entropy

MDPI

Article

A Dual-Stage Attention Model for Tool Wear Prediction in Dry Milling Operation

Yongrui Qin [1,†], Jiangfeng Li [2,*,†], Chenxi Zhang [2], Qinpei Zhao [2,*] and Xiaofeng Ma [3]

[1] Faculty of Engineering, The University of Sydney, Sydney, NSW 2006, Australia
[2] School of Software Engineering, Tongji University, Shanghai 201804, China
[3] College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
* Correspondence: lijf@tongji.edu.cn (J.L.); qinpeizhao@tongji.edu.cn (Q.Z.)
† These authors contributed equally to this work.

Abstract: The intelligent monitoring of tool wear status and wear prediction are important factors affecting the intelligent development of the modern machinery industry. Many scholars have used deep learning methods to achieve certain results in tool wear prediction. However, due to the instability and variability of the signal data, some neural network models may have gradient decay between layers. Most methods mainly focus on feature selection of the input data but ignore the influence degree of different features to tool wear. In order to solve these problems, this paper proposes a dual-stage attention model for tool wear prediction. A CNN-BiGRU-attention network model is designed, which introduces the self-attention to extract deep features and embody more important features. The IndyLSTM is used to construct a stable network to solve the gradient decay problem between layers. Moreover, the attention mechanism is added to the network to obtain the important information of output sequence, which can improve the accuracy of the prediction. Experimental study is carried out for tool wear prediction in a dry milling operation to demonstrate the viability of this method. Through the experimental comparison and analysis with regression prediction evaluation indexes, it proves the proposed method can effectively characterize the degree of tool wear, reduce the prediction errors, and achieve good prediction results.

Keywords: tool wear prediction; attention mechanism; signal analysis; machine learning; gated recurrent unit

Citation: Qin, Y.; Li, J.; Zhang, C.; Zhao, Q.; Ma, X. A Dual-Stage Attention Model for Tool Wear Prediction in Dry Milling Operation. *Entropy* **2022**, *24*, 1733. https://doi.org/10.3390/e24121733

Academic Editor: Markus Pauly

Received: 13 October 2022
Accepted: 23 November 2022
Published: 28 November 2022

1. Introduction

With the continuous improvement and optimization of sensor technology, internet of things technology, and deep learning algorithms, the development of industrial intelligent manufacturing systems is more rapid, and constantly moving toward the integration of various emerging technologies. In the industrial intelligent manufacturing environment, most of the machining process is the cutting process, which inevitably causes tool wear. Tool wear refers to a process in which the metal material on the tool surface is continuously disappearing and the surface morphology is continuously changing due to the mechanical, chemical, and thermal effects of the cutting process [1].

The tool wear has an important impact on the machining process. When the tool is worn to the scrap state, but the machining process has still not stopped, it will damage the workpiece and even break down the machine tool, which may directly affect the processing efficiency, product quality, and production cost [2,3]. The traditional coping mode is to change tools at regular intervals, which can cause the waste of materials. In that case, the coping mode has gradually updated to intelligent tool changing based on the prediction of tool wear. Changing the severely worn tools through online monitoring and real-time prediction can not only improve tool utilization rate and processing quality, but also reduce safety accidents and shutdown rate.

In recent years, many scholars have conducted a lot of work on tool monitoring. Most of them are committed to the online monitoring of tool wear status and the prediction of remaining tool life usefulness. According to different measurement methods, the automatic monitoring solutions in tool wear can be mainly divided into two types, direct method and indirect method. The mainstream method is to use different sensors indirectly to collect digital signals, such as vibration, force, current, and acoustic emission. These signals can reflect the changes with time about the process of tool wear and can realize online intelligent monitoring without interrupting the machining process to establish a correlation between signal data and wear status, which is able to obtain the wear degree of the tool. Obviously, extracting useful features from original signals is critical for tool wear monitoring, and it can directly affect the prediction results of the model. Commonly used feature processing methods are time domain analysis, frequency domain analysis, and time–frequency domain analysis [4].

With the wide use of deep learning and neural network in various fields, the current research on tool wear mainly focuses on the combination of feature engineering and deep learning. Kong et al. [5] optimized the time–frequency statistic features and developed an artificial neural network model to predict the degree of tool wear. Tao et al. [6] presented a novel method based on the long short-term memory network (LSTM) and hidden Markov model to track the flank wear and predict the remaining useful life. Wu et al. [7] proposed a tool wear prediction model based on singular value decomposition (SVD) and bidirectional long short-term memory neural network (Bi-LSTM). SVD was used to extract the signal features from the reconstructed signal matrix. An et al. [8] used sparse auto-encoder and Pearson correlation coefficient to extract the sensitive features of the original cutting force signal and used these features to train back propagation neural network.

Although these studies have made great progress to improve the accuracy and reliability of the tool wear prediction model, there are still some limitations. In the feature extraction process of different sensors, only a few common features are extracted from experts' domain knowledge, which can not be applied to more scenes or adjust different tools. In the process of feature selection, the influence degree between different features and target wear value is not reflected or ignored. The design of many neural network structures may cause the gradient attenuation between layers. Due to the instability and variability of the collected signal, the designed tool wear prediction model is not stable and accurate enough.

Hence, to solve these problems, and to accurately reflect the change in tool wear status with time during machining process, a dual-stage attention model is proposed to predict tool wear online. When the set threshold is reached, the machine will be stopped intelligently and change new tools. The main contributions of this paper include three parts:

Firstly, to obtain comprehensive signal characteristics, features are extracted from the collected sensor signals during milling cutter processing through three-domain analysis. Selecting highly relevant features by using the maximum information coefficient can reduce the redundant features and improve modelling efficiency.

