IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 April 2024, accepted 24 May 2024, date of publication 30 May 2024, date of current version 11 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406568

RESEARCH ARTICLE

An Augmented AutoEncoder With Multi-Head Attention for Tool Wear Prediction in Smart Manufacturing

CHUNPING DONG[1] AND JIAQIANG ZHAO[2]
[1]School of Computer Engineering, Weifang University, Weifang 261000, China
[2]School of Physics and Electronic Information, Weifang University, Weifang 261000, China
Corresponding author: Jiaqiang Zhao (zhaojiaqiang@wfu.edu.cn)

ABSTRACT Computer numerical control (CNC) machine tools play a crucial role in the manufacturing industry, and cutting tools, as key functional components, directly impact the quality of the machining process. An improved autoEncoder with multi-head attention for tool wear prediction is proposed. MultiCNN-Attention-GRU (MCAG) consists of an encoder and decoder. The encoder contains multiple sets of Convolutional Neural Networks (CNNs) and CNNs adaptively extract signal features. The decoder includes Multi-Head Attention (MHA) and Gated Recurrent Unit (GRU), which can adaptively enhance the relevant feature weights and extract long-term, deep different features. For the model training, a monotonicity loss function is defined. The proposed method is validated on the 2010 PHM Data Challenge (PHM2010) public dataset. The original dataset is dimensionally reduced and then resampled. The experimental results demonstrate the effectiveness of the proposed algorithm, achieving an Mean Absolute Error (MAE) of 6.15 and Mean Square Error (MSE) of 79.6, which is approximately 1.6 and 13 lower than the second-place algorithm. The result validates the superior performance of the proposed model compared to other deep learning algorithms in predicting tool wear.

INDEX TERMS Deep learning, multi-head attention, milling machine tool, remaining life prediction.

I. INTRODUCTION

With the progress of science and computer technology, the advanced manufacturing industry is developing rapidly in the direction of intelligence and information technology [1], [2], [3]. In the contemporary world, all industrialized developed countries carry out strategic layouts. Germany has proposed the national development plan of ''Industry 4.0'', and China has put forward the national strategic plan of ''Made in China 2025'' [4], [5]. CNC machine tools, serving as the ''industrial mothership'' of manufacturing, not only represent the degree of industrial development of the country but also represent the core competitiveness of the country [6]. Within the machining process of CNC machine tools, milling cutters play a vital role as machining tools [7]. The quality and wear status of tools directly influence the machining accuracy, efficiency, and remaining useful life [5], [8]. Therefore, it is of great significance for national development and social development to accurately realize the monitoring of milling machine tool wear condition and wear life prediction, improve the processing efficiency of CNC machine tools, reduce production costs, and improve manufacturing quality.

After long-term research and analysis, the tool monitoring techniques classified into direct and indirect methods [9], [10]. The direct method uses contact sensors or precision optical measuring instruments to directly obtain the tool wear shape for further analyze, with computer vision being a commen approach [11], [12]. This method provides accurate and intuitive wear information but requires stopping the machining process, which cannot be monitored in real-time [13]. Indirect measurement is a method to infer the tool state by measuring the sensor signal data, like vibration, force [14], secondary electron signals [15], [16]. By measuring

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

their changes, the tool wear condition of the tool can be inferred [17]. This real-time method can be monitored without stopping the machine and improve production efficiency [18]. With the development of signal monitoring technology, monitoring accuracy and reliability are also improving, which will further promote the development and application of indirect measurement methods.

The indirect method uses the monitoring signal collected by the sensors to establish the feature map relationship between the sensor data and the different tool wear stages [19]. Feature extraction is to transform the original monitoring signal into feature variables, and there are usually traditional methods and modern methods. Traditional feature extraction methods extract statistical features based on time domain and frequency domain signals, such as mean, variance, peak, etc, [20]. Modern methods use deep learning algorithms, such as Convolutional Neural Network (CNN) [21] and Recurrent Neural Network (RNN) [22], [23], to adaptively extract feature information of original signal data and automatically learn the feature map relationship between tool wear state and monitoring signals.

