



An approach for tool wear prediction using customized DenseNet and GRU integrated model based on multi-sensor feature fusion

Xianli Liu¹ · Bowen Zhang¹ · Xuebing Li¹ · Shaoyang Liu¹ · Caixu Yue¹ · Steven Y. Liang²

Received: 19 March 2021 / Accepted: 23 August 2021 / Published online: 26 May 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

An accurate prediction of the machining tool condition during the cutting process is crucial for enhancing the tool life, improving the production quality and productivity, optimizing the labor and maintenance costs, and reducing workplace accidents. Currently, tool condition monitoring is usually based on machine learning algorithms, especially deep learning algorithms, to establish the relationship between sensor signals and tool wear. However, deep mining of feature and fusion information of multi-sensor signals, which are strongly related to the tool wear, is a critical challenge. To address this issue, in this study, an integrated prediction scheme is proposed based on deep learning algorithms. The scheme first extracts the local features of a single sequence and a multi-dimensional sequence from DenseNet incorporating a heterogeneous asymmetric convolution kernel. To obtain more perceptual historical data, a “dilation” scheme is used to extract features from a single sequence, and one-dimensional dilated convolution kernels with different dilation rates are utilized to obtain the differential features. At the same time, asymmetric one-dimensional and two-dimensional convolution kernels are employed to extract the features of the multi-dimensional signal. Ultimately, all the features are fused. Then, the time-series features hidden in the sequence are extracted by establishing a depth-gated recurrent unit. Finally, the extracted in-depth features are fed to the deep fully connected layer to achieve the mapping between features and tool wear values through linear regression. The results indicate that the average errors of the proposed model are less than 8%, and this model outperforms the other tool wear prediction models in terms of both accuracy and generalization.

Keywords Heterogeneous asymmetric convolution kernel · DenseNet · Depth-gated recurrent unit · Feature extraction · Tool wear prediction · Dilated convolution kernel

Introduction

With the continuous development of information technology, increasingly complex and intelligent machinery and equipment are being employed in the manufacturing industry and workshops (Karomati et al., 2020). To meet the fast-growing consumer demand of products, the efficiency and quality of products are particularly important. In the production workshop, tools are an essential part of the cutting process, and their wear seriously restricts the quality and accuracy of

workpieces. Continuous tool wear is often accompanied by changes in the contact state between the cutting edge and the workpiece, and if the tool is not changed in time, it can even cause the workpiece to be scrapped and the equipment to stop, which may eventually lead to substantial economic losses and heavy casualties. According to the statistics, tool failure accounts for 7–20% of the milling machine’s total downtime, and the costs of tools and tool changes account for 3–12% of the total processing cost. Reasonable and timely tool changes can increase the production efficiency by 10–60% (Yang et al., 2019; Zhou & Xue, 2018). Therefore, the tool condition monitoring (TCM) technology is vital for reducing the equipment downtime, improving production efficiency, and lowering processing costs. Commonly used TCM methods are categorized into direct and indirect measurement methods. In the direct measurement methods, the flank wear is measured through a specific technique after shutdown, while in the indirect measurement methods, a mapping relationship

✉ Xianli Liu
Xianli.liu@hrbust.edu.cn

¹ Key Laboratory of Advanced Manufacturing and Intelligent Technology, Ministry of Education, Harbin University of Science and Technology, Harbin 150080, People’s Republic of China

² George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta 30332, USA

is established between the tool wear and wear-related signals (Mohanraj et al., 2020). Obviously, the data-driven indirect measurement methods do not need to stop the machine, and they are more suitable for real-time monitoring of tool status.

The data-driven TCM methods are based on advanced sensor technology (Peltier and Buckley, 2020), industrial big data (Rousseaux, 2017), and digital twin (Qiao et al., 2019) to achieve online monitoring. Different sensors are installed during processing, and the collected signals are often analyzed in the time-domain, frequency-domain, and time–frequency domain to extract features that are strongly related to the tool wear such as mean, variance, kurtosis, spectrum, cepstrum, and wavelet coefficients (XuTing et al., 2020); the extracted features are based on a certain wear recognition method, such as Kalman filter (Tiwari et al., 2018), hidden Markov model (Li & Liu, 2019), support vector machine (Jurkovic et al., 2018; Qu et al., 2020) and other machine learning models. However, such methods have some drawbacks. On the one hand, these statistical features require appropriate feature extraction methods, which are often time-consuming and labor-intensive, and they entail professional knowledge and skills in related fields. In the era of big data manufacturing, feature extraction based on experience is still challenging (Wang et al., 2019). On the other hand, the features obtained by signal processing methods only contain shallow information. To adapt to the stricter conditions and realize a more accurate monitoring during processing, shallow information is often not enough, and in-depth information may be needed to characterize the complex structure in the original data. Further, the signals collected during the machining process are sequential. If they are ignored, a large amount of information related to tool wear can be lost. The above limitations can be surmounted by using deep learning methods.

Deep learning achieves autonomous feature learning on input signals, transforms the initial "low-level" feature representation into "high-level" features, and avoids manual feature extraction, so it has received widespread attention, especially in tasks such as speech recognition (Dokuz & Tufekci, 2021), image classification (Choudhary et al., 2019; Pin et al., 2021), and natural language processing (Alireza et al., 2020). Typical deep learning techniques have been used to monitor the state of machinery and equipment. Jaini et al. (2020) proposed a multi-layer perceptron (MLP) feedforward neural network (FF-NN) model for multi-class classification of tool status, Cao et al. (2019) combined derived wavelet frames and convolutional neural network (CNN) to establish a method for identifying tool wear status using machine spindle vibration signals. Cooper et al. (2020) used a two-dimensional deep CNN to extract information from acoustic signals for TCM. To better detect bearing faults, Liu et al. (2021) proposed a variational autoencoding generative adversarial network, which expanded the data

of a small number of fault categories. The results showed that the generated data was similar to the original sample, and the data-enhancing ability and diagnostic performance of the proposed method was better than those of traditional methods. The above studies realized TCM based on deep learning methods. However, compared with monitoring the state of machinery and equipment, predicting the state of equipment is more meaningful for big data-driven intelligent manufacturing (Huibin et al., 2019).

In the recent years, deep learning methods have also been applied for the prediction of tool wear and remaining useful life. Chen et al., (2018) proposed a deep belief network (DBN) based on deep learning methods to predict the flank wear of cutting tools. Aghazadeh et al., (2018) proposed a hybrid feature extraction method using wavelet time–frequency transformation and spectral subtraction to extract the features from current signals and used CNNs to predict the tool wear. Huang et al. (2019) proposed a multi-domain feature fusion method based on a deep CNN to predict tool wear. This method fused multiple sensors features and achieved better overall performance than the methods based on the extraction of single sensor features. Compared with traditional machine learning methods, the above methods exhibited significantly improved prediction accuracy, but the time-varying features of tool wear were not considered, which could lead to information loss. For capturing the nonlinear dynamical features in time series data, it is necessary to retain the memory of historical information. For the prediction of time series data, recurrent neural network (RNN) is introduced to solve the storage problem of historical input data in the model, which can quickly reflect the dynamical changes in the input data. Yu et al. (2020) used an autoencoder scheme based on bidirectional RNN for the prediction of remaining useful life (RUL). The results showed that the proposed method provided more robust and reliable RUL predictions than any independent algorithm on all the studied datasets. However, with the advent of big data, predictive models should be able to capture long-term memory. Due to the single internal structure of the conventional RNN, a large amount of data is gradually forgotten, so it cannot capture long-term dependencies in the data. To overcome these limitations of conventional RNN, long short-term memory network (LSTM) has been developed as a variant of RNN for modeling temporal sequences (Hochreiter & Schmidhuber, 1997). As a simplified version of LSTM, the GRU has a more straightforward structure (Cho et al., 2014). Compared with LSTM, GRU has fewer parameters, so it can be trained faster. Youdao et al. (2020) reviewed the performance of RNN and its variants in the prediction of RUL. It was observed that the LSTMs and gated recurrent unit (GRU) networks were superior to the basic RNNs. LSTM and GRU are specially designed to deal with the disappearance of gradients, and it effectively captures the long-term dependencies

in the sequential data, making it quite effective in various prediction tasks. Cai et al. (2020) proposed a hybrid information system based on LSTM in which stacked LSTM was used to extract deep features contained in the time-series of multi-sensor acquisition signals for tool wear prediction. This system exhibited good performance under different working conditions. Wang et al., (2019) proposed a hybrid prediction scheme to predict the tool wear, which was realized by a newly developed deep heterogeneous GRU model and local feature extraction. The experimental results showed that the method achieved significant performance in the multi-step prediction task and the long-term prediction task.

Deep learning methods have achieved great success in predicting the state of equipment (Liu & Zhu, 2020; Skordilis & Moghaddass, 2020; Xia et al., 2020) but some issues still need to be resolved. For example, although CNNs can effectively extract the local features of signals, they have been rarely utilized to capture historical information. Richer historical features are usually extracted by increasing the receptive fields, but the conventional convolution operation has limited receptive fields under the same parameters. To increase the perceptual view in the historical field, only a large convolution kernel can be used, which leads to an increase in the computational complexity. Compared with a single sensor signal, the multi-sensor signal can reveal the variation trend of tool wear more accurately and completely (Zhang et al., 2015). It has been confirmed that the convolution kernels of different receptive fields can extract richer features (Szegegy et al., 2014), but few researchers have used them in multi-sensor information fusion. Further, RNNs and variants can handle time series problems, but they do not consider spatial dependence. The above problems can cause the loss of information, so the model cannot be fully utilized. Therefore, the effective mining of the fusion features of multiple sensors is still a challenging problem.

In this study, a deep learning integrated model (HACDNet-GRU) based on DenseNet stacked with heterogeneous asymmetric convolution kernel and GRU is proposed for tool wear prediction. With its jump connections and dense connections, the DenseNet network can “widen” the filters of each layer of the model while learning fewer parameters and exhibits excellent performance at the same time. Therefore, to deeply mine the fusion features of sensors, different heterogeneous asymmetric convolution kernels are used in the dense blocks of different signal lengths to extract the features of single signal and multi-dimensional signal. For extracting the features of a single signal, two convolution kernels with different expansion rates are used to increase the perceptual historical field. For extracting multi-dimensional signal features, different one-dimensional (1D) and two-dimensional (2D) convolution kernels are used. Finally, they are stitched together, which indicates the fusion of features at different scales. Through adaptive learning, the loss of information in

the traditional manual feature extraction process is avoided to overcome the challenges in modeling large-scale datasets. During the process of tool wear prediction, the wear value that has a strong correlation with the current moment must be near this moment, i.e., the closer it is to the current moment, the more valuable the wear data. Therefore, to capture the temporal pattern in the extracted feature sequence, the adaptively extracted fused features are used as the input of the GRU, and the number of hidden units corresponds to the amount of information memorized between the time steps (hidden state). The time-series features are captured through the conversion of update and reset gates and selective input. Finally, the tool wear values are output by linear calculation of different hidden units in the fully connected layer. The experimental results indicate that the proposed model is superior to the traditional methods for tool wear prediction.

