coatings

Article

Multi-Scale Dual-Attention Feature Network with Bidirectional Temporal Constraints for Tool Wear Monitoring

Youqiang Xu [1], Rongyi Li [1,*], Xianli Liu [1], Haotuo Liu [1], Ying Wang [2], Xiaohua Liu [3] and Yuqiang Gan [1]

1 Key Laboratory of Advanced Manufacturing and Intelligent Technology, Ministry of Education, Harbin University of Science and Technology, Harbin 150080, China; 18846022204@163.com (Y.X.); xianli.liu@hrbust.edu.cn (X.L.); liuhaotuo@hrbust.edu.cn (H.L.); 19189424398@163.com (Y.G.)
2 School of Mechanical Engineering, North University of China, Taiyuan 030051, China; y_wang0813@163.com
3 Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong; xiaohualiu@mae.cuhk.edu.hk
* Correspondence: lirongyi@hrbust.edu.cn; Tel.: +86-136-0451-6610

Abstract

Accurate tool wear monitoring plays a decisive role in machining efficiency, product quality and reliability in modern manufacturing systems. Existing deep learning methods struggle to balance the high-frequency transient features and low-frequency evolution trends in tool wear signals, often losing key temporal evolution details when processing long-range degradation data. Therefore, this paper proposes an online prediction method of tool wear value that combines multi-scale convolution and dual-attention temporal features. This method extracts local mutation and trend features in wear signals through multi-scale convolution, captures wear evolution features through bidirectional cyclic network, and adaptively fuses local detail information and global trend through dual attention mechanism SWGC-DA to generate a multi-scale time series feature-driven prediction model. The ablation experiment based on the PHM2010 public data set verifies the effectiveness of the network structure design and demonstrates the model's superior predictive ability. Experiments on the self-built TiAl alloy milling dataset achieved a stable prediction of R^2 up to 99.1%, with MAE and RMSE of 2.29 and 2.47, respectively. The results show that this method significantly improves the accuracy and robustness of wear prediction.

Keywords: tool wear prediction; multi-scale convolution; dual-attention mechanism; Bidirectional Long Short-Term Memory (Bi-LSTM); multi-sensor information fusion

1. Introduction

In the context of modern manufacturing advancing toward intelligence, cutting tools serve as core fundamental components in high-end equipment manufacturing, and their wear condition directly affects machining efficiency, machining quality, and manufacturing costs [1]. During the machining process, flank wear is a complex and inevitable progressive process. Severe tool wear can lead to tool breakage and workpiece rejection, and may even affect the normal operation of machinery [2]. In actual production, traditional tool management relies excessively on manual experience; however, premature replacement results in 15%–20% waste of production costs [3]. Delayed replacement, on the other hand, leads to decreased machining accuracy and equipment damage. Therefore, establishing an accurate and reliable online tool wear monitoring system in the manufacturing field is of

Academic Editor: Mark Baker

Received: 26 January 2026
Revised: 13 February 2026
Accepted: 25 February 2026
Published: 27 February 2026

 https://doi.org/10.3390/coatings16030291

great significance for improving machining quality stability and reducing manufacturing costs [4].

Early research on tool wear monitoring generally employed direct observation and measurement of tool wear using equipment such as optical microscopes and scanning electron microscopes [5]. Although the direct method can provide more accurate tool wear information, it requires stopping the equipment for inspection, making real-time online monitoring difficult to achieve. Consequently, it has gradually been replaced by indirect monitoring methods. Indirect monitoring methods refer to the use of various sensor devices installed on machine tools to collect signals related to tool wear, such as cutting force signals, vibration signals, and acoustic emission signals, and then infer the tool wear condition through signal processing [6]. To overcome the high cost and low stiffness of conventional intelligent tool holders, Xue et al. [7] developed a mountable monitoring ring. This hardware solution integrates triaxial vibration sensing with Variational Mode Decomposition (VMD) and manifold learning for efficient feature reduction and online monitoring. Among these relevant signals, cutting force signals can directly reflect the dynamic characteristics during the machining process, with cutting force gradually increasing as wear intensifies. Vibration signals are more sensitive to tool wear and can reflect changes in the dynamic characteristics of tool wear. However, characterizing wear using a single signal is limited. In contrast, multi-sensor information fusion captures tool condition from multiple perspectives, integrating complementary data to enhance monitoring robustness and accuracy [8]. Xu et al. [9] proposed a parallel convolutional neural network (CNN) architecture that leverages multi-scale feature fusion and a channel attention mechanism. By incorporating residual connections to adaptively optimize feature map weights, this framework effectively enhances the robustness of tool wear prediction under complex cutting scenarios.

In terms of signal processing and feature extraction, traditional methods mainly rely on time-domain statistical features, frequency-domain features, and time-frequency domain features. These types of features have certain limitations: First, feature selection and extraction rely excessively on expert prior knowledge. Additionally, manual features can only capture shallow statistical characteristics of signals and have difficulty mining deep correlation patterns within the data. Most importantly, when fusing multi-sensor signals, how to determine the weights of different features and fusion strategies remains a challenging problem. Current research has extensively utilized machine learning algorithms to map sensor signals to wear states. For instance, Chen et al. [10] employed LS-SVM, Ganeshkumar et al. [11] utilized random forest and K-nearest neighbor regression, and Heitz et al. [12] incorporated physical models into XGBoost. However, these methods rely heavily on expert knowledge for manual feature extraction and often struggle to capture deep nonlinear correlation patterns within the data.

Among deep learning methods, CNN has been widely applied due to its powerful self-learning capability. As a representative architecture of deep learning, CNN can effectively extract spatial features from signals through local connectivity and weight sharing mechanisms. It reduces the complexity of neural network models while significantly decreasing the number of neural network parameters, thereby preventing overfitting. Ahmadzadeh Mdengd et al. [13,14] applied one-dimensional CNN to time-series signal processing, progressively extracting high-level abstract features by stacking multiple convolutional layers. Li R et al. [15] converted one-dimensional time-series signals into two-dimensional time-frequency images through continuous wavelet transform, and then utilized two-dimensional CNN for feature extraction. However, such methods require additional time-frequency transformation steps, increasing computational complexity, and the selection of time-frequency transformation parameters also affects the final prediction performance. To fully leverage the advantages of CNN, some researchers have proposed hybrid architectures. Al Mamun A et al. [16] designed a multi-channel parallel CNN structure that simultaneously processes multi-source information from time domain, frequency domain, and time-frequency domain, achieving integration of complementary information through feature fusion strategies. Huang et al. [17] proposed a multi-scale CNN model based on attention mechanism, utilizing convolutional kernels of different sizes to capture multi-scale features and adaptively adjusting the weights of features at each scale through attention modules.

Recurrent neural networks (RNN) and their improved variant, long short-term memory networks (LSTM), have been introduced into tool wear monitoring due to their advantages in processing sequential data. The LSTM proposed by Hochreiter et al. [18] effectively addresses the gradient vanishing problem of traditional RNN and is well-suited for processing time-series data. To optimize temporal feature extraction, Cui Z et al. [19] constructed a multi-layer stacked LSTM network to model complex patterns under multiple working conditions, while Marani et al. [20] systematically evaluated network depth and width, identifying a two-layer architecture with eight units as the optimal configuration. Furthermore, Wang G et al. [21] proposed an LSTM model based on encoder–decoder structure, transforming the tool wear prediction problem into a sequence-to-sequence mapping task and improving the accuracy of multi-step prediction. Cai et al. [22] constructed a deep learning framework based on stacked Long Short-Term Memory (LSTM) networks. By fusing multi-sensor time-series features with processing parameters, this approach achieves robust tool wear prediction under complex operating conditions. However, standard LSTM employs a unidirectional information propagation mechanism, which can only utilize information from historical moments to model the current state. For progressive degradation processes with obvious trends such as tool wear, information from future moments also contains important clues for state evolution. Xu Z et al. [23] applied Bi-LSTM to tool wear state recognition, and experimental results demonstrated that the bidirectional structure can more accurately capture the bidirectional dependencies of wear evolution. Li X et al. [24] proposed a tool wear trend prediction model based on Bi-LSTM, significantly improving prediction accuracy by fusing forward and backward hidden state information. However, overall, research on the application of Bi-LSTM in tool wear monitoring remains relatively limited, and its potential requires further exploration. The introduction of attention mechanisms provides deep learning models with a dynamic weighted feature selection approach. In tool wear monitoring, global attention can identify time periods that have greater influence on wear states, improving model interpretability. Global attention, as utilized by Sun M et al. [25] and Xue Z et al. [26], effectively identifies critical wear stages but involves all temporal moments, making it computationally expensive and susceptible to noise. To alleviate this, researchers proposed local [27,28] or sparse [29] attention mechanisms to reduce complexity and focus on short-term dynamic changes. Building on this, Li B et al. [30] introduced soft-thresholding to reinforce trend recognition. However, since tool wear signals contain both low-frequency trends and high-frequency mutations, single mechanisms often fail to adequately capture complex features. Consequently, recent studies have developed hybrid architectures: Zhou Y et al. [31] designed a multi-resolution network (MRDAN), while Hou K et al. [32] proposed the Swinsion framework integrating CNN and Transformer. Additionally, spatiotemporal [33] and channel-spatial [34] attention mechanisms have been employed to refine feature selection. Moreover, to address specific machining challenges, Zhang et al. [35] utilized meta-learning for small-sample scenarios, while Zhang et al. [36] incorporated physics constraints for cross-condition adaptation. Nevertheless, existing methods still struggle to adaptively balance global and local features to mitigate interference from complex machining dynamics.

Despite existing advancements, current data-driven methods still face three critical challenges: First, fixed-scale convolutional structures often fail to simultaneously capture instantaneous impact features and long-term degradation trends, limiting the model's receptive field. Second, traditional temporal modeling relies heavily on unidirectional propagation, ignoring the bidirectional dependencies inherent in the continuous wear process. Third, existing attention mechanisms struggle to balance global context with local details, making it difficult to adaptively mitigate interference from complex machining dynamics.