In recent years, the adaptive feature extraction ability of deep learning has attracted much attention, so it has also been widely used in the field of tool life prediction. An et al. [24] proposed a hybrid model CNN-SBULSTM, which superimposed CNN with bidirectional and unidirectional LSTM (SBULSTM) networks, for sequence data in tool remaining useful life prediction tasks. Xu [25] proposed a multi-scale convolutional GRU network (MCGRU) to process raw sensory data, designed six parallel independent branches of different kernel sizes to form a multi-scale convolutional neural network, and then input these features of different scales extracted from the raw data into the depth-GRU network for tool wear prediction. Wang and Zhang [26] propose an end-to-end deep learning model that uses attention mechanisms and RNN to monitor and predict tool wear. Hu and Tang [27] designed a feature encoder based on bidirectional LSTM network, combined residual GRU network with self-attention, and proposed a ResGRUA model to realize intelligent prediction of tool wear. The combination of typical machine learning algorithm and neural network algorithm can also achieve better results. Yao et al. [28] based on K-means clustering, recurrent fuzzy neural network (RFNN) and genetic algorithm (GA), proposed a recurrent fuzzy neural network (CFRFNN) based on clustering features for prediction of tool wear. Marei and Li [29] proposed a CNN-LSTM hybrid model for predicting cutting tool RUL based on embedded transfer learning mechanism.

To sum up, traditional methods can leverage specific prior knowledge to describe tool wear characteristics, thereby achieving higher accuracy. Recent research on modern approaches indicates that they are more effective. In particular, deep learning offers a promising way for tool wear prediction without the need for manual feature extraction. However, there is currently a lack of methods that integrate prior knowledge with deep neural networks. Moreover, compared with the existing fixed-weight methods, none of them can adaptively adjust the weights between different features to better capture the key information. In this paper, we propose a novel MultiCNN-Attention-GRU(MCAG) for tool wear life prediction. MCAG is trained and tested on the 2010 PHM Data Challenge dataset after downsampled. The main contributions of this paper include the following:

1) MCAG consists of an encoder and decoder. The encoder contains multiple CNNs and CNNs adaptively extract signal features. The decoder includes MHA and GRU, which can adaptively enhance the relevant feature weights and extract long-term, deep different features.
2) A new monotonicity loss function. To optimize the model training process, a monotonicity loss function is defined. The final loss function MNTLoss is composed of MSELoss loss function and MNT monotone loss function.
3) Three different evaluation experiments are performed on the PHM2010 reconstructed dataset to prove the validity of the proposed model.

The rest of the paper is organized as follows: the basic structure of AutoEncoder, the MHA, and GRU are introduced and described in Section II; Section III presents a novel MultiCNN-Attention-GRU(MCAG) for tool wear life predication. In more detail, MCAG consists of an encoder and decoder, where the encoder contains CNNs and the decoder includes MHA and GRU. Section IV presents the details of the dataset, evaluation experiments, results, and a discussion. Finally, Section V presents the conclusion and future work.

II. REVIEW OF RELATED WORK

A. AUTOENCODER

Autoencoder(AE) [30] is an unsupervised learning model based on neural network, as shown in Figure 1. Feature representations are extracted from the input data and these features are used to reconstruct the input data. The Autoencoder is composed of an encoder and a decoder. The encoder extracts features from input data while simultaneously compressing the data's dimensions. The decoder maps the vectors in the latent space back to the original data's dimensionality, thereby reconstructing the original data.

$$y = f\left(Wx + b\right) \tag{1}$$

$$x' = f\left(W'y + b'\right) \tag{2}$$

where x represents the input data, W is the weight matrix of the encoder, b is the bias coefficient, $f(.)$ Is the activation function, y is the output of the encoder, which can also be seen as a representation in the potential space, x' represents the output of the decoder, that is, the data reconstructed by the decoder, W' is the weight matrix of the decoder, and b' is the decoder bias coefficient.

