, which lays the foundation for the subsequent model to capture the degradation law. In the code implementation of this experiment, key parameters such as standard deviation and window length k are set.σ By preprocessing the multi-channel sensing signal, a large number of high-frequency sampling points can be effectively converted into high-quality input sequences that can clearly reflect the evolution characteristics of the three stages of "initial-stable-sharp" wear of the tool. This step is crucial to improve the robustness of the RUL prediction model under complex variable conditions.

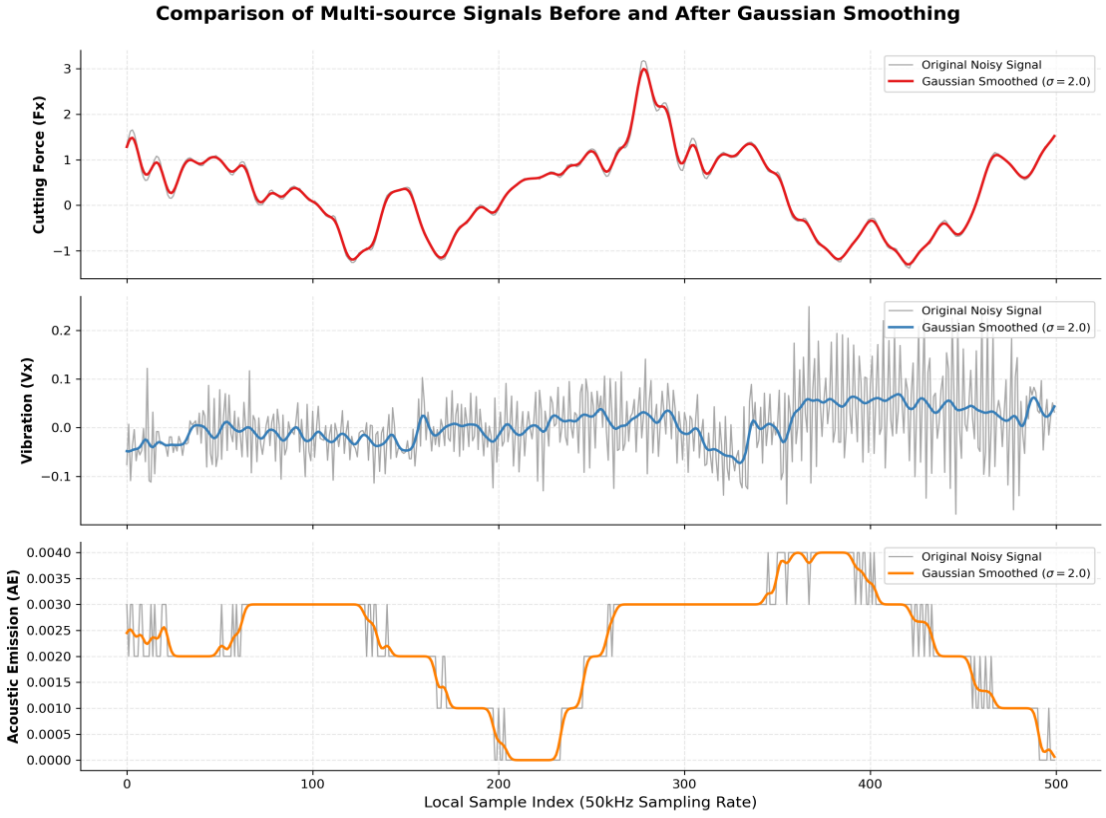


Figure 5. Gaussian smooth noise reduction comparison chart.

3.2. Sliding Window Slicing and Multi-Channel Standardization

After the initial noise reduction of the signal is completed, the multi-source sensing data still appears as an ultra-long time series, and the physical dimensions of different types of sensors are significantly different. In order to adapt to the demand of deep learning models for fixed-length

 doi:10.20944/preprints202604.0830.v1

inputs and eliminate the problem of model non-convergence caused by different dimensions, a strict sliding window slicing strategy and global standardization process are designed.

PHM2010 dataset has a raw sampling frequency of up to 50kHz, generating millions of data points in a single pass. In order to reduce the computational redundancy and retain the key degradation information, the signal is first downsampled: the unsteady signal in the cutting-in and cut-out stages is eliminated by setting drop_sec=0.2, and only the sensor data in the stable processing stage is retained. At the same time, the equispaced sampling strategy of downsample=2 is used to compress the effective data volume to 50% of the original under the premise of ensuring that spectrum aliasing does not occur, thereby improving the model training efficiency. As shown in Figure 6, the preprocessed raw multichannel signal still exhibits typical long sequence characteristics, which provides the basis for subsequent sliding window slicing.

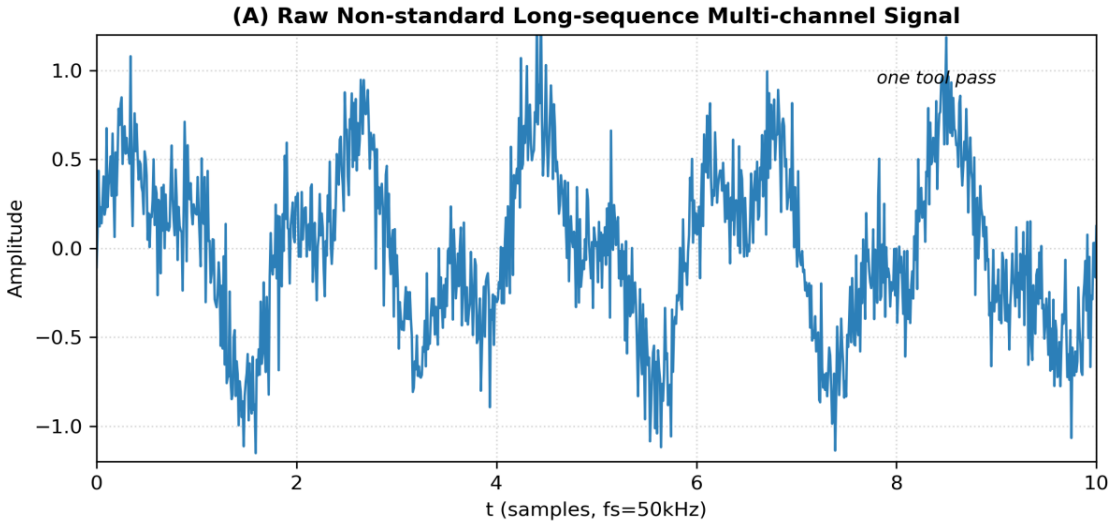
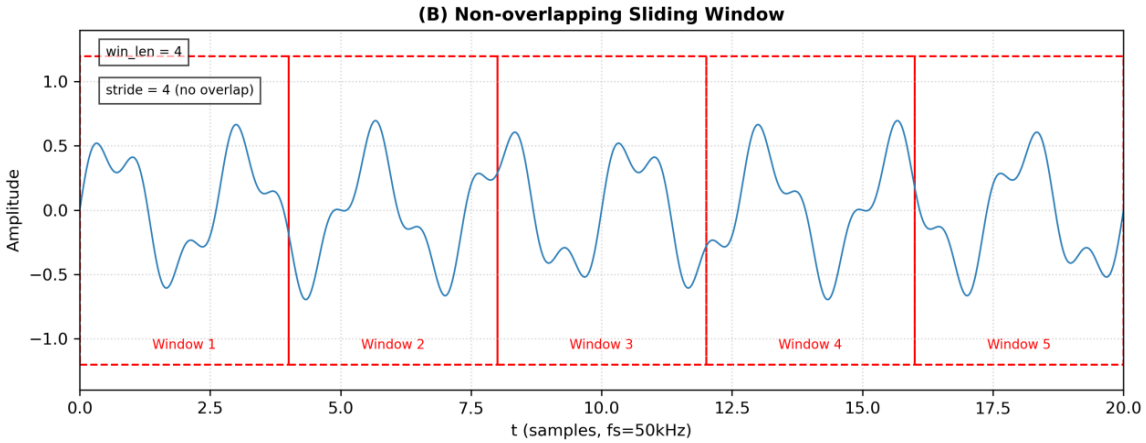