In summary, the main contributions of this article are as follows:

- To meet the requirements of adaptability and autonomous decision-making in intelligent manufacturing in the case of big data, an integrated model based on deep learning methods is developed to predict tool wear in real-time, which can be used to maintain the safety of operators and reduce production accidents.
- The multi-signal fusion method is adopted to reflect the tool wear status. Considering the information loss that may be caused by manual feature extraction, the original time-domain signal is collected as the input of the model during the cutting process. A heterogeneous asymmetric convolution kernel is added to extract the features of single sequence signal and multi-dimensional signal, which are merged to indicate the influence of multi-dimensional signals on tool wear. At the same time, a GRU is added for mining the sequence’s time pattern.
- To evaluate the model’s predictive performance, the data from American Society of Prediction and Health Management (PHM) is used as the experimental dataset. The results verify that the model is superior to other traditional models.

Framework for tool wear prediction

In this study, an effective method is proposed for predicting the tool wear during machining, focusing on the time-varying trend of tool wear, and its framework is shown in Fig. 1. It is extremely challenging to observe tool wear online during processing, but processing parameters such as force, vibration, and acoustic emission signals can be acquired online. Let the signal collected at time t be X_t . After the tool pass

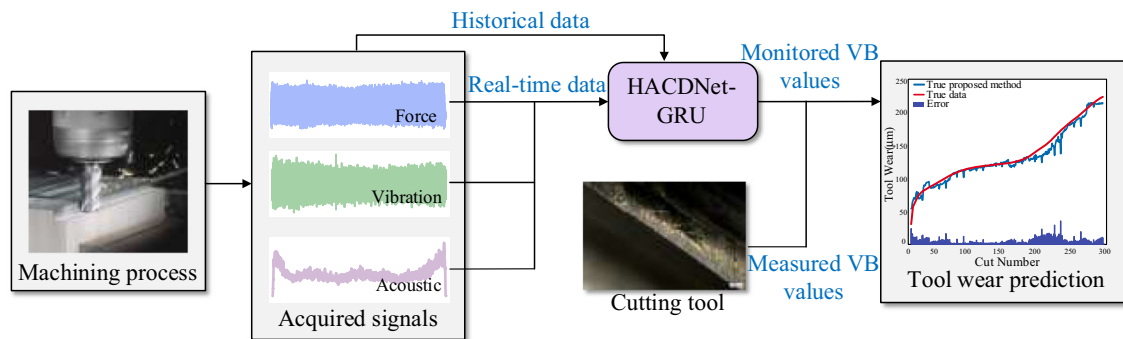


Fig. 1 Tool wear prediction framework

is completed, the tool wear is measured and marked offline by a microscope. The predicted tool wear value after the tool pass is denoted by Y_t . X_t and Y_t can be expressed as follows:

$$X_t = [X_t^1, X_t^2, X_t^3, \dots, X_t^i]^T \quad (1)$$

$$Y_t = F(X_t) \quad (2)$$

where $F(\cdot)$ is the mapping from the input value to the output value of this model, X_t^i is a $(1 \times d)$ tensor, i is the dimension of the cutting parameter, and d is the length of the sequence signal.

In the initial state of HACDNet-GRU, the weight of each parameter is not determined. Therefore, the historical data is used to train the model. According to the model's performance in predicting tool wear value, the model's parameters are determined through continuous iteration and adjustment. To improve the model's accuracy and generalization ability, the number of hidden layer neurons and the time step, the convolutional layer parameters, the pooling layer, the fully connected layer, and the model depth are reasonably optimized. The key features of HACDNet-GRU are described in detail in the next section.

Proposed models

In this section, the structure of HACDNet-GRU is described. In this model, DenseNet incorporates heterogeneous asymmetric convolution kernels to extract single signal features and multi-dimensional signal features related to the tool wear and then fuses them together. The deep GRU is used as the back end to extract sequential time-series features. Finally, the features are mapped to the tool wear values through the regression layer. The workflow and structural framework of the model are shown in Fig. 2.

The system comprises three parts: the data acquisition module, the deep learning feature extraction module, and

the tool wear prediction module. The implementation steps are summarized as follows:

Step 1: The signal collected from the multiple sensors is transformed into a 7×1000 time-domain signal after pre-processing.

Step 2: The preprocessed time-domain signal extracts the local features of a single signal and multi-dimensional signals through DenseNet integrating heterogeneous asymmetric convolution kernels (HACDNet), and then the extracted features are connected in series to achieve information fusion of multi-channel sensors. Based on the structural advantages of DenseNet, high-dimensional and low-dimensional features are continuously extracted as the next layer of the network's input.

Step 3: The designed deep GRU extracts the time-series features from the multi-sensor fusion information.

Step 4: A deep, full-connected regression layer is designed to realize the mapping between the features and tool wear values.

Data preprocessing

If the complete time-domain signal of each sample is used as the model's input, the amount of data will be too large, and the model training efficiency will be significantly reduced. Therefore, the length of the time-domain signal must be reduced. Figure 3b shows the time-domain signals of milling force in three directions with an average flank wear value of $91.8 \mu\text{m}$. The data collected by the sensor fluctuates due to the feed and retraction of the tool during the milling process. The milling force at this stage is unstable. However, the signals in the other stages show periodic changes (see Fig. 3a). To obtain a highly reliable signal of cutting process parameter, we remove this part (Ezugwu & Wang, 1997). Next, we further reduce the data sampling length. The window function is used to process the remaining intercepted signal. The number of windows selected in this article is 1000. Because different tool wear values correspond to different lengths of

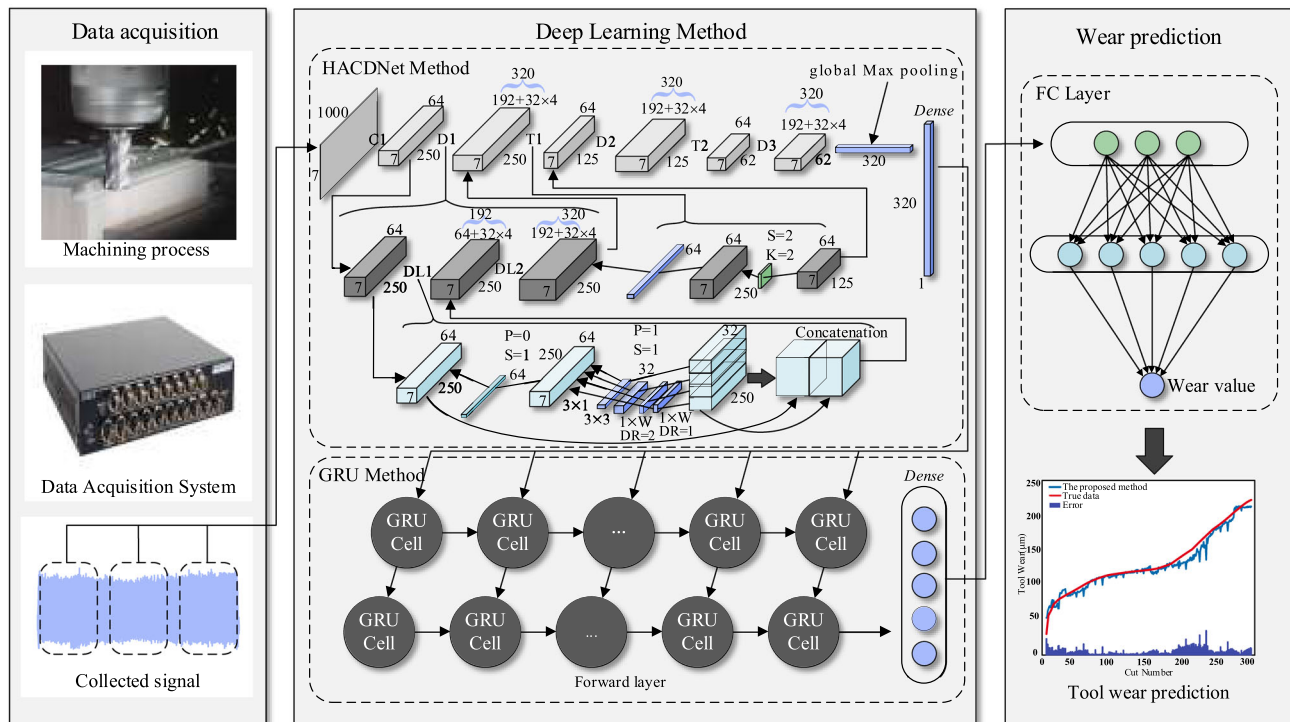


Fig. 2 Model structure and prediction process

collected signals during the cutting process, the window size is selected according to the length of the intercepted signal. Finally, the sampled data of each window is averaged to obtain the signal length (An et al., 2020), which is a tensor with 1000 values, as shown in Fig. 3c.

Local feature extraction module

DenseNet: dense convolutional neural network

The proposed model is based on the DenseNet (Huang et al., 2016) framework. DenseNet is one of the recently improved networks following the milestone innovation of ResNet (He et al., 2016). It incorporates the essential part of the ResNet framework and includes additional innovative characteristics. DenseNet introduces two gradients to enable down-sampling in model. The first one is called densely connected block, which is multiple densely connected, as shown in Fig. 4, and the other one is called the transition layer, which is used for convolution and aggregation. Therefore, each layer of the DenseNet can be directly connected to all the preceding layers to facilitate feature reuse. Simultaneously, each layer in the network has a narrow width, which can deepen the network while learning fewer feature maps to reduce redundancy. Since each layer is connected to all the previous layers in the channel dimension and used as the

input of the next layer, the total number of dense block connections with the L layer is $l(l + 1)/2$. This can be expressed as follows:

$$x_l = H_l([x_0, x_1, x_2, \dots, x_{l-1}]) \tag{3}$$

where $H_l(\cdot)$ is a nonlinear combination transformation function, including a series of BN, ReLU, and convolution operations.