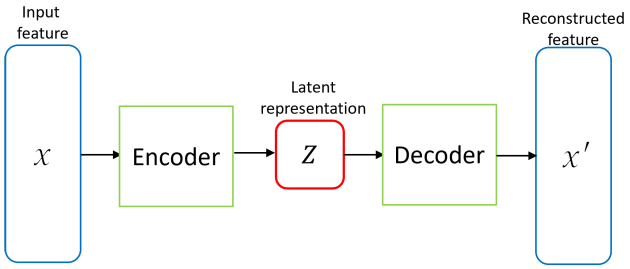

FIGURE 1. Schematic diagram of autoencoder.

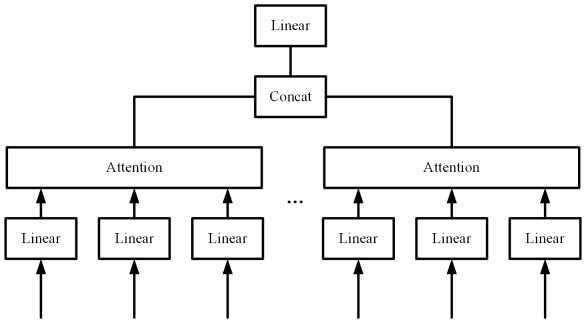

FIGURE 2. Schematic diagram of MHA.

B. MULTI-HEAD ATTENTION

The Multi-Head Attention(MHA) [31] mechanism is a model structure based on attention mechanisms. The main purpose of MHA is to increase the expressiveness and learning ability of the model, especially for the processing of long-series data, as shown in Figure 2. The core idea of MHA is to map the input sequence into different spaces, and then by weighted summation of these mapped spaces, obtain the final output vector. The equation for calculating the attention weight is shown as 3. Specifically, the multi-head attention mechanism consists of multiple heads, each of which is an independent linear mapping that transforms the input sequence into a new vector space. Then, each head calculates the correlation between the input sequence and its query vector and normalizes it to get the attention distribution of each head. Finally, the output vectors of all the heads are concatenated or weighted to get the final output vector. Each head can learn different aspects of the input sequence information, and weigh the different information to improve the expression ability of the model.

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K, V are the Query vector, Key vector, and Value vector matrices, respectively, $\sqrt{d_k}$ represents the scaling factor, *softmax* indicates the normalization process, and T denotes the transpose operation.

C. GATED RECURRENT UNIT

The Gated Recurrent Unit (GRU) [32] is a variant of the Recurrent Neural Network (RNN) model that avoids the shortcomings of the standard RNN. Compared with traditional RNN models, GRUs have fewer parameters and better memory, and can effectively process long time series data. At the core of GRU is a gating mechanism that flexibly controls the flow and update of information to achieve better results in sequence data modeling tasks. GRU mainly consists of two Gate control units (Reset Gate and Update Gate), as shown in Figure 3. In addition, compared with LSTM neural network [33], GRU has better performance and fewer parameters, which can inhibit overfitting.

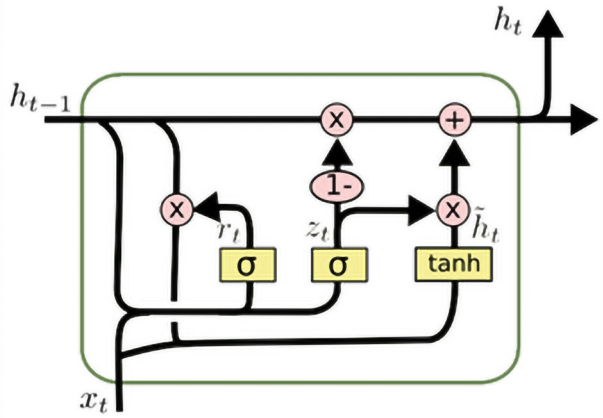

FIGURE 3. Schematic diagram of GRU [32].

At each time step, the GRU receives input x_t and the hidden state $h_(t-1)$ of the previous time step, and outputs the hidden state h_t of the current time step.