Secondly, to extract deep features and reflect influence degree of different features, a network based on convolutional neural network (CNN) and bidirectional gated recurrent unit (BiGRU) with attention (CBGA) is proposed to encode feature vectors by applying the self-attention to assign different weights.

Finally, to improve the stability of tool wear prediction model, an IndyLSTM-attention network is proposed to predict the wear values.

The rest of the paper is organized as follows. Section 2 introduces related works about the application of deep learning in the tool wear prediction. A dual-stage attention prediction model is specifically described in Section 3. Section 4 evaluates the performance of the proposed model by a case study for tool wear prediction in a dry milling operation. Section 5 is the conclusion section of this paper.

2. Related Works

With the development of deep learning in the last few years, many deep learning methods based on predictive analysis have been widely adapted for the process of tool condition monitoring and tool wear prediction [9–11]. Dai, Zhang, and Meng used a stacked sparse auto-encoder network to reduce the feature vectors and build a least squares support vector machine prediction model based on cuckoo optimization parameters [12]. An adaptive method was developed by Cao, Sun, and Zhang by using a deep network to replace manual feature extraction from signals and proposed an on-line tool wear monitoring model based on convolution neural networks [13]. Wu, Jennings, and Terpenny introduced a method based on random forests for tool wear prediction [14]. To realize a real-time and accurate monitoring of the tool wear in machining process, Kong, Dong, and Chen presented a model based on the integrated radial basis function with kernel principal component analysis (KPCA_IRBF) and relevance vector machine (RVM) [15].

Considering the characteristics of the time series and dynamic changes of input data, the recurrent neural network introduces a cyclic structure, which can model dynamic time series data better than other neural networks. Therefore, RNN and its variations, such as LSTM and GRU, have been widely applied to this field of tool wear. For example, a recurrent neural network based on health indicator (RNN-HI) for RUL prediction of bearings was proposed by Guo, Li, and Jia [16]. Zhu, Xie, and Li established a tool wear monitoring model on the basis of long-term and short-term memory neural networks [17]. A deep neural network structure named convolutional bi-directional long short-term memory (CBLSTM) has been designed to address raw sensory data [18]. It presented a hybrid prediction scheme to solve long-term prediction problems by a newly developed deep heterogeneous GRU model, along with local feature extraction [19]. Inspired by the success of deep learning methods that redefine representation learning from raw data, Zhao, Wang, and Yan [20] proposed a network named local feature-based gated recurrent unit (LFGRU). It was a hybrid approach that combined handcrafted feature design with automatic feature learning for machine health monitoring.

Both GRU and LSTM are special RNN structures, which are proposed to solve the problems of gradient disappearance in RNN. Although these structures solve the gradient problems to some extent by using the tanh and sigmoid function as activation function, they will also cause gradient attenuation between layers. IndyLSTMs (independently recurrent long short-term memory cells) [21] were proposed on the basis of IndRNN [22], which adds the gate structure of LSTM. Compared with the traditional LSTM, the cyclic weight is no longer a full matrix, but a diagonal matrix. In each layer of IndyLSTM, the number of parameters and nodes shows a linear relationship, while the traditional LSTM is quadratic. This feature makes the model smaller and faster, and the accuracy of this model is better than the LSTM model in most cases. Therefore, the IndyLSTM is introduced to build a new network to solve the problems of gradient attenuation between layers and to obtain a stable and accurate model for tool wear monitoring.

Moreover, most tool wear prediction models based on the recurrent neural network mainly focus on the selection of input features, ignoring the influence degree of input features on the tool wear. The attention mechanism is widely used in various types of deep learning tasks, such as natural language processing, image recognition, and speech recognition. As a resource allocation mechanism, it can assign different weights to input features so that different features containing important information will not disappear with the increase in time steps, which can highlight the impact of more important information. In this way, full use of the network can help to study and improve the prediction quality for a longer period of stability [23].

In summary, existing works have used deep networks instead of traditional methods, such as manual or machine learning methods, to extract features from signals to improve the prediction accuracy. However, existing models, such as IndyLSTM, ignore the difference in the degree of influence of selected input features on tool wear. In addition, existing methods do not make full use of prior knowledge to improve model performance.

3. Dual-Stage Attention Prediction Model

The framework of the tool wear prediction model based on dual-stage attention is shown in Figure 1. The whole model mainly includes three layers: the feature engineering layer, the deep feature extraction layer, and the model prediction layer. After the initial feature engineering process to the raw signals, the CBGA network will be used to extract deep features. Finally, applying the IndyLSTM-Attention model to train and output the wear values generates a stable model to realize real-time prediction of tool wear.

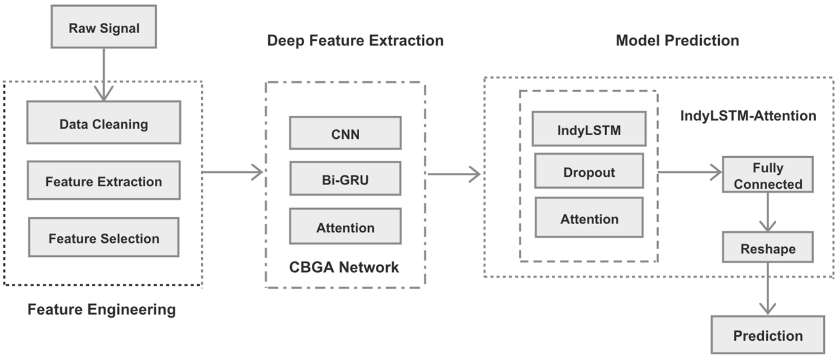

Figure 1. A dual-stage attention model for tool wear prediction.