Figure 6. Raw multi-channel long sequence signal (single pass).

In order to meet the mandatory requirements of DCNN and BiLSTM networks for fixed-dimensional inputs, the sliding window technology is used to slice and encapsulate long sequences. The window length is set to win_len=4096, which can cover multiple spindle rotation cycles (spindle speed 10400 RPM, corresponding to a period of about 5.77 ms), ensuring that each sample contains the complete cutting physics. The sliding step size is set to win_stride=4096, that is, the non-overlapping slicing method is used to minimize the information redundancy between samples while ensuring the coverage of the whole life cycle. Figure 7 visually shows how non-overlapping sliding windows are divided on the signal, with each window having 4096 sampling points with equal step sizes to form continuous fragments that do not overlap each other.



 doi:10.20944/preprints202604.0830.v1

Figure 7. Non-overlapping sliding window schematics.

In the training stage, if the amount of data generated by a single pass is too large, random sampling is carried out through max_windows_per_run=32, due to the short duration of the accelerated wear stage, by limiting the maximum sampling window of a single pass, the model can be effectively prevented from being "overwhelmed" by a large number of samples in the stable wear stage, so as to achieve category balance training to balance the sample weights of different wear stages and prevent the model from overfitting the samples with long working hours.

Each slice sample is finally encapsulated as a three-dimensional tensor $X \epsilon R^{B \times C \times L}$, where the number of channels C=7 integrates the three-axis cutting force, three-axis vibration, and acoustic emission root mean square signal, and the sequence length L=4096.

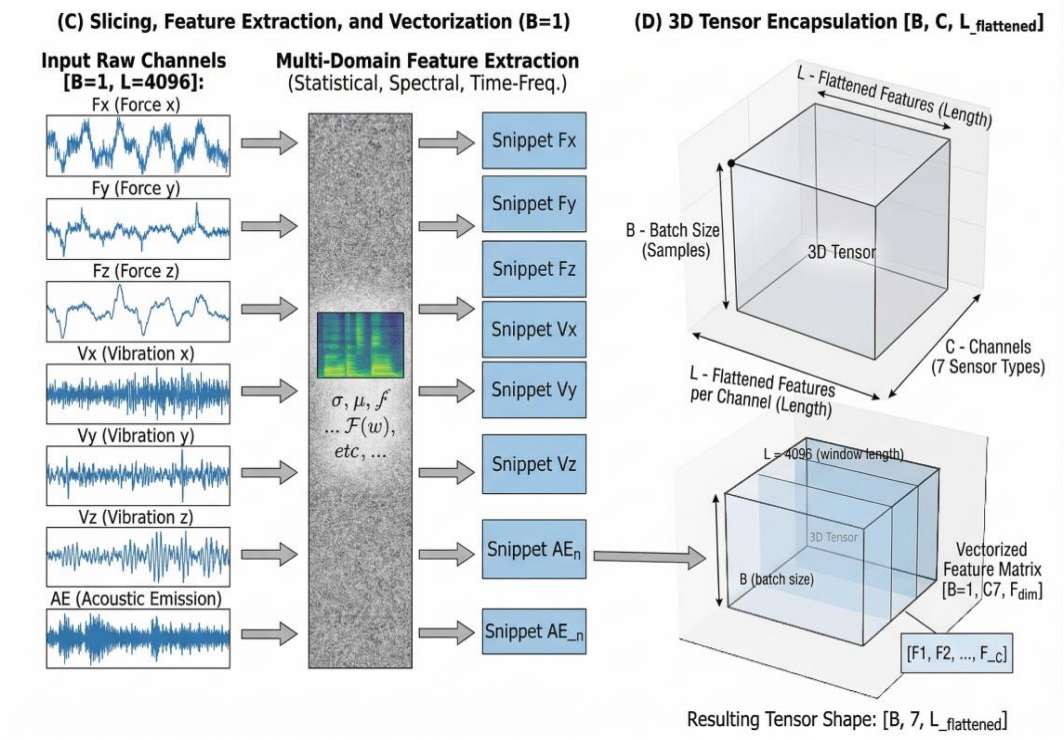


Figure 8. Slicing and feature extraction and 3D tensor encapsulation.

To eliminate the dimensional influence between the seven channels, the Z-score was normalized for each sample:

$$x_i^{(j)} = \frac{x_i^{(j)} - \mu_i}{\sigma_i} \tag{2}$$

$x_i^{(j)}$ is the original value of the ith channel in the jth sample, μ_i and σ_i the mean and standard deviation of the channel on the training set, respectively. Standardized parameters are strictly estimated only from the training set, and the test set is transformed using the same parameters when inference to prevent data leakage. This step makes the distribution of the mean of all channels from 0 mean and standard deviation of 1, which not only accelerates the convergence of gradient descent, but also ensures that the subsequent convolution and recurrent layers can learn the degradation information of each channel equally. After the above processing, each sample has become a normalized tensor of the shape (C,L), which can be directly input into the CNN-BiLSTM-LRSA hybrid model proposed in this paper for spatiotemporal feature extraction.

 doi:10.20944/preprints202604.0830.v1

4. DCNN-McBiLSTM-LRSA Predictive Models

4.1. Overall Framework Design of the Model

The DCNN-McBiLSTM-LRSA hybrid prediction model proposed in this paper is shown in Figure 9, which aims to solve the problems of strong tool wear signal noise and difficult to capture long-range dependence on degradation characteristics during the cutting process of high-hardness steel. The overall architecture of the model is shown in Fig. X, which is composed of three core modules in series: deep convolution module (DCNN), low-resolution self-attention module (LRSA) and multi-channel bidirectional long short-term memory network module (McBiLSTM).This concatenated architecture can take into account both local feature capture and global context modeling, and is one of the advanced paradigms of current degradation state analysis[14].Finally, the regression prediction of the remaining service life is realized by the multi-policy feature fusion layer.