The update gate is used to control whether the status information of the last time and the current input information should be updated. The sigmoid activation function in the update gate weights and compresses the input to a value between 0 and 1, indicating how open the update gate is. When the output of the update gate is close to 1, it means that the status information of the current moment needs to be updated. When the output of the update gate is close to 0, it means that the status information of the previous time is retained.

The reset gate is used to control how the status information of the previous moment is combined with the current input information. The sigmoid activation function in the reset gate similarly weights and compresses the input to a value between 0 and 1, indicating how open the reset gate is. When the output of the reset gate is close to 1, it means that the status information of the previous moment is ignored, and only the input information of the current moment is used. When the output of the reset door is close to 0, it indicates that the status information of the previous moment is retained and historical information is paid more attention.

$$z_t = \sigma_g\left(W_z x_t + U_z h_{t-1} + b_z\right) \tag{4}$$
$$r_t = \sigma_g\left(W_r x_t + U_r h_{t-1} + b_r\right) \tag{5}$$
$$\tilde{h}_t = tanh\left(W_h x_t + U_t\left(r_t \odot h_{t-1}\right)\right) \tag{6}$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{7}$$

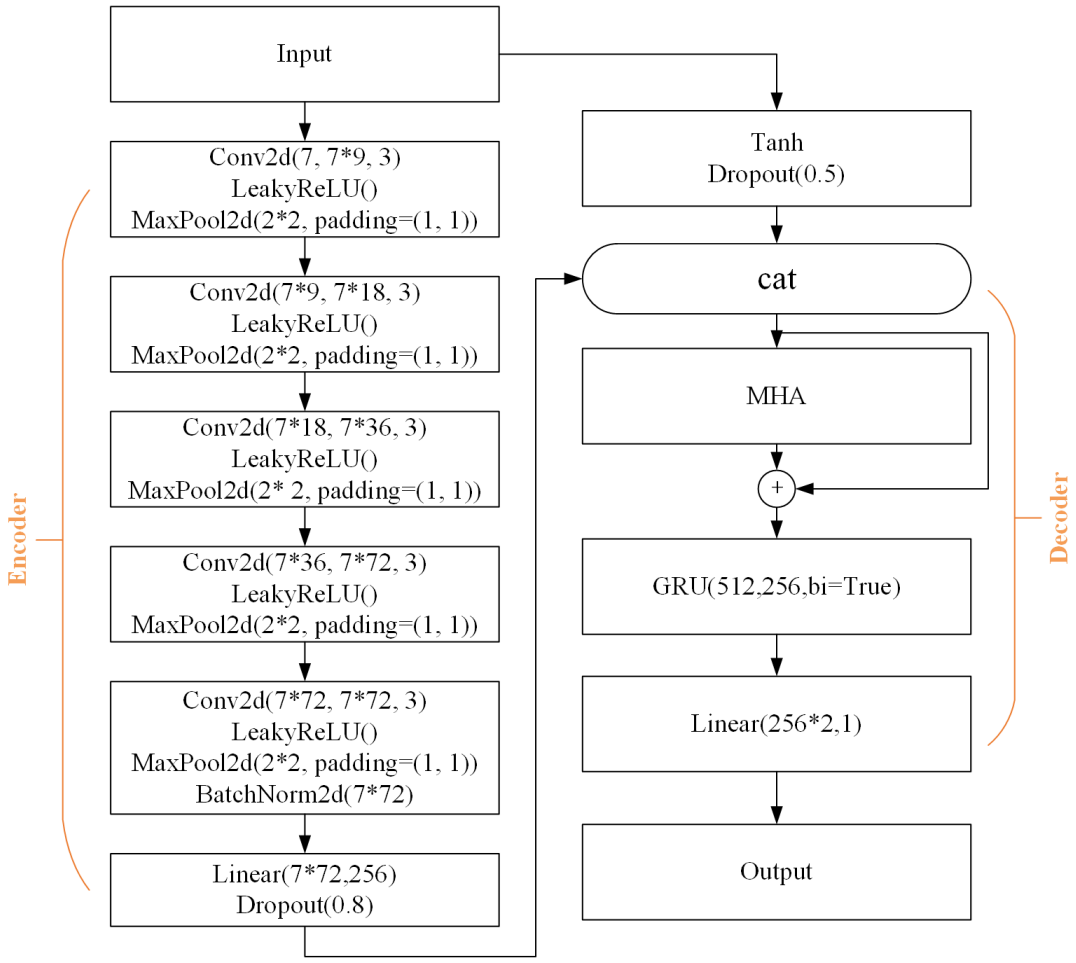

FIGURE 4. Framework of the proposed method.