3.1. Feature Engineering

The feature engineering layer consists of data cleaning, feature extraction, and feature selection. In this part, data cleaning mainly includes zero-averaging, removing the trend term, and normalizing signals. Meanwhile, according to the wavelet packet decomposition theory, high-frequency noise will be filtered out. Using the common signal analysis methods on the cleaned data, the statistical features of signal data are extracted from three domains. Combined with the existing research, this paper integrates time domain, frequency domain, and time-frequency to analyze the sensor signals comprehensively. After feature fusion, a preliminary feature selection is conducted on the extracted features. The flow of feature engineering is shown in Figure 2.

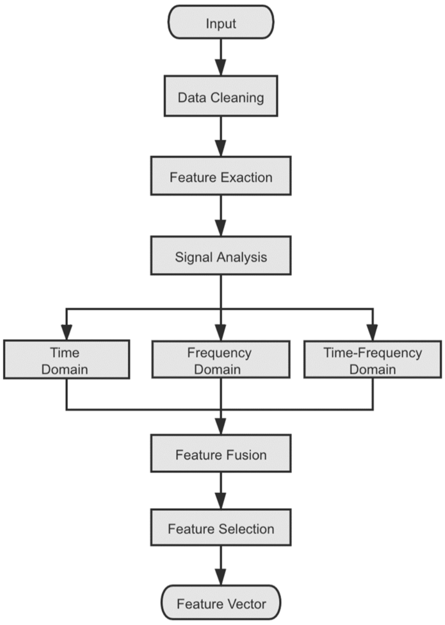

Figure 2. The flow of feature engineering.

3.1.1. Feature Exaction: Signal Analysis

Time domain analysis uses time axis as the coordinate to express the relationship between dynamic signals. It can effectively improve the signal-to-noise ratio and find the similarity and correlation of signal waveform transformations at different times. These obtained features can reflect the operating status of mechanical equipment.

Frequency domain analysis transforms the signals to the frequency axis. This method based on frequency characteristics makes up for the shortcomings of time domain analysis. It indirectly reveals the time domain performance of signals and easily displays the effect of system parameters on system performance. In this paper, spectrum analysis is used to analyze the signals after fast Fourier transform and to extract frequency domain features.

Wavelet analysis is a common time–frequency domain analysis method, which takes the signal information in both time domain and frequency domain into account. By analyzing frequency spectrum of sampled signal, the level of wavelet decomposition about signal is determined. The energy of each frequency band and the total energy entropy are taken as the time–frequency features after decomposition. In the Formula (1), F_s is the sampling frequency, n is the number of layers, and f_{min} is the minimum frequency band.

$$f_{min} = F_s/2^{n+1} \tag{1}$$

The sampling frequency is 50 kHz, so a five-layer wavelet packet decomposition is performed. One then takes 2^5 = 32 frequency band energy and energy entropy as time–frequency domain features. Thus, the signal analysis selects 13 features in the time domain, 3 features in the frequency domain, and 33 features in the time-frequency domain. In that case, 49 different features of each sensor channel will be extracted. The main extracted features are shown in Table 1.

Table 1. The information of extracted features.

Domain	Feature Name	Formula	Feature Name	Formula
Time	mean	$\overline{y} = 1/n \sum\limits_{i=1}^{n} y_i$	form factor	$\dfrac{\sqrt{(1/n\sum_{i=1}^{n} {y_i}^2)}}{1/n\sum_{i=1}^{n}\lvert y_i\rvert}$
	root mean square	$\left(\dfrac{1}{n}\sum\limits_{i=1}^{n} {y_i}^2\right)^{1/2}$	spectral kurtosis	$\dfrac{E\left[\{(y-\overline{y})/\sigma\}^4\right]}{\left(\sqrt{1/n\sum_{i=1}^{n} {y_i}^2}\right)^4}$
	maximum	$\mathrm{Max}\{\lvert y_i\rvert\}$	spectral skewness	$\dfrac{E\left[\{(y-\overline{y})/\sigma\}^3\right]}{\left(\sqrt{1/n\sum_{i=1}^{n} {y_i}^2}\right)^4}$
	variance	$\frac{1}{n}\sum\limits_{i=1}^{n} (y_i-\overline{y})^2$	impulse index	$Max\{\lvert y_i\rvert\}/\overline{y}$
	skewness	$E\left[\{(y-\overline{y})/\sigma\}^3\right]$	clearance index	$\dfrac{Max\{\lvert y_i\rvert\}}{\left(1/n\sum_{i=1}^{n}\lvert y_i\rvert\right)^2}$
	kurtosis	$E\left[\{(y-\overline{y})/\sigma\}^4\right]$	standard deviation	$[\dfrac{1}{n}\sum\limits_{i=1}^{n} (y_i-\overline{y})^2]^{1/2}$
	crest factor		$\dfrac{max(y)-min(y)}{\sqrt{1/n\sum_{i=1}^{n} (y_i)^2}}$	
Frequency	centroid frequency	$FC = \dfrac{\sum_{f=1}^{n/2} fw(f)}{\sum_{f=1}^{n/2} w(f)}$	frequency variance	$\dfrac{\sum_{f=1}^{n/2}(f-FC)^2 w(f)}{\sum_{f=1}^{n/2} w(f)}$
	mean square frequency		$MSF = \dfrac{\sum_{f=1}^{n/2} f^2 w(f)}{\sum_{f=1}^{n/2} w(f)}$	
Time-Frequency	wavelet energy	$E_{i=1,2,\ldots 32} = wt_{\varnothing}^2(i)/n$	wavelet entropy	E = $\sum_{i=1}^{32} wt_{\varnothing}^2(i)/n$