Multi-information fusion feature extraction

The raw time-domain signal X_t collected during the processing is preprocessed. Then, through the convolutional layer and the maximum pooling layer. The maximum pooling layer can reduce the network scale and improve the model’s robustness and generalization ability. The reduced tensor uses DenseNet to further extract features. To reflect the influence of a single signal and a multi-dimensional signal on the tool wear, a heterogeneous asymmetric convolution kernel is added to the dense block to extract the features of sequential signal and multi-dimensional signal. Here, we use the convolution kernel of $F_{filter} \in R^{H \times W \times C}$. The heterogeneous asymmetric convolution kernel structure is $H \times 1, 1 \times W$, and $H \times H$. The input is $I_{input} \in R^{U \times V \times C}$, which represents the feature map with the number of channels as C and the resolution as $U \times V$, and the output is $O_{output} \in R^{Z \times T \times D}$. Since

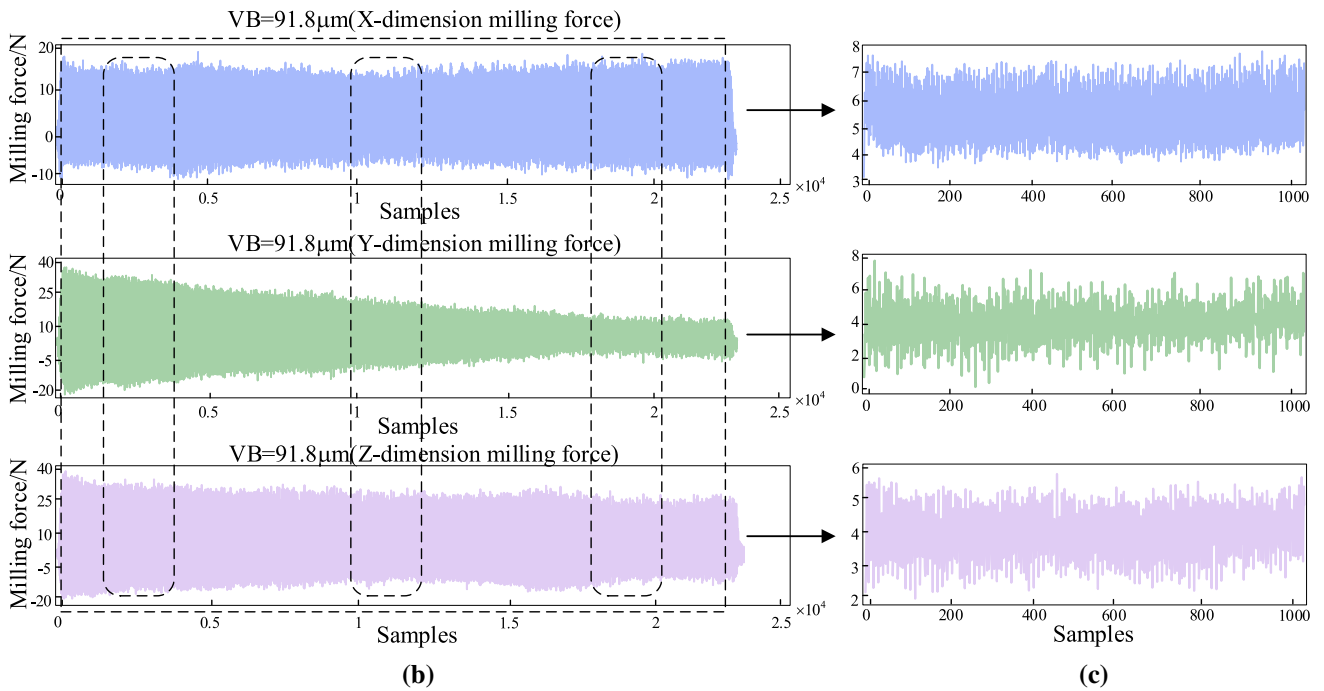
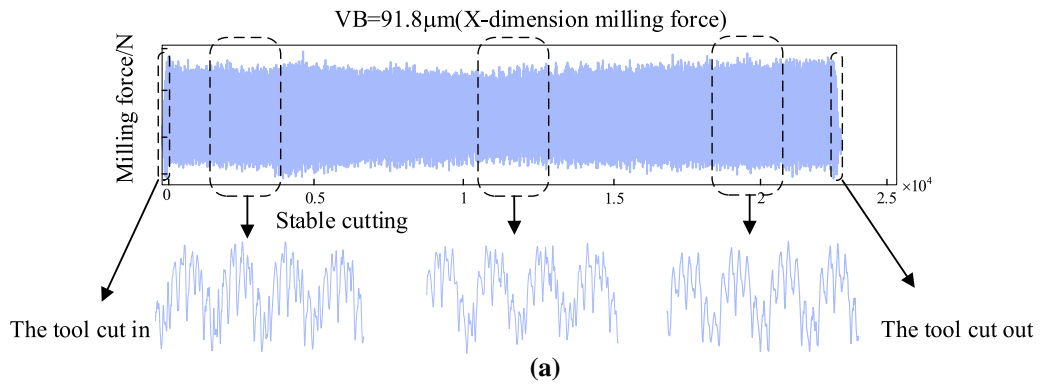
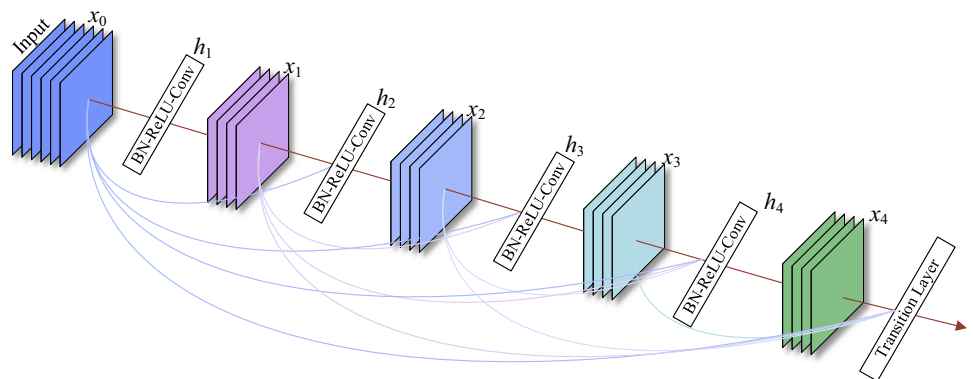


Fig. 3 Original signal of the milling force in X-dimension with $VB = 91.8 \mu m$ and the processed signal

Fig. 4 Dense block



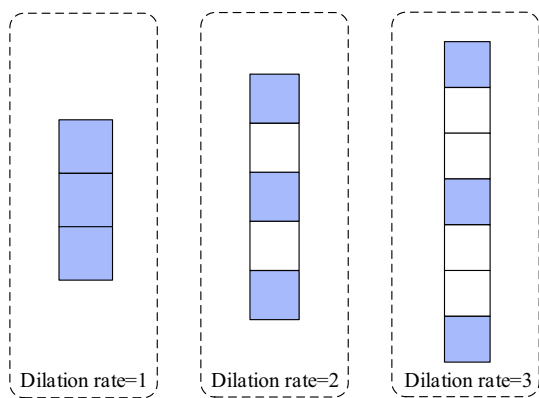


Fig. 5 Dilation kernels with different dilation rates (where rate = 1 is the original convolution operation)

the mining of long-history domain information is essential for the accurate prediction (Chadha et al., 2021), the “dilation” scheme is adopted on the 1D convolution kernel to obtain more historical information. Yu and Koltun (2016) initially proposed to extend the convolution filter. The dilated convolution filter can capture a large context while preserving the details, so it is often used for semantic segmentation. The WaveNet model was proposed in 2016 (Oord et al., 2016). This network analyzes time series of thousands of samples through 1D dilated convolution. In this paper, two dilated convolution kernels with different dilation rates are used to mine the features of the 1D time series. W is the dilation length and W decreases with the shortening of the extracted feature length. For the feature length of 3, 5, and 7, the dilation rate is 1, 2, and 3, respectively. Figure 5 shows the dilated convolution kernels with different dilation rates for univariate time series. For multi-dimensional signals, $H \times 1$ and $H \times H$ convolution kernels are used. The use of different sizes of convolution kernels implies different sizes of receptive fields, and the final stitching indicates the fusion of different scale features.

For the j th filter, the structure of the heterogeneous asymmetric convolution kernel feature extractor is shown in the Fig. 6, and the convolution operation is shown in Eq. (4)

$$Output_{:,i} = f \left(\sum_{k=1}^C Input_{:,k} * F_{filter_{:,k}}^i + \beta_i \right) \tag{4}$$

where $*$ is the convolution operation, $Input_{:,k}$ is the input feature map of the k th channel, and $F_{filter_{:,k}}^i$ is the convolution kernel of the i th filter on the k th channel, $Output_{:,i}$ is the output feature map of the i th filter, β_i is the bias of the i th filter, $f(\cdot)$ is the ReLU function, which can be expressed as follows:

$$ReLU(x) = \begin{cases} 0, & x < 0 \\ x, & x > 0 \end{cases} \tag{5}$$

Convolution operation reduces the spatial dimensionality. Extracting in-depth information can also cause a loss of input features. To ensure that the resolution of the multi-dimensional feature map after convolution operation is consistent with the resolution of the input feature map, the boundary of the input feature matrix is filled with zeros, and the length of the output characteristic matrix (O) after filling is shown in Eq. (6)

$$O = \frac{(w - k + 2p)}{s} + 1 \tag{6}$$

where s is the sliding step size, w is the length of the input feature matrix, p is the size of zero paddings, and k is the filter size. Further, to accelerate the learning process and prevent the gradient from disappearing and exploding, batch normalization is performed on the extracted feature matrix, as shown in Eq. (7).

Input: Values of x over a mini-batch: $B = \{x_1 \dots x_N\}$;
 Parameters to be learned: γ, β
 Output: $\{y_i = BN_{\gamma, \beta}(x_i)\}$

$$\delta_B \leftarrow \frac{1}{N} \sum_{i=1}^N x_i \tag{7}$$

$$\sigma_B \leftarrow \frac{1}{N} \sum_{i=1}^N (x_i - \delta_B)^2 \tag{8}$$

$$\hat{x}_i \leftarrow \frac{x_i - \delta_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{9}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv BN_{\gamma, \beta}(x_i) \tag{10}$$

where δ_B and σ_B are the mean and deviation of the small batch, respectively.

Figure 7 shows the structure of each transition layer and the third densely connected layer in HACDNet. In the densely connected block, the number of output feature maps used for each convolutional layer is very small, and it is 128 after the fusion of different scales. Since fewer parameters are learned, the network based on the DenseNet framework has a regularization effect, i.e., it has a certain inhibitory effect on overfitting. The Concat operation is used to concatenate all the previous layer outputs as the input of the next layer, i.e., the features of each layer are reused. The transition layer is used between two densely connected layers to reduce the number of channels passed to the next densely connected layer. After three densely connected blocks, each dense block contains two dense layers. Finally, through global Max pooling, a feature vector of dimension 320 is generated for each multi-dimensional input signal, which is used for subsequent processing.