where W_r, U_r, and b_r represent the weight, parameter vector, and bias vector of the reset gate r_t, respectively; W_z, U_z, and b_z correspond to the weight, parameter vector, and bias vector of the update gate z_t; h_t and $\tilde{h}_t$ denote the output state and candidate state at time t, respectively; z_t and r_t represent the update gate and reset gate, respectively; *tanh* refers to the hyperbolic tangent activation function, and σ_g represents the sigmoid activation function.

III. THE PROPOSED METHOD

A. OVERALL FRAMEWORK

Deep learning(DL) technology plays an increasingly important role in the field of time series data prediction. Based on the existing DL technology, a new algorithm network of MCAG(MHA and GRU-based AutoEncoder) is proposed for tool life prediction. The overall framework of MCAG is shown in 4. The network architecture based on the AE is mainly designed as follows:

1) Signal data collection. The signal data of multiple sensors in milling machine cutting work are collected, the corresponding tool wear values are measured, and the values are used to train the model.
2) Encoder based on CNN. The collected signal data is input to CNN, and the features are extracted from the signals collected by different sensors using CNNs.
3) Decoder based on MHA and GRU. In the input features, the key feature information is selected by MHA to achieve comprehensive feature extraction. GRU is used to better capture long dependencies in sequences and avoid gradient disappearing or exploding. The residual connection effectively extracts the latent features.
4) Monotonicity Loss Function. To optimize the model training process, a monotonicity loss function based on sequential output is proposed. This loss function is combined with the MSELoss loss function to form the final loss function.

B. ENCODER BASED ON CNN

Signal data from different sensors reflect different aspects of tool wear and have different capabilities in predicting future tool life. As shown in Figure 5, CNNs are used to encode the signal from M sensors into M vectors. CNN adaptively extracts features without the need for additional

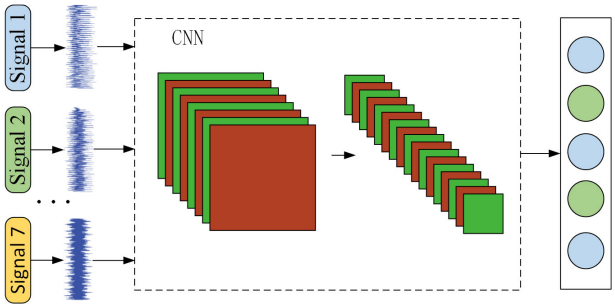

FIGURE 5. Encoder based on CNNs network.

expert knowledge. The convolution operation can be written as:

$$x_j^l = \sum_{k \in M_j} x_k^{l-1} * \omega_{kj}^l + b_j^l \tag{8}$$

where x_j^l represents the j^{th} feature map; x_k^{l-1} is the k^{th} output feature map; ω_{kj}^l means the convolution operation; M_j is the input feature size, and b_j^l is the bias; l represents the layer.

The CNN encoder model consists of 5 hidden layers and 1 linear layer. Each hidden layer contains convolutional layer, activation function, max pooling layer. Maxpool2d and LeakyReLU activation functions are used. In the first hidden layer, the conv2d layer input is 7 because the signal data has 7 different channels. After several convolution operations of kernel size 3*3, the feature map is finally flattened to 256 by a linear layer. Normalization speeds up the training process and enhances the accuracy of the model.