3.1.2. Feature Selection: Based on MIC

Unnecessary features will reduce training speed and generalization performance of the test set. In this paper, the features are selected and reduced by the maximum information coefficient (MIC). MIC can express various linear and non-linear relationships, and its value range is between 0 and 1. The higher the value, the stronger the correlation, so it has been widely used to select features in machine learning [24]. The basic principle of MIC utilizes the concept of mutual information, which is used to measure the degree of interdependence between two random variables. The mutual information can be explained as the following equation.

$$I(x;y) = \int p(x,y)log_2^{\frac{p(x,y)}{p(x)p(y)}}dxdy \tag{2}$$

where $p\ (x,\ y)$ is the joint probability density function of x and y, $p\ (x)$ is the marginal probability density function of x, and $p\ (y)$ is the marginal probability density function of y. The calculation formula of MIC is shown in Formula (3):

$$MIC(x;y) = \begin{matrix} max \\ a*b<B \end{matrix} \frac{I(x;y)}{log_2^{min(a,b)}} \tag{3}$$

In the above formula, a and b are the number of dividing grids in the x and y directions, B is a variable, and the size of B is generally set to 0.5 or 0.6 power of the total amount of data. Calculate the maximum information coefficient of tool statistical features and choose the target number of features according to the correlation and, at last, return the feature vectors after feature selection.

3.2. Deep Feature Extraction: CBGA Network

In this section, a CNN-BiGRU-attention (CBGA) network is proposed to encode and mine the deep features. It consists of CNN, Bi-GRU, and the attention mechanism, expanding the features in two dimensions of space and time.

CNN is a neural network with a deep structure that includes convolution calculations. It uses local connection and weight sharing to perform a higher-level and more abstract process on original data and can effectively extract local features of the data. CNN is mostly used for static output and is difficult for obtaining dynamic characteristics, especially when the data fluctuate or are unstable.

Bi-GRU can capture long-term dependencies and describe the continuous state output in time. It is suitable for analyzing time series data because of its memory function. The bidirectional structure can make full use of historical information and learn the dynamic laws of both positive and negative directions at the same time.

Simultaneously, self-attention as one of the attention mechanisms can discover the internal characteristics of sequence data and highlight the important features. Thus, the self-attention layer is added to obtain final deep features. Based on the above theory, the CBGA network is designed to further process the feature vectors. Figure 3 shows the whole structure of the CBGA. As shown in this figure, the CBGA network can be divided into four parts: the multi-channel convolution layer, the max-pooling layer, the bidirectional GRU layer, and the attention layer.

Assume that the number of initially extracted features is n. The input data of this module is a n-dimensional feature vector that uses $\mathrm{I}[i_0, i_1, \ldots i_n]$ to represent this input vector. Firstly, multi-channel convolution will be performed, it will connect the sequence results of the output, and it will obtain a t-dimensional vector $\mathrm{C}[c_0, c_1, \ldots c_t]$, where $t = k{*}f$, k is the number of convolution layers, and f is the number of filters of the convolution neural network.

The max-pooling operation will be carried out. After that, the full sequence feature extraction will be conducted on the input through the bidirectional GRU. This part output $\mathrm{G}[g_0, g_1, \ldots g_m]$ is the concatenation of the results of forward GRU and backward GRU, where $m = 2*h$, and h is the number of Bi-GRU's hidden units. At last, the attention value of each GRU node will be calculated in the attention layer. A weighted feature vector $Z[z_0, z_1, \ldots z_m]$ is obtained as the final deep feature encoding vector.

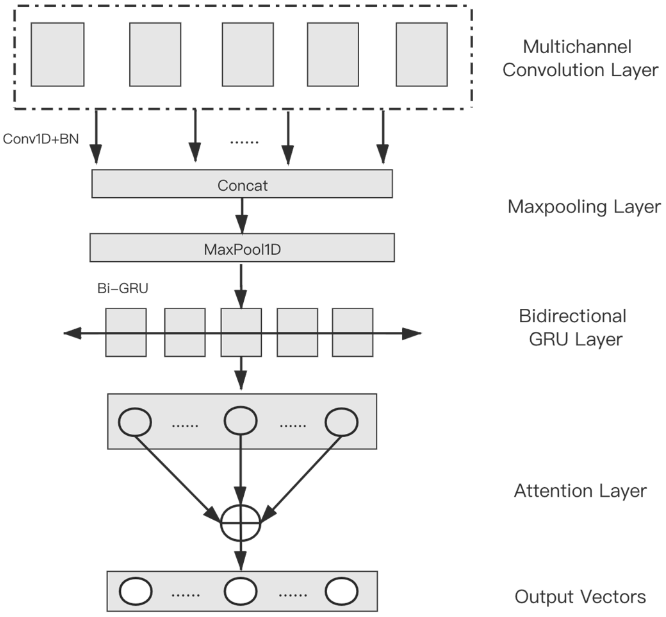

Figure 3. The structure of the CBGA.

Using mathematical formulas to express the self-attention mechanism, the input sequence from the Bi-GRU layer is $G[g_0, g_1, \ldots g_m]$, and the output sequence is $Z[z_0, z_1, \ldots z_m]$. Obtain three sets of vector sequences through linear transformation:

$$Q = W_Q G \in \mathbb{R}^{d_3 \times N} \tag{4}$$

$$K = W_K G \in \mathbb{R}^{d_3 \times N} \tag{5}$$

$$V = W_V G \in \mathbb{R}^{d_2 \times N} \tag{6}$$

where Q, K, and V are query vector sequence, key vector sequence, and value vector sequence, respectively. W_Q,W_K, and W_V are the parameter matrix that can be learned, respectively. In the definition of self-attention, $Q = K = V = G$, so the output vector z_i is calculated as:

$$z_i = att((K, V), q_i) = \sum_{j=1}^{N} \alpha_{ij} v_j = \sum_{j=1}^{N} softmax\left(s\left(k_j, q_i\right)\right) \tag{7}$$

where $i, j \in [1, m]$ are the positions of the output and input vector sequences, and $s\left(k_j, q_i\right)$ is a function to calculate the similarity between two vectors.

