



# Tool wear prediction based on a fusion model of data-driven and physical models in the milling process

Chang Fan<sup>1</sup> · Zhao Zhang<sup>1,2</sup> · Dinghua Zhang<sup>1,2</sup> · Ming Luo<sup>1,2</sup> 

Received: 19 February 2024 / Accepted: 27 May 2024 / Published online: 15 June 2024  
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2024

## Abstract

Precision tool wear prediction in the milling process is crucial in enhancing product quality and machining efficiency. Both data-driven models and physical models are vital for indirect tool wear prediction. However, data-driven models rely on appropriate model structures and extensive datasets to achieve high prediction accuracy; physical models face challenges when adapting to complex cutting conditions, in which case accurate modeling is difficult. Aiming at employing the advantages of both methods for accurate tool wear prediction, a fusion model is developed by integrating both the data-driven model and the physical model by constructing an indirect prediction layer and a parameter constraint layer. The indirect prediction layer incorporates domain knowledge of tool wear, while the parameter constraint layer utilizes priori knowledge from accumulated data. Validation results show that with the introduction of domain knowledge and prior knowledge as constraints, the range and shape of the fusion model's prediction result confidence intervals are effectively constrained to more reasonable zones, the area of the confidence intervals is reduced by 73.7%, and the average prediction accuracy of the fusion model is improved by 11.5%.

**Keywords** Tool wear · Milling · Data-driven model · Physical model · Fusion model

## 1 Introduction

With the continuous improvement of the performance and life requirements of key components and high-end equipment in the aerospace field, the difficult-to-cut materials represented by the nickel-based superalloy GH4169 are widely used in the manufacture of heat-resistant parts, such as casings [1], blisks [2], and blades [3], of aero-engines, due to their excellent mechanical properties, thermal fatigue strength, oxidation resistance [4], and corrosion resistance [5]. Based on the difficult machinability of GH4169, such as poor thermal conductivity, high strength, and high stiffness, the machining process of the parts has a tendency for

severe surface hardening, large cutting force, and rapid tool wear. This presents challenges to both the machining cost [6] and the machining quality [7] of the parts. The state of tool wear significantly influences the stability of the machining process [8]. Research indicates that tool failure contributes to over 20% of the total machine tool downtime [9], and the cost associated with tool use and replacement constitutes 3–12% of the overall machining cost [10]. Achieving timely and accurate prediction of tool wear in the machining process facilitates the replacement of severely worn tools. This approach contributes to reducing production costs, optimizing production efficiency, and enhancing surface quality [11]. Furthermore, it plays a crucial role in realizing intelligent manufacturing.

With an increase in cutting time, the tool continuously loses material during the interaction with the workpiece [12], and this process is irreversible. Consequently, wear leads to changes in the geometry of the tool, such as tool diameter and tool length [13]. These changes further impact force, heat, vibration, and stress in the cutting region [14]. Additionally, these alterations affect machine tool current, power, torque, and other relevant information [15]. Collecting information related to tool wear at the

✉ Ming Luo  
luoming@nwpu.edu.cn

<sup>1</sup> Ministry of Industry and Information Technology, Key Laboratory of High Performance Manufacturing for Aero Engine, Northwestern Polytechnical University, Xi'an 710072, China

<sup>2</sup> Engineering Research Center of Advanced Manufacturing Technology for Aero Engine, Northwestern Polytechnical University, Ministry of Education, Xi'an 710072, China

machining site allows for the analysis of the wear state of the tool. Tool wear monitoring methods are categorized into two types: direct monitoring and indirect monitoring [16], based on the correlation of information affected by tool wear. Typically, direct monitoring methods utilize equipment to collect information directly associated with the loss of tool material and offer high recognition accuracy under ideal conditions. For instance, CCD cameras can capture the shape of the wear area [17]. Other methods, such as machine vision [18], digital image processing [19], geometric descriptor techniques [20], resistive methods [21], and radiographic methods [22], can also be employed. However, the application of the direct monitoring methods is more restricted. Indirect monitoring methods employ sensors to collect information indirectly associated with the loss of tool material. Since the indirect monitoring method has less impact on the machining process, it has garnered significant attention from both academia and industry. General monitoring signals include force, vibration, acoustic emission, feed current, and spindle power signals [16]. As tool wear progresses, the ability of the cutting edge to shear the workpiece material decreases, and this inefficiency usually leads to an increase in cutting forces. As the tool geometry changes, which can alter the dynamic balance of the cutting process and potentially introduce or amplify vibrations, it can create a feedback loop where increasing vibrations lead to faster tool wear, which then leads to more vibrations.

The tool wear state cannot be directly inferred from the raw monitoring signals. Therefore, researchers have developed various data-driven models, including artificial neural networks, genetic algorithms, fuzzy logic, support vector machines, and Markov models [23]. Its purpose is to establish a mapping relationship between the signals and the tool wear. Bao et al. [24] proposed a data model of machine and tool, communication framework, and access strategy based on OPC UA and established a BP neural network model reflecting the relationship between machine condition and tool parameters for tool health prediction. Wang et al. [25] proposed a multi-scale principal component analysis (MSPCA) method to construct statistical indicators and corresponding control limits for tool wear monitoring, realizing online monitoring of tool wear in the milling process. Shi et al. [26] proposed a PCA method for extracting features from multiple sensor signals in the machining process and constructed a tool wear prediction model based on a least squares support vector machine by learning the correlation between the extracted features and the actual tool wear. Data-driven models in the mechanical field are mostly derived from the computer and mathematical fields. These models, which rely on sensor data collected at the machining site [27], often fail to leverage the abundant domain knowledge already available.

Scientific models invariably involve some degree of idealization, abstraction, or fictionalization of their target system [28]. In contrast to data-driven models, which mainly rely on machining data and lack a physical background, physical models are built based on domain knowledge in the mechanical field. These models use mathematical or empirical formulas to capture the primary factors of the tool wear process while ignoring secondary factors. Physical models exhibit applicability and generalizability across various cutting conditions, offering greater interpretability and physical significance. Tyler [29] pioneered the expression of tool life as a function related to cutting parameters such as depth of cut and feed rate, where the parameters to be determined in function are constants for a given tool-workpiece material pair, and the constants can be obtained through cutting machining experiments. Müller [30] was the first to employ an empirical function containing a unitary linear term and an exponential term to establish a mapping relationship between tool wear and cutting time, serving as a means to characterize the evolving pattern of flank wear. According to Pálmai [31], the most effective empirical function of wear-time is the physical model proposed by Sipos [32], which contains exponential and polynomial terms. Zhang et al. [33] validated a generalized wear model with adjustable coefficients based on experimental data and compared it with other celebrated wear models and then further improved its adaptability and generalization ability. Fan et al. [34] analyzed the tool wear process through an evolutionary cluster analysis method and proposed a physical model of tool wear jointly constructed by three sub-equations, which has better fitting accuracy and generalizability relative to existing models. Nevertheless, physical models simplify complex manufacturing systems by making reasonable assumptions. It requires a comprehensive and in-depth understanding of the tool wear mechanism and process. Model parameter calibration must be conducted for the specific working conditions during application. The calibration process for model parameters needs to be repeated when facing new working conditions. This complexity makes it challenging to calibrate model parameters for scenarios involving complex working conditions.

As a problem with a clear application background and physical significance, the tool wear process follows certain laws. Consequently, when predicting tool wear values, the continuous predictions should exhibit certain trends and align with the domain knowledge of tool wear. However, the data-driven model often relies solely on applying data labels as constraints for dealing with the tool wear problem, but cannot be compatible with the domain knowledge, that is, employing the physical model as the constraints of the model. Therefore, combining the advantages of data-driven model and physical model while mitigating the shortcomings of a single prediction method has become an important

direction for enhancing the predictive ability and interpretability in the tool wear prediction problem. In 2017, Stewart et al. [35] introduced a method to supervise neural networks and constrain the output space by incorporating known laws of physics as domain knowledge, which was experimentally verified to significantly reduce the need for labeled data, but poses a new challenge for encoding priori knowledge into an appropriate loss function. Wang et al. [36] proposed a novel physics-guided neural network model for tool wear prediction in the form of cross physics-data fusion (CPDF) scheme as a modeling strategy to fuse the hidden information explored by the physics-based and data-driven models, which eliminates the physical inconsistencies present in the traditional data-driven models through the physics guided loss function. Hua et al. [37] introduced a weighted-loss PINN (PNNN-WLs) based on the evaluation of uncertainty, which quantifies the prediction error of the variance, proposed a novel weight allocation strategy based on uncertainty evaluation, and established an improved PINN framework for accurate and stable prediction of manufacturing systems. However, on the one hand, the fitting accuracy and generalizability of the physical models of tool wear selected for existing model fusion studies are poor. On the other hand, while existing model fusion studies have improved prediction accuracy, they still lack the ability to intuitively explain how the physical model affects the prediction process of the data-driven model and how it constrains the data distribution of tool wear prediction results.

To address the challenge of underutilizing tool wear domain knowledge and existing data priori knowledge in current studies, this study proposes a novel tool wear model fusion method, which employs an indirect prediction layer and a parameter constraint layer to effectively incorporate the data-driven model with the physical model. The contributions of this study are in two aspects.

- a) The physical model of tool wear, which represents the domain knowledge of tool wear, is transformed into an indirect prediction layer that is integrated into the network structure of the fusion model. This addition facilitates the derivation of the loss function, ensuring that the prediction results of the fusion model are bound by both data and physical model constraints, thereby applying tool wear domain knowledge to tool wear prediction to improve predictive ability and interpretability;
- b) Designing the parameter constraint layer based on the saturation function, we utilized the physical model of tool wear to solve model parameters on accumulated tool wear data. This process utilizes the priori knowledge of the accumulated tool wear data to determine the parameter ranges of the physical model under specific working conditions, forming the foundation for configuring the parameter constraint layer attributes. This configuration

aims to improve prediction accuracy and expedite the model training efficiency.

The rest of the paper is organized as follows: the “**Tool wear model fusion method**” section presents the model fusion framework; the “**Experiment design**” section details the design of the milling experiment and the collection of wear-related data; the proposed model fusion methodology is validated in the “**Analysis and validation**” section, and the public dataset PHM2010 is introduced to further validate the generalization and interpretability of the model fusion methodology. Finally, this study is summarized, and the future research direction is prospected in the “**Conclusion**” section.

## 2 Tool wear model fusion method

In this study, based on the data-driven model, the network structure of the fusion model is designed, and the indirect prediction layer representing domain knowledge and the parameter constraint layer representing priori knowledge are included. The framework of the proposed tool wear model fusion method is illustrated in Fig. 1, including six parts: data preprocessing, attention mechanism, one-dimensional convolutional neural network, indirect prediction layer based on the physical model, parameter constraint layer based on the saturation function, and fusion model. Firstly, the tool wear data is preprocessed, and a data-driven model is formulated by substituting the correlation analysis process with the attention mechanism. Secondly, the indirect prediction layer is devised based on the physical model, incorporating domain knowledge into tool wear prediction. Thirdly, the parameter constraint layer is designed using the saturation function. Mathematical analyses and actual tool wear data are employed to determine the constraint range of the parameter constraint layer positioned before the indirect prediction layer. Finally, the indirect prediction layer and the parameter constraint layer are integrated with the multi-column neural network structure, devised based on the network structure of the data-driven model, to form a fusion model. In the model, its loss function is constrained by the data labels and the physical model.

### 2.1 Data-driven model

The data-driven model establishes the mapping relationship between data and labels through iterative training, and the data-driven model represented by the neural network has strong nonlinear mapping ability and can fit any nonlinear function. In the field of tool wear, sensor signals are difficult to be directly used in data-driven models. Therefore, data preprocessing and attention mechanisms are used to

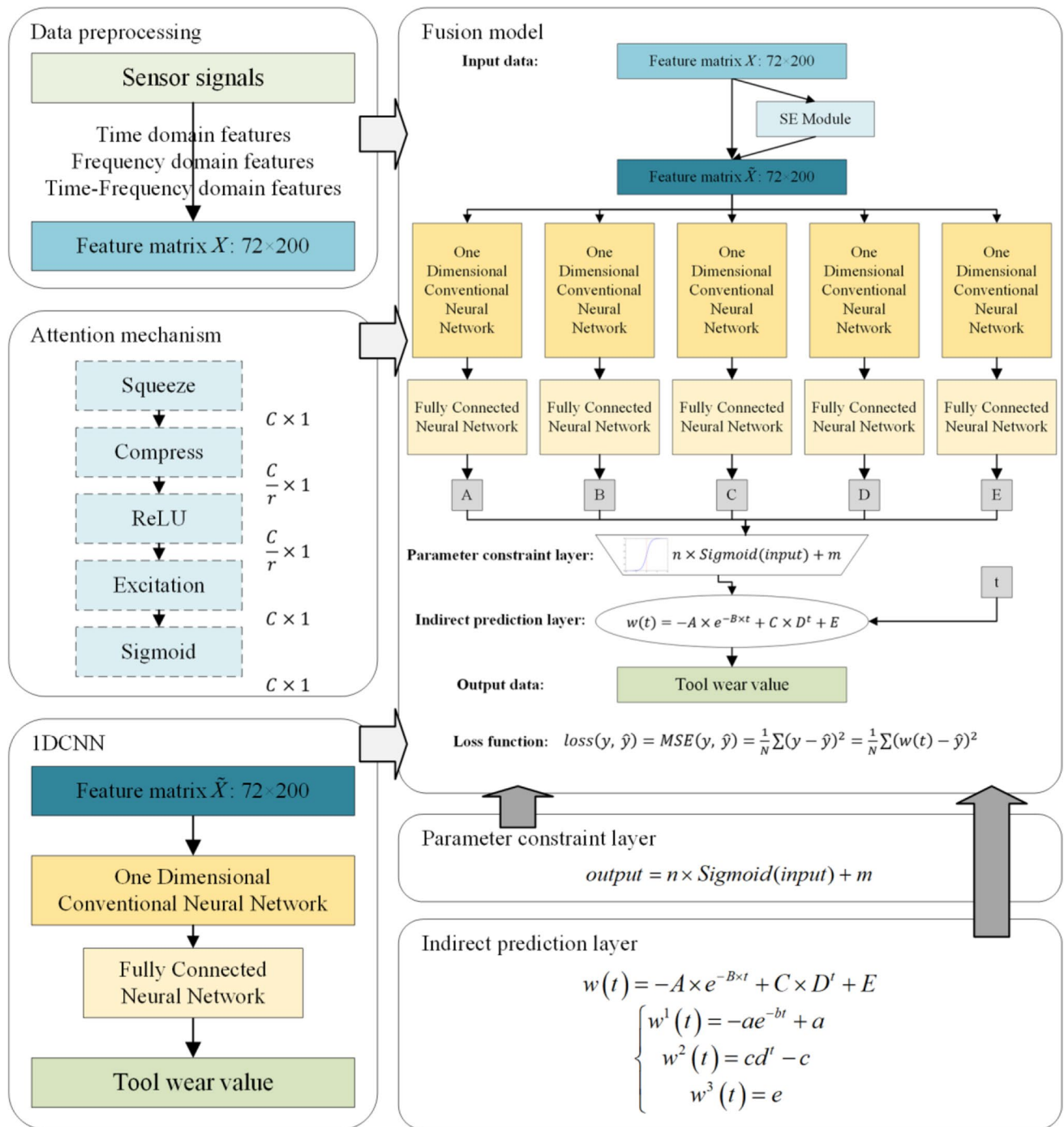


Fig. 1 Model fusion methodology framework

transform the sensor signals into a feature matrix suitable for the tool wear prediction.