Fig. 6 Feature extraction based on heterogeneous asymmetric convolution kernel

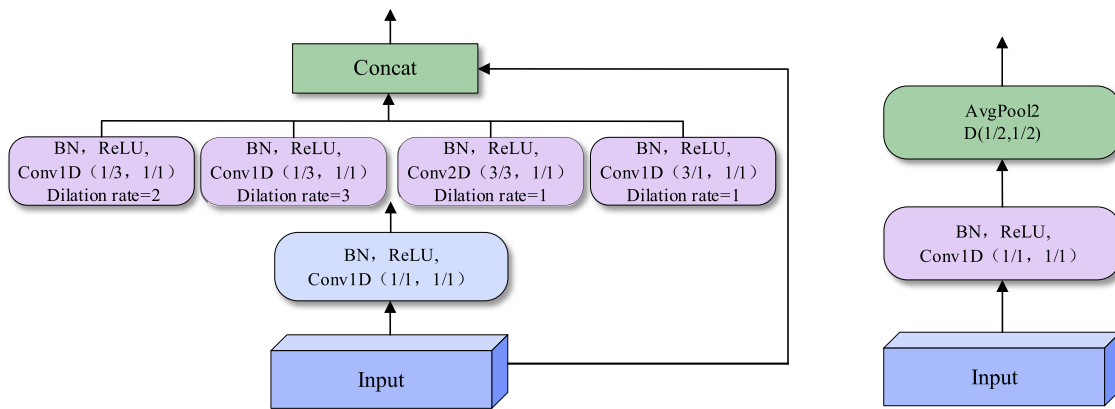
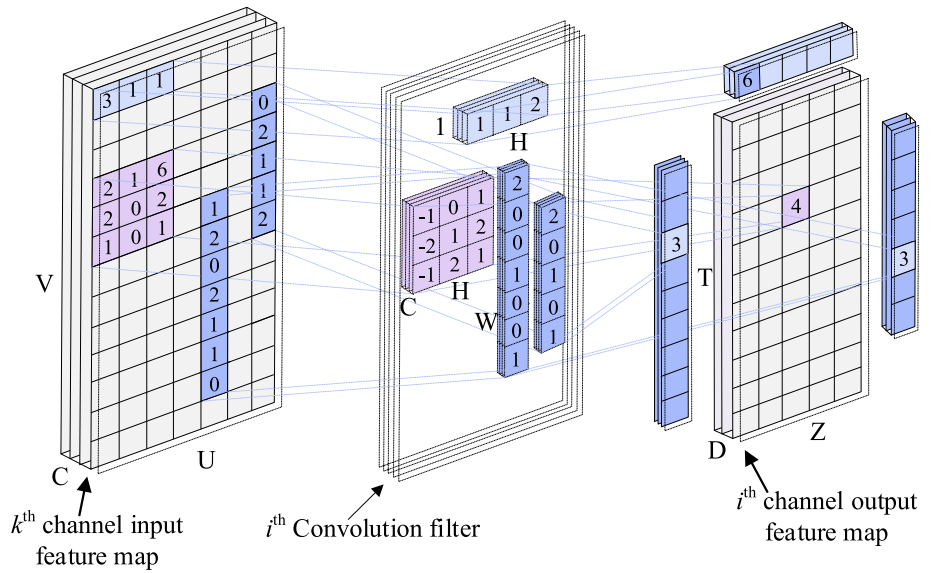


Fig. 7 Structure of the third dense block layer and transition layer

Extraction module of time-series features

Gated recurrent unit

RNNs can use previous information to process current information for capturing the sequence’s time pattern. In theory, RNN can capture all the previous information. However, the principal eigenvalue of the weight matrix of the RNN in the backpropagation process is less than 1, and the gradient can disappear (Sepp Hochreiter, 1998). Therefore, RNN cannot capture the long-term dependencies in the information. To solve this problem, LSTM and GRU have been developed. However, GRU has one less "gate unit" than LSTM, and the parameters are reduced, while the function is equivalent to LSTM, which significantly reduces the training time of the model. As shown in Fig. 8, the GRU has two gate control units: the reset gate and the update gate. Specifically, the reset gate determines how to combine newly entered information with previous memories. When the reset gate is close

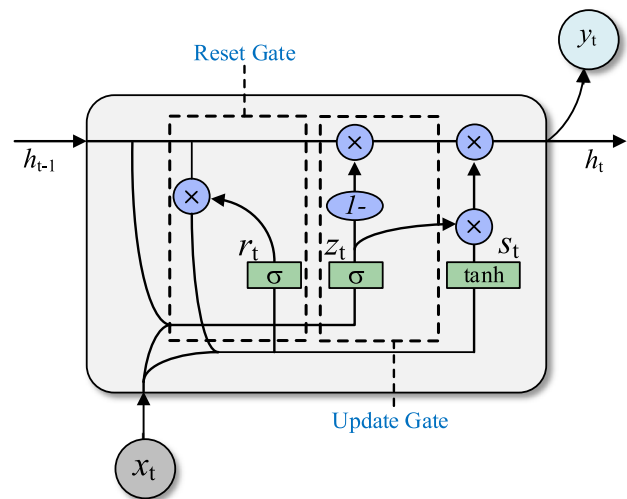


Fig. 8 A GRU cell

to 0, the hidden state is forced to ignore the previous hidden state and is reset with the current input only. The update gate determines how much previous memory is retained, i.e., it controls how much information from the previous hidden state is transferred to the current hidden state.

Update gates, reset gates, memory cells, and new hidden states are calculated as follows:

$$u_{update\ gate} = \text{sigmoid}(W_{ug}X_t + U_{ug}h_{t-1}) \quad (11)$$

$$r_{est\ gate} = \text{sigmoid}(W_{rg}X_t + U_{rg}h_{t-1}) \quad (12)$$

$$\tilde{h}_t = \tanh((r_{est\ gate})_t \otimes U h_{t-1} + W X_t) \quad (13)$$

$$h_t = (u_{update\ gate})_t \otimes h_{t-1} + (1 - (u_{update\ gate})_t) \otimes \tilde{h}_t, \quad (14)$$

where W_{ug} , W_{rg} , W , U_{ug} , U_{rg} , U are the weight matrices of different gates. $u_{update\ gate}$ controls how much h_{t-1} is retained and converted to h_t , i.e., it decides which information is discarded and which new information is added. $r_{est\ gate}$ determines the contribution of h_{t-1} to \tilde{h}_t . If $r_{est\ gate}$ is set to zero, it reads the input sequence and forgets the previously calculated state. \tilde{h}_t is the new information at the current moment, and $r_{est\ gate}$ is the control to retain the previous memory. h_t contains the relevant information of the previous node, and it is obtained by adding h_{t-1} and \tilde{h}_t , and their weights are controlled by $u_{update\ gate}$.

Time-series features extraction

Since tool wear is sequential, the current state of the tool is related to the past state. Therefore, the proposed model incorporates deep GRUs based on adaptive feature extraction. The extraction process of sequential features is shown in Fig. 9. At time t , the fusion features extracted by the HACDNet are fed into the GRU of the first layer. Each feature is non-linearly expressed, stored, and selected by the GRU through its gate mechanism. Then, the hidden state is sequentially passed among hundreds of GRU cells. The information is appropriately selected to ensure that the important information is transmitted and redundancy is avoided. According to Eqs. (11)–(14), the hidden state vector propagating forward from left to right can be obtained. The output hidden state vector of each cell continues to be the input of the second layer GRU, where the GRU uses the last hidden state as the final extracted feature vector. Finally, the hidden state vector of the last GRU that includes all previous GRU information is output, which means that the GRU effectively captures the time-series features related to tool wear. At different times, this vector can be expressed as follows:

$$\vec{F} = [\vec{h}_{T_1}, \vec{h}_{T_2}, \vec{h}_{T_3} \dots \vec{h}_{T_i}] \quad (15)$$

Full-connected regression layer

The features extracted by the GRU network are used as the input of the fully connected network to realize the mapping between the features and the tool wear value. Each neuron in this network is linked with all the neurons in the previous layer. The fully connected neural network has three layers, and the output can be expressed as follows:

$$\hat{y}_i = w_k A_{i-1} + b_k \quad (16)$$

where A_{i-1} is the input feature vector; w_k and b_k are the weight matrix and bias vector of the fully connected layer, respectively. The parameters are updated through back propagation.

Model training and evaluation indicators

The various parameters in the HACDNet-GRU model are obtained through model training. In theory, a deeper network means higher accuracy, but it also means that the network needs to train more parameters, which increases the amount of calculation, and the generalization will also be affected. So, a lot of experimentation is needed to find the right number of network layers. To evaluate the performance difference between different networks, the performance of the model is evaluated by two performance indicators. They are the root mean square error (RMSE) and the mean absolute error (MAE). The expected value of MAE and RMSE is zero, which is the optimal model. For a signal of length n , the mean square error and root mean square error formulas are shown in Eqs. (17) and (18).

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (17)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (18)$$

where y_i and \hat{y}_i denote the actual value and prediction value, respectively.

Deep learning model training relies on large-scale datasets, but in TCM, it is difficult to obtain a large number of datasets for training because tool wear during processing is a long process. Here, the Adam optimizer (Kingma & Ba, 2014) is used to minimize the loss function. To avoid overfitting in the training model, L2 regularization technique is

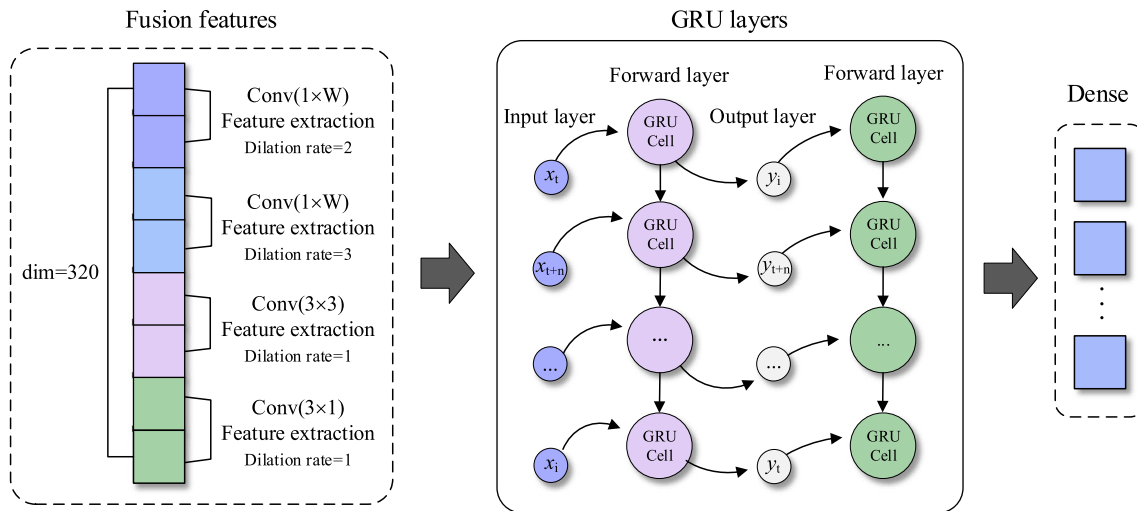


Fig. 9 GRU feature extraction

implemented in the model. The loss function for introducing L2 regularization is defined as follows:

$$loss = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 + \frac{1}{2} \lambda w^T w \quad (19)$$

where w is the weight vector and λ is the regularization factor.

Experiments

Experimental setup

To evaluate the performance of the proposed HACDNet-GRU model, the data of tool wear competition held by the American PHM Association in 2010 (PHM Society Conference Data Challenge; see <https://www.phmsociety.org/competition/phm/10>) is used. A schematic of the experimental platform is shown in Fig. 10. The high-speed CNC machine tool runs under dry milling condition, where the spindle speed is 10,400 rpm, and the feed speed in the X-axis direction is 1555 mm/min. The cutting depths in the Y-axis and Z-axis directions are 0.125 and 0.2 mm, respectively. A 6 mm three-flute ball nose tungsten carbide cutter is used in the milling process, and the workpiece material is stainless steel (HRC52).