To address the aforementioned issues, this paper proposes a tool wear monitoring method based on the Multi-Scale Convolution and Dual Attention Temporal Feature Network (MS-DATFNet). The main contributions of this method include:

(1) A parallel multi-scale convolution module is constructed to synchronously extract high-frequency impact features and low-frequency degradation trends. This design aims to enhance the signal-to-noise ratio and capture intrinsic degradation patterns robust to environmental noise.
(2) A Bidirectional LSTM (Bi-LSTM) architecture is employed for modeling comprehensive temporal dependencies. By integrating both historical accumulation and future context, this module is designed to precisely track the nonlinear evolution characteristics of the entire wear lifecycle.
(3) A global-local dual-stream attention mechanism is integrated to mitigate interference from complex machining dynamics. Through the synergy of macroscopic critical moment identification and microscopic periodic pattern capture, this objective ensures the model maintains stable prediction accuracy across varying cutting conditions.

2. Methods

This section introduces a multi-source information fusion network, termed MS-DATFNet, designed to address the limitations of fixed-scale feature extraction, unidirectional temporal modeling, and the insufficient adaptability of existing attention mechanisms. The architecture is designed to effectively fuse multi-sensor signals and conduct in-depth mining of wear-related features. Specifically, the model is composed of three collaborative modules, dedicated to multi-scale feature extraction, temporal dependency modeling, and adaptive feature fusion, respectively. First, the multi-scale convolution module is employed to extract discriminative key features from raw force signals and vibration signals, followed by max pooling layers to compress feature dimensions and remove redundant data. Subsequently, a Bidirectional Long Short-Term Memory (Bi-LSTM) module is introduced, leveraging its bidirectional temporal modeling capability to capture both forward and backward temporal dependency patterns in signal sequences. Furthermore, global attention and local attention mechanisms are incorporated, enhancing the representation capability of core temporal features through dual attention constraints. Finally, multiple fully connected layers are constructed at the top of the network to increase output nonlinearity, and a regression layer is utilized to generate the target tool wear values. Figure 1 illustrates the overall flowchart of tool wear prediction.

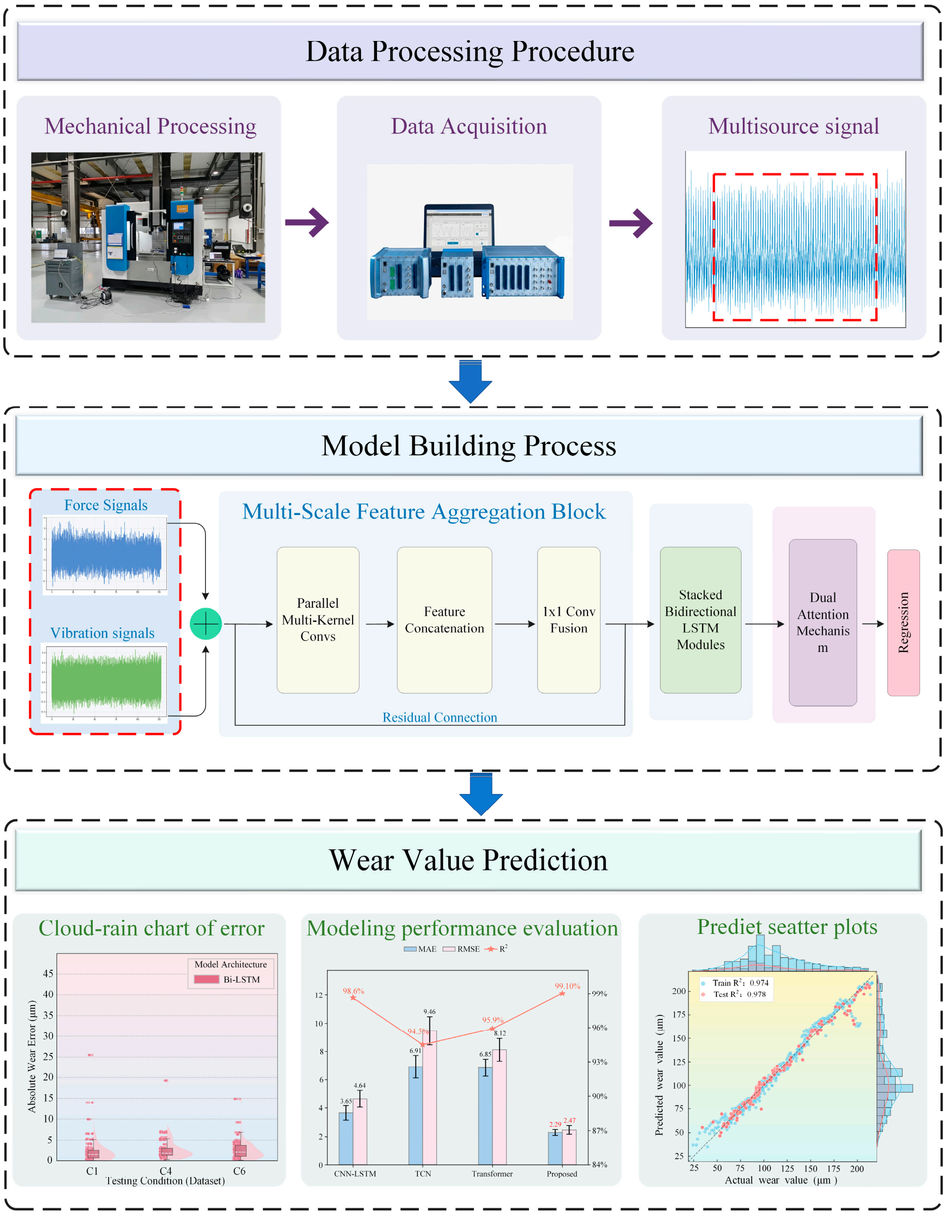

Figure 1. Tool wear prediction based on the MS-DATFNet.

2.1. Parallel Convolution Fusion

Convolutional Neural Networks (CNNs) utilize local receptive fields and parameter sharing to model the spatial correlations of multi-channel sensor signals. Based on the dimensionality of input variables, CNNs can be primarily categorized into 1D-CNNs for processing one-dimensional temporal signals and 2D-CNNs for two-dimensional image analysis. Many existing studies convert raw signals into two-dimensional time-frequency spectrograms through Fourier transform or wavelet transform to meet the input requirements of 2D-CNNs. Although such methods can improve the prediction performance of models, they rely heavily on prior knowledge and involve high computational complexity. In contrast, 1D-CNNs can directly perform end-to-end modeling on raw time-domain signals. Tool wear signals exhibit distinct multi-time-scale characteristics, encompassing both local abrupt features and slowly evolving trends. Convolution kernels with a single scale and fixed receptive field struggle to simultaneously capture these heterogeneous temporal patterns. Therefore, this paper designs a multi-scale convolution module consisting of three parallel one-dimensional convolution branches, with the structure shown in Figure 2. The module incorporates three distinct kernel sizes: 1×3, 1×5, and 1×7. This multi-resolution structure is designed to decouple the vibration signals into different frequency domains for comprehensive feature extraction. The smaller kernel specializes in capturing high-frequency local variations. In contrast, the larger kernel focuses on low-frequency global patterns, effectively characterizing the gradual degradation trend of the tool. The intermediate kernel serves to bridge these scales, ensuring semantic continuity across the feature maps. By limiting the receptive fields to this specific range, the model achieves a balance between detailed local sensitivity and global trend awareness. This selection provides sufficient coverage for tool wear dynamics, whilst effectively avoiding the computational redundancy and potential overfitting risks associated with excessively large kernels. The feature extraction process of the three branches along the temporal dimension can be expressed as:

$$F^{(s)} = \sigma(BN(Conv1d^{(s)}(X))),\ k_s = 3 \tag{1}$$

$$F^{(m)} = \sigma(BN(Conv1d^{(m)}(X))),\ k_m = 5 \tag{2}$$

$$F^{(l)} = \sigma(BN(Conv1d^{(l)}(X))),\ k_l = 7 \tag{3}$$

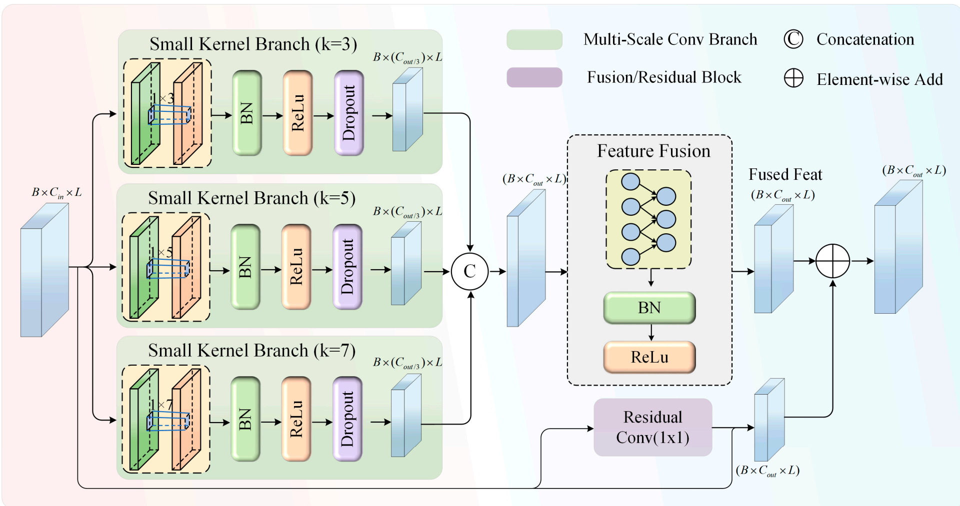

Figure 2. Parallel convolution module structure.