2.1.1 Data preprocessing

For data-driven models and fusion models, tool wear-related data collected from the machining site, such as raw cutting

force signals, vibration signals, and acoustic emission signals, commonly exhibit drawbacks such as large data volume, high data dimensionality, and redundancy. These issues pose significant challenges to the iterative training process and the convergence of model parameters. To improve data quality and reduce data sparsity, raw signals require preprocessing steps involving the removal of air cut data,

outliers, and trend terms. Additionally, to further diminish the data volume and convert the data dimensionality from high-dimensional space to low-dimensional space, research typically employs time domain analysis, frequency domain analysis, and time–frequency domain analysis methods [38] to extract data features.

Time domain analysis involves calculating the features of the signal with time as the independent variable. For the raw signals in this study, the time domain features include maximum value, average value, peak and valley values, variance, standard deviation, root mean square, skewness, kurtosis, waveform factor, peak factor, impulse factor, and margin factor. Simultaneously, alterations in the tool wear state often induce modifications in the frequency structure of the raw cutting force signal and vibration signal. Therefore, exploring the frequency domain of the original signal becomes essential. This involves employing mathematical methods to transform the signal from the time domain to the frequency domain, calculating the power spectral density function  $S(f)$  with frequency as the horizontal coordinate and power as the vertical coordinate. Relevant frequency domain features, including frequency center, mean square frequency, and frequency variance, are then derived from the power spectrum. To address the limitations of the Fourier transform in time and frequency domain analysis, a time–frequency domain analysis method based on wavelet packet transform is employed. In this method, the signal undergoes a three-layer wavelet packet decomposition, resulting in its complete orthogonal decomposition into eight independent sub-frequency bands. Subsequently, wavelet packet energy values are extracted from these bands, serving as the time–frequency domain features of the signal [39]. The time domain features, frequency domain features, and time–frequency domain features are used together as the relevant features of tool wear to jointly construct the feature matrix for tool wear prediction. The definitions of all the features are detailed in Table 1.

Due to the variation in magnitudes and dimensions among different features, the impact of features with lower orders of magnitude may be ignored. To maintain the trend of the selected features and mitigate the impact of their varying magnitudes and dimensions on the prediction results, it is necessary to normalize the selected features by compressing their values into the range of [0, 1].

### 2.1.2 Attention mechanism

Multi-source features related to tool wear, obtained from different sensors through various feature extraction methods [40], contain different tool wear information. To enhance the feature matrix quality and mitigate the impact of redundant features on the tool wear prediction results, correlation analysis is commonly employed for feature

selection based on Pearson’s correlation coefficient [41] or the mutual information coefficient [42]. This aims to identify features that exhibit sensitivity to tool wear. Varied selection thresholds in different studies for different cutting conditions result in differences in the types of features obtained through correlation analysis selection. To eliminate uncertainty generated from the feature selection step and improve the generalizability of the model fusion method, this study employs the attention mechanism to process the feature matrix instead of using correlation analysis for feature selection. Each channel within the feature matrix contains varying amounts of information about tool wear. The attention mechanism enables the prediction model to autonomously learn to allocate attention, automatically adjust the influence weights of each channel, diminish the weights of redundant features, and emphasize the impact of important features on the prediction results [43]. The attention mechanism serves as a module that is positioned between data preprocessing and the model’s network structure. The attention mechanism operates as follows: Utilizing the squeeze-and-excitation block [44], weights for each channel of the feature matrix are calculated. These weights are then assigned to the feature matrix, as illustrated in Fig. 2, to obtain an enhanced attention feature matrix.

The feature matrix  $X$ , consisting of  $C$  channels extracted from the monitoring signal, is inputted into the attention mechanism module. This module outputs the feature matrix  $\tilde{X}$ , containing a total of  $C$  channels, to the prediction model. The squeeze global information embedding module corresponds to a global average pooling operation. It compresses global channel information into a channel descriptor by reducing the feature matrix, with the number of channels and vector length  $C \times H$ , into a  $C \times 1$  weight vector. This vector serves as a channel descriptor, encapsulating the global information of the feature set:

$$z = F_{sq}(X_c) = \frac{1}{H} \sum_{i=1}^H X_c(i) \quad (1)$$

The excitation adaptive recalibration part contains two fully connected layers. After squeeze global pooling, a convolution of length 1 is applied to the  $C$ -dimensional vector, resulting in a  $C/r$ -dimensional vector. Where  $r$  is a down-sampling scale factor, the original number of channels  $C$  is divided by  $r$  to get the number of compressed channels  $C/r$ . It is common to choose a power of 2 as the value of  $r$ . After experimental comparisons, the value of  $r$  in this study is 4. This is followed by ReLU activation. Subsequently, another convolution of length 1 transforms the  $C/r$ -dimensional vector back to a  $C$ -dimensional vector, and a Sigmoid activation is applied to ensure values range between 0 and 1, yielding the weight vector:

**Table 1** Feature definition

Domain	Significance	Feature	Definition		
Time domain	Statistical properties of signal over time	Maximum	$\max = \max(f_i)$		
		Average	$\text{avg} = \frac{1}{T} \sum_{i=1}^T f_i$		
		Peak-to-valley	$\text{pv} = \max(f_i) - \min(f_i)$		
		Variance	$\text{var} = \frac{1}{T} \sum_{i=1}^T (f_i - \bar{f})^2$		
		Standard deviation	$\text{std} = \sqrt{\frac{1}{T} \sum_{i=1}^T (f_i - \bar{f})^2}$		
		Root mean square	$\text{rms} = \sigma \sqrt{\frac{1}{T} \sum_{i=1}^T f_i^2}$		
		Skewness	$\text{ske} = \frac{1}{T} \sum_{i=1}^T \frac{(f_i - \bar{f})^3}{\sigma^3}$		
		Kurtosis	$\text{kur} = \frac{1}{T} \sum_{i=1}^T \frac{(f_i - \bar{f})^4}{\sigma^4}$		
		Waveform factor	$\text{wf} = \frac{\text{rms}}{\frac{1}{T} \sum_{i=1}^T  f_i }$		
		Crest factor	$\text{cf} = \frac{\text{pv}}{\text{rms}}$		
		Impulse factor	$\text{if} = \frac{\text{pv}}{\frac{1}{T} \sum_{i=1}^T  f_i }$		
		Margin factor	$\text{mf} = \frac{\text{pv}}{\left(\frac{1}{T} \sum_{i=1}^T \sqrt{ f_i }\right)^2}$		
		Frequency domain	Utilizing the spectral information of the signal, such as amplitude or phase at different frequencies	Frequency center	$\text{fc} = \frac{\sum_{\omega} \omega S(\omega)}{\sum_{\omega} S(\omega)}$
				Mean square frequency	$\text{msf} = \frac{\sum_{\omega} \omega^2 S(\omega)}{\sum_{\omega} S(\omega)}$
Frequency variance	$\text{vf} = \frac{\sum_{\omega} (\omega - \text{fc})^2 S(\omega)}{\sum_{\omega} S(\omega)}$				
Time–frequency domain	Takes into account information about the signal in both the time and frequency domains	Wavelet packet energy	wpe – band1 wpe – band2 wpe – band3 wpe – band4 wpe – band5 wpe – band6 wpe – band7 wpe – band8		

$$s = F_{\text{ex}}(z, \text{Conv}) = \text{Sigmoid}(\text{Conv}_2 \text{ReLU}(\text{Conv}_1 z)) \quad (2)$$

By multiplying the weight vector  $s$  with  $X$ , the feature weight-adjusted  $\tilde{X}$  is obtained:

$$\tilde{X} = F_{\text{scale}}(X_c, s_c) = s_c X_c \quad (3)$$

### 2.1.3 One-dimensional convolutional neural network

Different types of neural networks can process tool wear data with different data structures, such as cutting-edge pictures, signal time–frequency maps, and signal data. This study focuses on the widely used signal data in tool wear monitoring, selecting the one-dimensional convolutional neural network as the network structure of the data-driven model. This structure is also a part of the network

structure of the fusion model and is used for comparison with the fusion model.

The one-dimensional convolutional neural network (1DCNN) is a deep learning architecture that effectively reduces the number of trainable parameters through local perception and weight sharing, resulting in computational efficiency and excellent performance. The 1DCNN model consists of six parts: the input layer, the convolutional layer, the activation layer, the pooling layer, the fully connected layer, and the output layer. The structure and parameters of the 1DCNN are detailed in Table 2. The output layer calculates predicted values  $\hat{y}$  of the network, and the loss function measures both model performance and the inconsistency between the actual values  $y$  and the predicted values  $\hat{y}$ . This inconsistency serves as the basis for calculating the gradient during back propagation to update the network parameters, indicating that the actual

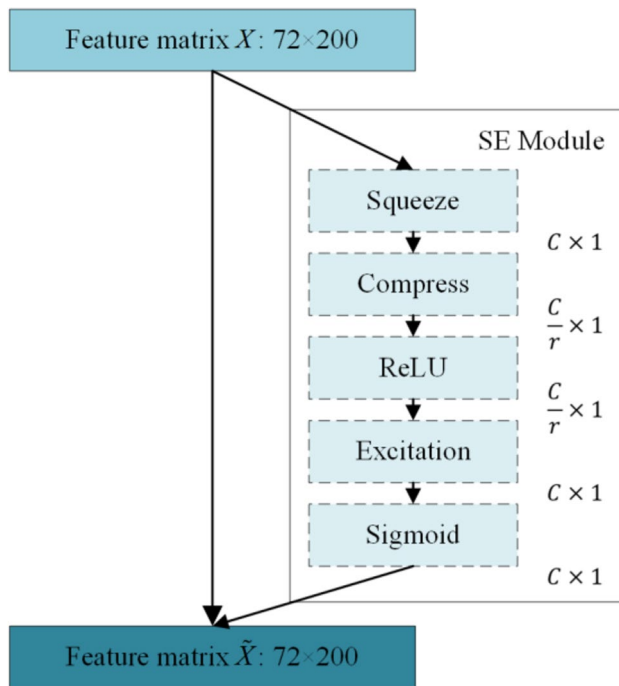


Fig. 2 Attention mechanism module

values are used as a constraint for the data-driven model. The loss function of the data-driven model is

$$loss(y, \hat{y}) = MSE(y, \hat{y}) = \frac{1}{N} \sum (y - \hat{y})^2 \tag{4}$$

The structure of the data-driven model is illustrated in Fig. 3.

### 2.2 Indirect prediction layer based on the physical model

Tool wear increases irreversibly during machining, and a large amount of domain knowledge on tool wear has been accumulated. Current research primarily focuses on establishing a physical model of tool wear to describe the tool wear process based on wear mechanisms, field experience, or mathematical derivations, which is used to map the relationship between the machining parameters and the tool wear values to describe the growth law of the tool wear values. However, due to the high complexity and nonlinearity of the machining process, there is no generally accepted physical model for the general tool wear process. Considering the physical characteristics of the tool wear process, the physical model should satisfy specific mathematical constraints.

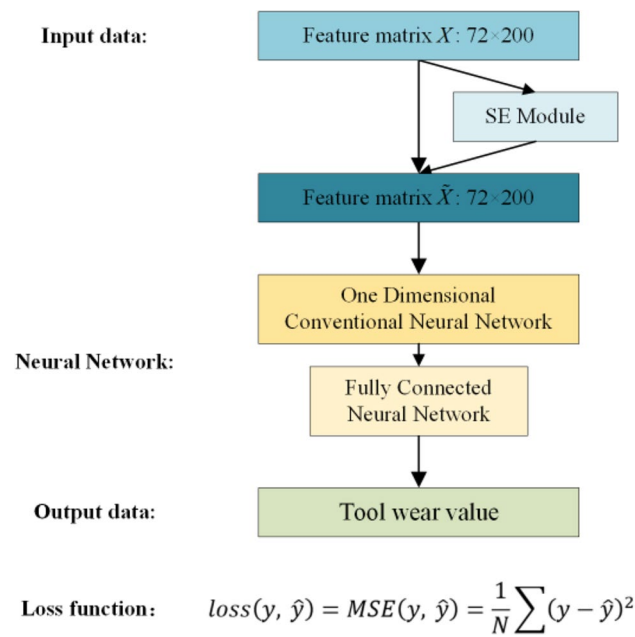
$$\left\{ \begin{array}{l} \lim_{\Delta t \rightarrow 0} \frac{\Delta VB}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{VB(t_0 + \Delta t) - VB(t_0)}{\Delta t}, t_0 \in t \\ \lim_{t \rightarrow t_0} w''(t) = w''(t_0) \\ w(t) \geq 0 \\ w'(t) > 0 \\ t \geq 0 \end{array} \right. \tag{5}$$

where  $t$  is the cutting time,  $t_0$  is any specific cutting time,  $\Delta t$  is the amount of change in cutting time,  $VB$  is the tool wear value,  $\Delta VB$  is the amount of change in the tool wear value,  $w$  is the physical model,  $w'$  is the first-order derivative of the physical model, and  $w''$  is the second-order derivative of the physical model.

Altintas [45] summarized the tool wear process into three stages: initial wear, normal wear, and severe wear, and different stages have unique evolutionary rules. Since the physical model serves as a reasonable simplification of the complex

Table 2 1DCNN network structure and parameters

No	Network layer	Out channels	Kernel size/Step size	Activation function
1	Conv1d	32	5 × 1/1 × 1	LeakyReLU
2	AvgPool1d		3 × 1/2 × 1	
3	Conv1d	48	5 × 1/1 × 1	Sigmoid
4	AvgPool1d		3 × 1/2 × 1	
5	Conv1d	64	5 × 1/1 × 1	LeakyReLU
6	AvgPool1d		3 × 1/2 × 1	
7	Conv1d	80	5 × 1/1 × 1	LeakyReLU
8	AvgPool1d		3 × 1/2 × 1	
9	Conv1d	96	5 × 1/1 × 1	LeakyReLU
10	Reshape	288		
11	Linear	96		
12	Linear	48		
13	Linear	4		
14	Linear	1		



**Fig. 3** Data-driven model structure

machining process, adapting it to the complex working conditions and tool wear process requires a physical model composed of various sub-functions, using different sub-functions to describe the different stages of the tool wear process. It should meet specific requirements:

- 1) different sub-functions control different stages
- 2) different sub-functions have less effect on other stages.