To obtain tool wear-related data, a Kistler three-component dynamometer is installed between the worktable and the workpiece to measure the cutting force. Further, three Kistler piezoelectric accelerometers and a Kistler acoustic emission sensor are installed on the workpiece to measure the vibration signal in the three directions and the ultrahigh frequency stress wave pulse signal released by cutting deformation. By reasonably arranging the sensors, the force,

vibration, and acoustic emission signals collected during the cutting process are amplified by the Kistler signal amplifier, and then a data acquisition device (NI DAQ, PCI1200) is used to collect the original time-domain signals. The signal sampling frequency is 50 kHz. Finally, a seven-dimensional signal is obtained. During the machining process, the tool cuts 108 mm in the X direction, and then a microscope (LEICA MZ12) is used to measure the side wear of the tool.

Data preparation

Three tool life tests are conducted on the test bench, which are denoted as C1, C4, and C6. Each test data contains 315 data samples. Simultaneously, the average wear value of the flank surface of the tool corresponding to each sample is used as a label, as shown in Fig. 11. Here, any two datasets are used for model training, and the other is used as a test set. The data used for testing does not appear in the training set. Finally, three sets of data are used to evaluate the performance of the model. The details of the datasets are shown in Table 1.

Experimental results and discussion

To examine the performance of the model, the same dataset is used to compare five methods: support vector regression (SVR), heterogeneous asymmetric convolution combined with GRU (HACNN-GRU), GRU, CNN (DenseNet), and the combination of CNN and GRU (DenseNet-GRU). Here, SVR is a traditional machine learning method, while the other four are deep learning methods. The initial learning rate, maximum number of iterations, and batch size of the proposed method are set to 0.001, 100, and 256, respectively; the regularization coefficient is 0.0008. The essential parameters of the comparison model are set as follows:

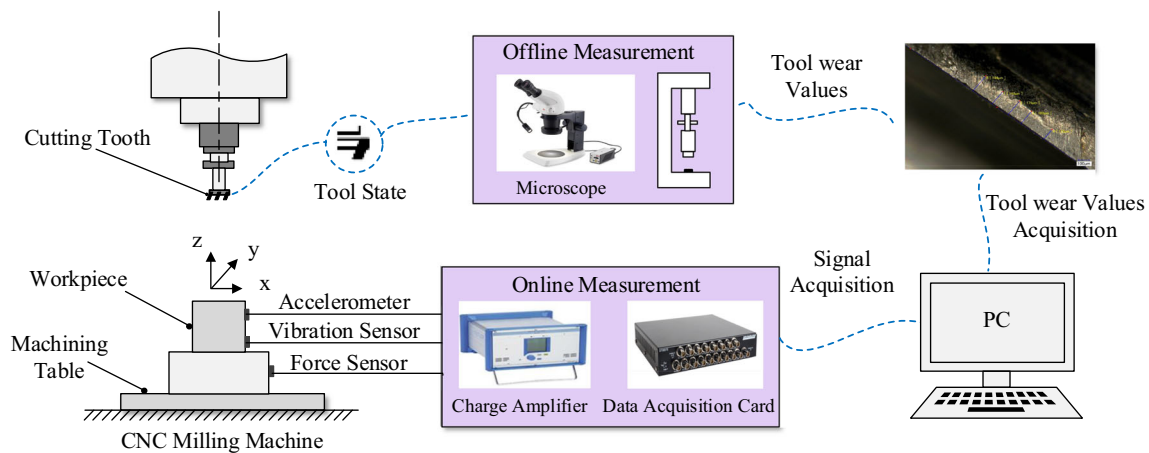


Fig. 10 Schematic diagram of the experimental setup for tool wear prediction

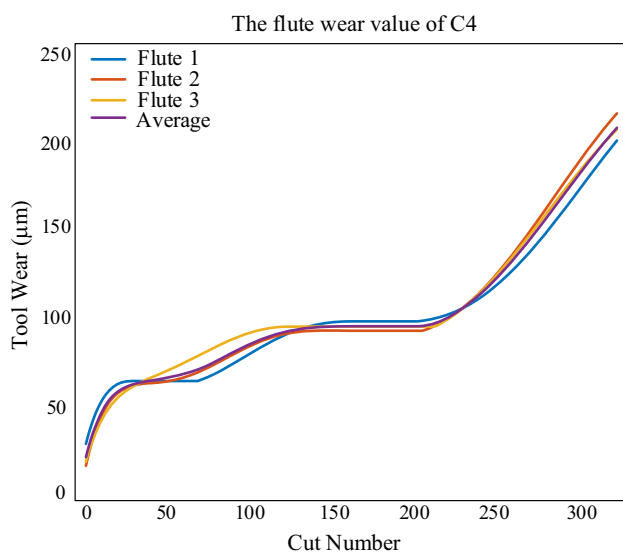


Fig. 11 Flank wear of the cutting tool C4

Table 1 Parameter settings of training set and test set

Symbol	Training set	Test set
S1	C4, C6	C1
S2	C1, C6	C4
S3	C1, C4	C6

- SVR: The time-domain, frequency-domain, and time–frequency domain features of the original signal are extracted. The time-domain analysis reflects the changes in the waveform of the cutting process parameter along the time axis, and the features of the time-domain signal are more comprehensive. As shown in Table 2, ten time-domain statistical features are calculated such as mean, root mean square, peak value, and so on. The time-domain signal

is converted to the frequency-domain signal through fast Fourier transformation, which can reflect other features related to the tool wear. There are ten frequency-domain statistical features, which include mean value, skewness, kurtosis, etc., as shown in Table 3 (XuTing et al., 2020). The time–frequency domain signal reflects the relationship between signal frequency and time. In this study, the wavelet packet energy entropy index is constructed through wavelet packet decomposition to examine the variations in the wavelet packet’s energy in different frequency bands (Li et al., 2008). The energy entropy of the wavelet packet is defined in Eq. (20).

$$H = - \sum_{k=1}^{2^j} p(k) \log_{2^j} p(k) \tag{20}$$

where $p(k) = E(k) / \sum_{k=1}^{2^j} E(k)$, $E(k) = \sum_{i=1}^n (x(i)^2)$, k is the frequency band number, i is the number of discrete points contained in each frequency band signal, and $x(i)$ is the amplitude of the discrete points of the signal.

A matrix of all the statistical features in the time-domain, frequency-domain, and time–frequency domain is constructed as $E'_f = (E'_{f_1}, E'_{f_2}, E'_{f_3}, \dots, E'_{f_{147}})$. However, all the extracted features are not sensitive to the tool wear. To reduce redundant features, E'_f is converted to a low-dimensional matrix through principal component analysis (PCA). In this study, the feature matrix includes ten principal components. The features that reflect tool wear are extracted by PCA and recorded as: $F_p = (F_{p1}, F_{p2}, F_{p3}, \dots, F_{p10})$. The SVR kernel function is linear. Finally, F_p is used as the input of the SVR model to predict tool wear.

- GRU: The model uses a GRU with two hidden layers of 100 neurons to predict the tool wear. The model structure is

Table 2 Time-domain features

Feature	Formula	Feature	Formula
Mean	$X_{mean} = \frac{1}{N} \sum_{i=1}^N x_i$	RMS	$X_{rms} = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right)^{1/2}$
Standard deviation	$X_{sd} = \left(\frac{\sum_{i=1}^N (x_i - X_{mean})}{N-1} \right)^{1/2}$	Shape factor	$X_{sf} = \frac{X_{rms}}{\frac{1}{N} \sum_{i=1}^N x_i }$
Kurtosis value	$X_{kv} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - X_{mean})^4}{\left(\frac{1}{N} \sum_{i=1}^N (x_i - X_{mean})^2 \right)^2}$	Skewness	$X_{skew} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - X_{mean})^3}{\left(\frac{1}{N} \sum_{i=1}^N (x_i - X_{mean})^2 \right)^{3/2}}$
Peak value	$X_{pv} = \max x_i $	Crest factor	$X_{cf} = X_{pv} / X_{rms}$
Impulse factor	$X_{if} = \frac{X_{pv}}{\frac{1}{N} \sum_{i=1}^N x_i }$	Clearance factor	$X_{cf} = \frac{X_{pv}}{\left(\frac{1}{N} \sum_{i=1}^N \sqrt{ x_i } \right)^2}$

Table 3 Frequency-domain features

Feature	Formula	Feature	Formula
Mean of frequency	$X_{meanf} = \frac{1}{L} \sum_{l=1}^L y_l$	Root variance of frequency	$X_{rvf} = \left(\frac{1}{L} \sum_{l=1}^L (y_l - X_{meanf})^2 \right)^{1/2}$
Maximum of frequency	$X_{maxf} = \max(y_l)$	Minimum of frequency	$X_{minf} = \min(y_l)$
RMS of frequency	$X_{rmsf} = \left(\frac{1}{L} \sum_{l=1}^L y_l^2 \right)^{1/2}$	Kurtosis value of frequency	$X_{kvf} = \left(\frac{1}{L} \sum_{l=1}^L \left(\frac{y_l - X_{meanf}}{X_{rvf}} \right)^2 \right)^4$
Skewness value of frequency	$X_{svf} = \left(\frac{1}{L} \sum_{l=1}^L \left(\frac{y_l - X_{meanf}}{X_{rvf}} \right)^3 \right)$	Frequency center	$X_{cf} = \sum_{l=1}^L (f_l y_l) / X_{meanf}$
Skewness factor of frequency	$X_{skfff} = X_{kvf} / X_{rvf}^3$	Kurtosis factor of frequency	$X_{skfff} = X_{kvf} / X_{rvf}^3$

$Input \rightarrow GRU(100) \rightarrow GRU(100) \rightarrow FC(4096) \rightarrow FC(4096) \rightarrow \bar{y}$.

- HACNN-GRU: The model structure is the same as the method proposed in this article, but the connection between the reused features is removed. The traditional CNN with heterogeneous asymmetric convolution kernel is used to predict the tool wear. The model structure is $Input \rightarrow HACNN \rightarrow GRU(128) \rightarrow GRU(128) \rightarrow FC(4096) \rightarrow FC(4096) \rightarrow \bar{y}$.
- DenseNet: The structure of this model is similar to that of the local feature extraction module of the proposed model. The DenseNet does not include a heterogeneous asymmetric convolution kernel. The model structure from top to bottom is $Input \rightarrow Conv(64) \rightarrow Maxpool \rightarrow DenseBlock \rightarrow Transition \rightarrow GMax Pooling \rightarrow FC(4096) \rightarrow FC(4096) \rightarrow \bar{y}$.
- DenseNet-GRU: The structure of this model is $Input \rightarrow DenseNet \rightarrow GRU(128) \rightarrow GRU(128) \rightarrow FC(4096) \rightarrow FC(4096) \rightarrow \bar{y}$.

All the above deep learning models use MSE as the loss function and Adam as the optimization function. L2 regularization technique is used to prevent overfitting and improve the generalization ability of the model.

Figures 12, 13 and 14 shows the prediction results and the error between the actual value and the predicted value for the six models. MAE and RMSE are used to illustrate the prediction performance of each method. The analysis results are shown in Table 4.