In the formula, $Conv1d^{(\cdot)}$ denotes one-dimensional convolution, where different convolution kernel sizes correspond to distinct time windows. $\sigma(\cdot)$ represents the ReLU activation function, and BN stands for batch normalization. To ensure the stability of feature extraction, all three branches follow the standardized processing pipeline: "convolution-batch normalization-activation-dropout". The temporal features of the three branches are concatenated along the channel dimension, and 1×1 convolution is further employed for feature integration to reduce computational complexity.

$$F_{concat} = Concat([F^{(s)}, F^{(m)}, F^{(l)}], dim = 1) \tag{4}$$

$$F_{fused} = \sigma(BN(Conv1d_{1\times1}(F_{concat}))) \tag{5}$$

Finally, the fused multi-scale temporal features are added to the input signal via residual connection:

$$Hbilstm = [h1 \oplus h1, h2 \oplus h2, \ldots, hT' \oplus hT'] \tag{6}$$

This design serves two purposes: it preserves the integrity of the original temporal signal while mitigating the gradient vanishing problem common in deep networks. Consequently, the model can simultaneously capture multi-scale feature representations in a single forward pass, thereby providing rich temporal features for subsequent modeling.

2.2. Temporal Encoding Module

Due to the strong temporal dependencies inherent in sensor signals, tool wear assessment requires consideration of information from each time step. To capture the temporal dependencies of signal features, given the feature sequence after multi-scale convolution processing and dimension transposition $H \in \mathbb{R}^{B \times T' \times D}$, where B denotes the batch size $T' = T/4$, represents the temporal length after pooling, and D is the feature dimension. The feature vectors extracted from the 1D-CNN, $X_i = [X_1, X_2, X_t, \ldots, X_N]$ are fed into a Bidirectional Long Short-Term Memory network (BiLSTM) to achieve deep temporal modeling. The stacked BiLSTM processes the input sequence simultaneously through forward and backward directions using two independent LSTM structures. The forward LSTM considers the influence of historical information on the current hidden state, thereby capturing the trends of tool wear. The sequence is processed chronologically from time step $t = 1$ to $t = T'$:

$$\overrightarrow{h}_t = LSTM_{forward}(h_t, \overrightarrow{h}_{t-1}) \tag{7}$$

$$\overrightarrow{h}_t = \sigma_h(W^f_{hh}\overrightarrow{h}_{t-1} + W^f_{ih}h_t + b^f_h) \tag{8}$$

The backward LSTM primarily utilizes future information to update the current hidden state. Simultaneously considering both historical and future information further enhances the capability of the LSTM unit to perceive dynamic variations in feature sequences. The sequence is processed in reverse chronological order from time step $t = T'$ to $t = 1$:

$$\overleftarrow{h}_t = LSTM_{backward}(h_t, \overleftarrow{h}_{t+1}) \tag{9}$$

$$\overleftarrow{h}_t = \sigma_h(W^b_{hh}\overleftarrow{h}_{t+1} + W^b_{ih}h_t + b^b_h) \tag{10}$$

where $W^{f/b}_{hh}$ denotes the hidden state weight matrix of the forward LSTM, $W^{f/b}_{ih}$ represents the input weight matrix of the backward LSTM, $W^{f/b}_{h}$ denotes the bias vector, and σ_h represents the tanh activation function.

$$H_{bilstm} = [\overrightarrow{h}_1 \oplus \overleftarrow{h}_1, \overrightarrow{h}_2 \oplus \overleftarrow{h}_2, \ldots, \overrightarrow{h}_{T'} \oplus \overleftarrow{h}_{T'}] \tag{11}$$

where $\oplus$, denotes vector concatenation. While deeper LSTM layers help address non-linear features in temporal prediction, they also increase computational resource consumption and weaken the model's generalization ability. To balance network performance and overfitting, this paper adopts a two-layer bidirectional LSTM structure and introduces a dropout mechanism to prevent overfitting.

$$H^{(l)}_{bilstm} = Dropout\left(BiLSTM^{(l)}\left(H^{(l-1)}_{bilstm}\right)\right), l \in \{1, 2\} \tag{12}$$

The output bidirectional LSTM feature $H_{bilstm} \in \mathbb{R}^{B \times T' \times 2D_{hidden}}$ contains both forward and backward contextual information at each time step, where D_{hidden} denotes the number of LSTM hidden units (Figure 3).

It is important to note that the bidirectional mechanism utilizes information from the entire input window, which necessitates a buffered processing strategy for online monitoring. In practice, the prediction for the current timestamp is generated once the sliding window is fully populated. Given the high sampling frequencies employed in our experiments, this buffer introduces a latency only on the scale of tens of milliseconds. Since tool wear is a relatively slow degradation process, this micro-latency is negligible for process control, allowing the system to benefit from the enhanced accuracy of bidirectional modeling without compromising operational effectiveness.

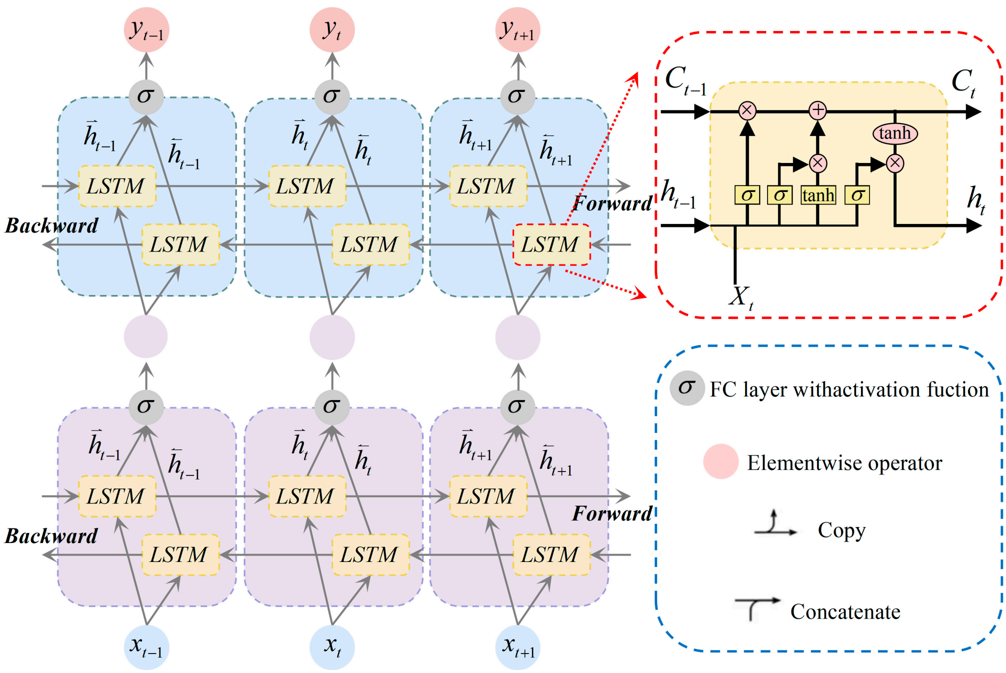

Figure 3. Schematic diagram of the overall architecture and internal cell structure of the stacked bidirectional LSTM network.

2.3. *A Dual-Attention Mechanism Integrating Sliding Windows with Global Context*

During the tool wear process, long-sequence data contains not only long-term wear patterns but also local abnormal features that indicate critical states. Effectively capturing and fusing these multi-scale temporal features is a key challenge. To overcome the limitations of traditional single attention mechanisms, this paper proposes a dual attention mechanism, which employs a simplified global attention and local adaptive attention to achieve effective fusion of multi-scale temporal features. The hidden sequence states output by the bidirectional LSTM are denoted as $H \in \mathbb{R}^{B\times T\times d}$, and the formal definition of the dual attention mechanism is as follows: the attention energy value of each time step is calculated via non-linear transformation,

$$e_t = W_2 \cdot tanh(W_1 \cdot h_t + b_1) + b_2 \tag{13}$$

where $W_1 \in \mathbb{R}^{d\times d'}$ and $W_2 \in \mathbb{R}^{1\times d'}$ are learnable weight matrices, b_1 and b_2 are bias terms, d' denotes the dimension of the intermediate layer. A two-layer neural network is used to map the hidden states to scalar energy values, where the tanh activation function introduces non-linearity, enabling the model to learn complex attention patterns. Then, the energy values are normalized using softmax to obtain the attention weights for each time step:

$$\alpha_t = \frac{expe_t)}{\sum\limits_{i=1}^{T} exp(e_i)} \tag{14}$$

This ensures that the sum of weights across all time steps equals 1 ($\sum\limits_t \alpha_t = 1,\ \alpha_t \in [0,1]$), and that each weight value lies within the interval [0, 1]. The attention weight α_i reflects the importance of time step i to the final prediction. Based on these attention weights, the global context vector is calculated as:

$$c_{\text{global}} = \sum_{t=1}^{T} \alpha_t \cdot h_t \tag{15}$$

By aggregating information across all time steps via weighted summation, c_{global} encapsulates the global semantic information of the entire sequence—time steps with larger weights contribute more to the global representation.

To capture local detailed features, the sequence is divided into k local windows with a window size of w. For the k-th window, the window feature is first constructed through a concatenation operation:

$$W_k = H[:, k \cdot s : k \cdot s + w, :] \tag{16}$$

where $H \in \mathbb{R}^{B \times T \times d_h}$ is the hidden state matrix, s denotes the stride, k represents the feature slice of the window extracted from the complete sequence, and w is the window size—this ensures that the local attention only focuses on adjacent time steps.

The local attention energy is calculated for each local window:

$$e_{k,t}^{\text{local}} = \text{MLP}_{\text{local}}(w_{k,t}) \tag{17}$$

where $\text{MLP}_{\text{local}}$ is a lightweight multi-layer perceptron, and $w_{k,t}$ denotes the hidden state at time step t within window k.

This network learns the local dependency relationships inside the window; its parameter count is much smaller than that of the global attention, which reduces computational complexity.

$$\beta_k = softmax(MLP_{global}(c_k^{local}))c_{local} = \sum_{k=1}^{K} \beta_k \cdot c_k^{local} \tag{18}$$

where c_k^{local} denotes the local context vector of the k-th window, and β_k is the cross-window attention weight. This two-level attention structure first aggregates information within the window to obtain c_k^{local}, then performs selective fusion across windows via β_k, enabling the model to focus on local regions containing key wear features. The window size w is adaptively adjusted based on the sequence length T, typically set as $w = [T/K]$, where K is the preset number of windows.