The selection of sub-equations should satisfy both mathematical constraints and meet requirements (1) and (2). After comparative analysis, the tool wear model generated by the evolutionary cluster analysis method [34], comprising a combination of exponential and constant functions, is chosen from existing physical models of tool wear. The natural exponential function is used to control the slope of the curve from large to small in the early stage, and then the slope converges to zero, which has no effect on the slope of the subsequent curve; the exponential function is used to control the slope of the subsequent stages from stable to gradually increasing, with relatively little effect on the slope of the earlier stage of the curve; the constant function is used to adjust the initial value of the wear curve to control the overall position of the curve, which are given as

$$\begin{cases} w^1(t) = -ae^{-bt} + a \\ w^2(t) = cd^t - c \\ w^3(t) = e \end{cases} \quad (6)$$

where  $a$ ,  $b$ ,  $c$ ,  $d$ , and  $e$  are the fitting coefficients of the model.

The final tool wear model is shown in the following equation:

$$w(t) = w^1(t) + w^2(t) + w^3(t) = -A \times e^{-B \times t} + C \times D^t + E \quad (7)$$

where  $A$ ,  $B$ ,  $C$ ,  $D$ , and  $E$  are the fitting coefficients of the model, and  $A, B, C > 0$ ,  $D > 1$ ,  $E \geq A - C$ . When no cutting is performed, and the tool is not worn, then  $E = A - C$ .

One of the basic assumptions in the tool wear modeling problem is that the tool wear process is continuous, so the first-order and second-order derivatives of this tool wear model  $w(t)$  are respectively

$$w'(t) = AB \times e^{-B \times t} + C \ln D \times D^t \quad (8)$$

$$w''(t) = -AB^2 \times e^{-B \times t} + C \ln^2 D \times D^t \quad (9)$$

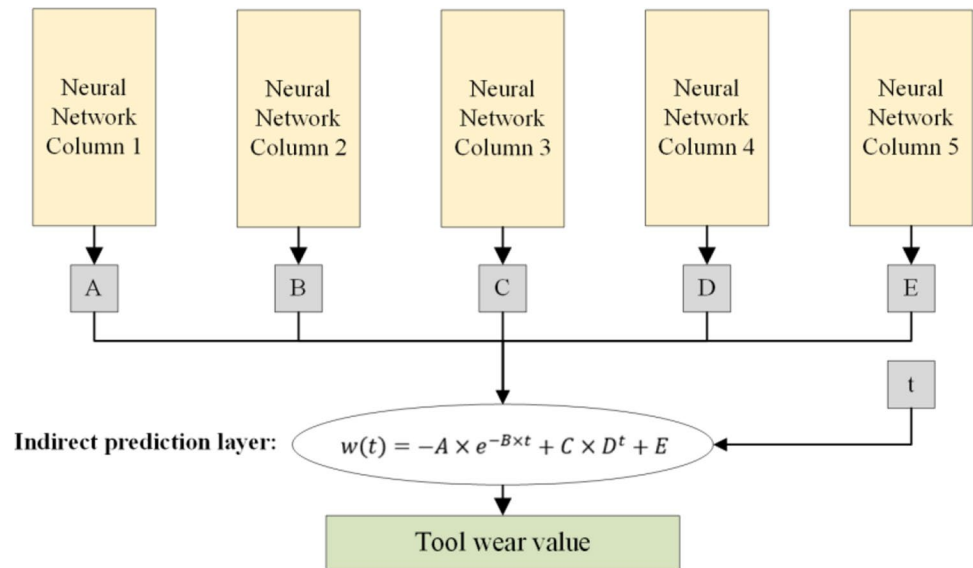
As  $t$  approaches 0, the value of  $-ae^{-bt} + a$  is relatively large, and the value of  $cd^t - c$  is small, at which point the function is monotonically increasing, and its value is dominated by  $w^1(t)$ .

As  $t$  increases, the rate of increase of  $w^1(t)$  gradually slows down while the rate of increase of  $w^2(t)$  gradually speeds up, at which point the value of the function is jointly dominated by  $w^1(t)$  and  $w^2(t)$ . In this process, when the second-order derivative changes from negative to positive, the function  $w(t)$  changes from convex to concave at some point, which means that there is an inflection point for the function. This inflection point is the point at which the second-order derivative changes from negative to positive, which is the solution to the equation  $w'''(t) = 0$ .

After the inflection point, the function starts to accelerate its growth, gradually dominated by  $w^2(t)$ . The critical cutoff point can be defined as the point at which the rate of growth of the function changes significantly. This point can be determined by observing the behavior of the first-order derivative  $w'(t)$ . The critical cutoff point can be considered to be reached when the rate of growth of the first-order derivative begins to accelerate significantly. The value of  $t$  corresponding to this cutoff point can be estimated numerically or by observing the image of the function.

The indirect prediction layer, shown in Fig. 4, is constructed based on the selected tool wear model. It consists of multiple input nodes, an output node, and a function. The number of input nodes equals the number of model parameters of the physical model plus one, corresponding to the model parameters and machining time, respectively. The output node corresponds to the tool wear value, while the function represents the mapping relationship of the physical model from the model parameters and machining time to the tool wear value. The indirect prediction layer is located



**Fig. 4** Indirect prediction layer structure

before the output layer of the fusion model and after the multi-column neural network structure. In contrast to the data-driven model, which directly calculates the tool wear value, the fusion model first calculates the model parameters of the physical model and then obtains the final predicted tool wear value through the indirect prediction layer.

### 2.3 Parameter constraint layer based on saturation function

The parameters of the tool wear model, which are constructed based on the actual machining process, exhibit a relatively certain range under different machining conditions. In contrast, the hidden parameters of the neural network, which have not undergone iterative training, are random. The model parameters calculated during the initial iterative training process may exceed the value range of the actual problem, thus affecting the training speed of the model. To address this, the theoretical distribution range of the model parameters can be roughly inferred using mathematical methods. Simultaneously, leveraging the accumulated industrial data from the actual production and machining process allows for solving the parameters of the selected tool wear model. Based on the similarity between the processing conditions of the industrial data and the target conditions, the approximate distribution range of the model parameters is determined. This information serves as priori knowledge to determine the range for the parameter constraint layer. With the continuous accumulation of tool wear data for the target conditions, the constraint range of the parameter constraint layer can be refined.

The saturation function serves as a nonlinear activation function capable of mapping any real number to a probabilistic value that tends to converge if the input is

too large or too small. Therefore, the saturation function can map the input value with an uncertain distribution range to the constraint interval with a certain distribution range. However, the output interval of the basic saturation function is fixed, which is inconsistent with the distribution range of the tool wear model parameters. Consequently, the saturation function needs to be further modified to achieve a parameter constraint layer with an adjustable output range. The structure of the parameter constraint layer based on the saturation function is given by  $output_j = n_j \times f(input_j) + m_j$ , where  $f$  represents either the Sigmoid function or Tanh function,  $n_j$  is the amplitude, and  $m_j$  is the phase. These parameters ( $n_j$ ,  $m_j$ , and  $f$ ) jointly constrain the distribution range of the model parameters.

Available tool wear data that can serve as priori knowledge act as a reference for the distribution of tool wear values under unknown working conditions. In this study, a physical model  $w(t)$  is introduced, which can be applied to fit the known tool wear curves, allowing the solution for the parameters and obtaining the distribution range of each parameter. Due to the similar but different assumptions of the known and unknown working conditions in the tool wear prediction problem, the parameter values of the tool wear curve for the unknown working conditions can be referenced to the range of distribution of the parameters of the tool wear curve for the known working conditions. Before model training, it is necessary to divide all the samples into training set and test set. The tool wear data of the training set are used to solve for the physical model parameters, and then the range of the parameter constraint layer constraints on each parameter is determined. In this study, the Sigmoid function is chosen as the  $f$ , and the finalized  $n_j$  and  $m_j$  are detailed in Table 3.

**Table 3** Parameter constraint layer properties

Layer $j$	$n_j$	$m_j$
A	0.1	0.000001
B	1	0.000001
C	20	0.000001
D	0.01	1.000001
E	5	0.000001

## 2.4 Tool wear fusion model

The tool wear fusion model is constructed using a data-driven model, an indirect prediction layer based on the physical model, and a parameter constraint layer based on the saturation function. The input layer of the multi-column neural network is the feature matrix, and the output layer is the tool wear value. The network structure consists of multiple columns with the same neural network structure, with the number of columns equal to the number of parameters in the physical model of tool wear. The neural network structure of each column is the same as that of the data-driven model. A parameter constraint layer is independently added at the end of each column of neural network structures. Additionally, an indirect prediction layer is added before the output layer and after the parameter constraint layer of the multi-column neural network, resulting in the complete tool wear fusion model. Derive the loss function of the fusion model as  $loss(y, \hat{y}) = MSE(y, \hat{y}) = \frac{1}{N} \sum (y - \hat{y})^2 = \frac{1}{N} \sum (w(t) - \hat{y})^2$ . The loss function reveals that the fusion model is constrained by both the data labels and the tool wear model during iterative training. The structure of the fusion model is illustrated in Fig. 5.

## 3 Experiment design

To fully apply the domain knowledge and priori knowledge of tool wear, this study proposes a fusion model of tool wear based on data-driven and physical constraints. The loss function analysis reveals that the network structure of the fusion model is constrained by both the data labels and the physical model during iterative training. To validate the feasibility of using physical model constraints as a new solution to the tool wear problem proposed in this study, the improvement in prediction accuracy and robustness of the fusion model with respect to the data-driven model is validated. In this study, several sets of milling experiments of Ni-based superalloy were designed to collect signals in the milling process as well as its corresponding tool wear value labels. Based on this, cross-validation experiments were designed between the data-driven model and the fusion model.

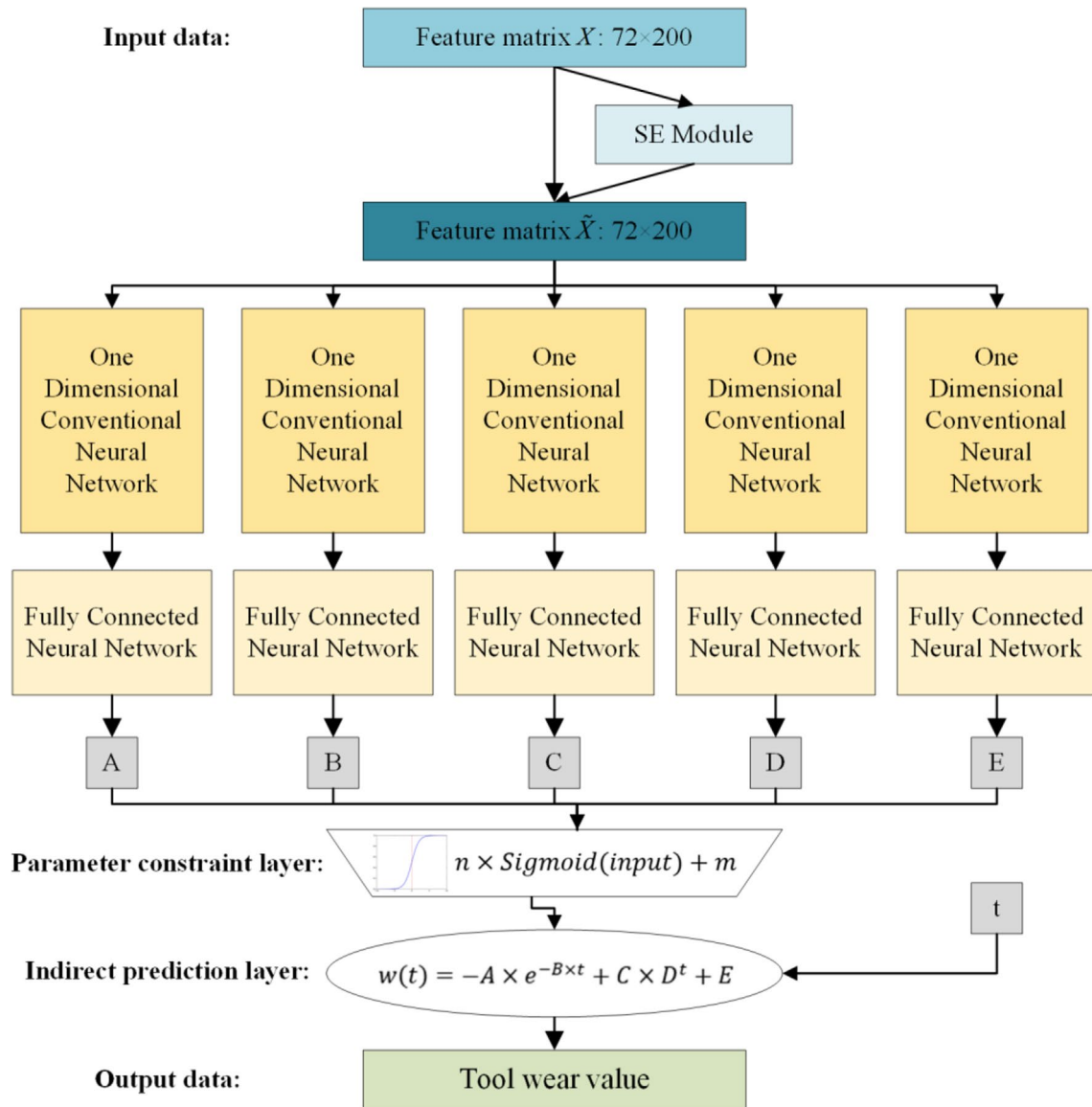
## 3.1 Experiment platform construction

The milling experiment of Ni-based superalloy in this study was carried out on YHVT850Z four-axis machining center; the workpiece material is Ni-based superalloy GH4169 with the dimensions of 196 mm × 120 mm × 16 mm; the nano-coated insert APKT11T304-APF with ultrafine cemented carbide and the toolholder EMP01-016-G16-AP11-02 with 16 mm diameter were selected for down milling and cooled by using the cutting fluid, with one insert installed on the toolholder; a total of 10 cases of milling experiments with variable parameters were designed in this study with different combinations of the cutting widths and cutting depths, and the experiment parameters are detailed in Table 4. In the experiment procedure, a total of 10 new inserts were used. The milling tool cut 120 mm in the X direction for each cycle, and the inserts were measured every three cuts. Tool wear values were measured and recorded using an Alicona G4 device, as illustrated in Fig. 6. The process was repeated until the tool wear value exceeded 100 μm, and a new insert was substituted for the subsequent set of experiments, continuing until all 10 sets of milling experiments were completed. In each experiment set, the acquisition of sensor signals was synchronized. KISTLER 9255B dynamometer was employed to collect milling force signals in the x, y, and z directions during milling, and the Dytran 5225F1 acceleration sensor was used to collect vibration signals. These signals were gathered through the multi-channel data acquisition device UEI DNR-12-1G, with a sampling frequency of 20 kHz. The experiment scheme is shown in Fig. 7.

A total of 185 experiment samples were gathered from 10 milling experiments. Each experiment sample included two channels of force signals, one channel of vibration signals and the corresponding tool wear values. As the inserts used in the experiments were new, the tool wear value of the first sample in each experiment group was set to 0 μm by default. The number of samples for each working condition is detailed in Table 5.