Figure 15 shows the average RMSE and MAE of different models for the three experiments. It is clear that the networks that incorporate CNN exhibit good prediction performance. In particular, the effectiveness of convolution is verified through the RMSE and MAE values of 7.82, 9.87, 10.36, 10.15 and 6.27, 7.91, 8.31, 7.75 for the proposed method, DenseNet-GRU, HACNN-GRU, and DenseNet, respectively. There is no significant difference between the DenseNet and DenseNet-GRU in terms of RMSE and MAE. Although GRU can capture local features extracted from the DenseNet, the features extracted by single 1D convolution operation cannot fully reflect the tool wear changes. Therefore, the combination of DenseNet and GRU does not capture too much time-series information related to tool wear. The DenseNet-GRU network with heterogeneous asymmetric convolution kernel can mine single signal and multi-dimensional signal features, and the fused features can comprehensively reflect the changes in tool wear. The mined fusion features fed into the GRU can capture more time-series

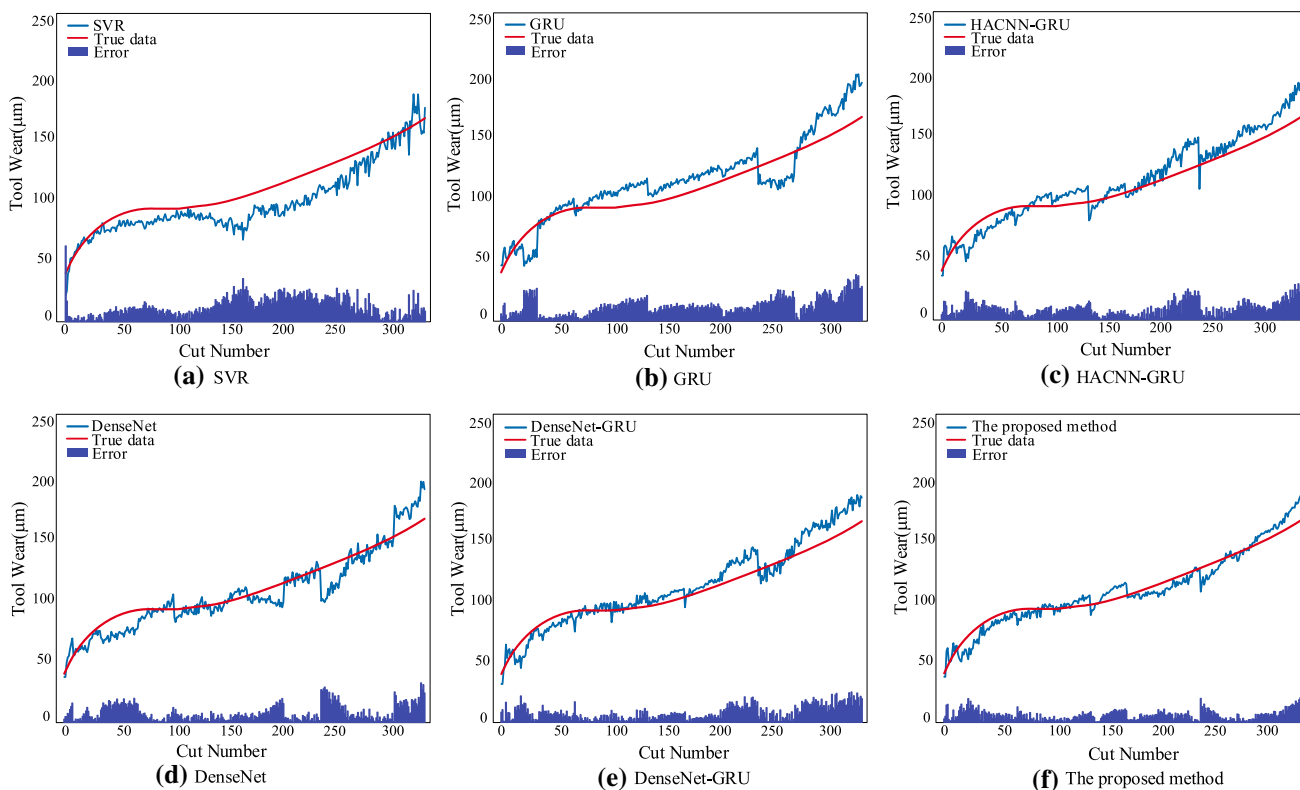


Fig. 12 Prediction results of six different models for experiment number S1

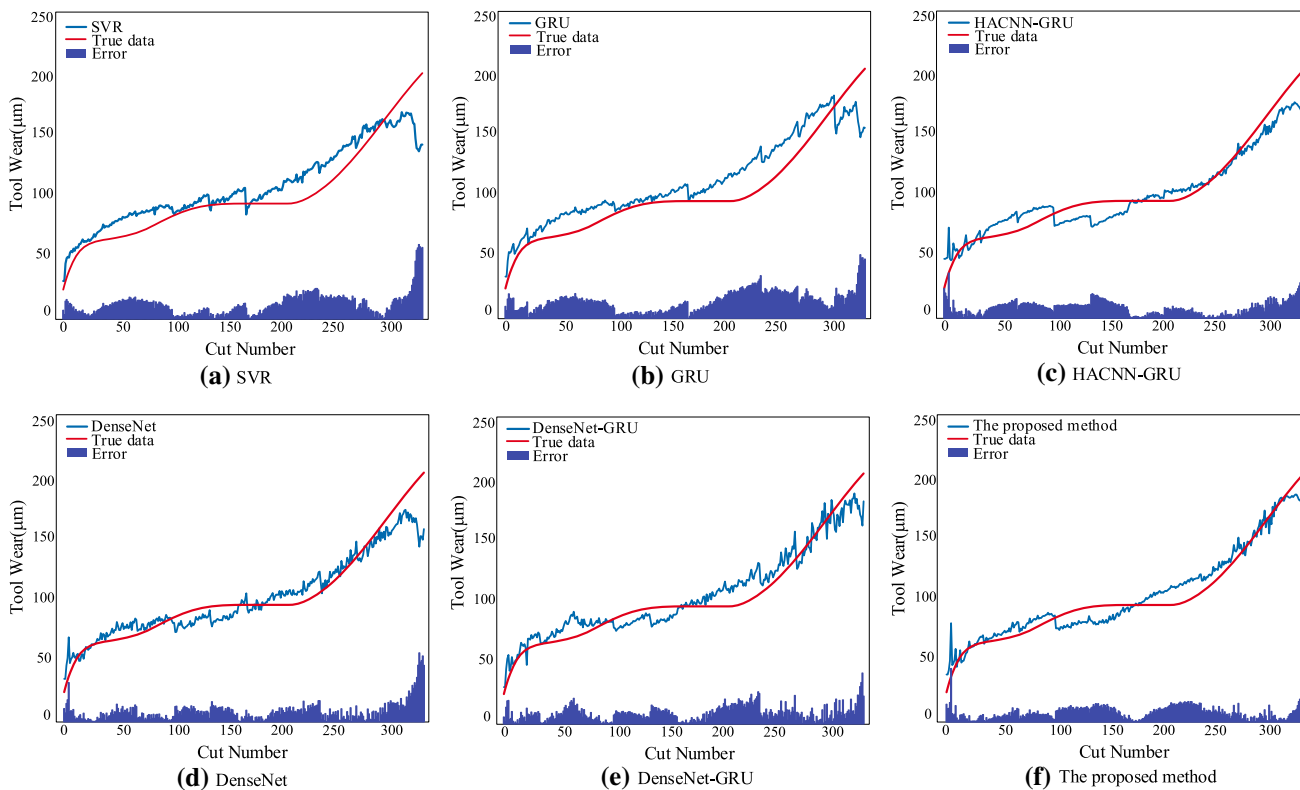


Fig. 13 Prediction results of six different models for experiment number S2

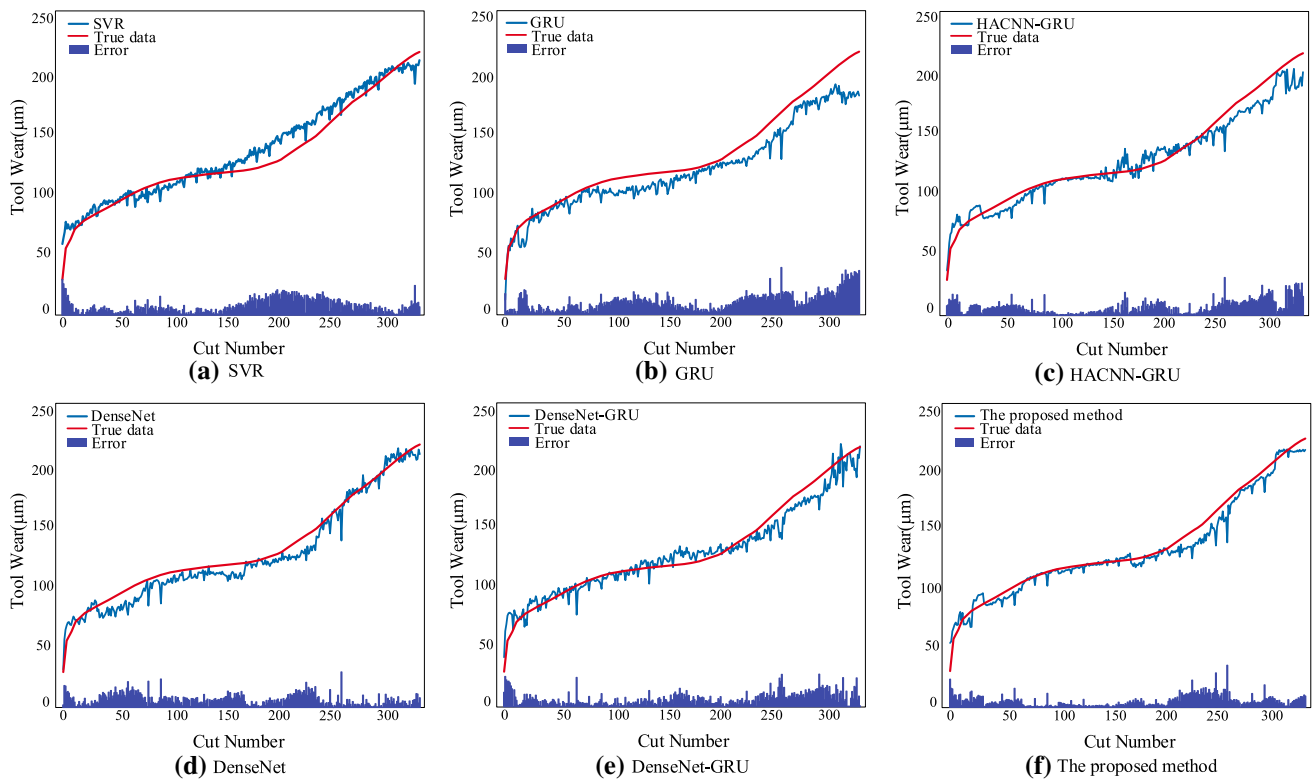


Fig. 14 Prediction results of six different models for experiment number S3

Table 4 RMSEs and MAEs of the compared methods

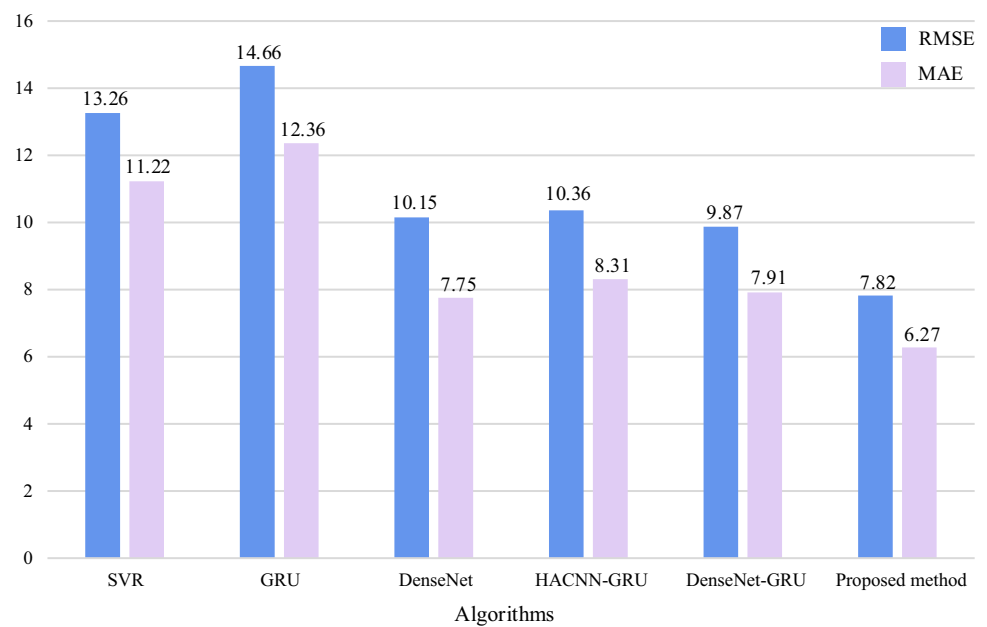
	Test data					
	S1		S2		S3	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
SVR	14.42	12.87	15.97	13.10	9.63	7.71
GRU	14.11	12.04	16.66	14.24	13.23	10.82
DenseNet	10.47	7.98	11.84	8.64	8.14	6.64
HACNN-GRU	11.06	9.23	10.76	8.64	9.27	7.06
DenseNet-GRU	10.02	8.04	10.89	8.95	8.70	6.75
Proposed method	7.21	5.86	9.31	7.80	6.94	6.33

features, so the proposed model outperforms the other models, and its RMSE and MAE are the lowest. Table 4 show that the traditional machine learning algorithm, i.e., SVR is better in some specific situations, but its generalization ability is poor. The GRU without local feature extraction do not work well. The average values of RMSE and MAE are 14.66 and 12.36, respectively. This is because it is difficult to capture the timing information for a large amount of data without considering the multi-sensor spatial information.

Further, we compared the accuracy of HACNN-GRU and HACDNet-GRU. It can be seen that the method based on DenseNet network structure is more effective than the traditional CNN. This can be attributed to the following reasons.

Firstly, one of the main advantages of DenseNet is that the network is narrower, and there are fewer parameters. This may be due to the design of this dense block. In this paper, the number of output feature maps of each convolutional layer in the dense block is very small (128), not hundreds or thousands of widths like other networks. Therefore, the network has a regularization effect, so it has a certain inhibitory effect on overfitting. Overall, since the parameters are reduced, the overfitting phenomenon is alleviated. Secondly, the DenseNet network utilizes the connection of features on the channel to reuse the features, which allows the network to perceive more features related to the tool wear. At the same time, this connection method makes the transfer

Fig. 15 Average values of RMSE and MAE of the different models in the three experiments



of features and gradients more effective, and the network is easier to train. However, the CNN network with the same number of layers cannot achieve feature reuse, and the number of convolution kernels should be increased to learn more features. Therefore, it can easily cause overfitting.

Inspired by Wang et al., (2019), to further verify the superiority of this method, the training set and test set established in Table 1 are used, and the observation model is the HACDNet network. At the same time, the actual tool wear value in the time ranges from start to time $t-n$ is used as the input, and the actual wear value after n time steps is used as the output to establish a deep GRU network for predicting the future tool wear. To verify the long-term prediction performance of the model, only the 30-step advance prediction is considered here, and the result is shown in Fig. 16. Table 5 shows a comparison between the five deep learning methods in terms of RMSE and MAE. It is evident that in addition to the good fitting and advanced prediction performance of the proposed method, the accuracy of the tool wear prediction is further improved. Although the RNN and its variants can capture the hidden information of time series, a single network structure cannot reveal the hidden information in the spatial structure data, and the valuable information is lost.