Finally, the integration of dual attention features is achieved through a fusion network:

$$c_{fused} = MLP_{fusion}\left(\left[c_{global}; c_{local}\right]\right) \tag{19}$$

Here, $[;]$ represents the feature concatenation operation, and MLP_{fusion} is the fusion network, which learns the optimal combination weights of global and local features. By non-linear transformation, this network adaptively balances the contributions of the two attention mechanisms, allowing the fused features to retain both the global wear trend and local abnormal information.

For global attention, the attention weights are first computed, followed by normalization operations to obtain global information. For adaptive local attention, the sequence is partitioned according to window size, local attention is computed within each window, and then fused through cross-window attention. The window size w is adaptively adjusted according to the sequence length T. Finally, the integration of dual attention features is achieved through the fusion network.

As shown in Figure 4, the input features are simultaneously fed into two parallel modules: the global attention mechanism and the adaptive local attention mechanism. The resulting global and local attention features are fused through concatenation operations and further optimized via a multi-layer perceptron. Finally, a fusion network is employed to learn the optimal combination weights of the two attention mechanisms. By explicitly modeling the global-local dependencies, the Dual Attention Mechanism (DAM) allows the model to leverage complementary multi-scale wear-related information while maintaining computational efficiency. This facilitates the mining of temporal evolution patterns that are essential for high-precision tool wear prediction. The fused attention features derived from this method encapsulate the deep semantics of tool wear information across multiple time scales, which not only preserve the global information reflecting the long-term degradation trend but also capture the local detailed features indicating abrupt state changes.

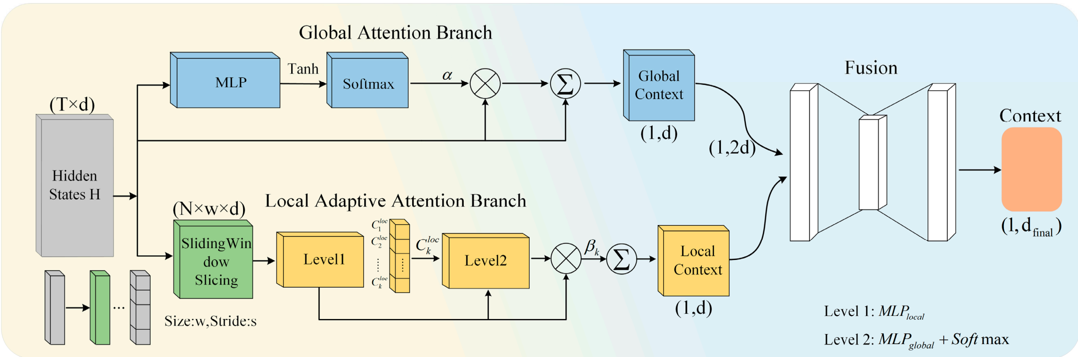

Figure 4. Architecture of the dual-branch global-local attention mechanism.

3. Experiments

To comprehensively evaluate the effectiveness and generalization ability of the proposed model, dual verification was performed using two datasets: one provided by the Prognostics and Health Management Society (PHM) based in New York, NY, USA [37], and the other independently constructed from TiAl alloy milling experiments. This section elaborates on the experimental data acquisition setup, machining parameters, and monitoring system configuration.

3.1. Publicly Available Dataset

This study utilizes the PHM2010 public dataset for model validation. The experimental setup is shown in Figure 5. The experiments were conducted on a Roders Tech RFM 760 high-speed CNC milling machine (Röders GmbH, Soltau, Germany). A three-flute cemented carbide ball-end milling cutter with a diameter of 6 mm was employed, and dry side milling was adopted to machine Inconel 718 alloy. The relevant experimental parameters are as follows: spindle speed: 10,400 r/min, feed rate: 1555 mm/min, axial depth of cut: 0.2 mm, radial depth of cut: 0.125 mm, and feed per tooth: 0.001 mm. The sampling frequency was 50 kHz. The data acquisition equipment included a Kistler 9265 B three-dimensional dynamometer (Kistler Group, Winterthur, Switzerland), a Kistler 5019 A multi-channel charge amplifier (Kistler Group, Winterthur, Switzerland), an NI DAQ data acquisition card, and a Leica MZ 12 metallographic microscope (Leica Microsystems, Wetzlar, Germany). In this experiment, three-axis cutting force data were acquired using a Kistler 9265B three-dimensional dynamometer, while vibration signals were captured via a Kistler 5019 A multi-channel charge amplifier; acoustic emission signals were collected with an acoustic emission sensor. After gathering signals from seven channels (encompassing force and vibration data for the x, y, and z axes, plus acoustic emission data), the signals were conditioned using the Kistler multi-channel charge amplifier—this step ensured the signal-to-noise ratio and accuracy of data acquisition via the NI PCI 1200 system. To obtain ground-truth labels for tool wear, the flank wear width (VB) was measured offline using a LEICA MZ 12 microscope. Six full-life-cycle tests were conducted using up-milling of the material. The duration of each pass was equal, and the wear on the flank face of the cutting tool was measured after each pass. Machining was completed after 315 cuts. Among the six datasets obtained, the C1, C4, and C6 datasets were used to measure the wear of the milling cutter. Since wear assessment based solely on a single cutting edge is prone to local bias, the average wear value of the three cutting edges across 315 sample sets was designated as the reference standard for characterizing the tool wear state.

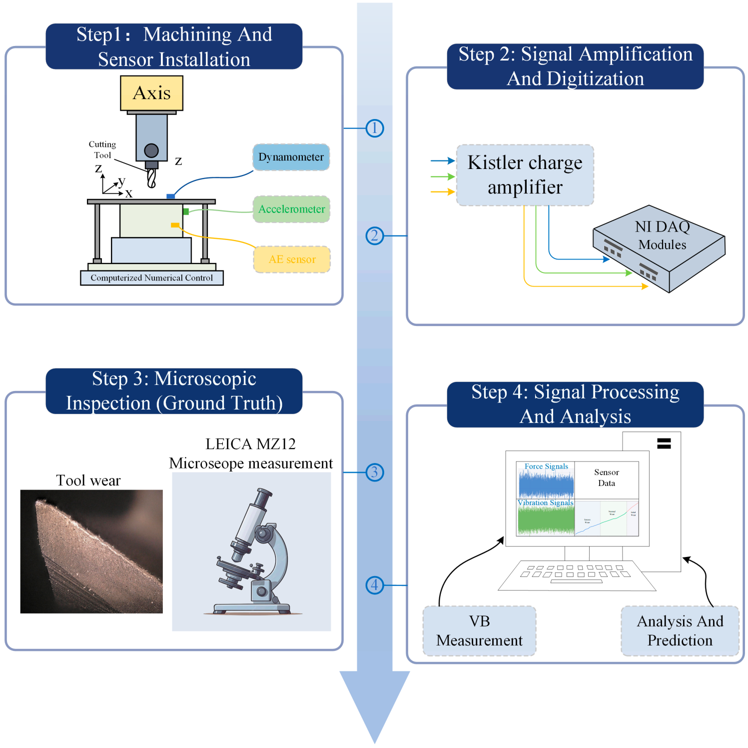

Figure 5. Experimental setup of PHM2010.

3.2. *Self-Built TiAl Alloy Milling Dataset*

To further verify the reliability and generalizability of the model, this study conducted a complete tool wear cycle milling experiment through TiAl milling tests. The primary objective of this experiment is to evaluate the prediction accuracy and adaptability of the proposed method under distinct machining scenarios. The machining equipment employed herein was a CAMUMC 850 ultrasonic-assisted three-axis milling machine. (The workpiece material was a cylindrical TiAl alloy, with dimensions specified as φ74.8 mm × 199 mm. A four-flute cemented carbide end mill was used as the cutting tool. During the experiment, force signals were recorded by a Kistler dynamometer, and vibration signals were monitored and recorded by an IGTech A26 F100 T01 C system. The experimental parameters are shown in Table 1.

Table 1. Milling Parameters of TiAl Alloy.

Experimental Parameter	Value
Spindle Speed	3000 rpm
Feed per Revolution	0.2 mm/r
Depth of Cut	0.1 mm
Width of Cut	0.15 mm
Sampling Frequency	20 khz

Upon completion of each tool pass, the flank wear width of the four cutting edges was measured using a super-depth-of-field microscope. The average of these four measurements was adopted as the tool wear value for that pass. Figure 6 depicts the specific experimental setup for the TiAl alloy milling. A total of 56 cutting tests were conducted throughout the complete tool life cycle. The resulting wear progression curve is shown in Figure 7. Based on the wear curve, tool wear was categorized into three distinct stages—initial wear, steady-state wear, and accelerated wear—using the K-means clustering algorithm. The dataset was subsequently divided in an 8:2 ratio for model training and validation, respectively.

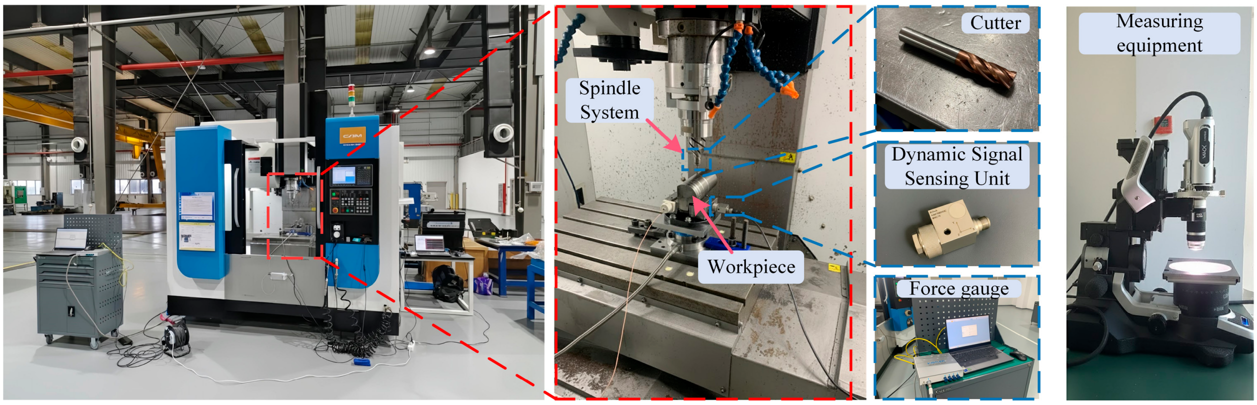

Figure 6. Experimental Setup of TiAl Alloy Milling.