## 3.2 Parametric sensitivity analysis of the physical model

Multi-column neural networks share the same learning rate for each column when not preset. However, model parameters with the same learning step exhibit varying degrees of influence on the model output. Thus, it is essential to analyze the sensitivity of each parameter of the physical model  $w(t)$  to adjust the learning rate for each column of the network. Sensitivity analysis is performed through sub-equations, that is, the degree of change in the model output caused by changing the model parameters. The relative error is equal to the absolute error divided by the variable; the absolute error has a positive or negative value, and the relative error has no positive or negative value. Therefore, all equations are default to absolute values.



**Loss function:**  $loss(y, \hat{y}) = MSE(y, \hat{y}) = \frac{1}{N} \sum (y - \hat{y})^2 = \frac{1}{N} \sum (w(t) - \hat{y})^2$

Fig. 5 Fusion model structure

Three sub-equations of the model (6) can be simplified as

$$\begin{cases} w^1(t) = ae^{bt} \\ w^2(t) = cd^t \\ w^3(t) = e \end{cases} \quad (10)$$

The absolute error is calculated from the error transfer equation  $\epsilon(y) = \sum \frac{\partial y}{\partial x_i} \epsilon(x_i)$ :

$$\begin{cases} \epsilon(w^1(t)) = e^{bt} \times \epsilon(a) + ate^{bt} \times \epsilon(b) \\ \epsilon(w^2(t)) = d^t \times \epsilon(c) + ct d^{t-1} \times \epsilon(d) \\ \epsilon(w^3(t)) = \epsilon(e) \end{cases} \quad (11)$$

The relative error is derived as

$$\begin{cases} \delta(w^1(t)) = \frac{1}{a} \times \epsilon(a) + t \times \epsilon(b) \\ \delta(w^2(t)) = \frac{t}{c} \times \epsilon(c) + \frac{t}{d} \times \epsilon(d) \\ \delta(w^3(t)) = \frac{1}{e} \times \epsilon(e) \end{cases} \quad (12)$$

**Table 4** Experiment parameters

Experiment case	Feed per tooth (mm/tooth)	Spindle speed (rpm)	Cutting speed (m/min)	Radial depth (mm)	Axial depth (mm)
C1	0.1	796	40	0.7	0.4
C2	0.1	796	40	0.7	0.6
C3	0.1	796	40	0.7	0.8
C4	0.1	796	40	0.6	0.4
C5	0.1	796	40	0.6	0.6
C6	0.1	796	40	0.6	0.8
C7	0.1	796	40	0.5	0.6
C8	0.1	796	40	0.5	0.8
C9	0.1	796	40	0.4	0.6
C10	0.1	796	40	0.4	0.8

**Table 5** Number of samples for each working condition

Experiment case	Sample number
C1	17
C2	13
C3	15
C4	16
C5	15
C6	15
C7	15
C8	15
C9	15
C10	15

Taking the relative error of sub-equation  $w^1(t)$  as an example, the first term in the equation is only related to the absolute error of  $a$  and the magnitude of  $a$ . The smaller the relative error of  $a$ , the smaller the relative error of  $w(t)$  ( $a$  is a constant term); the second term is related to the absolute error of  $b$  and the magnitude of  $t$ . The larger  $t$  is, the larger the relative error of  $w(t)$  is.

$$\begin{cases} \delta(w^1(t)) = \frac{1}{a} \times \varepsilon(a) + t \times \varepsilon(b) = \delta(a) + t \times \varepsilon(b) \\ \delta(w^2(t)) = \frac{1}{c} \times \varepsilon(c) + \frac{t}{d} \times \varepsilon(d) = \delta(c) + \frac{t}{d} \times \varepsilon(d) \\ \delta(w^3(t)) = \delta(e) \end{cases} \quad (13)$$

The independent variable  $t$  represents cutting time, cutting length, or the number of cuts in the model. To maintain the

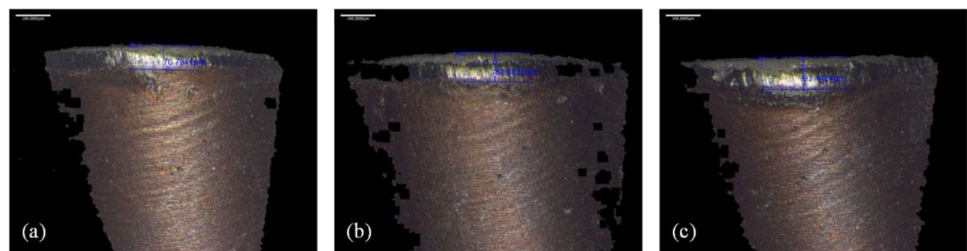
balanced influence of the learning rate on the convergence speed of each model parameter during network training, the learning rate LR2 for the model parameters  $B$  and  $D$  in the network column should be lower than that of the learning rate LR1 of the model parameters  $A$ ,  $C$ , and  $E$ . The ratio of the learning rates is set as

$$\frac{LR1}{LR2} = 10 \quad (14)$$

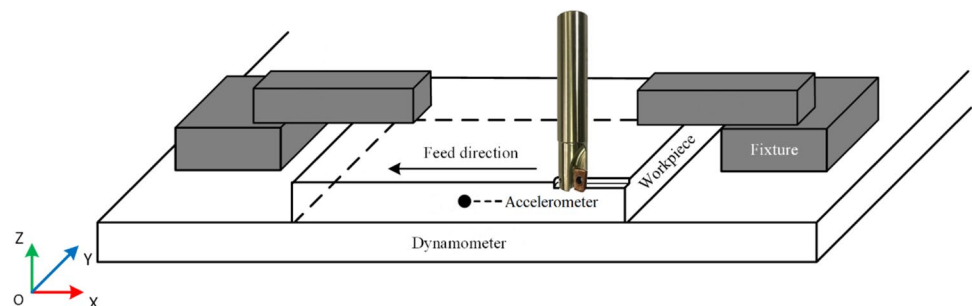
### 3.3 Model training setup

To evaluate the variance between the data-driven model and the fusion model regarding prediction capability, the 10 milling experiments are grouped into G1 (C1, C2, C3), G2 (C4, C5, C6), and G3 (C7, C8, C9, C10). The study employed

**Fig. 6** Tool wear value measurement  $a_p=0.4$  mm and  $a_e=0.7$  mm: (a) cutting length=5400 mm, (b) cutting length=10,800 mm, and (c) cutting length=16,200 mm



**Fig. 7** Experiment scheme



the cross-validation strategy to construct the training and test samples, as detailed in Table 6. To evaluate and optimize the performance of the model, 80% of the samples are randomly selected from the training samples as the training set and 20% of the samples as the validation set. The data-driven model and the fusion model take the features extracted from monitoring signals as inputs. Utilizing the sliding window method, 12 time domain features, 4 frequency domain features, and 8 time–frequency domain features are extracted from the monitoring signals of the three channels, resulting in a final feature matrix shape of (185, 72, 200). The network output of the data-driven model is the predicted tool wear value, which can be directly used to calculate the loss function with the labeled values. On the other hand, the network output of the fusion model is the 5 parameters of the physical model, which need to be inputted into the indirect prediction layer along with the machining time  $t$  to obtain the predicted tool wear value and then calculate the loss function with the labeled values. The learning rates for the 5-column network in the fusion model are set to 0.0001, 0.00001, 0.0001, 0.00001, 0.0001, and 0.0001, and the learning rate of the network in the data-driven model is set to 0.00001. Both models employ the Adam gradient descent algorithm [46] with 50 network iterations.

To quantitatively compare the predictive ability of the model, three indicators are chosen to evaluate the prediction results of the model. The accuracy indicator is used to evaluate the accuracy of the predicted value and the labeled value; the closer to 100%, the better. Root mean square error (RMSE) is used to measure the deviation between the predicted value and the labeled value, which is sensitive to the outliers in the data; the smaller, the better. Mean absolute error (MAE) is used to represent the average of the absolute error between the predicted value and the labeled value; the smaller, the better. Each training situation is repeated three times, and the corresponding indicator takes the average of the three times. The expressions of the three indicators are as follows:

$$\left\{ \begin{array}{l} \text{Accuracy} = 1 - \frac{\sum_{i=1}^N |w_i - \hat{w}_i|}{\sum_{i=1}^N |w_i|} \\ \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (w_i - \hat{w}_i)^2} \\ \text{MAE} = \frac{1}{N} \sum_{i=1}^N |w_i - \hat{w}_i| \end{array} \right. \quad (15)$$

**Table 6** Division of training and test samples

Case study	Training sample	Test sample
Case 1	G1 + G2	G3(C7, C8, C9, C10)
Case 2	G1 + G3	G2(C4, C5, C6)
Case 3	G2 + G3	G1(C1, C2, C3)

where  $w_i$  represents the labeled value of tool wear,  $\hat{w}_i$  represents the predicted value of tool wear, and  $N$  represents the number of test samples.

## 4 Analysis and validation

### 4.1 Cross-validation

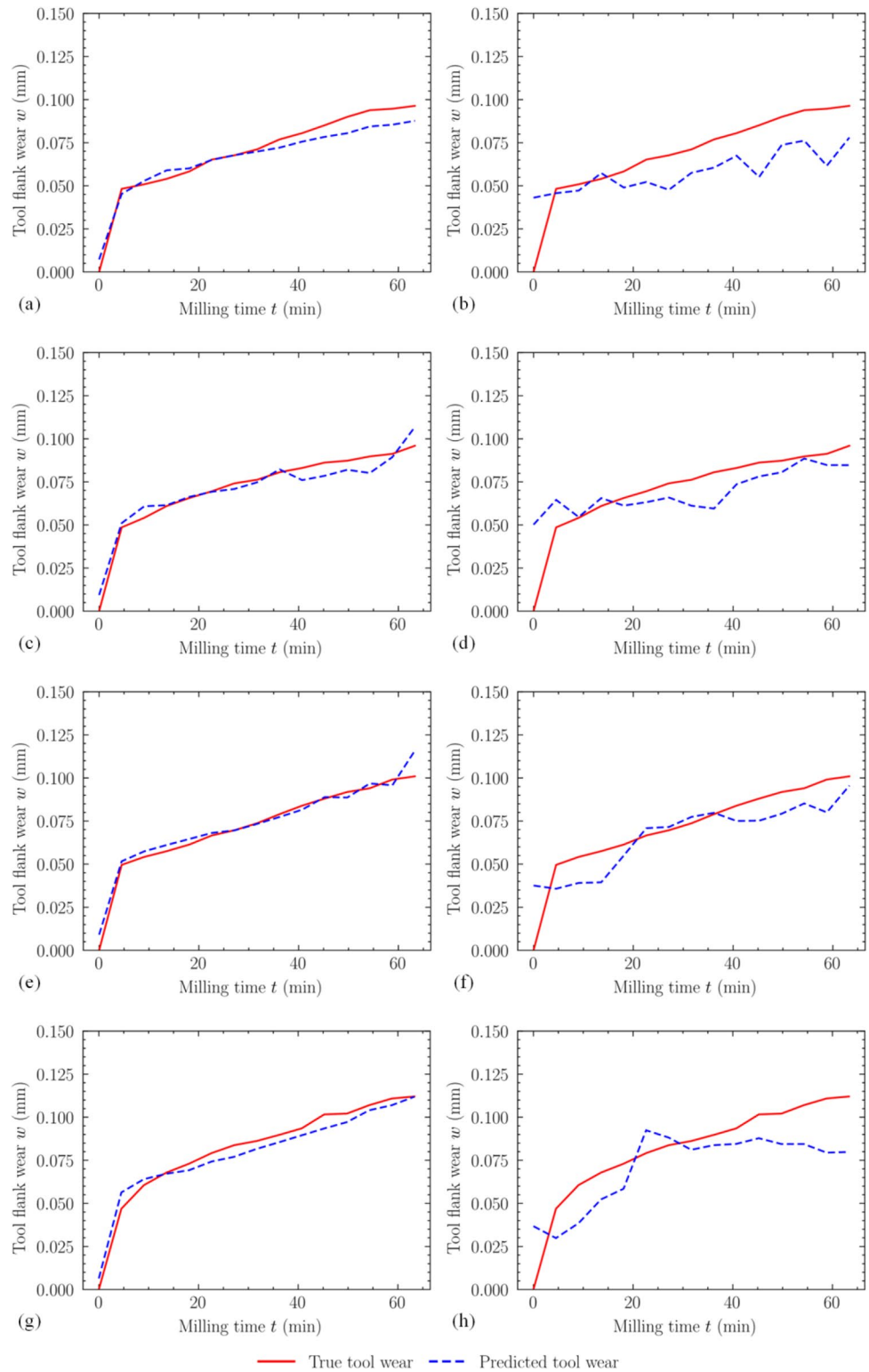
To compare the prediction results of using only data to constrain the network with those of using both data and physical model to constrain the network, three training cases (G1G2-G3, G1G3-G2, and G2G3-G1) are devised. These cases are created using the cross-validation strategy, controlling the input data, network structure, and hyperparameter settings for both the data-driven and fusion models, which are the same. The training results are shown in Figs. 8, 9, and 10.

The evaluation indicators of the training results are illustrated in Fig. 11, where the fusion model is compared with the data-driven model, and the accuracy is improved by 11.5%, with an improvement rate of 14.2%; the RMSE is reduced by 0.00690, with an improvement rate of 52.4%; and the MAE is reduced by 0.00835, with an improvement rate of 61.4%. These indicators reveal a significant improvement in prediction performance with the addition of physical constraints. Upon observing the training results of the two models, it is evident that the prediction results of the fusion model align more closely with the actual tool wear trend and are in line with the design trend of the physical model.