Through the above comparison, it is evident that the proposed model is superior to the other models. The excellent performance of the model can be ascribed to the following aspects. Firstly, the hidden information of the spatial structure data is effectively extracted by adding a heterogeneous asymmetric convolution kernel to the DenseNet network. Particularly, dilated convolution kernels with different dilation factors are used to perceive more historical

features of a single sequence signal, and 1D and 2D convolution kernels with different receptive fields are used to extract multi-dimensional sequence signal features. The features extracted from different domains are fused, and these fused features can fully reflect the changes in the tool wear. While reducing the sequence length to mine more information about the tool wear, it also reduces the noise of the original signal through a large number of nonlinear operations, thus the features that are strongly related to the tool wear can be easily captured by subsequent GRU units.

Secondly, considering the time-series of the sequence signal, GRU is added to control the update of important information, and finally more timing information related to tool wear is captured. Overall, the proposed model can better predict tool wear by combining local fusion features and time-series features.

Figure 17 shows the calculation time of the four models: DenseNet, DenseNet-GRU, HACNN-GRU, and HACDNet-GRU. In this experiment, the CPU of the computer running the deep learning algorithm is AMD Ryzen7 5800X. The graphics processor of the computer is Nvidia GeForce RTX 3080. It can be seen that the running time of the combination of DenseNet network and GRU is nearly similar to that of the DenseNet, while the running time of the HACNN-GRU network and the DenseNet-GRU model with the heterogeneous asymmetric convolution kernel is approximately 3–4 times higher than that of the DenseNet-GRU network. It may be noted that the main computational time of the model is spent on mining multi-dimensional signal fusion information. The HACDNet-GRU network takes a longer time than the HACNN-GRU network because it needs to learn the reusable features.

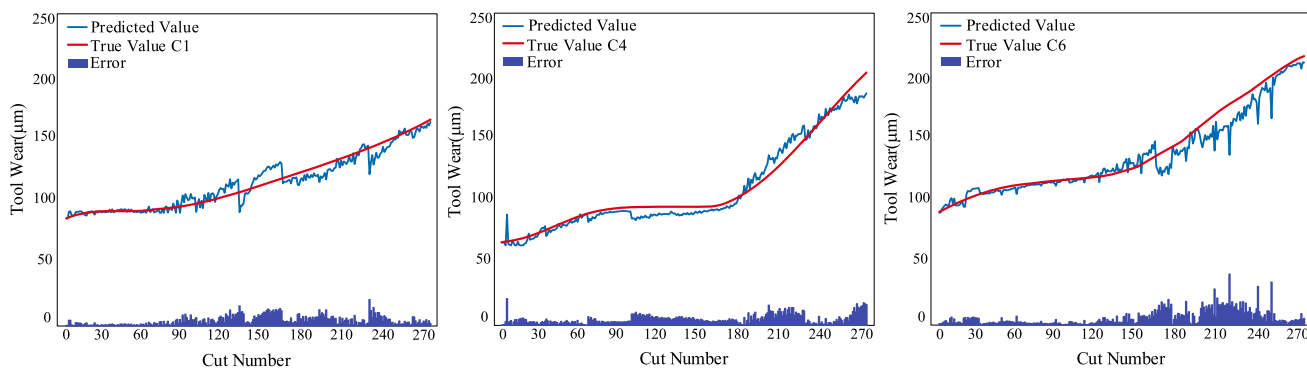
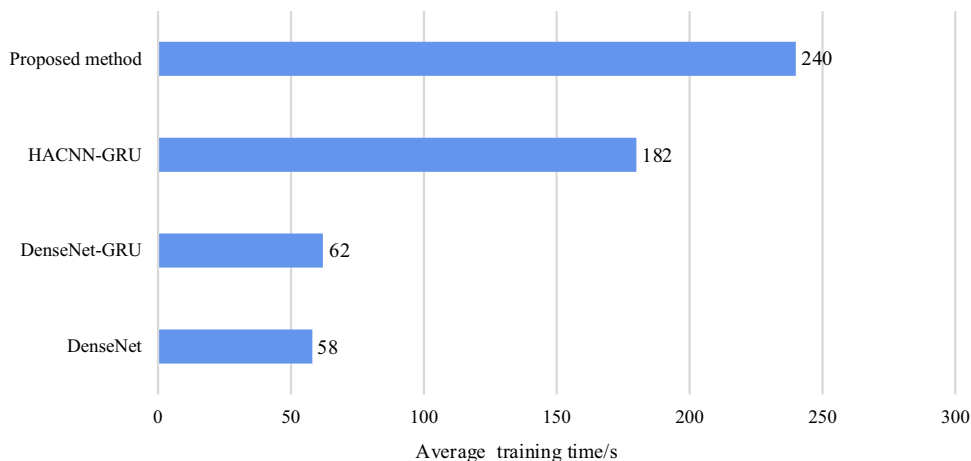


Fig. 16 30-steps-ahead prediction results of the proposed method for the test datasets C1, C4 and C6

Table 5 Performance analysis of different algorithms under 30 steps-ahead prediction

	Test data					
	S1		S2		S3	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
GRU	14.05	11.86	27.86	21.39	11.95	9.05
DenseNet	10.12	6.87	14.31	9.54	11.17	8.17
HACNN-GRU	9.90	7.02	8.55	8.09	8.40	7.31
DenseNet-GRU	7.07	5.39	12.72	8.45	9.49	7.34
DH-GRU(Wang et al., 2019)	6.48	5.18	9.09	6.05	10.16	7.84
Proposed method	5.81	4.50	6.47	5.19	8.61	5.96

Fig. 17 Running times of different prediction models



Conclusion

In this study, a unique model, called HACDNet-GRU, was proposed for tool wear prediction. The efficacy of the model was validated through a large number of experiments and comparison with conventional prediction models. The main results of the study are summarized as follows:

- The multi-dimensional signal can reflect the variation trend of tool wear to a certain extent. Compared with

the 1D convolutional DenseNet, the DenseNet with a heterogeneous asymmetric convolution kernel can facilitate effective mining of fusion features of multi-dimensional signals.

- Tool wear is a dynamic and time-varying process. Compared with the original signal, by dilating the convolution kernel to perceive the historical information of different fields and combining the depth of different receptive fields to mine the spatial information of the multi-dimensional signal, the GRU can easily capture the fused information,

which can provide a more accurate prediction of the tool wear trend.