C. DECODER BASED ON MHA AND GRU

A decoder architecture based on GRU and MHA is designed. This structure can not only predict the current tool wear value, but also predict the tool wear value of the next K steps. As shown in Figure 6, the dashed arrow indicates the forward calculation flow of MHA. Different sensor data reflect different characteristics of tool wear. The characteristics of some signals may not contribute much to predicting the current value or the predicted value of the sustained K steps, so MHA is introduced in the decoder structure. MHA fuses different features generated by different attention heads to form the input vector of the GRU. GRU solve the problem of common RNN gradients disappearing or exploding, and better capture long-term feature in sequences. The residual connection effectively extracts the latent features. The input size of GRU is 512 and 256 is the output size.

As shown in Figure 4, the deep feature extracted from CNNs in the encoder is used as the input feature of the decoder, and MHA is employed here. The primary feature is obtained through the calculation of the original signal data, and the calculation formula is shown as Formula 9, and the feature is taken as the Q (query) and K (key) of MHA. This reduces the complexity and computational cost of the model, and can better preserve the local dependence of the

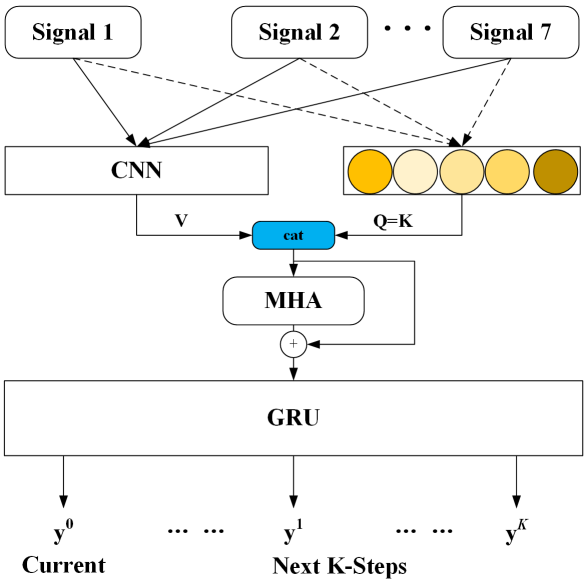

FIGURE 6. Decoder based on GRU-MHA network.

input sequence. The feature input by the encoder is taken as V(Value) of MHA, and Q, K, and V are weighted according to Formula 3. For specific weighting details, see Section II-B.

$$\bar{x}_m = \tanh\left(W_m^d \hat{x}_m + b_m^d\right), m = 1, 2, \ldots, M \tag{9}$$

where $\hat{x}_m$ represents the one-dimensional feature of the raw data x_m from the m^{th} sensor; W_m^d and b_m^d are the weight and bias coefficients of the layer, respectively.

The decoder produces a series of sequential outputs that can simultaneously obtain the prediction values of the current tool and the next multi-step (K-Step). Therefore, the first wear value of the MCAG output can be used to predict the current tool wear value. The remaining K-Step prediction value can be used in monotonicity loss function.

D. MONOTONICITY LOSS FUNCTION

In previous model training, the mean square error loss function (MSELoss) is often used to predict tool wear, and the equations is as follows:

$$\text{MSELoss} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K}\left(\hat{y}_i^k - y_i^k\right)^2 \tag{10}$$

where $\hat{y}_i^k$ is the k^{th} predicted value output by the deep learning model, and y_i^k is the true value, N represents the number of samples.

A basic prior knowledge in the tool wear process is that the tool wear is a monotone function in the best case. The amount of tool wear accumulates monotonously in the continuous cutting process of the machine tool. Therefore, monotonicity is an important feature to describe the trend of tool wear degradation, and the changing trend of the milling cutter state can be analyzed through the change of monotonicity of wear