### 4.2 Effects of indirect prediction layer

When a feature matrix is input into the trained model, the data-driven model predicts a tool wear value, while the fusion model directly predicts a set of parameter values from the physical model. By inputting the parameter values and the corresponding cutting time  $t$  into the indirect prediction layer, the fusion model indirectly obtains a predicted tool wear value. This tool wear value is positioned on a tool wear curve jointly constructed by the physical model  $w(t)$  and the predicted physical parameter values. Its specific position uniquely corresponds to the cutting time  $t$ . The tool wear curve co-constructed from the physical model  $w(t)$  and the physical parameter values predicted by the fusion model is termed a reconstructed tool wear curve, in the sense that it is a confidence curve of the tool wear over the entire life cycle inferred by the trained fusion model from the feature matrix, where the known cutting time  $t$  corresponds to the predicted tool wear value and the entire confidence curve corresponds to the complete predicted tool wear curve. Obtaining one tool wear confidence curve requires one feature matrix, and multiple confidence curves obtained from

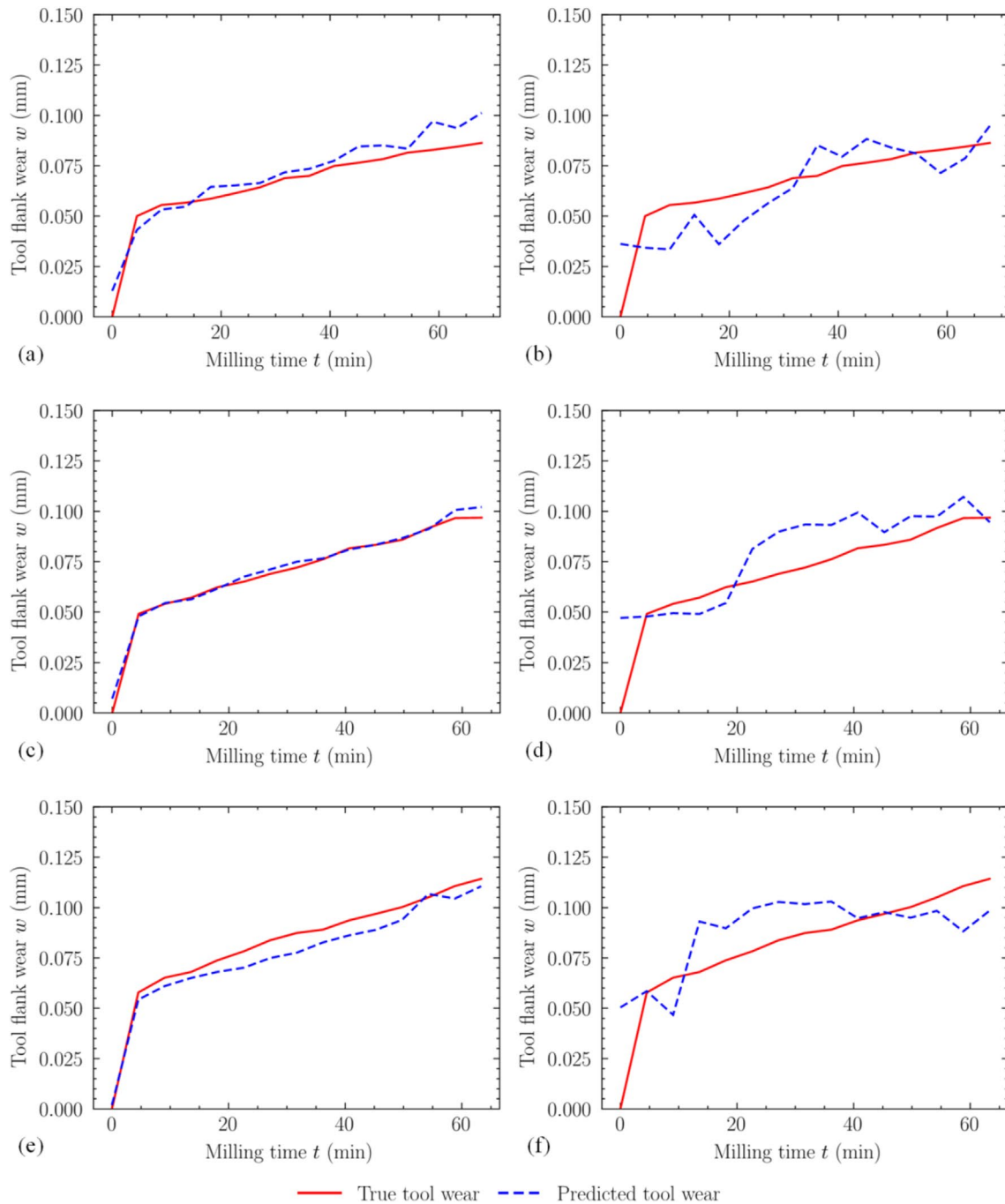
**Fig. 8** Case 1 comparison of training results: (a) fusion model G1G2-C7, (b) data-driven model G1G2-C7, (c) fusion model G1G2-C8, (d) data-driven model G1G2-C8, (e) fusion model G1G2-C9, (f) data-driven model G1G2-C9, (g) fusion model G1G2-C10, and (h) DDM G1G2-C10



multiple feature matrices together constitute the confidence interval for the tool wear prediction results. This confidence interval acts as the tool wear prediction solution space of the fusion model, and its trend is constrained by the physical model  $w(t)$ , which is related to  $t$ . Therefore, the confidence

interval is curved, and the confidence curve within it exhibits monotonically increasing properties, aligning with the trend of the physical model.

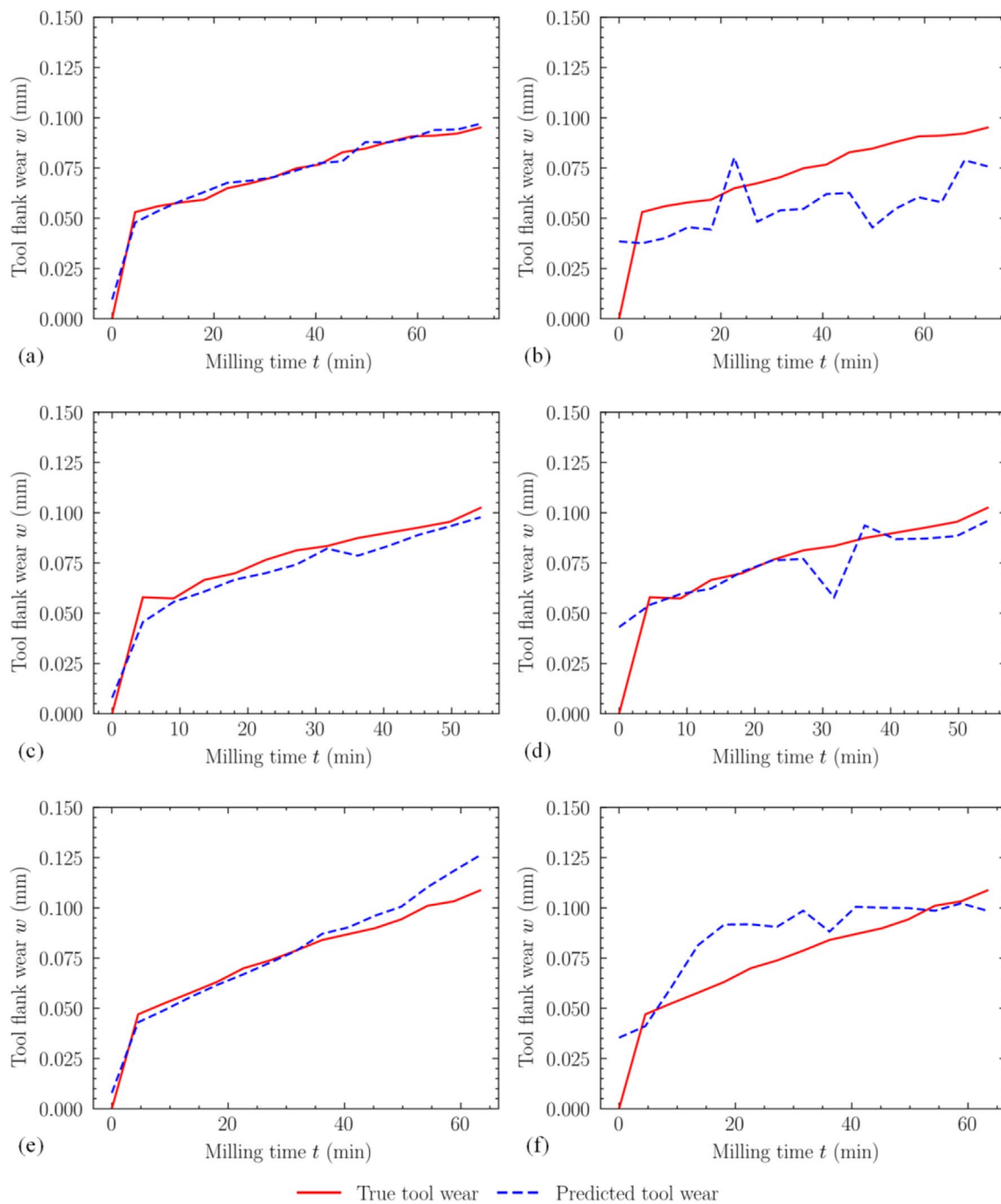
Comparing the network structure of the data-driven model with that of the fusion model, the data-driven model



**Fig. 9** Case 2 comparison of training results: (a) fusion model G1G3-C4, (b) data-driven model G1G3-C4, (c) fusion model G1G3-C5, (d) data-driven model G1G3-C5, (e) fusion model G1G3-C6, and (f) data-driven model G1G3-C6

can be regarded as the fusion model with only the constant term indirect prediction layer. The equivalent structure of the data-driven model is illustrated in Fig. 12. Reconstructing with the constant term indirect prediction layer yields multiple confidence curves, forming the confidence interval of the data-driven model prediction results. This interval represents the solution space for the tool wear prediction of the data-driven model, which is independent of  $t$ . Consequently,

the confidence interval is rectangular, and the confidence curve is no longer monotonically increasing. This leads to a much larger range of confidence intervals compared to that of the fusion model. The addition of the physical model as a constraint significantly reduces the area of the confidence interval of the fusion model, and the confidence interval is accurately concentrated around the predicted value. By comparing the confidence intervals of the fusion model and



**Fig. 10** Case 3 comparison of training results: (a) fusion model G2G3-C1, (b) data-driven model G2G3-C1, (c) fusion model G2G3-C2, (d) data-driven model G2G3-C2, (e) fusion model G2G3-C3, and (f) data-driven model G2G3-C3

the data-driven model, reconstructed using the training case of G1G2-C9 as an example, it can be found that the confidence interval of the fusion model, after incorporating the physical model  $w(t)$ , is confined to a relatively small range, as shown in Fig. 13. This not only aligns with the trend of the physical model  $w(t)$  but also closely resembles the real tool wear curve. The improvement in prediction accuracy

of the fusion model is attributed to the constraints on the area of the confidence interval. Calculating the confidence interval area reveals that, in the training case G1G2-C9, the confidence interval area of the fusion model is reduced by 73.7% compared to the data-driven model in the results of tool wear curve reconstruction. The corresponding accuracy is improved by 10.1%, and the improvement rate is 12.1%.



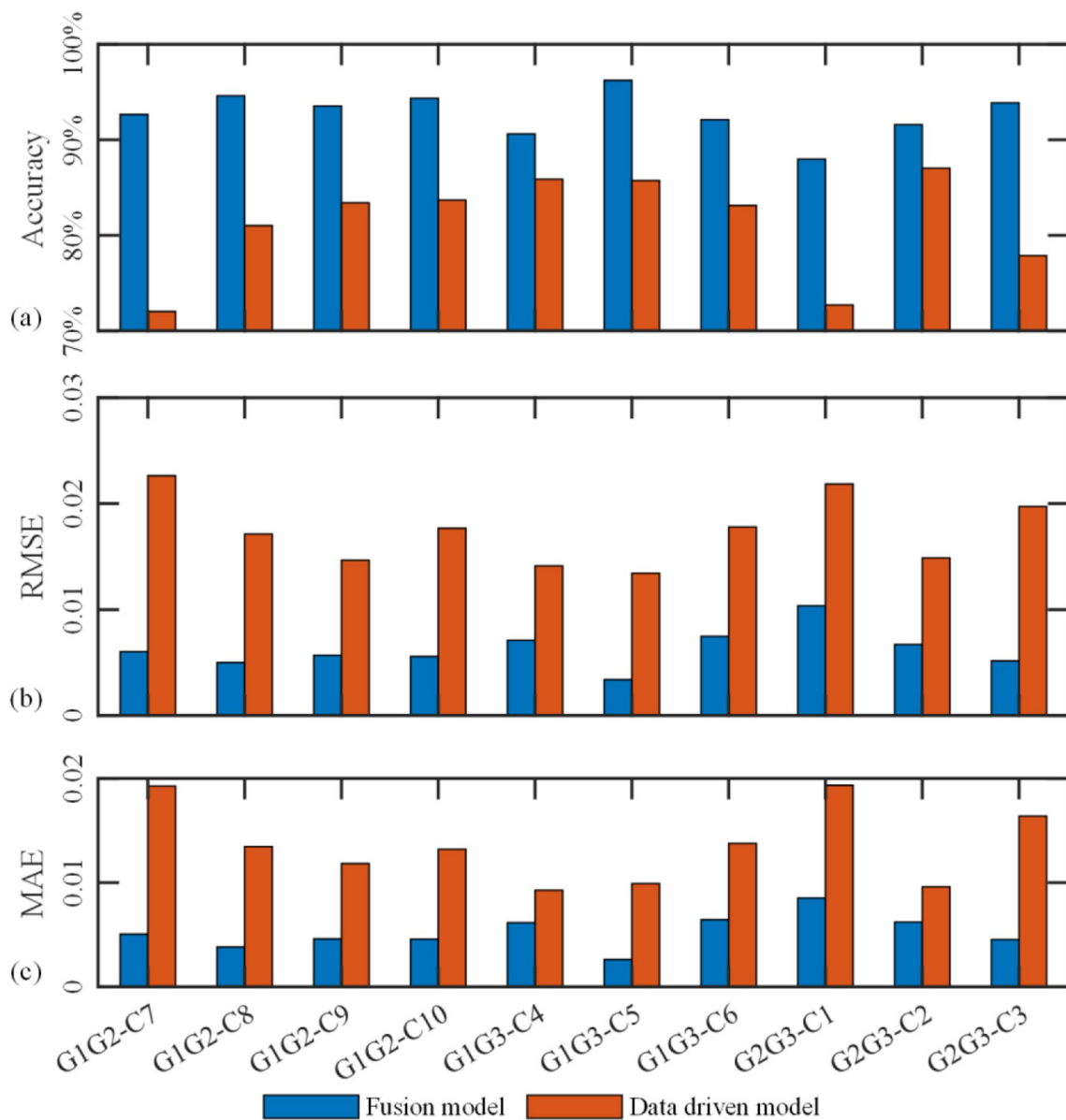


Fig. 11 Comparison of evaluation indicators for fusion and data-driven models: (a) accuracy, (b) RMSE, and (c) MAE