3.3. Model Prediction: IndyLSTM-Attention

The IndyLSTM-attention model is used in this paper to train and output the prediction. The model consists of the IndyLSTM network, the attention network, and the fully connected network, and these decode the feature sequences and output the required prediction. In the common RNN decoding unit, generally only the last sequence is taken from the output result as the prediction result. However, other sequences in the network structure are also meaningful. Combining other sequences through the attention mechanism may help us achieve a better fitting effect.

Therefore, the Bahdanau attention [25] is added after the IndyLSTM layer. The formula of Bahdanau attention mechanism is as follows:

$$att((K,V),q) = \sum_{i=1}^{N} \alpha_i v_i = \sum_{i=1}^{N} \frac{exp(s(k_i,q))}{\sum_j exp(s(k_j,q))} v_i \tag{8}$$

Input the deep feature results $Z\,[z_0, z_1, \ldots z_m]$ extracted by the CBGA network into the model. The IndyLSTM-Attention network will output the vector $H[h_0, h_1, \ldots h_u]$, where u is equal to the number of hidden units of the IndyLSTM cell. At the fully connected network, two layers of fully connected networks are used to get the prediction. The architecture of the IndyLSTM-Attention model is shown in Figure 4.

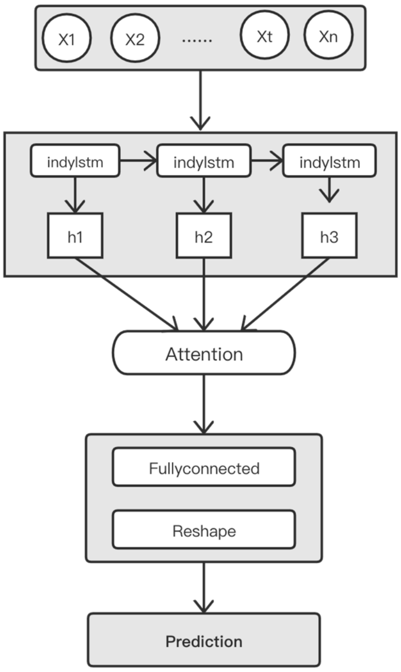

Figure 4. The architecture of the IndyLSTM-Attention model.

4. Experiments

In this section, an empirical evaluation is conducted to test the performance of the proposed model. The descriptions of the datasets and experimental setup are introduced in detail. The proposed model is compared with other common prediction methods to form the comparison results and discussion.

4.1. Descriptions of Datasets

Open datasets were used to verify the predictive performance of the model, which were collected from the ball end carbide milling cutter of a high-speed (CNC) machine operated under dry milling operations [26]. Each training record contains one "wear" file that lists the flank wear values measured for three cutting edges after each cut in 10^{-3} mm and a folder with approximately 300 individual data acquisition files (one for each cut). The data acquisition files have seven columns of dynamometer, accelerator, and acoustic emission data. The main equipment and cutting parameters are specifically listed in Table 2.

In the experiment, six independent milling cutters (c1~c6) were used for the full tool life test. The force sensor, acceleration sensor, and acoustic emission sensor were used to collect signals, and each tool collected seven channel signals, which include force in three directions (X, Y, Z), vibration in three directions (X, Y, Z), and AE–RMS. During the experiment, all the sensor data were collected on a data acquisition card, and the data acquisition card transmitted all the information to the computer. Meanwhile, in the process of cutting the workpiece, the tool would be stopped in every cutting and used the microscope to measure the wear in the X, Y, and Z directions. The test was terminated when the tool was severely worn out and could not work anymore. 315 samples were obtained during the test, taking the average of the flank wear in three directions as the true value of tool wear estimation, and the unit of the tool wear is 10^{-3} mm.

Table 2. The main equipment and cutting parameters of milling experiment.

Machine Tool	Roders Tech RFM760	Spindle Speed	10,400 r/min
Cutter	Ball end Carbide milling	Feed rate	1555 mm/min
Milling material	Stainless steel HRC52	Cutting width	0.125 mm
Dynamometer	Kistler 9265B	Cutting depth	0.2 mm
Vibration sensor	Kistler 8636C	Sampling frequency	50 kHZ
Collector	NI DAQ PCI 1200	Charge amplifier	Kistler 5019A
Wear gauge	Microscope LEICA MZ12	Cooling condition	Dry Machining

In the given datasets, c1, c4, and c6 are training data with corresponding wear values, and c2, c3, and c5 are test data without wear values. Therefore, during the process of model verification, c1, c4, and c6 were selected as the training dataset. The leave-one-out method was adopted to achieve cross-validation by using two datasets as training set and using the rest one for verification. Therefore, three different test cases can be created, denoted as C1, C4, and C6. The partition of datasets is shown in Table 3.

Table 3. The setup of training and testing data.

Group	Training Set	Testing Set
C1	c4, c6	c1
C4	c1, c6	c4
C6	c1, c4	c6

4.2. Evaluation Index

In order to quantify the performance of all comparison methods, three commonly used evaluation indicators were selected to evaluate the regression loss, including mean absolute error (MAE), root mean square error (RMSE) and the coefficient of determination (R^2 score). Among the selected functions, both the MAE and R^2 scores are relatively robust and insensitive to outliers and noise. On the contrary, RMSE can integrate the advantages and disadvantages of MSE and MAE, is very sensitive to extremely large or small errors, and can make the model tend to be optimal.