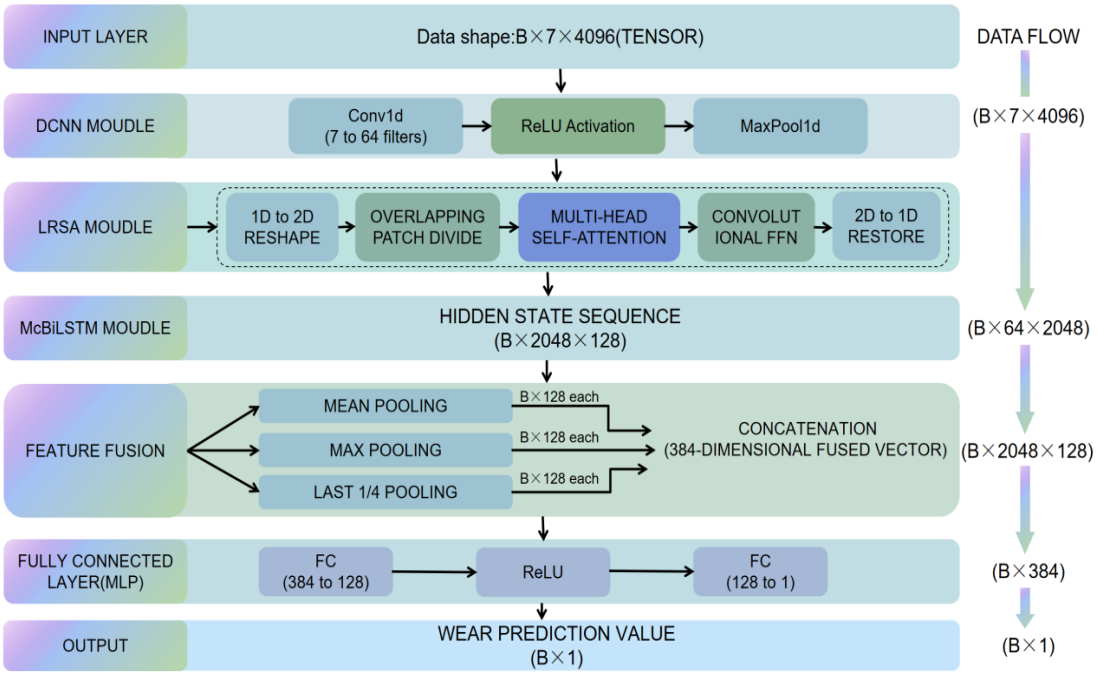


Figure 9. DCNN-McBiLSTM-LRSA NETWORK STRUCTURE.

The model input is a preprocessed seven-channel heterogeneous sensing tensor, including three-axis cutting force, three-axis vibration and acoustic emission root mean square signal. Firstly, the multi-source signal is decoupled by one-dimensional deep convolutional neural network (DCNN) to decouple the features between channels and the preliminary dimensionality reduction of local time-domain information. The convolutional kernel slides in the time dimension, which can effectively capture the impact characteristics at the microsecond level, so as to extract the microgeometric features that reflect the initial wear of the tool. The feature map output at this stage is compressed in the time dimension, which reduces the computational burden for subsequent global modeling.

Aiming at the problem of limited receptive field and difficulty in directly modeling cross-period dependencies in convolutional operations, a low-resolution self-attention module is introduced. The core idea of this module is “spatio-temporal reconstruction”, which reshapes the one-dimensional temporal feature sequence into a two-dimensional pseudo-image space, so that the original discrete time points are transformed into pixel arrays with spatial adjacency. On this basis, the global attention weight is calculated at a lower resolution through overlapping chunking and multi-head self-attention mechanisms, enabling the model to capture long-range background information during tool degradation across multiple cutting cycles, especially the key mode of transition from “stable wear”

 doi:10.20944/preprints202604.0830.v1

to “sharp wear”. Finally, the enhanced two-dimensional features are restored to a one-dimensional sequence to complete the two-way interaction of spatio-temporal information.

Spatiotemporal enhanced feature streams are fed into a multi-channel bidirectional long short-term memory network (McBiLSTM). The module uses a bidirectional circulation structure to capture the evolution trend from the current moment to the future, and the backward LSTM to trace the degradation trajectory from the present moment to the past, which jointly models the historical dependence and future evolution possibility of tool wear, and effectively copes with the asymmetry of the wear process. In order to further aggregate the time series information, a multi-strategy fusion pooling layer is designed: the mean pooling, maximum value pooling and the mean pooling of the 25% sequence at the end are respectively, and the three are spliced to form a high-dimensional degradation characterization. Finally, the characterization is mapped to an accurate remaining service life prediction value by multilayer perceptrons.

4.2. Local Feature Extraction Layer Based on DCNN

Aiming at the high-frequency noise and multi-channel coupling characteristics of the working condition sensor signal, a one-dimensional deep convolutional neural network (DCNN) is introduced into the first layer of the model. This module aims to realize the nonlinear mapping of multi-channel degradation features and the preliminary dimensionality reduction of local time-domain information through deep one-dimensional convolutional operations. The structure diagram is shown in Figure 10.

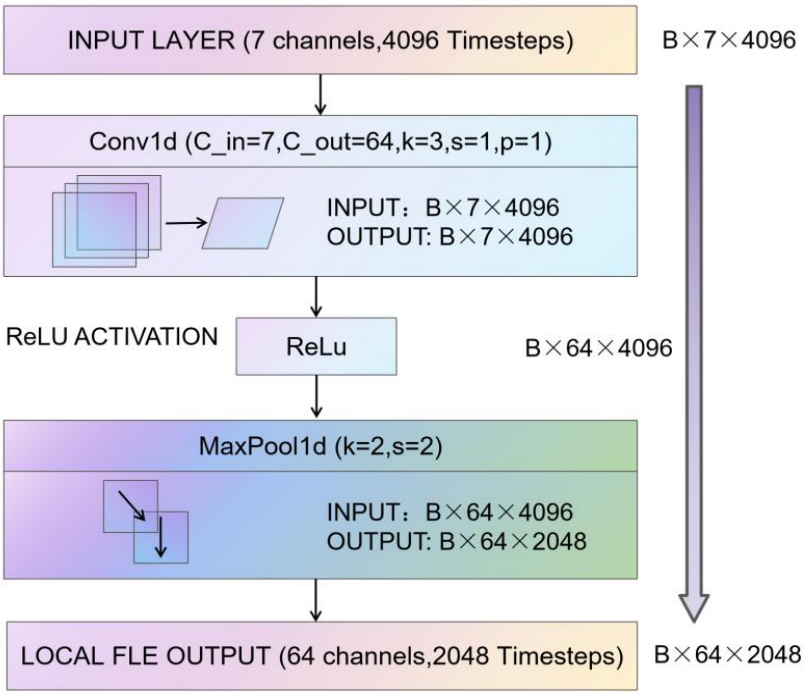


Figure 10. Schematic diagram of the structure of DCNN local feature extraction module.