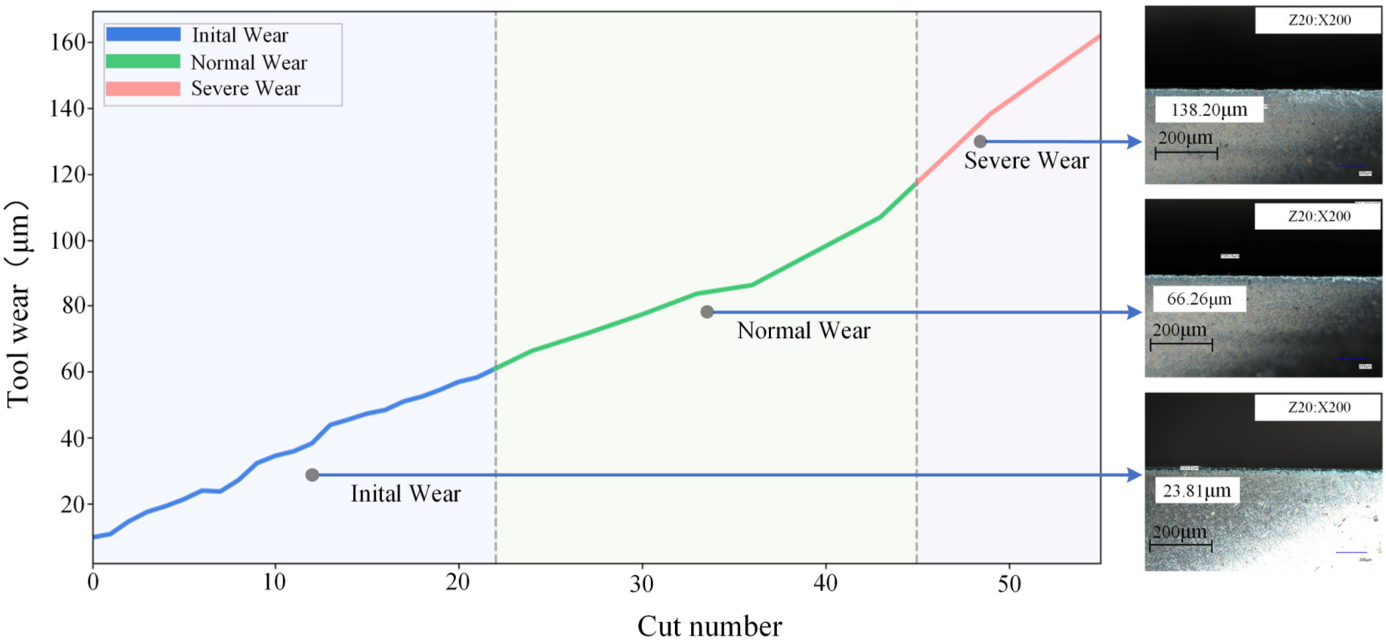

Figure 7. Tool Wear Curve for Tool Life in TiAl Alloy Milling Experiment.

3.3. Training and Validation of the Model

The model was trained on the Windows 11 operating system, with an NVIDIA RTX 4090 graphics card utilized for acceleration. The training environment was built on PyTorch 2.0 paired with CUDA 11.8. During the training process, the AdamW optimizer was adopted to perform iterative optimization of model parameters, while the cosine annealing with warm restarts strategy was integrated to dynamically adjust the learning rate. The detailed hyperparameter configuration of the network is presented in Table 2.

Table 2. Hyperparameter Configuration for Model Training.

Hyperparameter	Setting Value
Total Training Epochs	160
Batch Size	18
Initial Learning Rate	0.0002
Weight Decay Coefficient	0.0005
Dropout Rate	0.45

3.4. Performance Evaluation

The mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2) were selected as the evaluation metrics, and their definitions are given in Equations. Herein, $\hat{y}_i$ and y_i represent the predicted value and actual value, respectively, $\overline{y}$ denotes the mean value of the actual values and N is the total number of samples. These commonly used metrics were employed to evaluate the performance of the proposed method.

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left|\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}}\right| \tag{20}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}}\right)^2} \tag{21}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}}\right)^2}{\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{y}_{\mathrm{i}} - \overline{\mathrm{y}}\right)^2} \tag{22}$$

4. Ablation Experiment of Key Modules

To verify the effectiveness of each core component of the model proposed in this paper, ablation experiments were designed for validation. Three core modules were mainly analyzed as follows: (1) Multi-scale convolution, which was used to evaluate the contribution of different convolution kernels to feature extraction; (2) Bidirectional LSTM (Bi-LSTM), which was adopted to verify the performance improvement of bidirectional temporal modeling compared with unidirectional temporal modeling; (3) Dual attention mechanism, which was utilized to assess the necessity of fusing local attention and global attention. To ensure the rigor and reliability of the experiments, the principle of a single variable was followed in each group of ablation experiments. Except for the module to be verified, the remaining network structure, computer hardware environment, and training configurations were kept unchanged.

4.1. Ablation Experiment of the Parallel Convolution Fusion Module

To verify the effectiveness of the parallel convolution fusion module proposed in this paper for capturing different frequency-domain features during the tool wear process, multiple groups of comparative experiments were designed in this section. Specifically, the experiments included three single-scale convolutional layer control groups (Baseline A: Kernel = 3, Baseline B: Kernel = 5, Baseline C: Kernel = 7), two dual-scale combination control groups (Baseline D: Kernel = 3, 5 and Baseline F: Kernel = 5, 7), and the multi-scale fusion method proposed in this paper (Proposed). The experimental results are presented in Table 3 and Figure 8.

Table 3. Performance Comparison Under Different Convolution Kernel Configurations.

Experiment ID	Kernel Configuration	MAE	RMSE	Overall R^2	C1 R^2	C4 R^2	C6 R^2
Baseline A	Kernel = 3	6.309	8.676	0.952	0.941	0.929	0.970
Baseline B	Kernel = 5	4.767	7.321	0.966	0.979	0.943	0.975
Baseline C	Kernel = 7	5.765	8.191	0.957	0.966	0.925	0.976
Baseline D	Kernel = 3, 5	5.257	8.294	0.957	0.925	0.981	0.955
Baseline F	Kernel = 5, 7	4.702	7.005	0.969	0.979	0.942	0.983
Proposed	Multi-Scale	3.657	5.840	0.978	0.978	0.968	0.983

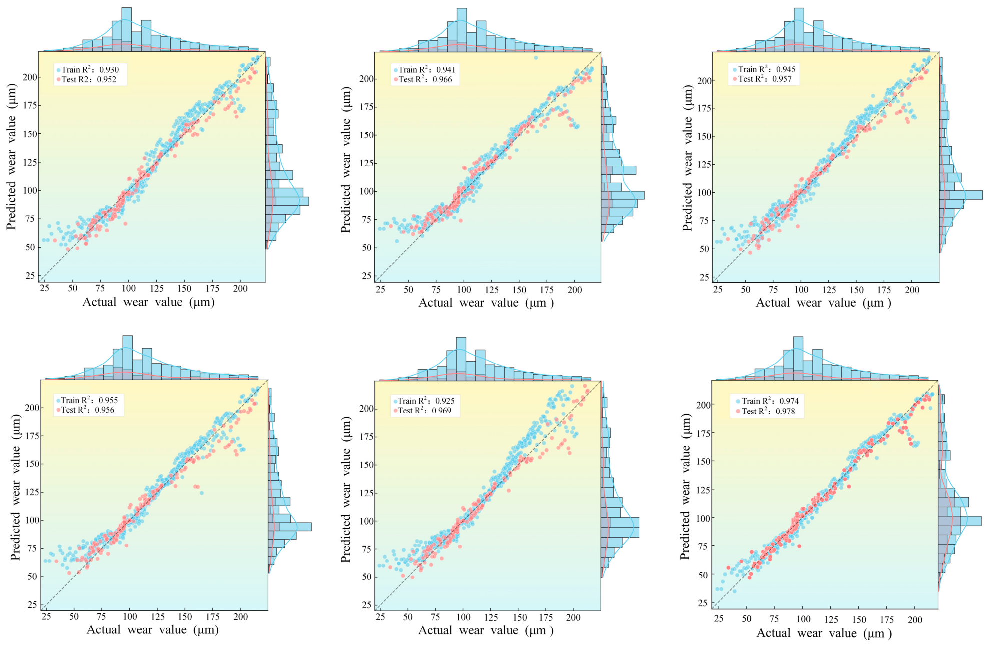

Figure 8. Wear Prediction Scatter Plots of Training and Test Sets for Different Models.

From the data in Table 3, it can be analyzed that the prediction accuracy of single-scale models (Baseline A, B, C) is significantly affected by the convolution kernel size, and they cannot simultaneously achieve an optimal balance between local sensitivity and global stability. For example, although Baseline B (Kernel = 5) performed best among the single-scale group, its MAE remained at 4.767. In contrast, the multi-scale convolution fusion method proposed in this paper effectively overcomes the limitations of a single receptive field. By extracting features at different scales through parallel convolution and then fusing them, all evaluation metrics were improved. The MSE and RMSE were reduced to 3.657 and 5.840, respectively, and the overall R^2 reached 0.978, demonstrating the best performance among all control experiments.