MAE is the average of absolute values of errors. RMSE is the square root of mean of all squared errors. These two metrics are calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\overline{y_i} - y_i| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\overline{y_i} - y_i)^2} \tag{10}$$

where y_i and $\overline{y_i}$ are true and predicted tool wear values.

The R^2 score is the coefficient of determination, reflecting how much of the fluctuation of y can be described by the fluctuation of x. The value range of the R^2 score is between 0 and 1. The closer the value is to 1, the higher the degree of interpretation of the variable. The expression is as follows:

$$R^2 = 1 - \frac{\sum (y_i - y_i)^2}{\sum (y_i - \overline{y})^2}, where\ \overline{y} = \frac{1}{n}\sum\nolimits_{i=1}^{n} y_i \qquad (11)$$

4.3. Evaluation Setup

The vibration sensor signals were chosen for modeling. According to the part of feature exaction, the signal analysis was performed from three directions (X, Y, and Z) of vibration signals. Therefore, 147-dimensional (49*3) features were obtained in total. Then, the 40 best features were selected with high correlation by MIC, and the feature vectors were obtained after preprocessing. Meanwhile, the average of the wear value (mm) in three directions after each cutting was taken as true wear value of the tool. The initial values of the parameters of the experimental models are set with reference to the pre-trained models. In the experiment, the parameters are adjusted one by one by fixing other parameters and fine-tuning one parameter until the optimal result is obtained. The specific structure and experimental parameter configurations of the model are shown in Table 4.

Table 4. Model structure and parameter configuration of this paper.

Structure	Layer
CBGA	Conv1D (filters = 128, kernel_size = 2) Batch Normalization
	Conv1D (filters = 128, kernel_size = 3) Batch Normalization Conv1D (filters = 128, kernel_size = 4) Batch Normalization
	MaxPool1D (pool_size = 5, strides = 1)
	Bi-GRU (hidden_units = 64)
	Self-Attention
Model	IndyLSTM (hidden_units = 128) Dropout (0.8) Attention (attention_length = 22) Fully Connected Reshape

The IndyLSTM-attention model was also compared with common neural networks including RNN, LSTM, GRU, IndRNN, and IndyLSTM. These different neural network methods are used as the model to output tool wear values after training. The input vectors of these models are extracted by CBGA network, and the size of hidden units in recurrent neural cells is unified to be the same as 128. The loss function used is mean squared error (MSE). Stochastic gradient descent (SGD) is adopted as an optimizer algorithm to train the models. During the training process, by using the estimator to build a deep recurrent network model, it can easily configure, train, and evaluate various machine learning models.

4.4. Experimental Evaluation

This section shows the results of comparison experiments and makes a brief analysis. This gives the prediction curves of the proposed method for three different test sets, as shown in Figure 5. In the figure, the broken line is the actual value of tool wear, the smooth curve is the predicted value of tool wear, and the bottom histogram is the error between the predicted value and the true value.

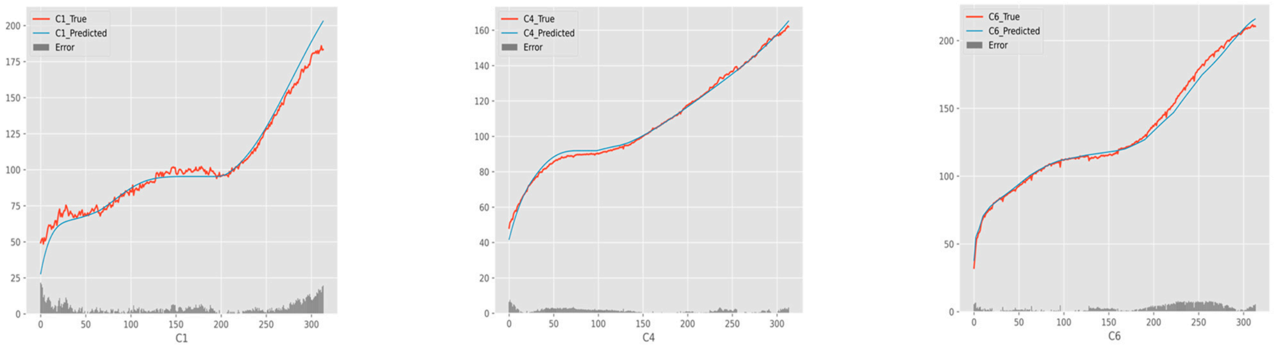

Figure 5. The tool wear prediction curve of a dual-stage attention model.

Table 5 shows all the results of common methods on three test cases, including the RMSE, MAE and R^2 score.

Table 5. The MAE, RMSE and R^2 for all methods for three test cases.

Models	C1			C4			C6		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
RNN	7.111	5.611	0.955	2.360	1.983	0.993	6.080	5.060	0.979
LSTM	6.535	4.391	0.959	2.414	2.024	0.992	5.757	4.130	0.981
GRU	6.378	4.855	0.963	2.020	1.731	0.994	5.568	4.150	0.983
IndRNN	6.216	5.129	0.964	2.037	1.702	0.995	5.728	4.537	0.982
IndyLSTM	6.180	4.499	0.965	1.977	1.632	0.995	4.485	3.745	0.988
IndyLSTM-Att	**5.796**	**4.242**	**0.970**	**1.822**	**1.406**	**0.996**	**3.681**	**2.776**	**0.992**

The bold face indicates the best performance.

Figure 6 shows the error area between the true wear values and the predicted values. The area chart can clearly display the error size of each prediction model.