- The obtained average values of RMSEs and MAEs are 7.82 and 6.27, which verified the feasibility and effectiveness of the model in tool wear prediction. Further, it was confirmed that the proposed prediction model is superior to the conventional deep learning and machine learning models.

During the cutting process, to avoid the loss caused by premature or late tool change, TCM has become a research hotspot in the machining field. Therefore, the data-driven indirect methods to predict tool wear will continue to be popular. With the advancement of sensor technology, cutting process will produce a large amount of information related to tool wear, and the integration of multiple models can provide in-depth information on the tool status. The existing tool wear prediction models are focused on constant working conditions. However, to obtain better interpretability and stability, data-driven methods should incorporate the characteristics of the tool wear process, such as the variable conditions. This forms the future scope of this study as it can boost the industrial applications of the prediction model.

Acknowledgements This research was funded by Projects of International Cooperation and Exchanges NSFC (Grant Numbers 51720105009), National Key Research and Development Project (Grant Numbers 2019YFB1704800), Outstanding Youth Fund of Heilongjiang Province (Grant Numbers YQ2019E029).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

- Aghazadeh, F., Tahan, A., & Thomas, M. (2018). Tool condition monitoring using spectral subtraction and convolutional neural networks in milling process. *The International Journal of Advanced Manufacturing Technology*, 98(9), 3317–3227.
- Alireza, B., Martin, M., David, S., Henrik, M., Muratoglu, O. K., & Varadarajan, K. M. (2020). Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation. *Computers in Biology and Medicine*, 129, 104140. <https://doi.org/10.1016/j.combiomed.2020.104140>
- An, Q., Tao, Z., Xu, X., Mansori, M. E., & Chen, M. (2020). A data-driven model for milling tool remaining useful life prediction with convolutional and stacked LSTM network. *Measurement*, 154, 107461.
- Cai, W., Zhang, W., Hu, X., & Liu, Y. (2020). A hybrid information model based on long short-term memory network for tool condition monitoring. *Journal of Intelligent Manufacturing*. <https://doi.org/10.1007/s10845-019-01526-4>
- Cao, X.-C., Chen, B.-Q., Yao, B., & He, W.-P. (2019). Combining translation-invariant wavelet frames and convolutional neural network for intelligent tool wear state identification. *Computers in Industry*, 106, 71–84. <https://doi.org/10.1016/j.compind.2018.12.018>
- Chadha, G. S., Panara, U., Schwung, A., & Ding, S. X. (2021). Generalized dilation convolutional neural networks for remaining useful lifetime estimation. *Neurocomputing*, 452, 182–199.
- Chen, Y., Jin, Y., & Jiri, G. (2018). Predicting tool wear with multi-sensor data using deep belief networks. *The International Journal of Advanced Manufacturing Technology*, 99(5), 1917. <https://doi.org/10.1007/s00170-018-2571-z>
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. *Computer Science*. <https://doi.org/10.3115/v1/D14-1179>
- Choudhary, M., Tiwari, V., & Venkanna, U. (2019). An approach for iris contact lens detection and classification using ensemble of customized DenseNet and SVM. *Future Generation Computer Systems*, 101, 1259–1270. <https://doi.org/10.1016/j.future.2019.07.003>
- Cooper, C., Wang, P., Zhang, J., Gao, R. X., Roney, T., Ragai, I., et al. (2020). Convolutional neural network-based tool condition monitoring in vertical milling operations using acoustic signals. *Procedia Manufacturing*, 49(C), 105–111. <https://doi.org/10.1016/j.promfg.2020.07.004>
- Dokuz, Y., & Tufekci, Z. (2021). Mini-batch sample selection strategies for deep learning based speech recognition. *Applied Acoustics*, 171, 107573. <https://doi.org/10.1016/j.apacoust.2020.107573>
- Ezugwu, E. O., & Wang, Z. M. (1997). Titanium alloys and their machinability—a review. *Journal of Materials Processing Technology*, 68(3), 262–274.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision & pattern recognition*. <https://doi.org/10.1109/CVPR.2016.90>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(2), 107–116. <https://doi.org/10.1142/S0218488598000094>
- Huang, G., Liu, Z., Laurens, V. D. M., & Weinberger, K. Q. (2016). Densely connected convolutional networks. <https://doi.org/10.1109/CVPR.2017.243>.
- Huang, Z., Zhu, J., Lei, J., Li, X., & Tian, F. (2019). Tool wear predicting based on multi-domain feature fusion by deep convolutional neural network in milling operations. *Journal of Intelligent Manufacturing*. <https://doi.org/10.1007/s10845-019-01488-7>
- Huibin, S., Jiduo, Z., Rong, M., & Xianzhi, Z. (2019). In-process tool condition forecasting based on a deep learning method. *Robotics and Computer-Integrated Manufacturing*, 64, 101924. <https://doi.org/10.1016/j.rcim.2019.101924>
- Jaini, S. N. B., Lee, D.-W., Lee, S.-J., Kim, M.-R., & Son, G.-H. (2020). Indirect tool monitoring in drilling based on gap sensor signal and multilayer perceptron feed forward neural network. *Journal of Intelligent Manufacturing*. <https://doi.org/10.1007/s10845-020-01635-5>
- Jurkovic, Z., Cukor, G., Brezocnik, M., & Brajkovic, T. (2018). A comparison of machine learning methods for cutting parameters prediction in high speed turning process. *Journal of Intelligent Manufacturing*. <https://doi.org/10.1007/s10845-016-1206-1>
- Karomati, B. D., Hsing, C. C., & Lihui, W. (2020). Systematic literature review on augmented reality in smart manufacturing: Collaboration between human and computational intelligence. *Journal of Manufacturing Systems*. <https://doi.org/10.1016/J.JMSY.2020.10.017>

- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- Li, B., Zhang, P., Liang, S., & Ren, G. (2008). Feature extraction and selection for fault diagnosis of gear using wavelet entropy and mutual information. In *2008 9th international conference on signal processing*.
- Li, W., & Liu, T. (2019). Time varying and condition adaptive hidden Markov model for tool wear state estimation and remaining useful life prediction in micro-milling. *Mechanical Systems and Signal Processing*, *131*, 689–702. <https://doi.org/10.1016/j.ymssp.2019.06.021>
- Liu, C., & Zhu, L. (2020). A two-stage approach for predicting the remaining useful life of tools using bidirectional long short-term memory. *Measurement*, *164*, 108029. <https://doi.org/10.1016/j.measurement.2020.108029>
- Liu, S., Jiang, H., Wu, Z., & Li, X. (2021). Rolling bearing fault diagnosis using variational autoencoding generative adversarial networks with deep regret analysis. *Measurement*, *168*, 108371. <https://doi.org/10.1016/j.measurement.2020.108371>
- Mohanraj, T., Shankar, S., Rajasekar, R., Sakthivel, N. R., & Pramanik, A. (2020). Tool condition monitoring techniques in milling process—A review. *Journal of Materials Research and Technology*, *9*(1), 1032–1042. <https://doi.org/10.1016/j.jmrt.2019.10.031>
- Oord, A., Dieleman, S., Zen, H., Simonyan, K., & Kavukcuoglu, K. (2016). *WaveNet: A generative model for raw audio*.
- Peltier, R. E., & Buckley, T. J. (2020). Sensor technology: A critical cutting edge of exposure science. *Journal of Exposure Science & Environmental Epidemiology*, *30*(6), 901–902. <https://doi.org/10.1038/s41370-020-00268-3>
- Pin, W., En, F., & Peng, W. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*, *141*, 61–67. <https://doi.org/10.1016/j.PATREC.2020.07.042>
- Qiao, Q., Wang, J., Ye, L., & Gao, R. X. (2019). Digital twin for machining tool condition prediction. *Procedia CIRP*, *81*, 1388–1393. <https://doi.org/10.1016/j.procir.2019.04.049>
- Qu, D., Zheng, W., Wang, B., Wu, B., & Yi, H. (2020). Nondestructive acquisition of the micro-mechanical properties of high-speed-dry milled micro-thin walled structures based on surface traits. *Chinese Journal of Aeronautics*, *34*, 438–451.
- Rousseaux, F. (2017). BIG DATA and data-driven intelligent predictive algorithms to support creativity in industrial engineering. *Computers & Industrial Engineering*, *112*, 459–465. <https://doi.org/10.1016/j.cie.2016.11.005>
- Skordilis, E., & Moghaddass, R. (2020). A deep reinforcement learning approach for real-time sensor-driven decision making and predictive analytics. *Computers & Industrial Engineering*, *147*, 106600. <https://doi.org/10.1016/j.cie.2020.106600>
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., & Rabinovich, A. (2014). *Going deeper with convolutions*. IEEE Computer Society.
- Tiwari, K., Shaik, A., & Arunachalam, N. (2018). Tool wear prediction in end milling of Ti–6Al–4V through Kalman filter based fusion of texture features and cutting forces. *Procedia Manufacturing*, *26*, 1459–1470. <https://doi.org/10.1016/j.promfg.2018.07.095>
- Wang, J., Yan, J., Li, C., Gao, R. X., & Zhao, R. (2019). Deep heterogeneous GRU model for predictive analytics in smart manufacturing: Application to tool wear prediction. *Computers in Industry*, *111*, 1–14. <https://doi.org/10.1016/j.compind.2019.06.001>
- Xia, M., Zheng, X., Imran, M., & Shoaib, M. (2020). Data-driven prognosis method using hybrid deep recurrent neural network. *Applied Soft Computing Journal*, *93*, 106351. <https://doi.org/10.1016/j.asoc.2020.106351>
- XuTing, M., Feng, Z., Gang, W., Yan, C., & KaiFu, Y. (2020). Semi-random subspace with Bi-GRU: Fusing statistical and deep representation features for bearing fault diagnosis. *Measurement*. <https://doi.org/10.1016/j.measurement.2020.108603>
- Yang, Y., Guo, Y., Huang, Z., Chen, N., Li, L., Jiang, Y., et al. (2019). Research on the milling tool wear and life prediction by establishing an integrated predictive model. *Measurement*, *145*, 178–189. <https://doi.org/10.1016/j.measurement.2019.05.009>
- Youdao, W., Sri, A., & Yifan, Z. (2020). Recurrent neural networks and its variants in remaining useful life prediction. *IFAC-PapersOnLine*, *53*(3), 137–142. <https://doi.org/10.1016/J.IFACOL.2020.11.022>
- Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *ICLR*.
- Yu, W., Kim, I. Y., & Mechefske, C. (2020). An improved similarity-based prognostic algorithm for RUL estimation using an RNN autoencoder scheme. *Reliability Engineering and System Safety*, *199*, 106926. <https://doi.org/10.1016/j.ress.2020.106926>
- Zhang, K. F., Yuan, H. Q., & Nie, P. (2015). A method for tool condition monitoring based on sensor fusion. *Journal of Intelligent Manufacturing*, *26*(5), 1–16.
- Zhou, Y., & Xue, W. (2018). Review of tool condition monitoring methods in milling processes. *The International Journal of Advanced Manufacturing Technology*. <https://doi.org/10.1007/s00170-018-1768-5>

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