The regression scatter plot in Figure 8 offers a more intuitive visualization of the aforementioned analysis. Specifically, the first row of subplots (from left to right) depicts Baseline A, Baseline B, and Baseline C, while the second row presents Baseline D, Baseline F, and the Proposed model (MS-DATFNet).In this plot, blue and red dots illustrate the correlation between predicted values and ground-truth values, while the histograms on both sides separately delineate the density distribution of the corresponding datasets. The best-performing single-scale model B exhibits significant instability in the high-wear stage, with a recorded maximum absolute error of 16.2 μm. This large local deviation indicates that single-scale features fail to capture rapid degradation signals. In contrast, the proposed MS-DATFNet demonstrates superior robustness under severe wear conditions. In the same high-wear stage, the maximum absolute error is drastically reduced to 7.5 μm. Such a significant error reduction highlights that the multi-scale fusion strategy successfully overcomes the limitations of single-scale models in handling extreme wear patterns. However, it is worth noting that a slight overestimation trend is observed in the initial wear stage (approx. 0–50 μm). This is primarily attributed to the unstable tribological behavior during the "run-in" period, where weak micro-wear signals are susceptible to background noise. Future improvements could address this via stage-specific denoising or weighted feature learning. Despite this localized deviation, the overall predictive capability remains robust. This performance improvement is mainly attributed to the feature complementarity of the multi-scale convolution module: small-scale convolution kernels sensitively capture local features reflecting abrupt wear changes, while large-scale kernels effectively characterize the overall degradation trend. The combination of both significantly enhances the model's capability for feature extraction and representation throughout the entire tool life cycle. In addition to this multi-scale fusion strategy, the internal residual connections are crucial for ensuring effective model training. Comparative analysis with a non-residual variant revealed that removing these connections resulted in optimization difficulties, requiring approximately 20 additional epochs to reach convergence and increasing the MAE from 3.657 to 3.924. This confirms that residual learning is indispensable for maintaining both gradient stability and high prediction accuracy in the deep architecture.

4.2. Ablation Experiment of the Temporal Encoding Module

Considering the deficiency of convolutional features in temporal modeling, a temporal encoding layer was introduced to improve prediction accuracy. To evaluate the effectiveness of the proposed module and determine its optimal structure, a stepwise ablation experiment involving four distinct configurations (as presented in Table 4) was designed and conducted in this section. The configurations are detailed as follows: Variant 1 (w/o LSTM): the recurrent neural network (RNN) layer was removed; Variant 2 (Uni-LSTM): a unidirectional LSTM structure was adopted; Variant 3 (GRU): the LSTM units were replaced with gated recurrent units (GRU); Proposed (Bi-LSTM): a bidirectional LSTM structure was implemented. The experimental results are presented in Table 4 and Figure 9.

Table 4. Prediction Accuracy Evaluation of Different Ablation Variants.

(Ablation Settings)	MAE	RMSE	Overall R^2	C1 R^2	C4 R^2	C6 R^2
Variant1 (w/o LSTM)	7.092	15.762	0.843	0.316	0.921	0.973
Variant2 (Uni-LSTM)	6.634	11.455	0.917	0.702	0.933	0.981
Variant3 (Bi-GRU)	4.950	8.120	0.955	0.910	0.952	0.982
Proposed (Bi-LSTM)	3.657	5.840	0.978	0.978	0.968	0.983

It can be clearly observed from the table that after removing the LSTM structure, the model exhibited the worst performance, with its RMSE value reaching as high as 15.762. Moreover, the overall R^2 on the C1 dataset dropped to as low as 0.316, indicating that relying solely on convolutional features cannot effectively handle the complex nonlinear wear process, resulting in near-complete model failure. With the introduction of temporal structures, the model's prediction performance improved in a stepwise manner. However, the Uni-LSTM and Bi-GRU structures demonstrated limited fitting capability on the C1 and C4 datasets. In contrast, the Bi-LSTM proposed in this paper achieved the best prediction performance, with its RMSE value reduced by as much as 62.9% compared to the w/o LSTM structure, and the R^2 improved to 0.978. This result fully demonstrates the decisive role of the bidirectional LSTM structure in capturing global temporal context and enhancing model generalization capability.

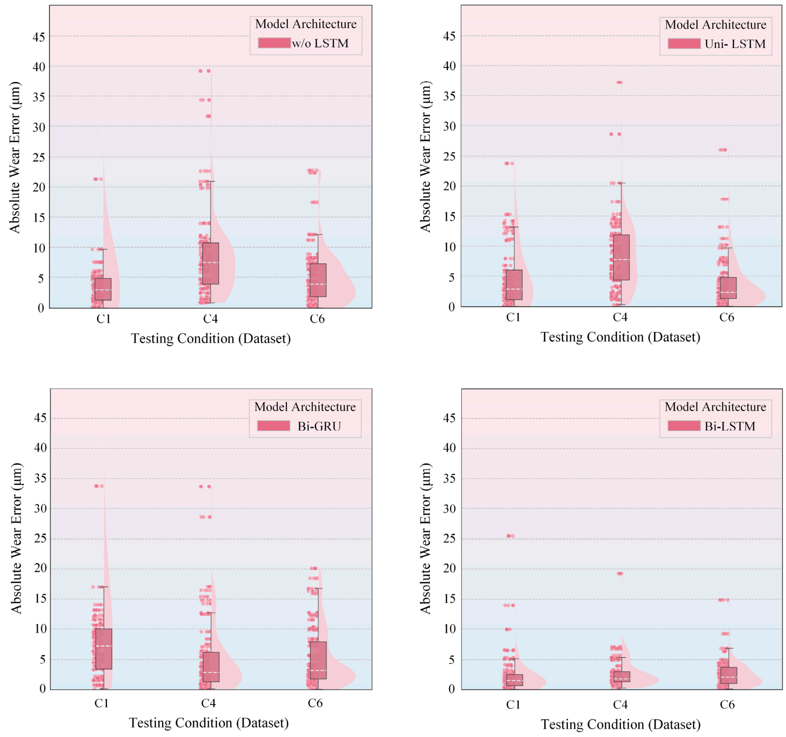

Figure 9. Rain Cloud Plot of Absolute Wear Error Distribution for Different Ablation Variants.

As shown in Figure 9, the raincloud plots containing probability density curves, box plots, and scatter plots comprehensively illustrate the absolute wear errors of different ablation variants. The extremely wide error distribution and low-density peak of Variant 1 intuitively indicate the model's failure in the temporal dimension. Although Variant 2 introduces temporal memory, long-tail extensions still persist. This is because the model can only utilize historical degradation trajectories, and when abrupt changes in wear rate occur, the lack of future information for reference prevents timely deviation calibration. Variant3 narrows the interquartile range through the bidirectional mechanism, but it is less effective than the Proposed model in filtering complex wear noise. The model proposed in this paper simultaneously integrates historical trajectories and future evolution information of tool wear, effectively capturing bidirectional dependencies in wear signals. This global perspective endows the model with exceptional robustness, enabling it to almost completely eliminate the long-tail phenomenon in both the C1 and C4 datasets, thereby maintaining prediction accuracy even in dynamic environments.

4.3. Ablation Experiment of Sliding Window Global Attention

Relying solely on the final time step of Bi-LSTM tends to cause the loss of degradation information in long sequences. In addition, tool wear involves both a gentle global trend and intense local fluctuations, which cannot be simultaneously captured by a single attention mechanism. To verify the effectiveness of the attention mechanism and the necessity of the dual-stream adaptive fusion mechanism proposed in this paper, four groups of comparative experiments were designed in this section, namely M0 (without attention mechanism), M1 (single global attention mechanism), M2 (single local attention mechanism), and M3 (dual attention mechanism).

As can be seen from Figure 10, although the M1 model incorporating the global attention mechanism can capture the overall degradation trend of the entire sequence and reduce the MAE value to 4.398, the RMSE value and overall R^2 did not improve but instead decreased. This phenomenon indicates that single global weighting causes over-smoothing, resulting in weak model sensitivity to high-frequency transient features that characterize abrupt wear changes. The M2 model with a single local attention mechanism significantly improved prediction accuracy through fine-scale temporal window partitioning, validating the primary role of local features in wear detection. However, on the C1 dataset, its R^2 was lower than that of the M0 model without attention mechanisms, indicating that local models lacking global perspective are more susceptible to noise interference and fail to adequately preserve critical long-term dependency information. The dual-stream adaptive attention model M3 proposed in this paper combines global trends and local details, achieving optimal performance in both MAE and RMSE values. On the C1 dataset, the R^2 was significantly improved to 0.978. This demonstrates that the M3 structure can adaptively adjust attention weights according to signal states, simultaneously capturing global trends and local abrupt changes, thereby achieving high-precision wear prediction.

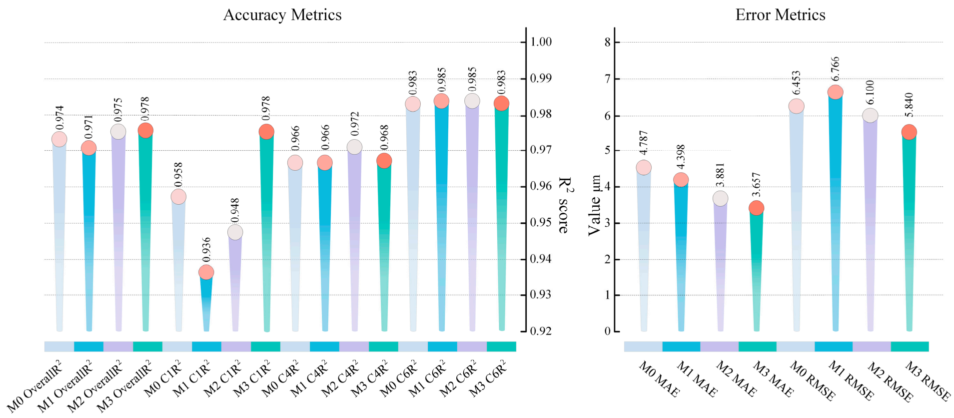

Figure 10. Visualization of Ablation Experiment Results for the Effectiveness of the Dual Attention Mechanism.