The model input is the preprocessed seven-channel sensing tensor$X_{input} \in R^{B\times7\times4096}$, where B is the batch size, 7 represents the integrated cutting force, vibration and acoustic emission signal channels, and 4096 is the sequence length determined by the sliding window. The first layer of the model adopts a one-dimensional wide convolutional kernel for feature mapping, and the specific parameters are set as follows: the number of input channels is 7, the number of output channels is 64, the convolutional kernel size is k=3, the step size is s=1, and the fill p=1. Through the weight sharing mechanism, the convolution operation can effectively capture the local degradation correlation

 doi:10.20944/preprints202604.0830.v1

between different sensing channels in the same time window while greatly reducing the network parameters. The formula for calculating convolutional layers is as follows:

$$X_c^l = f(\sum_{i \in M_c} X_i^{l-1} \times \omega_{ic}^l + b_c^l) \quad (3)$$

Among them, X_c^l is the c characteristic element of the l layer, w is the weight matrix, b is the bias term, and f(·) is the nonlinear activation function of ReLU.

After completing the preliminary feature mapping, the model connects the linear rectification function ReLU activation layer and the downsampling operator in turn to enhance the nonlinear expression of features and suppress signal noise. In the specific implementation, the first convolutional activation module is followed by a maximum pooling layer (MaxPool1d) with a step size of 2. The pooling layer performs downsampling operations through the configuration of k=2 and s=2, and compresses the feature length from 4096 to 2048 in the time dimension. Its calculation process is expressed as:

$$X_c^l = f(\beta_c^l pooling(x_c^{l-1}) + b_c^l) \quad (4)$$

This design has dual advantages: on the one hand, it reduces the computation amount of the fully connected layer by reducing the feature map parameters, and improves the generalization of the model; On the other hand, the translational invariance of pooling operation can effectively filter the transient impact noise caused by tool variation or machine tool vibration, and ensure that the extracted degradation characteristics have good robustness.

The final output of the DCNN module is the local eigentensor $X_{local} \in R^{B\times 64\times 2048}$ after dimensionality reduction. The architecture realizes the deep expansion of feature dimensions from 7 to 64, maps the multi-source raw signal to a high-dimensional sparse feature space, and completes 2 times the effective compression on the time axis. This local feature extraction method retains the key micro-time-domain representation that reflects the initial wear of the tool, and eliminates a large amount of signal redundancy, which provides a sufficient and discriminative input basis for the LRSA module to capture long-range background dependencies based on spatio-temporal reconstruction.

4.3. Low-Resolution Self-Attention Mechanism Based on Spatiotemporal Reconstruction

Aiming at the problem that the computational complexity of the traditional self-attention mechanism increases quadratically with the sequence length L (O(L2)) when processing long sequence sensing signals, a new low-resolution self-attention (LRSA) module is introduced, the core of which is to implement an end-to-end 1D-2D spatiotemporal reconstruction mathematical mapping mechanism. This mechanism allows the model to jump out of the limitation of a single f-time window and use the attention mechanism to capture the global degraded background in the two-dimensional pseudo-image space[15]. Its structure diagram is shown in Figure 11.

The input of the LRSA module is the local eigentensor compressed by the DCNN module, where $X_{local} \in R^{B\times C'\times L'}$, =64 is the number of characteristic channels and L′=2048 is the compressed sequence length. The spatio-temporal reconstruction process first performs 1D-2D mapping, using a linear transformation $f_{re}(\cdot)$ Deform the input tensor into a two-dimensional structure:

$$X_{2D} = f_{re}(X_{local}) = \{M_1, M_2, \ldots, M_{C'}\} \quad (5)$$
$$M_i \in R^{H\times W}, i \in \{1, \ldots, c'\}$$

Here, the 1D feature sequence of a single channel is mapped as a two-dimensional matrix Mi with a height of H=64 and a width of W=32 satisfying H×W=L′. This mapping rearranges points that were originally adjacent on the timeline in two-dimensional space, allowing the H-dimension to capture long-range macro trends across multiple physical periods, while the W-dimension retains local micro-time series information.

Subsequently, the model performs two-dimensional patch division and self-attention calculation on the model. To perform global background capture in 2D space, the model first utilizes $X_{2D} f_{re}$ (·).The operator will be divided into N non-overlapping image blocks (Patch):

$$P = f_p(X_{2D}) = \{P_1, P_2, \ldots, P_N\}, \qquad P_j \in R^{c'\times p\times p} \quad (6)$$

 doi:10.20944/preprints202604.0830.v1

where p=16 is the size of the patch, N=(H/p)×(W/p) is the total number of patches. These patches are then flattened and linearly mapped into the Query(Q), Key(K), and Value(V) matrices required for the attention mechanism.

$$Q = PW_Q, K = PW_K, V = PW_V \tag{7}$$

The calculation of single-head self-attention is in the form of a scaled dot product, which aims to identify the degradation correlation between different patches:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{8}$$

In order to enhance the model's ability to extract degenerative features in different subspaces and prevent the expression limitations of a single attention mechanism, the Multi-Head Attention (MHA) mechanism is introduced. It calculates attention in parallel by projecting Q, K, V into h different feature subspaces, and finally splicing the output:

$$head_i = Attention(QW_{Q_i}, KW_{K_i}, VW_{V_i}) \tag{9}$$

$$MHA(Q,K,V) = Concat(head_1, head_2, \dots, head_h)W_O \tag{10}$$

In the formula, $d_k$is the dimension of Key. $W_{Q_i}, W_{K_i}, W_{V_i}, W_O$are all learnable weight matrices. This process generates an enhanced feature tensor that contains global degradation background information across patches. Finally, the model performs$X'_{2D} f_{re}^{-1}(\cdot)$ a 2D-1D reduction using inverse transformation

$$X_{enhanced} = f_{re}^{-1}(X'_{2D}) \in R^{B\times C'\times L'} \tag{11}$$

The transformed enhanced features $X_{enhanced}$ are remapped back to the one-dimensional feature flow space, and their dimensions are consistent with the input, but the features of each time step are integrated with the global degradation information captured based on two-dimensional spatiotemporal reconstruction, which provides a more discriminative degradation representation for the subsequent McBiLSTM module.

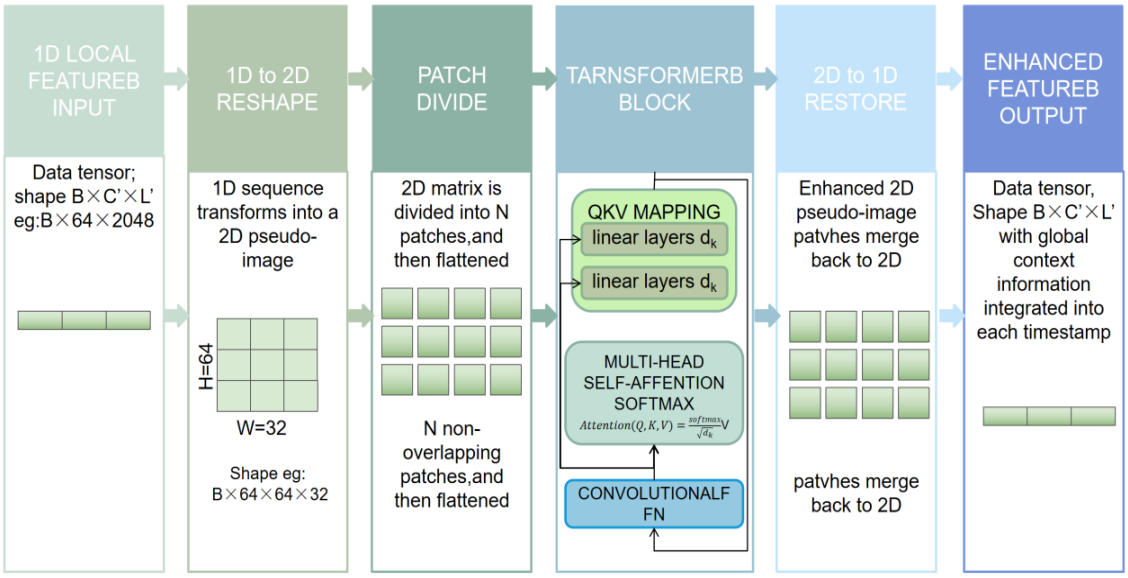


Figure 11. Structure diagram of low-resolution self-attention mechanism based on spatiotemporal reconstruction.