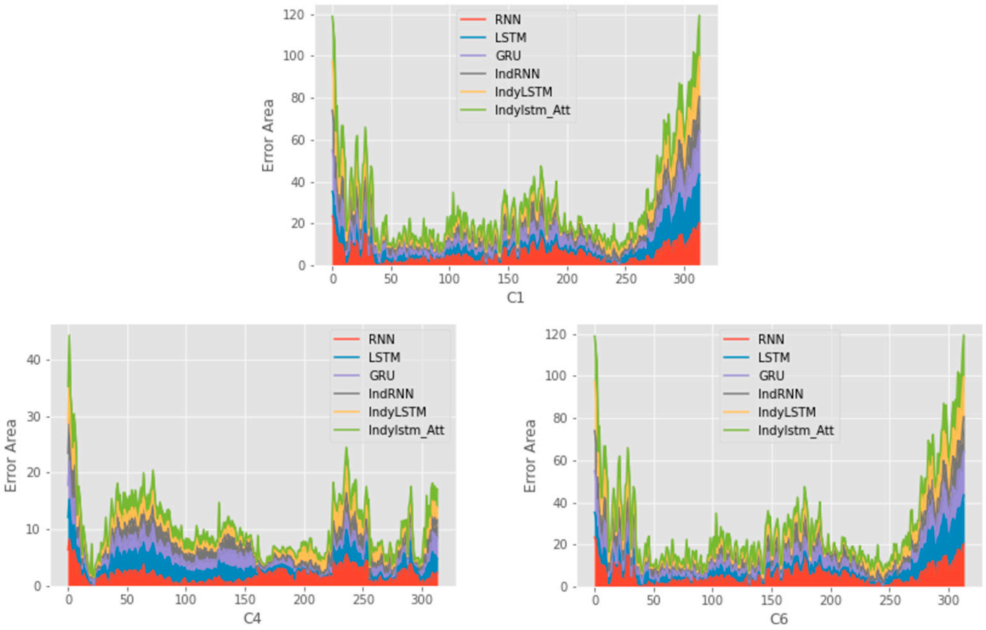

Figure 6. The area chart of error absolute value of six models on three cases.

In order to show the comparison results more intuitively, the average performance of six methods is calculated for three test cases, as shown in Table 6. The average performance comparison histogram of three cases is shown in Figure 7.

Table 6. The average performance of the three cases.

Models	RMSE	MAE	R^2
RNN	5.184	4.218	0.976
LSTM	4.902	3.515	0.977
GRU	4.655	3.579	0.980
IndRNN	4.660	3.789	0.980
IndyLSTM	4.214	3.292	0.983
IndyLSTM-Att	**3.766**	**2.808**	**0.986**

The bold face indicates the best performance.

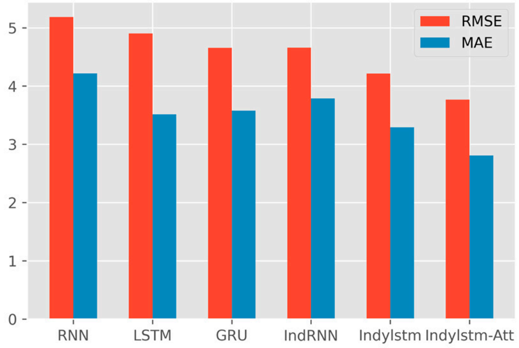

Figure 7. The area chart of error absolute value of six models for three cases.

Overall, the model proposed by this paper has a good fitting effect, and the curves basically match the true data from Figure 5. The prediction of C4 and C6 is better than C1. According to the comparison results of Tables 5 and 6, the IndyLSTM-attention in three indicators has the best performance. It can be observed that the IndyLSTM and IndyLSTM-attention outperforms the RNN, LSTM, and GRU neural network in three cases. This shows that independently recurrent long short-term memory networks perform better than traditional recurrent neural networks in this situation. Meanwhile, the comparison between the results of IndyLSTM and IndyLSTM-attention also shows that all indicators have been improved to a certain extent. Therefore, it can be concluded that the accuracy of prediction can be improved by adding an attention layer to the predictive model.

Further analysis is seen through the error area chart, the area where the prediction error of the proposed model is the smallest. At the same time, the model is found to have large fluctuations at the beginning and at the end of the tool prediction, and the error of other stages is small. Finally, through the histogram, the differences between different models can be observed. RMSE and MAE of the proposed model are much smaller than in the other models.

5. Conclusions

In this paper, an IndyLSTM model with a self-attention mechanism is proposed in order to solve the problem that existing deep learning methods ignore (the different influences of the degree of input features on tool wear in the process of intelligent tool wear monitoring). By using the 2010 PHM Society Conference Data Challenge open datasets, the proposed model has achieved better performance than common regression prediction methods in all three evaluation criteria (MAE, RMSE, and R^2 score). Through experimental verification, there are two main findings obtained.

(1) By applying the self-attention mechanism in the deep feature extraction and tool wear prediction model to assign different weights to different input features, performance of the prediction model for tool wear can be effectively improved.
(2) By combining prior experience, the feature selection method using the maximum information coefficient can effectively reduce redundant features, which shows an ability to improve modeling efficiency.

However, the features used in this paper are a combination of time domain, frequency domain, and deep learning features, wherein the time and frequency domain features are dependent on prior knowledge. The future work is to select better time and frequency domain features and better feature selection criteria to further improve the performance of the model.

Author Contributions: Conceptualization, Y.Q., J.L. and X.M.; methodology, Y.Q., Q.Z. and J.L.; experiment, Y.Q. and C.Z.; validation, Y.Q., Q.Z. and J.L.; formal analysis, C.Z. and J.L.; data curation, Y.Q., J.L. and X.M.; writing, Y.Q., Q.Z. and J.L.; supervision, C.Z.; project administration, C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Scientific Research Program of Science and Technology Commission of Shanghai Municipality, China (Grant Nos. 20ZR1460500, 22511104300, 21511101503), the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), and the Fundamental Research Funds for the Central Universities.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The dataset and parameter configuration used to support the findings of this study are included within the article.