4.4. Performance Comparison and Evaluation of Tool Wear Prediction Models on the TiAl Alloy Milling Dataset

To comprehensively evaluate the performance of the CNN-LSTM-DualAttention model proposed in this paper for tool wear prediction tasks, three representative and academically influential temporal modeling benchmark models were selected for comparison. (1) CNN-LSTM: This architecture integrates the local feature extraction capability of convolutional neural networks with the temporal modeling capability of long short-term memory networks, serving as a classic benchmark architecture in the field of industrial equipment condition monitoring [38]. (2) TCN (Temporal Convolutional Network): This model employs dilated causal convolution structures to efficiently capture long-range temporal dependencies and has demonstrated excellent performance in industrial time-series prediction tasks, making it one of the important reference standards in this field [39]. (3) Transformer: Based on the global self-attention mechanism to directly model dependency relationships between arbitrary positions in sequences, it represents the cutting-edge paradigm in current sequence modeling [40]. For a rigorous comparative analysis, the hyperparameters of all baseline models (CNN-LSTM, TCN, and Transformer) were systematically optimized via grid search on the validation set. Specifically, the fitting ranges for the initial learning rate were evaluated across $\{10^{-3}, 5 \times 10^{-4}, 2 \times 10^{-4}, 10^{-4}\}$, and the final value was determined to be 0.0002. The batch size was optimized within the range of {16, 32, 64}, with 16 selected as the optimal size. To ensure convergence and prevent overfitting, we employed an early stopping mechanism with a patience of 20 epochs, allowing for a maximum of 100 training epochs. Consistent with the dataset setup, an 8:2 training-to-validation split was maintained. All experiments were implemented on an NVIDIA RTX 4090 GPU. The proposed model demonstrates high computational efficiency, requiring only about 15 min to complete 100 training epochs. The detailed hyperparameter configurations are summarized in Table 5. While the core architectural configurations from the original literature were maintained, specific parameters such as kernel size, network depth, and attention heads were fine-tuned to adapt to the characteristics of the current datasets. The prediction curves and quantitative metrics of each comparative model are shown in Figures 11 and 12.

Table 5. Summary of optimal hyperparameter settings.

Parameter	Value
Total Training Epochs	100
Initial Learning Rate	0.0002
Dropout Rate	0.45
Batch Size	16

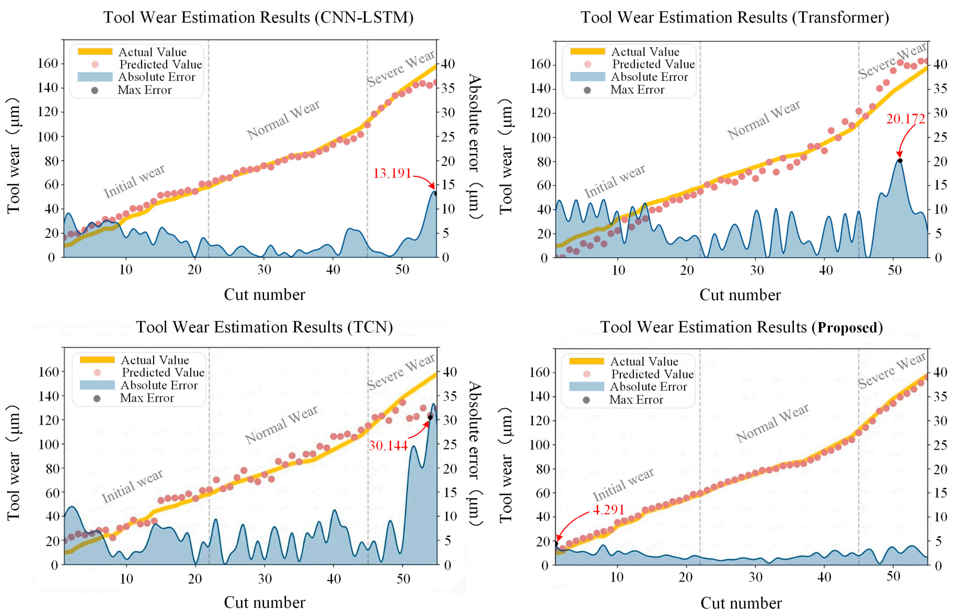

Figure 11. Comparison of tool wear prediction performance on the TiAl dataset.

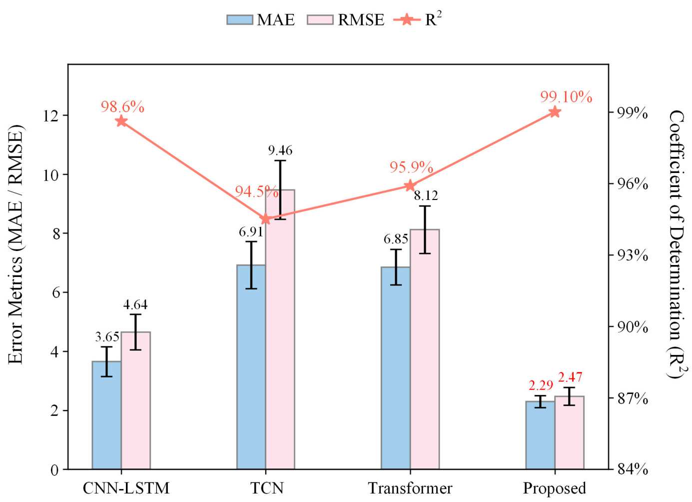

Figure 12. Statistical analysis of model prediction performance on the TiAl milling dataset.

In Figure 11, the yellow curve represents the actual wear values, the pink scatter points represent the predicted values, and the blue shaded area reflects the absolute error point by point. From the four groups of comparative experiments, it can be clearly observed that the predicted wear curve highly overlaps with the actual wear trajectory, and the absolute error band consistently remains at a very low level. The maximum errors of the first three models are all concentrated in the later stages of wear, indicating that these models struggle to effectively capture the nonlinear evolution characteristics during this wear phase. In contrast, the proposed model maintains stable prediction accuracy throughout the entire wear cycle. Figure 12 compares the MAE, RMSE, and R^2 metrics of each model in bar chart form. The experimental results demonstrate that the MAE (2.29 μm) and RMSE (2.47 μm) of the proposed model are significantly lower than those of CNN-LSTM (3.65 μm, 4.64 μm), Transformer (6.85 μm, 8.12 μm), and TCN (6.91 μm, 9.46 μm), with R^2 reaching 99.10%. This superior performance is primarily attributed to the high signal-to-noise ratio inherent in the controlled laboratory data collection environment, which facilitates clearer feature extraction. Furthermore, the application of a 0.45 dropout rate effectively regularizes the network structure. The consistent high accuracy observed on the independent test set demonstrates that the model has successfully learned intrinsic degradation patterns, rather than merely memorizing noise. The performance advantage of this model stems from the targeted design of each module: the parallel convolution fusion module synchronously extracts high-frequency impact and low-frequency trend features, effectively reducing the average prediction error. The bidirectional LSTM jointly models forward and backward temporal dependencies, enhancing the model's overall fitting capability for wear evolution patterns. Finally, the global-local dual-stream attention mechanism effectively mitigates interference from complex machining dynamics through the synergy of macroscopic critical moment identification and microscopic periodic pattern capture, ensuring that the model maintains stable prediction accuracy across all wear stages.

In addition to the accuracy of the prediction, computational efficiency is also crucial for practical applications. Quantitative analysis based on NVIDIA RTX 4090 confirms the efficiency of the proposed model. For the PHM2010 data set, the physical acquisition interval corresponding to the combination of 50 kHz sampling rate and 4096-point window is 81.92 ms. In this case, the average inference time is only 5.5 milliseconds, accounting for only 6.7% of the available time. For the 20 kHz self-built data set, the acquisition interval is extended to 204.8 milliseconds, further reducing the computational occupancy rate to less than 2.7%. These indicators confirm that the model processing speed can ensure a strict real-time response in the prediction task.

5. Conclusions

To address the issues of insufficient multi-scale temporal feature extraction, limited directionality in temporal modeling, and inadequate adaptability of attention mechanisms, this paper innovatively proposes an online tool wear prediction method fusing multi-scale convolution and dual-attention temporal features (MS-DATFNet). Through validation on both public and self-constructed datasets, the following conclusions are drawn:

(1) The parallel convolution fusion module overcomes the limitation of fixed receptive fields of a single convolution kernel by employing parallel kernels of different sizes. It significantly enhances the model's expression of different frequency-domain features during tool wear, providing a rich multi-grained feature foundation for subsequent temporal modeling.

(2) The bidirectional LSTM temporal encoding module fully integrates historical and future information. Through bidirectional temporal modeling, it strengthens the model's ability to perceive wear trends and respond dynamically. Ablation experi-

ment results show that this module outperforms models without temporal structures or with unidirectional temporal structures in terms of prediction accuracy and generalization capability.

(3) The dual attention mechanism combines global attention and adaptive local attention, effectively resolving the noise sensitivity and fixed window issues of single attention mechanisms in long-sequence processing. It enables adaptive weighting and enhancement of multi-scale features.

The proposed model achieves an R^2 of 0.978, an MAE of 3.657, and an RMSE of 5.840 on the public PHM dataset. On the self-constructed TiAl alloy milling dataset, it achieves an R^2 of 0.991, an MAE of 2.29, and an RMSE of 2.47. Performance comparisons with other models demonstrate that the proposed model achieves superior predictive performance and generalization capabilities across various evaluation metrics under diverse and complex operating conditions, validating the effectiveness and robustness of the proposed method.

This study still has limitations: (1) The generalization of the model to other materials needs further verification. (2) The deployment feasibility of the model on edge devices has not been evaluated; future work can achieve model lightweighting via network pruning techniques. (3) The backward information dependency of bidirectional LSTM limits its direct application in strictly real-time scenarios; subsequent research can explore alternative solutions based on causal convolution.