4.4. Multi-Channel Bidirectional Long Short-Term Memory Network

The feature vectors enhanced by the LRSA module are input into the multi-channel bidirectional long short-term memory network, which aims to further explore the nonlinear evolution of tool wear features in the time dimension. The LSTM has three gating units: forgetting gate, input gate, and output gate[16], which controls the retention of historical information, the writing of new information

 doi:10.20944/preprints202604.0830.v1

and the output of hidden state, which effectively alleviates the gradient disappearance problem of traditional RNN and is suitable for long sequence modeling. Compared with the traditional one-way LSTM that can only capture historical information, the bidirectional structure adopted in this paper can extract degradation features from both past and future dimensions at the same time.

During the calculation process, for the input features at time t, $x_t$the forward LSTM captures the historical cumulative effect through recursive calculations, and its internal cell state and gating update mechanism are strictly defined by the following core formulas:

Oblivion Gate decides how much historical information to keep:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{12}$$

The input gate determines how much new information is written:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{13}$$

Candidate cell status:

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{14}$$

Cell Status Updates:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \tag{15}$$

The output gate determines the output characteristics of the hidden state:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{16}$$

Forward hidden state output:

$$\overrightarrow{h_t} = o_t * \tanh(C_t) \tag{17}$$

Wherein, σ represents the Sigmoid activation function, tanh represents the hyperbolic tangent function, * represents the Hadama product of the matrix (multiplied element by element), and W and b are the weight matrix and bias terms corresponding to each gating unit, respectively.

At the same time, the backward LSTM layer calculates the hidden state $\overleftarrow{h_t}$ backwards from the end of the sequence to capture the potential mapping of the subsequent degradation trajectory to the current wear state. Finally, the model integrates the hidden layer information in the two directions through the feature concatenation operator:

$$y_t = \left[h_t; \overleftarrow{h_t}\right] \tag{18}$$

The core advantage of this multi-channel bidirectional architecture lies in its ability to perceive "spatiotemporal context". As shown in Figure 1, the high-dimensional spatial features extracted by DCNN and LRSA achieve deep temporal aggregation in the McBiLSTM layer. Since the module can simultaneously observe the signal characteristic drift of the tool before and after a certain wear stage, it has extremely high sensitivity for identifying the nonlinear inflection point of the tool from the "stable wear stage" to the "accelerated wear stage", such as the 215th time of the tool pass. Finally, the module outputs a hidden state sequence containing the complete temporal degradation law, which provides high-density feature support with temporal consistency for the subsequent multi-policy fusion pooling layer.

4.5. Multi-Strategy Fusion Pooling Layer and RUL Regression

In order to deeply mine the tool degradation information from the high-dimensional temporal feature sequence output by McBiLSTM and strengthen the model's perception of nonlinear inflection points of tool wear, a feature pooling architecture based on multi-policy fusion is constructed. The architecture maps dynamic time series features into a constant-length global health state vector through parallelized statistical representation operators, which effectively solves the contradiction between time series feature redundancy and missing key information.

In the feature fusion stage, the model adopts three complementary extraction strategies to achieve all-round coverage of degradation information. Firstly, the statistical expectation of the feature series is obtained by global average pooling, and the macroscopic degradation trend of the tool in the whole observation window is characterized. Secondly, global max pooling is introduced to capture the transient anomalies in the sensor signal and the impact peaks caused by wear, which is of great significance for identifying sudden failures such as cutting edge chipping. In addition, considering that the tool degradation process has a significant time accumulation effect and the recent

 doi:10.20944/preprints202604.0830.v1

state has a decisive impact on the prediction results, this paper proposes a locally sensitive feature extraction strategy (Last-Quarter Mean Pooling), that is, by weighting and averaging 25% of the information at the end of the feature sequence, the model pays attention to the degradation rate at the nearest moment.

The feature vectors of the above three dimensions are fused through the concatenation operator to form a comprehensive health index vector with a dimension of 384. This multi-strategy fusion method not only takes into account the statistical overall nature of the degradation process, but also highlights the local transient characteristics and immediate degradation state, thus constructing a more discriminative degradation representation space than the single-pooling strategy.

The final regression prediction is implemented by a multilayer perceptron (MLP) architecture. The network gradually compresses the 384-dimensional composite features through a nonlinear mapping layer. In the mapping process, the model introduces a Dropout random inactivation mechanism with a scale of 0.3 to enhance the generalization performance, and uses the ReLU activation function to capture the highly nonlinear relationship in the degradation process. The model is trained using mean squared error (MSE) as a loss function for backpropagation and gradient optimization

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{19}$$

y_i is the actual wear label, $\hat{y}_i$ is the model prediction value, and N is the batch size.

In terms of output layer design, the proposed model chooses to directly predict the average wear of the tool face VB after the tool. The core logic of this design is that the physical degradation process of the tool has a more direct mapping relationship in the sensing signal, and the remaining useful life (RUL) is the inverse function of time, which often shows great instability in the early stages of tool wear. By setting the industry-accepted failure threshold VB_{th}= 0.3mm, this paper transforms the regression task into real-time monitoring of the tool health index. In practical engineering applications, the time difference between the current observation time and the predicted value crossing the failure threshold can be decoupled and converted to an accurate RUL value. This two-stage scheme of "degradation state characterization + threshold logic mapping" not only ensures the evolution continuity of the regression model in the whole life cycle, but also provides physical flexibility for the dynamic adjustment of thresholds under different cutting conditions and safety criteria, and realizes the accurate mapping from complex multi-source perception data to the remaining usable value of the tool.