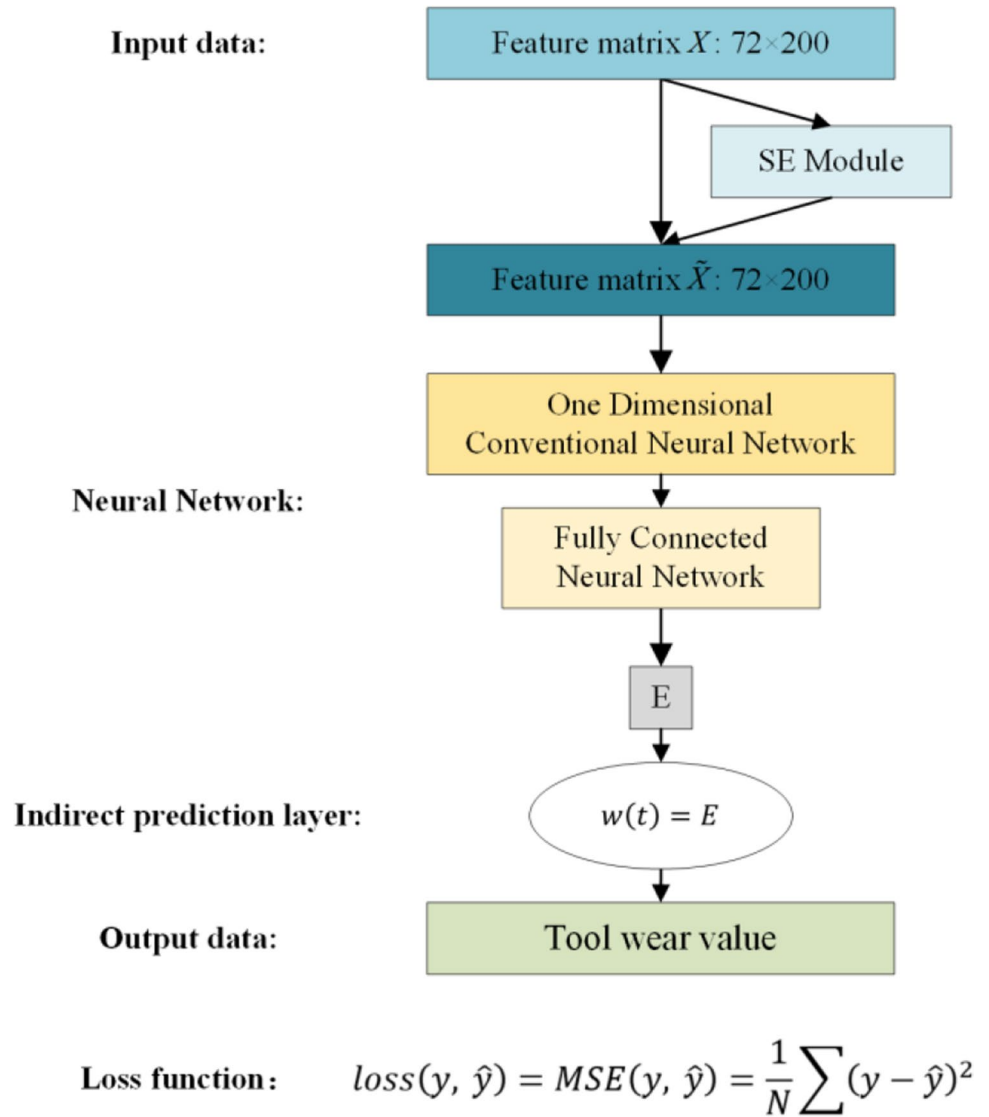
### 4.3 Effects of parameter constraint layer

To evaluate the effect of the parameter constraint layer on the fusion model, the layer is removed. To ensure the non-negativity of the inputs to the indirect constraint layer, the parameter constraint layer before the output layer of the 5-column network structure is replaced with the Relu function. While keeping other training conditions unchanged to control variables, the training results of the fusion model with parameter constraint layers and without parameter constraint layers are compared using G1G2-C9 as an example. The training results as well as the reconstructed

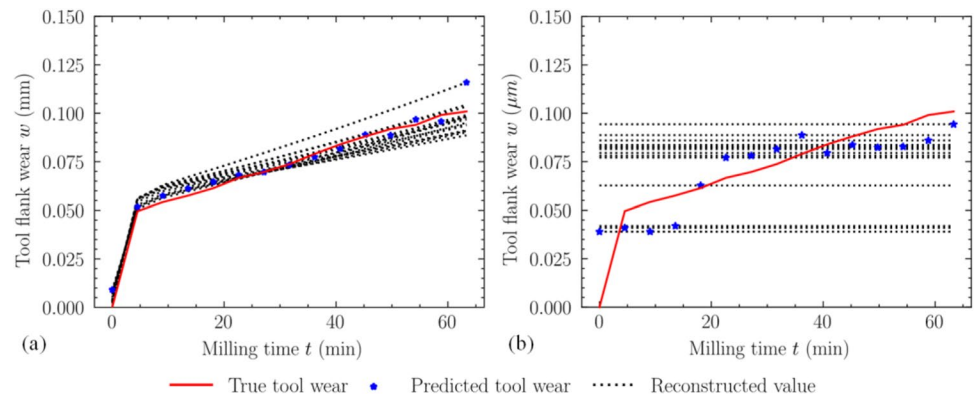
tool wear curves are illustrated in Fig. 14, where Fig. 14a is compared with Fig. 8e, which is the training results without parameter constraint layer. Figure 14b is compared with Fig. 13a, which is the reconstructed tool wear curve without parameter constraint layer, while the comparison of evaluation indicators is illustrated in Fig. 15. The addition of the parameter constraint layer reduces the area of the fusion model confidence interval, and its shape is closer to the theoretical physical model trend.

When comparing the fusion model with the parameter constraint layer against the one without the parameter constraint layer, the accuracy is improved by 2.49%, with

**Fig. 12** Data-driven model equivalent structure



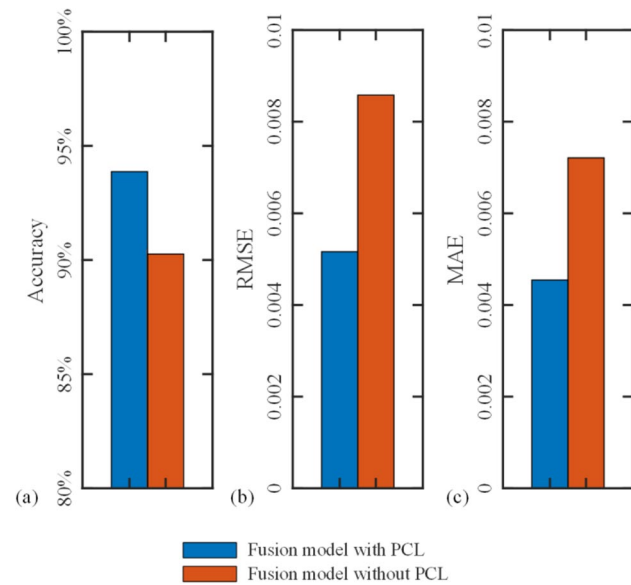
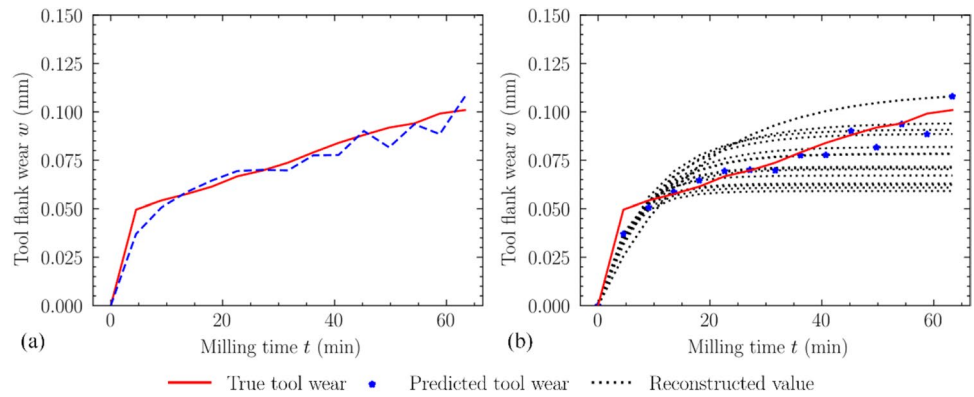
**Fig. 13** Reconstructed tool wear curve: (a) fusion model and (b) data-driven model



an improvement rate of 2.76%; the RMSE is reduced by 0.00233, with an improvement rate of 27.1%; the MAE is reduced by 0.00195, with an improvement rate of 27.1%.

The indicators reveal that incorporating the parameter constraint layer significantly improves the prediction performance of the fusion model.

**Fig. 14** (a) Training results without parameter constraint layer. (b) Reconstructed tool wear curve without parameter constraint layer



**Fig. 15** Comparison of evaluation indicators between fusion model with parameter constraint layer and fusion model without parameter constraint layer: (a) accuracy, (b) RMSE, and (c) MAE

### 4.4 Validation with the public data set

#### 4.4.1 Forward prediction validation

To analyze the effectiveness of the tool wear fusion model for predicting tool wear throughout the life cycle of the tool, as well as its generalisability for predicting tool wear in other cutting conditions, the public tool wear dataset, PHM2010, is introduced in this study for forward prediction experiment. The Prognostics and Health Management Society provided a public tool wear dataset PHM2010 for the Data Challenge in 2010; 315 cuts were performed for each experiment set, and a total of the following 6 sets of experiments were conducted: C1, C2, C3, C4, C5, and C6. In this study, we focused on the initial 300 samples of C1. These samples were collected using a sliding window,

**Table 7** Division of training and test samples

Case study	Training sample	Test sample
Case 1	1:50	51:100
Case 2	1:100	101:150
Case 3	1:150	151:200
Case 4	1:200	201:250
Case 5	1:250	251:300

extracting features from each of the 6 channels (the three directions of cutting force signals and three directions of vibration signals). The features included 12 time domain features, 4 frequency domain features, and 8 time–frequency domain features, resulting in a final feature matrix with dimensions (300,144,200). A forward prediction experiment was designed with 300 samples. The first 50 samples were employed for model training, and the subsequent 50 samples for testing. This procedure was iterated, combining all samples from previous training cases as training samples and using the next 50 samples as test samples. This cycle continued until all 300 samples were used, resulting in five training cases detailed in Table 7. To evaluate and optimize the performance of the model, 80% of the samples are randomly selected from the training samples as the training set and 20% of the samples as the validation set.

The forward prediction experiments involved 24 features distributed across the 6 channels of the public dataset. This leads to alterations in the channels of the input feature matrices for both the data-driven and fusion models. Considering the changes in the cutting conditions, the parameter constraint layer properties of the fusion model in the forward prediction experiments were redefined using the priori knowledge from the public dataset, that is, the tool wear data in the C4 and C6 sets; the new parameter constraint layer properties are shown in Table 8.

To compare the prediction results of using only data to constrain the network with the prediction results of

**Table 8** Parameter constraint layer properties

Layer $j$	$n_j$	$m_j$
A	100	0.01
B	2	0.01
C	20	0.01
D	0.2	1.01
E	5	0.01

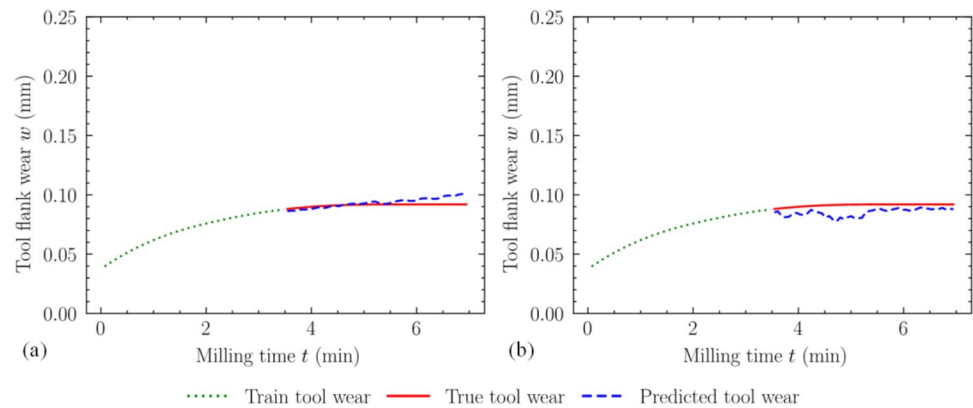
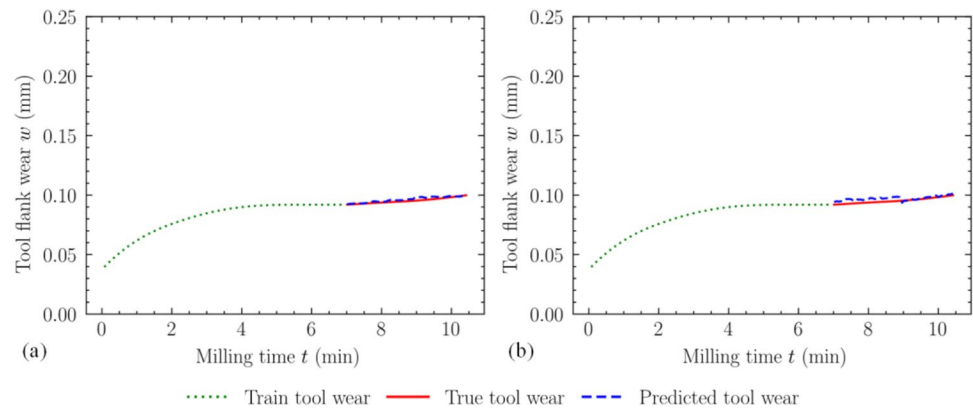
using both data and physical model to constrain the network throughout the tool lifecycle, after controlling the input data, network structure, and hyperparameter settings of the data-driven and fusion models to be the same, the five training cases designed using the forward prediction strategy are 1:50–51:100, 1:100–101:150, 1:150–151:200, 1:200–201:250, and 1:250–251:300, and the training results are illustrated in Figs. 16, 17, 18, 19, and 20.

The evaluation indicators of the training results are illustrated in Fig. 21, where the fusion model is compared with the data-driven model; the accuracy is improved by 1.84%, with an improvement rate of 1.92%; the RMSE is reduced by 0.00259, with an improvement rate of 45.4%; and the MAE is reduced by 0.00194, with an improvement rate of 41.9%. From the indicators, the addition of physical constraints can

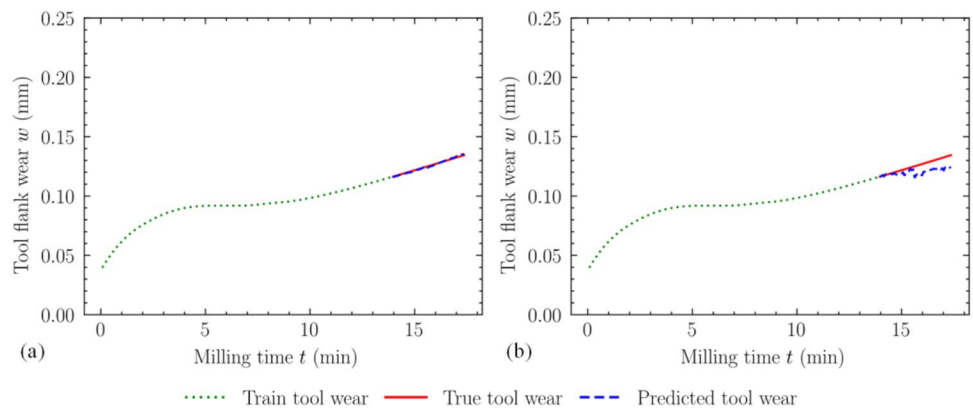
significantly improve the prediction performance of the network. Observing the training results of the two models, it is evident that the trend of the prediction results of the fusion model is closer to the trend of the actual tool wear and is in line with the design trend of the physical model.