Author Contributions: Conceptualization, Y.X.; methodology, Y.X.; software, R.L.; validation, X.L. (Xiaohua Liu); formal analysis, R.L.; investigation, Y.W.; data curation, Y.G. and Y.W.; writing—original draft preparation, Y.X.; writing—review and editing, R.L. and H.L.; visualization, Y.X.; supervision, H.L. and X.L. (Xianli Liu); project administration, X.L. (Xianli Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Heilongjiang Natural Science Foundation Outstanding youth project YQ2024E041, National Natural Science Foundation of China 52575487, "Chunyan" Team Support Plan Project of Heilongjiang (CYCX24001), the Opening Project of the Key Laboratory of Advanced Processing Technology and Intelligent Manufacturing (Heilongjiang Province), Harbin University of Science and Technology (KFKT202201).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is unavailable due to privacy or ethical restrictions.

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

References

1. Soori, M.; Arezoo, B. Cutting tool wear prediction in machining operations, a review. *J. New Technol. Mater.* **2022**, *12*, 15–26.
2. Das, M.; Naikan, V.N.A.; Panja, S.C. A review of cutting tool life prediction through flank wear monitoring. *Int. J. Qual. Reliab. Manag.* **2025**, *42*, 425–473.
3. Chukwunweike, J.; Anang, A.N.; Adeniran, A.A.; Dike, J. Enhancing manufacturing efficiency and quality through automation and deep learning: Addressing redundancy, defects, vibration analysis, and material strength optimization. *World J. Adv. Res. Rev.* **2024**, *23*, 1272–1295. [CrossRef]
4. Munaro, R.; Attanasio, A.; Del Prete, A. Tool wear monitoring with artificial intelligence methods: A review. *J. Manuf. Mater. Process.* **2023**, *7*, 129. [CrossRef]
5. Peng, R.; Pang, H.; Jiang, H.; Hu, Y. Study of tool wear monitoring using machine vision. *Autom. Control Comput. Sci.* **2020**, *54*, 259–270.
6. Zhang, P.; Gao, D.; Lu, Y.; Ma, Z.; Wang, X.; Song, X. Cutting tool wear monitoring based on a smart toolholder with embedded force and vibration sensors and an improved residual network. *Measurement* **2022**, *199*, 111520. [CrossRef]
7. Xue, Z.; Li, L.; Wu, Y.; Yang, Y.; Wu, W.; Zou, Y.; Chen, N. Study on tool wear state recognition algorithm based on spindle vibration signals collected by homemade tool condition monitoring ring. *Measurement* **2023**, *223*, 113787. [CrossRef]

8. Gao, K.; Xu, X.; Jiao, S. Measurement and prediction of wear volume of the tool in nonlinear degradation process based on multi-sensor information fusion. *Eng. Fail. Anal.* **2022**, *136*, 106164. [CrossRef]
9. Xu, X.; Wang, J.; Zhong, B.; Ming, W.; Chen, M. Deep learning-based tool wear prediction and its application for machining process using multi-scale feature fusion and channel attention mechanism. *Measurement* **2021**, *177*, 109254.
10. Chen, N.; Hao, B.; Guo, Y.; Li, L.; Khan, M.A.; He, N. Research on tool wear monitoring in drilling process based on APSO-LS-SVM approach. *Int. J. Adv. Manuf. Technol.* **2020**, *108*, 2091–2101. [CrossRef]
11. Ganeshkumar, S.; Deepika, T.; Haldorai, A. A Supervised Machine Learning Model for Tool Condition Monitoring in Smart Manufacturing. *Def. Sci. J.* **2022**, *72*, 712–720. [CrossRef]
12. Heitz, T.; He, N.; Ait-Mlouk, A.; Bachrathy, D.; Chen, N.; Zhao, G.; Li, L. Investigation on eXtreme Gradient Boosting for cutting force prediction in milling. *J. Intell. Manuf.* **2025**, *36*, 285–301.
13. Ravikumar, S.; Ramachandran, K.I. Tool Wear Monitoring of Multipoint Cutting Tool using Sound Signal Features Signals with Machine Learning Techniques. *Mater. Today Proc.* **2018**, *5*, 25720–25729. [CrossRef]
14. Ahmadzadeh, M.; Zahrai, S.M.; Bitaraf, M. An integrated deep neural network model combining 1D CNN and LSTM for structural health monitoring utilizing multisensor time-series data. *Struct. Health Monit.* **2025**, *24*, 447–465. [CrossRef]
15. Li, R.; Wei, P.; Liu, X.; Li, C.; Ni, J.; Zhao, W.; Zhao, L.; Hou, K. Cutting tool wear state recognition based on a channel-space attention mechanism. *J. Manuf. Syst.* **2023**, *69*, 135–149. [CrossRef]
16. Al Mamun, A.; Bappy, M.M.; Mudiyanselage, A.S.; Li, J.; Jiang, Z.; Tian, Z.; Fuller, S.; Falls, T.C.; Bian, L.; Tian, W. Multi-channel sensor fusion for real-time bearing fault diagnosis by frequency-domain multilinear principal component analysis. *Int. J. Adv. Manuf. Technol.* **2023**, *124*, 1321–1334.
17. Huang, Y.J.; Liao, A.H.; Hu, D.Y.; Shi, W.; Zheng, S.-B. Multi-scale convolutional network with channel attention mechanism for rolling bearing fault diagnosis. *Measurement* **2022**, *203*, 111935. [CrossRef]
18. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
19. Cui, Z.; Ke, R.; Pu, Z.; Wang, Y. Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values. *Transp. Res. Part C Emerg. Technol.* **2020**, *118*, 102674. [CrossRef]
20. Marani, M.; Zeinali, M.; Songmene, V.; Mechefske, C.K. Tool wear prediction in high-speed turning of a steel alloy using long short-term memory modelling. *Measurement* **2021**, *177*, 109329.
21. Wang, G.; Zhang, F. A Sequence-to-Sequence Model With Attention and Monotonicity Loss for Tool Wear Monitoring and Prediction. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3525611. [CrossRef]
22. Cai, W.; Zhang, W.; Hu, X.; Liu, Y. A hybrid information model based on long short-term memory network for tool condition monitoring. *J. Intell. Manuf.* **2020**, *31*, 1497–1510. [CrossRef]
23. Xu, Z.; Zhang, B.; Fan, L.L.; Yan, E.H.; Li, D.; Zhao, Z.; Yip, W.S.; To, S. Deep-learning-driven intelligent tool wear identification of high-precision machining with multi-scale CNN-BiLSTM-GCN. *Adv. Eng. Inform.* **2025**, *65*, 103234. [CrossRef]
24. Li, X.; Qin, X.; Wu, J.; Yang, J.; Huang, Z. Tool wear prediction based on convolutional bidirectional LSTM model with improved particle swarm optimization. *Int. J. Adv. Manuf. Technol.* **2022**, *123*, 4025–4039. [CrossRef]
25. Sun, M.; Han, Y.; Guo, K.; Sivalingam, V.; Huang, X.; Sun, J. A milling tool wear predicting method with processing generalization capability. *J. Manuf. Process.* **2024**, *120*, 975–1001. [CrossRef]
26. Xue, Z.; Chen, N.; Wu, Y.; Yang, Y.; Li, L. Hierarchical temporal transformer network for tool wear state recognition. *Adv. Eng. Inform.* **2023**, *58*, 102218. [CrossRef]
27. Guo, H.; Zhang, Y.; Zhu, K. Interpretable deep learning approach for tool wear monitoring in high-speed milling. *Comput. Ind.* **2022**, *138*, 103638. [CrossRef]
28. Li, W.; Fu, H.; Han, Z.; Zhang, X.; Jin, H. Intelligent tool wear prediction based on Informer encoder and stacked bidirectional gated recurrent unit. *Robot. Comput. Integr. Manuf. Int. J. Manuf. Prod. Process Dev.* **2022**, *77*, 102368. [CrossRef]
29. He, J.; Xu, Y.; Pan, Y.; Wang, Y. Adaptive weighted generative adversarial network with attention mechanism: A transfer data augmentation method for tool wear prediction. *Mech. Syst. Signal Process.* **2024**, *212*, 111288. [CrossRef]
30. Li, B.; Li, J.; Wu, X.; Tang, H. Gated recurrent unit and temporal convolutional network with soft thresholding and attention mechanism for tool wear prediction. *Measurement* **2025**, *240*, 115546. [CrossRef]
31. Zhou, Y.; Li, R.; Xia, T.; Zhang, F.; Wang, D.; Xi, L. Hierarchical temporal-modal synchronous fusion framework for tool wear recognition with multimodal monitoring data integrating uncertainty quantification. *J. Manuf. Process.* **2025**, *149*, 964–986. [CrossRef]
32. Hou, K.; Li, R.; Liu, X.; Yue, C.; Wang, Y.; Liu, X.; Xia, W. Swin-fusion: An adaptive multi-source information fusion framework for enhanced tool wear monitoring. *J. Manuf. Syst.* **2025**, *79*, 435–454.
33. Luo, J.; Zhang, X. Convolutional neural network based on attention mechanism and Bi-LSTM for bearing remaining life prediction. *Appl. Intell.* **2021**, *52*, 1076–1091. [CrossRef]

34. Wei, P.; Li, R.; Liu, X.; Gao, H.; Dai, M.; Zhang, Y.; Zhao, W.; Liu, E. Research on tool wear state identification method driven by multi-source information fusion and multi-dimension attention mechanism. *Robot. Comput.-Integr. Manuf.* **2024**, *88*, 102741. [CrossRef]
35. Zhang, B.; Liu, X.; Yue, C.; Liang, S.Y.; Wang, L.; Zhou, T. Physics-constrained domain-decoupled transfer method: Application to cross-condition tooth-wise monitoring in multi-tooth tools. *Mech. Syst. Signal Process.* **2026**, *242*, 113636. [CrossRef]
36. Zhang, B.; Liu, X.; Yue, C.; Liang, S.Y.; Wang, L. Meta-learning-based approach for tool condition monitoring in multi-condition small sample scenarios. *Mech. Syst. Signal Process.* **2024**, *216*, 111444. [CrossRef]
37. PHM Society. PHM Society Conference Data Challenge. 2010. Available online: https://www.phmsociety.org/competition/phm/10 (accessed on 20 December 2021).
38. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [CrossRef]
39. Bai, S. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271. [CrossRef]
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**. [CrossRef]