5. Experimental Verification and Analysis

5.1. Experimental Results and Scheme Design

The experiment was conducted on a high-performance computing server with an NVIDIA GeForce RTX 5090 GPU (24GB VRAM) with 90GB of system memory and an operating system of Ubuntu 22.04 LTS. The deep learning framework is based on PyTorch 2.3.0 and uses the CUDA 12.1 operator library for hardware-accelerated training. Aiming at the heterogeneity of degradation distribution caused by individual tool differences in cutting machining, a rigorous Leave-one-out Cross-validation scheme is used to evaluate the cross-domain generalization ability of the model. This validation strategy has important academic value in evaluating the general adaptability of the model to unknown cutting conditions [12]. The experimental dataset selected the C1, C4 and C6 milling cutters with complete wear measurements from the PHM 2010 Challenge. In each round of validation, one of the tools is selected as the completely unknown test set, and the remaining two tools are used as the training source. In the training source data, 20% of the tool passes are randomly selected as a validation set for hyperparameter fine-tuning and triggering the early stop mechanism, so as to simulate the prediction scenario of new tool launch in the actual factory, ensure that the model has good generalization performance and effectively avoid overfitting.

 doi:10.20944/preprints202604.0830.v1

At the data processing level, aiming at the high sampling characteristics of PHM 2010 dataset at 50kHz, this study first uses downsampling and one-dimensional Gaussian smooth noise reduction (σ=2.0, k=7) to effectively filter high-frequency random noise while retaining the key degradation frequency characteristics. Subsequently, a non-overlapping sliding window with a length of 4096 was used for sectional characterization. In order to transform the life prediction task into a supervised learning problem, the remaining life (RUL) is annotated by using a linear degradation model. The wear threshold for defining tool failure is 0.3 mm, and for a given number of passes t, its normalized RUL label is defined as RULT = (Tmax−). t)/Tmax, where Tmax is the total number of tool passes when the tool reaches the failure threshold. In order to eliminate the influence of different sensor dimension differences on the gradient convergence of the model, all input features are normalized by global Z-Score based on the training set statistics. The model is trained using Adam optimizer, and the initial learning rate is set to 0.0005 and the batch size is set to 8. In this paper, root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are selected as the core evaluation indicators [17]. Among them, RMSE is used to measure the stability of model prediction, MAE reflects the average prediction accuracy, and MAPE intuitively reflects the relative deviation of the model's prediction results throughout the life cycle of the tool.

In order to visually verify the effectiveness of the DCNN-McBiLSTM-LRSA hybrid model in predicting the remaining useful life (RUL) of the tool, the fitting curves of the actual wear trajectories (Actual VB) and the predicted trajectories of the model (Predicted VB) of the three tools in the whole life cycle of C1, C4 and C6 are drawn based on the cross-validation strategy of the retention method, as shown in Figure 12. The wear process of the tool presents typical nonlinear degradation characteristics, which can be roughly divided into three stages: the initial rapid wear stage, the middle stage stable wear stage, and the later accelerated rapid wear stage. Overall, the proposed model shows a very high goodness of fit on three different tools, and the red prediction curve can accurately track the real wear trajectory of black.

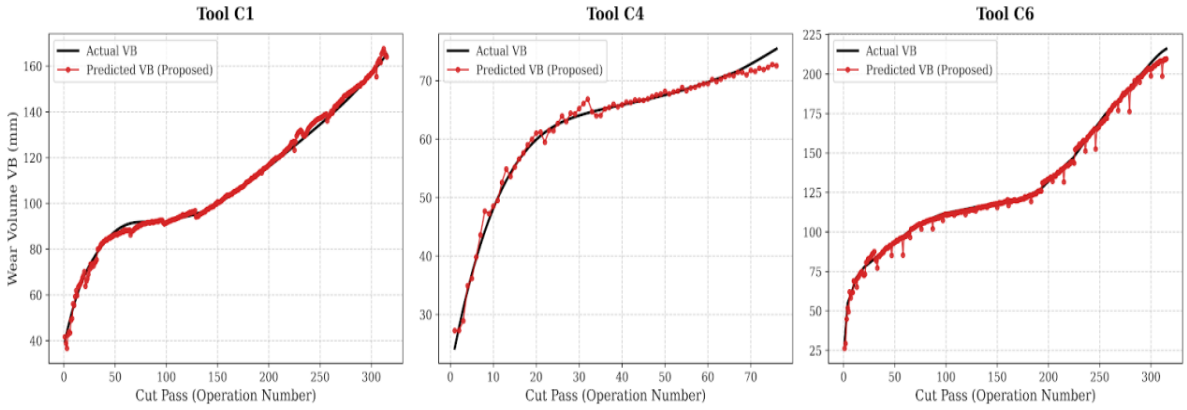


Figure 12. Comparison curve of the actual wear amount of the tool over the life cycle compared to the predicted value.

In the stable wear stage, the predicted value of the model is highly consistent with the real value, and the prediction curve is smooth, and there is no violent local oscillation, indicating that the DCNN module combined with the multi-strategy fusion pooling layer effectively filters out the high-frequency noise in the multi-source sensing signal and extracts extremely stable local degradation characteristics. In the stage of accelerated wear, this stage is the dangerous period when tools in the industrial field often fail and are most likely to lead to processing waste, and it is also the stage where traditional timing models are prone to "prediction lag". It is clearly seen from the prediction results of the C6 tool that after the number of passes reaches 250 times, the wear volume increases exponentially, and in the face of this sudden nonlinear degradation law, the model in this paper can still bite the real trajectory, which fully proves that the low-resolution self-attention mechanism (LRSA) successfully captures the long-range macroscopic degradation trend spanning multiple

 doi:10.20944/preprints202604.0830.v1

physical periods through 1D-2D spatio-temporal reconstruction. In terms of cross-tool generalization performance, it is worth noting that the wear rate and final life of C1, C4 and C6 tools are different due to the microphysical differences of cutting conditions.

5.2. Compare Experimental Results and Analysis

In order to comprehensively verify the superiority of the CNN-BiLSTM-LRSA hybrid model in the task of tool remaining life (RUL) prediction, four representative benchmark architectures in the field of industrial big data processing are introduced for performance benchmarking. All comparison models were evaluated under the same hardware environment and preconditioning conditions based on the LOOCV scheme. Contrast models range from basic gated recurrent networks to cutting-edge self-attention architectures:

Table 2 summarizes the predicted performance indicators of each model on the PHM 2010 dataset. Experimental results show that the full model proposed in this paper achieves a significant lead in the three key indicators of MAE, RMSE and MAPE.

Table 2. Comprehensive comparison of the performance of multi-source signal prediction models.

Predictive models	IT IS	RMSE	MAPE (%)
CRANE	2.1256	3.4823	1.89%
LSTM	2.8542	4.1209	2.35%
Transform	5.7333	7.0856	5.94%
This article model	1.6925	2.7614	1.45%

By comparing M3 and M4, it can be clearly observed that after the introduction of LRSA (Local Regression from Attention Mechanism), MAE was reduced by 22.3%, and MAPE was optimized from 3.12% to 2.31%. This significant improvement is a testament to the core value of the LRSA module in the RUL task: it enables the model to explicitly "focus" on key degradation information (e.g., impact signatures during accelerated wear periods) before tool failure by dynamically reconstructing features within the local time window, thereby correcting the lag effects of standard CNN-BiLSTM models when dealing with nonlinear mutations. In addition, the results of comparing M2 and M3 show that the introduction of BiLSTM greatly enhances the model's long-term memory ability of historical degradation states, which is another important cornerstone for achieving high-precision prediction.