Furthermore, the five training cases completely cover the three stages of tool wear: initial wear, normal wear, and severe wear. Notably, in case 3, the fusion model exhibits the least difference in prediction performance compared to the data-driven model. The difference in prediction performance steadily decreases steadily from case 1 to case 3, whereas it progressively increases from case 3 to case 5. The trend of the combined tool wear curves is used to analyze this phenomenon:

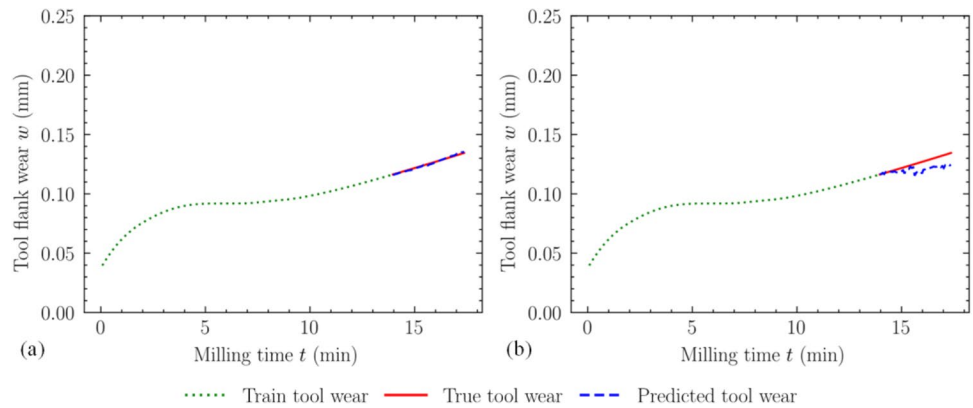
The 1:50 and 51:100 samples correspond to the transition process of tool wear from initial to normal wear, so the fusion model and the data-driven model in case 1 and case 2 learn the trend of tool wear from the initial and normal wear stages and predict it in the normal wear stage. From the prediction results, the difference in prediction performance comes from the fact that the predicted tool wear curve of the fusion model is smoother and more stable, while the predicted tool wear curve of the data-driven model is more fluctuating.

**Fig. 16** Case 1 comparison of training results: (a) fusion model and (b) data-driven model**Fig. 17** Case 2 comparison of training results: (a) fusion model and (b) data-driven model

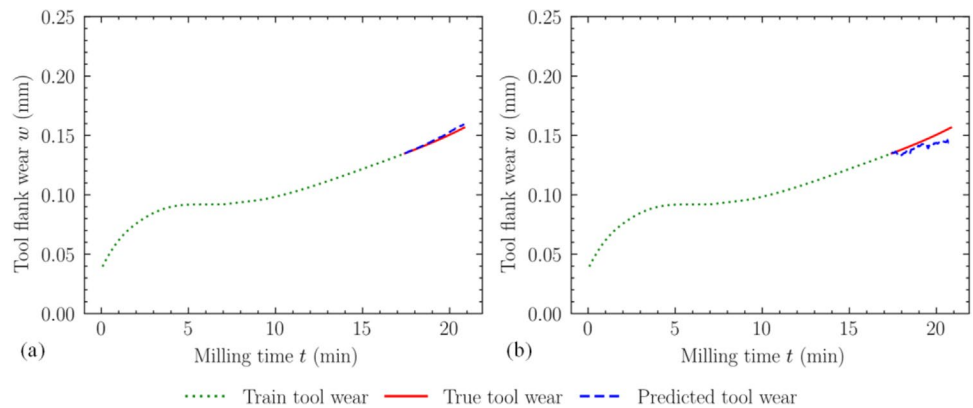
**Fig. 18** Case 3 comparison of training results: (a) fusion model and (b) data-driven model



**Fig. 19** Case 4 comparison of training results: (a) fusion model and (b) data-driven model

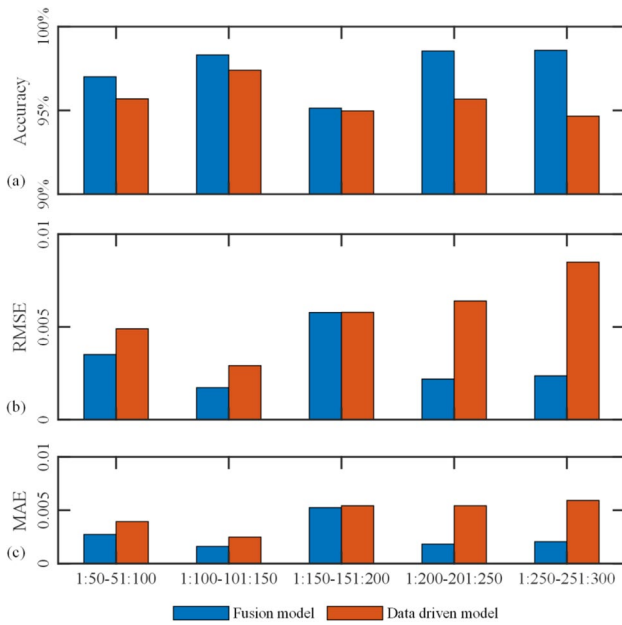


**Fig. 20** Case 5 comparison of training results: (a) fusion model and (b) data-driven model



The 101:150 sample corresponds to the transition from normal to severe tool wear, with an initial but insignificant tendency to accelerate wear. Therefore, the fusion model and the data-driven model in case 3 learn the trend of tool wear from the initial wear, normal wear, and initial severe wear stages and make predictions at the severe wear stage, which leads to a serious deviation of the predicted tool wear values from the actual tool wear values, and both underestimate the accelerated process of tool wear.

The samples of 151:200 and 201:250 correspond to the severe wear stage of the tool, and the prediction performance gap between the fusion model and the data-driven model increases in case 4 and case 5. From the prediction results, the reason that the fusion model possesses better prediction performance is that the fusion model is better able to learn the tool wear trend in the severe wear stage from the train samples, while the data-driven model



**Fig. 21** Comparison of evaluation indicators for fusion and data-driven models: (a) accuracy, (b) RMSE, and (c) MAE

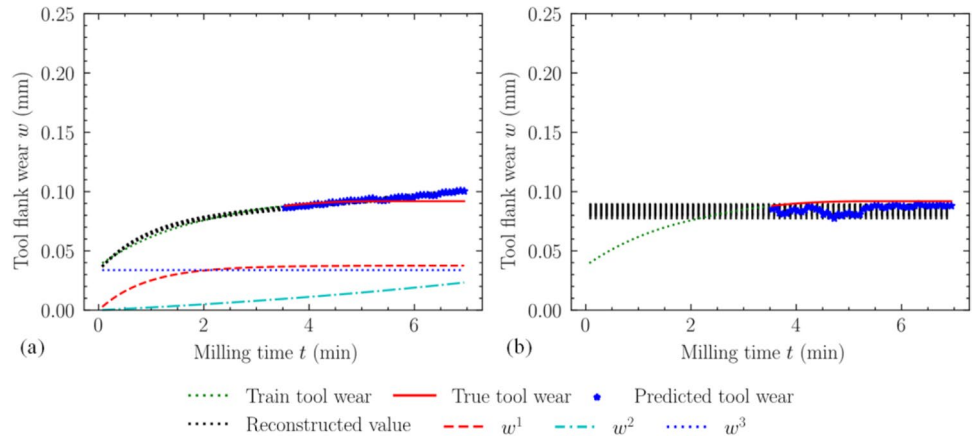
has difficulty in learning the accelerated wear trend in the severe wear stage.

#### 4.4.2 Reconstruction and decomposition of tool wear curves

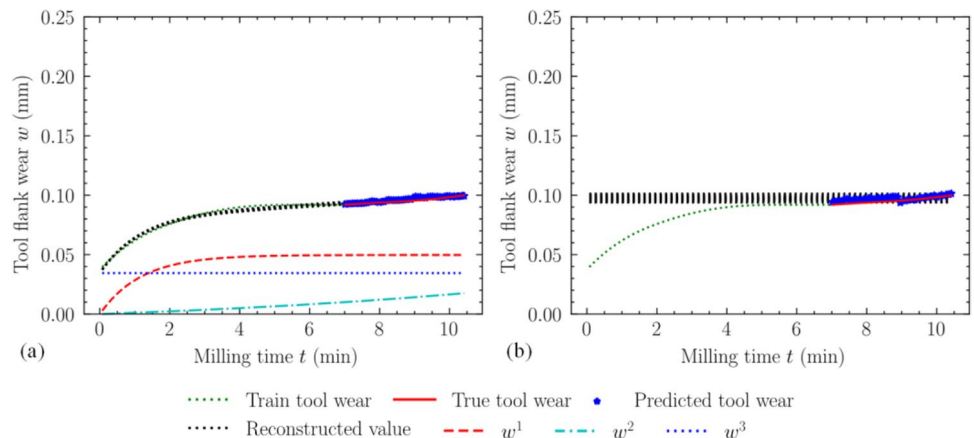
To further validate the effect of physical constraints on the tool wear prediction results, the tool wear curves are reconstructed using the same manner as in the “Effects of indirect prediction layer” section for the prediction results of the fusion model and the data-driven model in the forward prediction experiment, as illustrated in Figs. 22, 23, 24, 25, and 26. In particular, since the physical model consists of three sub-equations  $w^1(t)$ ,  $w^2(t)$ , and  $w^3(t)$  of  $w(t)$ , that is,  $w(t) = w^1(t) + w^2(t) + w^3(t)$ , the prediction results of the fusion model can be decomposed into the form of three sub-equations, which makes it possible to analyze the weight share of the different sub-equations in the prediction results of the fusion model more visually.

From the reconstructed tool wear curves of the fusion model, unlike the reconstructed tool wear curves of the cross-validation experiments, which are curved regions around the actual tool wear curves, the reconstructed tool wear curves of the forward prediction experiments are more

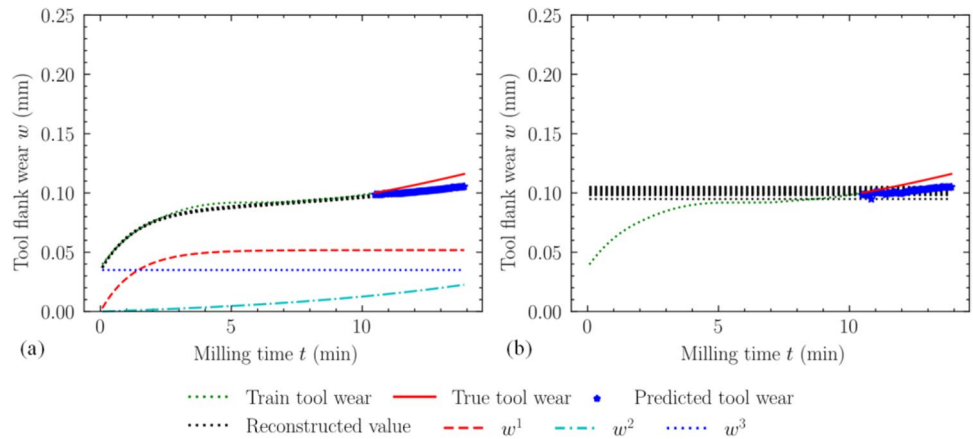
**Fig. 22** Case 1 comparison of reconstructed tool wear curves: (a) fusion model and (b) data-driven model



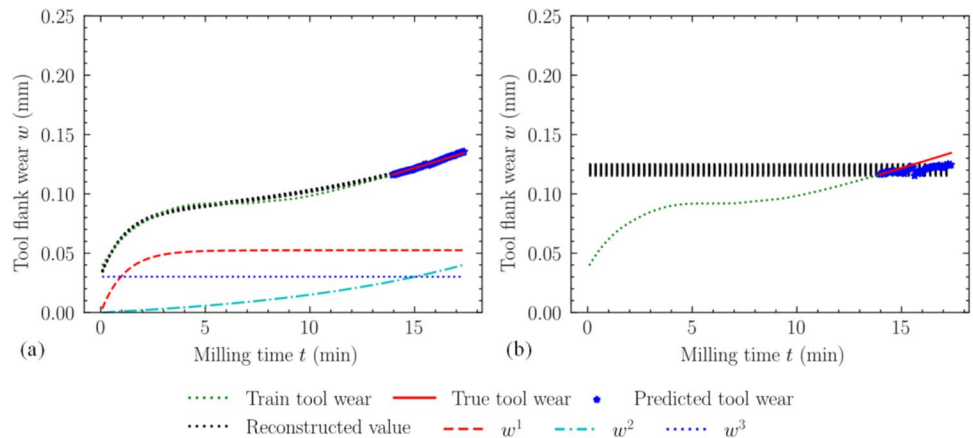
**Fig. 23** Case 2 comparison of reconstructed tool wear curves: (a) fusion model and (b) data-driven model



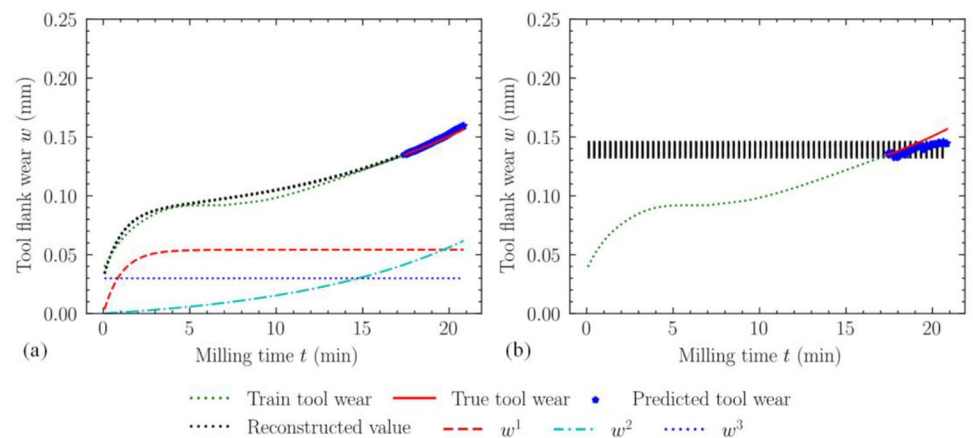
**Fig. 24** Case 3 comparison of reconstructed tool wear curves: (a) fusion model and (b) data-driven model



**Fig. 25** Case 4 comparison of reconstructed tool wear curves: (a) fusion model and (b) data-driven model



**Fig. 26** Case 5 comparison of reconstructed tool wear curves: (a) fusion model and (b) data-driven model



concentrated, close to a single curve. Therefore, it can be concluded that the reason why the fusion model can achieve better prediction performance is that, due to the constraints of the physical model, the fusion model can already have a better learning of the overall trend of the tool wear process from the training samples, and then, based on the existing tool wear trend, it can better obtain the prediction results that are close to the actual tool wear curve.

Observe the trends of the three sub-equations  $w^1(t)$ ,  $w^2(t)$ , and  $w^3(t)$  in the five training cases. Overall,  $w^1(t)$  controls the slope of the initial curve from large to small, and then the slope converges to zero, which has no effect on the slopes of the subsequent curves.  $w^2(t)$  controls the process of the slope of the subsequent stage from stable to gradually increasing, and has relatively little effect on the slope of the initial curve.  $w^3(t)$  controls the initial value of the tool wear curve

and affects the overall position of the curve. The prediction results of the fusion model achieve the modeling purpose of the physical model  $w(t)$  well.