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

References

1. 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. [CrossRef]
2. He, Z.; Shi, T.; Xuan, J.; Li, T. Research on tool wear prediction based on temperature signals and deep learning. *Wear* **2021**, *478–479*, 203902. [CrossRef]
3. Ambhore, N.; Kamble, D.; Chinchanikar, S.; Wayal, V. Tool Condition Monitoring System: A Review. *Mater. Today Proc.* **2015**, *2*, 3419–3428. [CrossRef]
4. Zeng, H.; Thoe, T.B.; Li, X.; Zhou, J. Multi-modal Sensing for Machine Health Monitoring in High Speed Machining. In Proceedings of the 2006 4th IEEE International Conference on Industrial Informatics 2006, Singapore, 16–18 August 2006; pp. 1217–1222. [CrossRef]
5. Liu, M.K.; Tseng, Y.H.; Tran, M.Q. Tool wear monitoring and prediction based on sound signal. *Int. J. Adv. Manuf. Technol.* **2019**, *103*, 3361–3373. [CrossRef]
6. Tao, Z.; An, Q.; Liu, G.; Chen, M. A novel method for tool condition monitoring based on long short-term memory and hidden Markov model hybrid framework in high-speed milling Ti-6Al-4V. *Int. J. Adv. Manuf. Technol.* **2019**, *105*, 3165–3182. [CrossRef]
7. Wu, X.; Li, J.; Jin, Y.; Zheng, S. Modeling and analysis of tool wear prediction based on SVD and BiLSTM. *Int. J. Adv. Manuf. Technol.* **2020**, *106*, 4391–4399. [CrossRef]
8. Shi, C.; Panoutsos, G.; Luo, B.; Liu, H.; Li, B.; Lin, X. Using Multiple-Feature-Spaces-Based Deep Learning for Tool Condition Monitoring in Ultraprecision Manufacturing. *IEEE Trans. Ind. Electron.* **2018**, *66*, 3794–3803. [CrossRef]
9. Li, X.; Ding, Q.; Sun, J.-Q. Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliab. Eng. Syst. Saf.* **2018**, *172*, 1–11. [CrossRef]
10. Xiao, P.; Zhang, C.; Luo, M. Modeling method for tool wear prediction based on ADNLSSVM. *China Mech. Eng.* **2018**, *29*, 842.
11. Roy, S.S. An Application of ANFIS-Based Intelligence Technique for Predicting Tool Wear in Milling. In *Intelligent Computing and Applications*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 343, pp. 299–306. [CrossRef]
12. Dai, W.; Zhang, C.; Meng, L. Support vector machine milling wear prediction model based on deep learning and feature re-processing. *Comput. Integr. Manuf. Syst.* **2020**, *26*, 2331–2343.
13. Cao, D.L.; Sun, H.B.; Zhang, J.D.; Mo, R. In-process tool condition monitoring based on convolution neural network. *Comput. Integr. Manuf. Syst.* **2020**, *26*, 74–80.
14. Wu, D.; Jennings, C.; Terpenny, J.; Gao, R.X.; Kumara, S. A Comparative Study on Machine Learning Algorithms for Smart Manufacturing: Tool Wear Prediction Using Random Forests. *J. Manuf. Sci. Eng.* **2017**, *139*. [CrossRef]
15. Kong, D.; Chen, Y.; Li, N.; Duan, C.; Lu, L.; Chen, D. Relevance vector machine for tool wear prediction. *Mech. Syst. Signal Process.* **2019**, *127*, 573–594. [CrossRef]

16. Guo, L.; Li, N.; Jia, F.; Lei, Y.; Lin, J. A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing* **2017**, *240*, 98–109. [CrossRef]
17. Zhu, X.; Xie, F.; Li, N. Tool wear state monitoring based on long-term and short-term memory neural network. *Manuf. Technol. Mach. Tools* **2019**, *10*, 112–117.
18. Zhao, R.; Yan, R.; Wang, J.; Mao, K. Learning to Monitor Machine Health with Convolutional Bi-Directional LSTM Networks. *Sensors* **2017**, *17*, 273. [CrossRef] [PubMed]
19. Wang, J.; Yan, J.; Li, C.; Gao, R.X.; Zhao, R. Deep heterogeneous GRU model for predictive analytics in smart manufacturing: Application to tool wear prediction. *Comput. Ind.* **2019**, *111*, 1–14. [CrossRef]
20. Zhao, R.; Wang, D.Z.; Yan, R.Q.; Mao, K.Z.; Shen, F.; Wang, J.J. Machine Health Monitoring Using Local Feature-Based Gated Recurrent Unit Networks. *IEEE Trans. Ind. Electron.* **2018**, *65*, 1539–1548. [CrossRef]
21. Gonnet, P.; Deselaers, T. Indylstms: Independently Recurrent LSTMS. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020, Barcelona, Spain, 4–8 May 2020; pp. 3352–3356. [CrossRef]
22. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5457–5466. [CrossRef]
23. Peng, W.; Wang, J.; Yin, S. Short term load forecasting model based on attention-lstm in power market. *Grid Technol.* **2019**, *5*, 1745–1751.
24. Kinney, J.B.; Atwal, G.S. Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3354–3359. [CrossRef] [PubMed]
25. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
26. PHM Society Conference Data Challenge. Available online: https://phmsociety.org/phm_competition/2010-phm-society-conference-data-challenge/ (accessed on 22 September 2022).