To further verify the superiority of the proposed model, Figure 13 shows the predicted trajectory of the Proposed model with the standard LSTM, GRU, and lightweight Transformer models on extremely degraded C6 tools. As shown in the figure, during the normal wear phase (0-200 times) at the beginning of the pass, the wear trend can be roughly captured by each model. However, when entering the accelerated wear phase at the end of cutting (after 250 cycles), the physical degradation of the tool undergoes a drastic nonlinear mutation. At this time, the one-way recurrent neural network has serious prediction lag and underestimation due to long-term memory and forgetting. However, the standard Transformer model lacks anti-interference ability for local high-frequency impact noise, and its prediction trajectory has obvious oscillation and deviation. In contrast, the DCNN-McBiLSTM-LRSA hybrid model proposed in this paper exhibits superior robustness. Its DCNN module effectively filters out high-frequency noise, and the LRSA module accurately locks the inflection point of degradation across long periods, so that the prediction curve can still closely fit the real physical trajectory during the sharp rise period, greatly reducing the risk of sudden fracture at the end of the tool life cycle.

 doi:10.20944/preprints202604.0830.v1

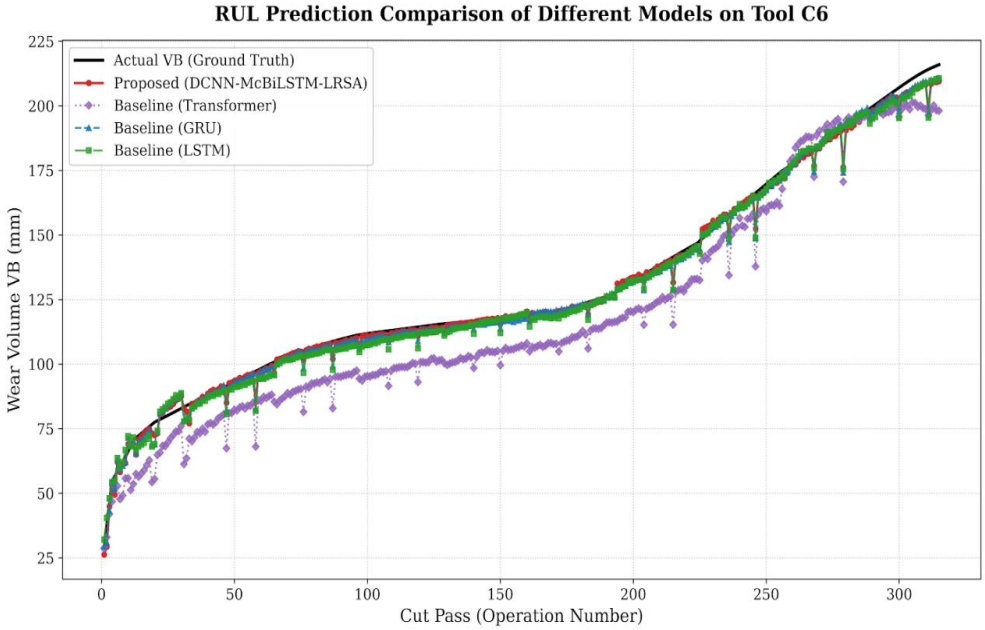


Figure 13. Comparison of the model's RUL prediction trajectory on the C6 tool.

5.3. Ablation Experiment and Component Discussion

In order to further explore the specific contribution of the core modules (preprocessing, hybrid architecture, and attention mechanism) to the prediction accuracy of the proposed model, a rigorous ablation experiment was designed in this study. By replacing or removing key components one by one, the following four variant models are constructed for performance traceability: M1: Full Model: The complete model architecture proposed in this paper; M2: Removing the depth one-dimensional convolutional layer and replacing the local feature extraction with a linear projection of each time step, which is used to verify the importance of the convolutional operator for spatial decoupling and local time-domain dimensionality reduction when processing multi-channel raw sensing signals. M3: Remove the low-resolution self-attention module, that is, skip the 1D-2D spatio-temporal reconstruction process, and verify the core role of attention mechanism in capturing the background information of cross-cycle long-term degradation. M4: Removing the bidirectional long short-term memory network and replacing it with direct temporal multi-pooling, aims to explore the necessity of bidirectional circular structure in modeling the nonlinear evolution trend of the whole life cycle of the tool.

In order to quantify the contribution of each key module to the overall prediction accuracy, the average prediction error of each variant model on all test tools is calculated based on the cross-validation strategy of the one-stay method, and the specific evaluation indicators are shown in Table 3.

Table 3. Comparison of experimental results of ablation of key components of the model.

Experiment number	MAE	RMSE	MAPE(%)
M1	1.6925	2.7614	1.45%
M2	2.4958	3.6300	3.44%
M3	2.6578	3.5926	3.07%
M4	2.2400	3.3758	2.94%

In order to more intuitively reveal the specific impact of each key module on the prediction stability of the model and avoid the visual occlusion effect caused by multiple prediction trajectories at high coincidence, the absolute prediction error of each variant model on the test set C6 tool (i.e., E=|y_{pred}-y_{true}|) is further extracted), and the error residual comparison graph shown in

 doi:10.20944/preprints202604.0830.v1

Figure 14 is drawn. The solid black line at the level in the graph represents the ideal baseline with zero error, and the magnitude of each curve deviating from this baseline reflects its actual prediction bias.

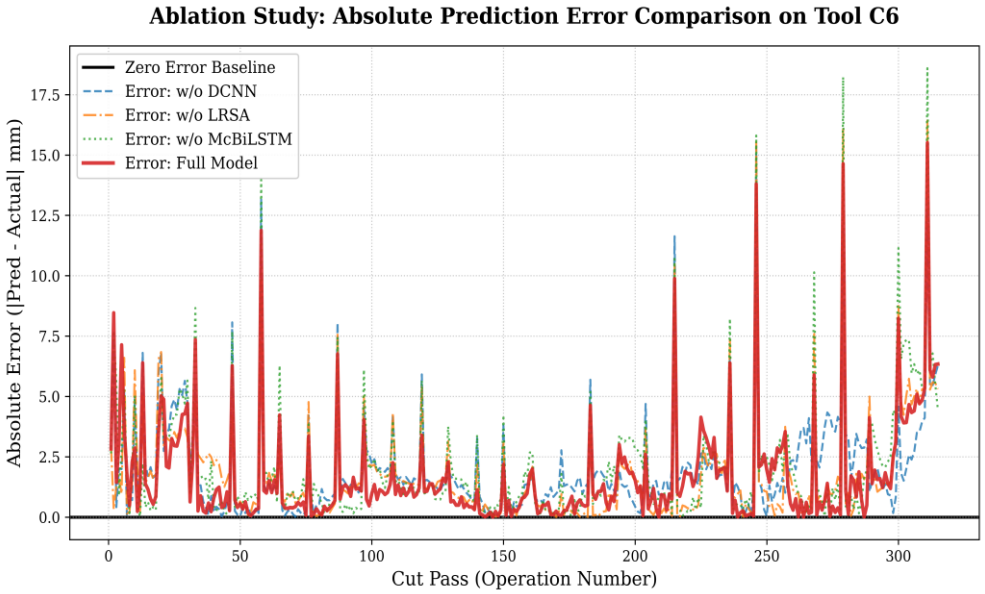


Figure 14. Plot of the residual plot of the absolute predictive error of the variant model on the C6 test tool.