Specifically, for each training case, the training samples for case 1 and case 2 focus on the initial wear and normal wear stages, with  $w^1(t)$  accounting for a large weighting of the prediction results,  $w^2(t)$  having negligible weighting, and  $w^3(t)$  being close to the initial tool wear value. The training samples for case 3 include transition samples from normal to severe wear, and the weight of  $w^2(t)$  begins to increase. The training samples of case 4 and case 5 cover the initial wear, normal wear, and severe wear stages, and the weight of  $w^2(t)$  increases significantly. Overall, based on the functional properties of the three sub-equations  $w^1(t)$ ,  $w^2(t)$ , and  $w^3(t)$  of the physical model  $w(t)$ , the fusion model can learn the trend of the tool wear process well from the training samples and thus can accurately predict the tool wear values in the forward prediction experiment.

## 5 Conclusion

In this paper, to apply domain knowledge in the tool wear field and knowledge from the accumulated data to the tool wear prediction, an indirect prediction layer and a parameter constraint layer are designed to transform the traditional data-driven model into a tool wear fusion model that is subject to both data-driven and physical constraints. The indirect prediction layer is designed from the physical model, whose inputs are the model parameters and corresponding time  $t$  computed by the multi-column network, and the outputs are the predicted tool wear value; the parameter constraint layer is designed, and the properties of the parameter constraint layer are determined from the accumulated tool wear data to constrain the distribution range of inputs to the indirect constraint layer. As a result, the predictions of the fusion model are constrained by both the data labels and the physical model, resulting in improved predictive ability and interpretability relative to the data-driven model. The effectiveness of the methodology is validated in the GH4169 milling and the dataset PHM2010. The main conclusions are as follows.

- (1) The results of cross-validation experiments based on the GH4169 milling experimental data show that the prediction performance of the fusion model exceeds that of the data-driven model, with an average prediction accuracy increase of 11.5%, by effectively utilizing the tool wear domain knowledge and the available prior knowledge.
- (2) Based on the physical model in the indirect prediction layer, the prediction results of the fusion model and the data-driven model are reconstructed to obtain con-

fidence intervals. Comparing the area and shape of the confidence intervals, the confidence interval area of the fusion model is reduced by 73.7% relative to that of the data-driven model, with the corresponding prediction accuracies improved by 10.1%. It is verified that the physical model introduced in this study can effectively constrain the range and shape of the confidence interval of the prediction results, thus improving the prediction accuracy.

- (3) By comparing the prediction results of the fusion model with a parameter constraint layer and the fusion model without a parameter constraint layer, the confidence interval area of the fusion model decreases and the shape of the fusion model is closer to the trend of the theoretical physical model after the addition of the parameter constraint layer, and the corresponding prediction accuracy is improved by 2.49%. The effectiveness of the priori knowledge from the accumulated tool wear data introduced in this study for improving the prediction accuracy is verified.
- (4) Through the reconstruction and decomposition of the tool wear curves with the public dataset, it is verified that the fusion model can learn the trend of the tool wear process well from the training samples, and thus, better prediction results can be obtained.

The approach proposed in this study, involving the modification of the data-driven model structure and the construction of the fusion model with the physical model as a constraint, holds the potential for enhancing various data-driven models in the current tool wear prediction. A more complex data-driven model can also be integrated into the following research.

**Acknowledgements** The authors sincerely thank Ms. Anran Zong for her help in the experiment.

**Author contribution** Chang Fan: methodology, writing, experiment, and data process. Zhao Zhang: supervision. Dinghua Zhang: supervision. Ming Luo: methodology, supervision, writing – review and editing, and supervision. All authors read and approved the final manuscript.

**Funding** This study was co-supported by the National Natural Science Foundation of China (Grant Nos. U2241249 and 52022082) and the Fundamental Research Funds for the Central Universities (D5000230081).

## Declarations

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Conflict of interest** The authors declare no competing interests.



## References

- Xi RT, Zhao YH, He J, Wang Y (2020) Effects of grain size on mechanical properties and machining deformation of GH4169 casing parts. *Foundry Technol* 41(6):518–521. <https://doi.org/10.16410/j.issn1000-8365.2020.06.003>
- Liu X, Kang X, Zhao W, Liang W (2013) Electrode feeding path searching for 5-axis EDM of integral shrouded blisks. *Procedia Cirp* 6:107–111. <https://doi.org/10.1016/j.procir.2013.03.041>
- Lu XD, Du JH, Deng Q, Zhuang JY (2014) Stress rupture properties of GH4169 superalloy. *J Market Res* 3(2):107–113. <https://doi.org/10.1016/j.jmrt.2014.03.003>
- Luo JT, Yu WL, Xi CY, Zhang CX, Ma CH (2019) Preparation of ultrafine-grained GH4169 superalloy by high-pressure torsion and analysis of grain refinement mechanism. *J Alloy Compd* 777:157–164. <https://doi.org/10.1016/j.jallcom.2018.10.385>
- An XL, Zhang B, Chu CL, Zhou L, Chu PK (2019) Evolution of microstructures and properties of the GH4169 superalloy during short-term and high-temperature processing. *Mater Sci Eng, A* 744:255–266. <https://doi.org/10.1016/j.msea.2018.12.019>
- Li HS, Niu S, Zhang QL, Fu SX, Qu NS (2017) Investigation of material removal in inner-jet electrochemical grinding of GH4169 alloy. *Sci Rep* 7(1):3482. <https://doi.org/10.1038/s41598-017-03770-1>
- Priarone PC, Rizzuti S, Rotella G, Settineri L (2012) Tool wear and surface quality in milling of a gamma-TiAl intermetallic. *Int J Adv Manuf Technol* 61:25–33. <https://doi.org/10.1007/s00170-011-3691-x>
- Fofana MS, Ee KC, Jawahir IS (2003) Machining stability in turning operation when cutting with a progressively worn tool insert. *Wear* 255(7–12):1395–1403. [https://doi.org/10.1016/S0043-1648\(03\)00356-9](https://doi.org/10.1016/S0043-1648(03)00356-9)
- Kurada S, Bradley C (1997) A review of machine vision sensors for tool condition monitoring. *Comput Ind* 34(1):55–72. [https://doi.org/10.1016/S0166-3615\(96\)00075-9](https://doi.org/10.1016/S0166-3615(96)00075-9)
- Zhou YQ, Xue W (2018) Review of tool condition monitoring methods in milling processes. *Int J Adv Manuf Technol* 96:2509–2523. <https://doi.org/10.1007/s00170-018-1768-5>
- Kwon YJ, Fischer GW (2003) A novel approach to quantifying tool wear and tool life measurements for optimal tool management. *Int J Mach Tools Manuf* 43(4):359–368. [https://doi.org/10.1016/S0890-6955\(02\)00271-7](https://doi.org/10.1016/S0890-6955(02)00271-7)
- Loizou J, Tian WM, Robertson J, Camelio J (2015) Automated wear characterization for broaching tools based on machine vision systems. *J Manuf Syst* 37:558–563. <https://doi.org/10.1016/j.jmsy.2015.04.005>
- Zhang Z, Jiang F, Luo M, Wu BH, Zhang DH, Tang K (2024) Geometric error measuring, modeling, and compensation for CNC machine tools: a review. *Chin J Aeronaut* 37(2):163–198. <https://doi.org/10.1016/j.cja.2023.02.035>
- Snr DED (2000) Sensor signals for tool-wear monitoring in metal cutting operations—a review of methods. *Int J Mach Tools Manuf* 40(8):1073–1098. [https://doi.org/10.1016/S0890-6955\(99\)00122-4](https://doi.org/10.1016/S0890-6955(99)00122-4)
- Doukas C, Stavropoulos P, Papacharalampopoulos A, Foteinopoulos P, Vasiliadis E, Chryssoulouris G (2013) On the estimation of tool-wear for milling operations based on multi-sensorial data. *Procedia Cirp* 8:415–420. <https://doi.org/10.1016/j.procir.2013.06.126>
- Nouri M, Fussell BK, Ziniti BL, Linder E (2015) Real-time tool wear monitoring in milling using a cutting condition independent method. *Int J Mach Tools Manuf* 89:1–13. <https://doi.org/10.1016/j.ijmachtools.2014.10.011>
- Kim JH, Moon DK, Lee DW, Kim JS, Kang MC, Kim KH (2002) Tool wear measuring technique on the machine using CCD and exclusive jig. *J Mater Process Technol* 130:668–674. [https://doi.org/10.1016/S0924-0136\(02\)00733-1](https://doi.org/10.1016/S0924-0136(02)00733-1)
- Szydłowski M, Powalka B, Matuszak M, Kochmański P (2016) Machine vision micro-milling tool wear inspection by image reconstruction and light reflectance. *Precis Eng* 44:236–244. <https://doi.org/10.1016/j.precisioneng.2016.01.003>
- Dutta S, Pal SK, Mukhopadhyay S, Sen R (2013) Application of digital image processing in tool condition monitoring: a review. *CIRP J Manuf Sci Technol* 6(3):212–232. <https://doi.org/10.1016/j.cirpj.2013.02.005>
- Castejón M, Alegre E, Barreiro J, Hernández LK (2007) On-line tool wear monitoring using geometric descriptors from digital images. *Int J Mach Tools Manuf* 47(12–13):1847–1853. <https://doi.org/10.1016/j.ijmachtools.2007.04.001>
- Murata M, Kurokawa S, Ohnishi O, Uneda M, Doi T (2012) Real-time evaluation of tool flank wear by in-process contact resistance measurement in face milling. *J Adv Mech Design, Syst, Manuf* 6(6):958–970. <https://doi.org/10.1299/jamdsm.6.958>
- Bhattacharyya SK, Jetley S, Ivkovic B (1978) Some uses of radioactive techniques in the tool wear process. *Wear* 48(2):399–407. [https://doi.org/10.1016/0043-1648\(78\)90235-1](https://doi.org/10.1016/0043-1648(78)90235-1)
- Pimenov DY, Bustillo A, Wojciechowski S, Sharma VS, Gupta MK, Kuntoğlu M (2023) Artificial intelligence systems for tool condition monitoring in machining: analysis and critical review. *J Intell Manuf* 34(5):2079–2121. <https://doi.org/10.1007/s10845-022-01923-2>
- Bao JS, Yuan GC, Zheng XH, Zhang JG, Ji X (2017) A data driven model for predicting tool health condition in high speed milling of titanium plates using real-time SCADA. *Procedia Cirp* 61:317–322. <https://doi.org/10.1016/j.procir.2016.11.191>
- Wang GF, Zhang YC, Liu C, Xie QL, Xu YG (2019) A new tool wear monitoring method based on multi-scale PCA. *J Intell Manuf* 30:113–122. <https://doi.org/10.1007/s10845-016-1235-9>
- Shi D, Gindy NN (2007) Tool wear predictive model based on least squares support vector machines. *Mech Syst Signal Process* 21(4):1799–1814. <https://doi.org/10.1016/j.ymsp.2006.07.016>
- Sayyad S, Kumar S, Bongale A, Kamat P, Patil S, Kotecha K (2021) Data-driven remaining useful life estimation for milling process: sensors, algorithms, datasets, and future directions. *IEEE Access* 9:110255–110286. <https://doi.org/10.1109/ACCESS.2021.3101284>
- Bokulich A (2011) How scientific models can explain. *Synthese* 180:33–45. <https://doi.org/10.1007/s11229-009-9565-1>
- Taylor FW (1906) On the art of cutting metals. *Trans ASME* 28:31–350
- Müller E (1962) Der Verschleiss von Hartmetallwerkzeugen und seine kurzzeitige Ermittlung. <https://doi.org/10.3929/ETHZ-A-000089004>
- Pálmai Z (2013) Proposal for a new theoretical model of the cutting tool's flank wear. *Wear* 303(1–2):437–445. <https://doi.org/10.1016/j.wear.2013.03.025>
- Sipos Z (1986) Investigation of cutting performance of coated HSS tools made in Hungary. *NME, Miskolc*
- Zhang Y, Zhu KP, Duan XY, Li S (2021) Tool wear estimation and life prognostics in milling: model extension and generalization. *Mech Syst Signal Process* 155:107617. <https://doi.org/10.1016/j.ymsp.2021.107617>
- Fan C, Zhang Z, Zhang DH, Luo M (2023) Research on tool wear modeling of superalloy based on evolutionary cluster analysis. *Int J Adv Manuf Technol* 129(1–2):143–166. <https://doi.org/10.1007/s00170-023-12251-x>
- Stewart R, Ermon S (2017) Label-free supervision of neural networks with physics and domain knowledge. *Proc AAAI Confer Artif Intell* 31(1). <https://doi.org/10.1609/aaai.v31i1.10934>

36. Wang JJ, Li YL, Zhao R, Cao RX (2020) Physics guided neural network for machining tool wear prediction. *J Manuf Syst* 57:298–310. <https://doi.org/10.1016/j.jmsy.2020.09.005>
37. Hua J, Li Y, Liu C, Wan P, Liu X (2023) Physics-informed neural networks with weighted losses by uncertainty evaluation for accurate and stable prediction of manufacturing systems. *IEEE Trans Neural Networks Learn Syst*. <https://doi.org/10.1109/TNNLS.2023.3247163>
38. Munaro R, Attanasio A, Del Prete A (2023) Tool wear monitoring with artificial intelligence methods: a review. *J Manuf Mater Process* 7(4):129. <https://doi.org/10.3390/jmmp7040129>
39. Dianfang MU, Liu XL, Yue CX, Liu Q, Bai ZY, Liang SY, Ding YP (2021) On-line tool wear monitoring based on machine learning. *J Adv Manuf Sci Technol* 1(2):20210023-1-2021002–9. <https://doi.org/10.51393/j.jamst.2021002>
40. Mohamed A, Hassan M, M'Saoubi R, Attia H (2022) Tool condition monitoring for high-performance machining systems—a review. *Sensors* 22(6):2206. <https://doi.org/10.3390/s22062206>
41. Zhang CJ, Yao XF, Zhang JM, Jin H (2016) Tool condition monitoring and remaining useful life prognostic based on a wireless sensor in dry milling operations. *Sensors* 16(6):795. <https://doi.org/10.3390/s16060795>
42. Zhou CA, Guo K, Sun J (2021) Sound singularity analysis for milling tool condition monitoring towards sustainable manufacturing. *Mech Syst Signal Process* 157:107738. <https://doi.org/10.1016/j.ymsp.2021.107738>
43. Xu XW, Li X, Ming WW, Chen M (2022) A novel multi-scale CNN and attention mechanism method with multi-sensor signal for remaining useful life prediction. *Comput Ind Eng* 169:108204. <https://doi.org/10.1016/j.cie.2022.108204>
44. Hu J, Shen L, Sun G (2018) Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition 7132–7141. <https://doi.org/10.48550/arXiv.1709.01507>
45. Altintas Y (2001) Manufacturing automation: metal cutting mechanics, machine tool vibrations, and CNC design. Cambridge University Press, New York
46. Kingma D, Ba J (2015) Adam: a method for stochastic optimization, *Comp Sci*. <https://doi.org/10.48550/arXiv.1412.6980>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.