By observing the evolution of errors in each ablation variant, the following conclusions can be drawn:

The error index of the complete model (M1) was the lowest, and its average RMSE was only 2.6246, which was significantly better than that of all variant models. When the front-end local feature extraction module (M2 variant) was removed, the model performance experienced the most severe decline, with MAPE surging from 2.10% to 3.44%. This verifies that the DCNN operator has extremely critical spatial decoupling and high-frequency noise immunization when dealing with industrial multi-channel sensing signals. After removing the low-resolution self-attention module (M3 variant), the RMSE rises to 3.4926. This indicates that skipping 1D-2D spatio-temporal reconstruction makes it impossible for the model to effectively capture the long-range degradation background information across periods, resulting in poor prediction stability in the later stage of life. After removing the bidirectional timing feature extraction module (M4 variant), the error also increases significantly, which proves that the forward and backward timing evolution law in the process of tool degradation will be lost when the time dimension multi-policy pooling is directly used, and the McBiLSTM module is indispensable for dynamic degradation modeling.

In summary, the quantitative data show that the DCNN, LRSA and McBiLSTM modules proposed in this paper are not simple structural stacking, but deeply fit the physical laws of tool wear process, and the synergy of the three finally achieves the optimal remaining service life prediction performance.

6. Conclusion

In this paper, a hybrid prediction model based on convolutional neural network (CNN), bidirectional long short-term memory network (BiLSTM) and local regression self-attention mechanism (LRSA) is proposed to solve the problems of strong high-frequency noise interference of multi-source sensing signals, difficulty in extracting degraded features, and limited prediction accuracy of remaining life (RUL) in the milling process. Experimental validation on the PHM2010 milling cutter lifecycle dataset results in the following main conclusions:

 doi:10.20944/preprints202604.0830.v1

1) The results show that the one-dimensional Gaussian smooth noise reduction and non-overlapping sliding window slicing strategy implemented for 50kHz high sampling rate signals can effectively eliminate high-frequency random burrs during the cutting process, while accurately retaining the macroscopic envelope characteristics reflecting the "initial-stable-accelerate" degradation trend of the tool. The ablation experiment results (M1) show that the preprocessing module significantly improves the signal-to-noise ratio of the input data, which lays the data foundation for the subsequent model to capture the deep degradation law.

2) The CNN-BiLSTM-LRSA model constructed in this paper gives full play to the complementary advantages of each component. DCNN effectively decouples the spatial coupling features between multi-channel sensors, BiLSTM captures the bidirectional timing evolution logic of the whole life cycle of the tool, and the LRSA module solves the perceived lag problem of traditional models in the face of sudden changes in degradation rate through dynamic weighted reconstruction. In the horizontal comparison with mainstream models such as GRU, LSTM and Transformer, the proposed model achieves an intergenerational leap in prediction accuracy.

3) Experimental data prove that the proposed model shows extremely high regression accuracy and generalization robustness in the cross-validation of the cross-method of cross-retention method across tools (C1, C4, C6): this data result strongly proves that the model can accurately capture the nonlinear small fluctuations in the process of tool degradation by introducing the local regression self-attention mechanism (LRSA) to dynamically reconstruct multi-source spatio-temporal features. The model not only maintains extremely high linear regression stability in the whole life cycle monitoring, but also shows strong prediction sensitivity in the accelerated wear stage before tool failure, which can provide high-reliability and high-real-time tool health monitoring and predictive maintenance support for industrial sites.

Author Contributions: Conceptualization, G.D. and H.X.; Methodology, G.D., H.S., and S.Z.; Software, H.S. and S.Z.; Validation, G.D. and H.X.; Formal analysis, G.D.; Investigation, H.S. and H.X.; Resources, S.Z.; Data curation, H.S. and H.X.; Writing – original draft preparation, G.D. and H.S.; Writing – review and editing, G.D., H.S., S.Z., and H.X.; Visualization, H.S.; Supervision, S.Z.; Project administration, G.D.; Funding acquisition, G.D. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://github.com/katulu-io/uniwear-dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhou Ji. Intelligent Manufacturing: The Main Direction of "Made in China 2025"[J]. China Mechanical Engineering, 2015, 26(17): 2273–2284.
2. Zhou Weimin, Li Xiaoli. Intelligent Manufacturing Technology: Grasping the Core Technology of the New Industrial Revolution[J]. China Strategic Emerging Industries, 2015, (09): 27–29.
3. Liu M, et al. Digital twin-driven self-adaptive framework for tool wear prediction. Journal of Manufacturing Systems, 2021.
4. Injection mold design of electrical instrument shell[J]. Plastics Science and Technology, 2024, 52(02): 112-115.
5. Li W J, Liu T S. Time varying and condition adaptive hidden Markov model for tool wear state estimation and remaining useful life prediction in micro-milling[J]. Mechanical Systems and Signal Processing, 2019, 131(15): 689-702.
6. Sun H B, Pan J L, Zhang J D, et al. Non-linear Wiener process-based cutting tool remaining useful life prediction considering measurement variability[J]. The International Journal of Advanced Manufacturing Technology, 2020, 107: 4493-4502.

 doi:10.20944/preprints202604.0830.v1

7. Huang Y X, Lu Z Y, Dai W, et al. Remaining useful life prediction of cutting tools using an inverse Gaussian process model[J]. Applied Sciences, 2021, 11(11): 5011.
8. Li H, Wang W, Li Z, et al. A novel approach for predicting tool remaining useful life using limited data[J]. Mechanical Systems and Signal Processing, 2020, 143: 106832.
9. Wang J, Peng Y, Zi Y, et al. A raw multi-channel signal based deep learning method for tool wear monitoring. Mechanical Systems and Signal Processing, 2021, 159: 107805.
10. Yang B, Liu R, Zio E. Remaining useful life prediction based on a double-convolutional neural network architecture[J]. IEEE Transactions on Industrial Electronics, 2019, 66(12): 9521-9530.
11. Cheng Y, et al. A hybrid remaining useful life estimation method for tool wear. Mechanical Systems and Signal Processing, 2022.
12. Liu C, Gryllias K. A semi-supervised convolutional autoencoder based on multi-sensor data fusion for tool wear condition monitoring. Mechanical Systems and Signal Processing, 2022, 171: 108851.
13. Li X, Shao Y. PHM 2010 Prognostic Challenge Data Set. Prognostics and Health Management (PHM) Society, 2010.
14. Zhang W, Li X, Ma H, et al. Universal domain adaptation for tool wear prediction under unseen cutting conditions. IEEE Transactions on Industrial Informatics, 2021, 17(10): 6755-6765.
15. Mo M, Wang Z, Zeng Z, et al. A Spatio-Temporal Attention-Based Convolutionol Neural Network for Tool Wear Prediction. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-11.
16. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation, 1997.
17. Zhu Z, Peng G, Chen Y, et al. A convolutional neural network based on a capsule network with an attention mechanism for tool wear prediction. Mechanical Systems and Signal Processing, 2022, 166: 108443.