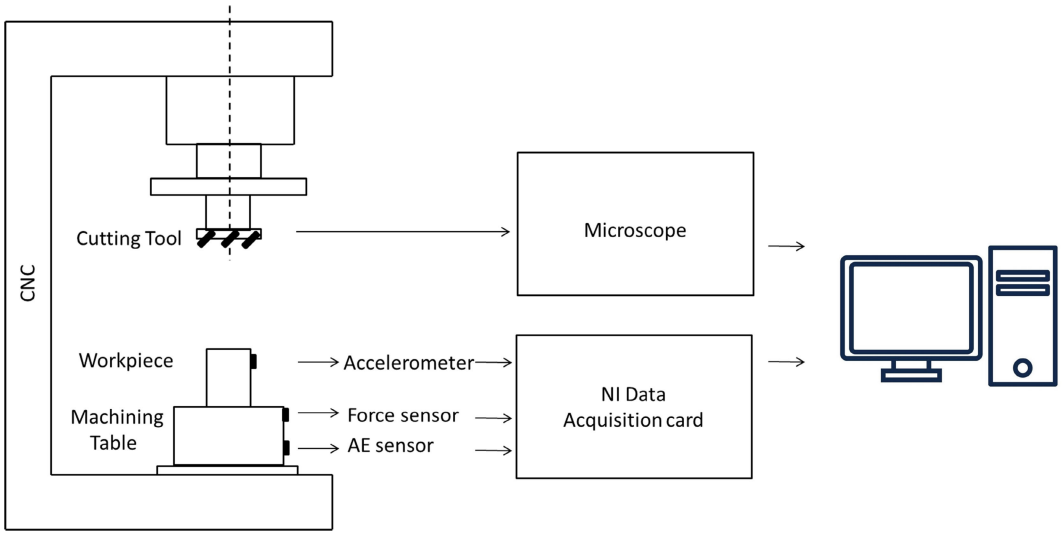

FIGURE 7. The detail of experimental setup and the data collected system.

amount. In this paper, the monotonic analysis of the tool wear is completed by using the K-step predicted value of the sequential output, and then the wear trend of the tool is analyzed. Therefore, based on the above theory, a monotone loss function is proposed in this paper, as shown as follows:

$$\Delta\hat{y}_i^k = \begin{cases} 0, & \left|\hat{y}_i^k - \hat{y}_i^{k+1}\right| \le \Delta m \\ \hat{y}_i^k - \hat{y}_i^{k+1}, & \left|\hat{y}_i^k - \hat{y}_i^{k+1}\right| > \Delta m \end{cases} \tag{11}$$

$$\mathrm{MNT} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left|\Delta\hat{y}_i^k\right| \tag{12}$$

where $\hat{y}_i^k$ is the k^{th} predictive value, $\hat{y}_i^{k+1}$ is the $(k+1)^{th}$ predictive value, Δm is the boundary value of cutting tool, N represents the number of samples.

In this paper, monotonicity loss function and the MSE loss function are used to construct the final loss function, as shown as follows.

$$\mathrm{MNTLoss} = \mathrm{MSELoss} + \mathrm{MNT} \tag{13}$$

IV. EXPERIMENTS AND RESULTS

A. DATASET DESCRIPTION

The experimental dataset is the PHM2010 [34]. When the wear condition of the CNC milling machine tool changes, the sensor signals will also change, such as cutting force signals, vibration signals, acoustic emission signals, etc., will fluctuate accordingly. In the experiment of wear degradation, the surface of the stainless steel workpiece is processed by using a carbide three-blade ball head tool. The cutting workpiece is an HRC52 square stainless steel plate with a surface length of 108mm, and the distance of each cutting of 108mm is marked as a complete cutting. After each cutting, the wear amount of the cutter's back tool face is measured and recorded by a professional microscope.

The CNC milling machine is equipped with a Kistler 8152 three-way dynamometer to capture cutting force signals in the X, Y, and Z directions. Additionally, a Kistler 8636C piezoelectric acceleration sensor is integrated to collect vibration data in the X, Y, and Z axes. Furthermore, an acoustic emission sensor from Kistler is employed to gather acoustic emission signal data. The visual representation of the experimental setup can be found in Figure 7.

In the wear experiment, a total of 6 carbide three-edge ball end milling cutters are completed independent life collection experiment, recorded as C1, C2, C3, C4, C5, C6. In this paper, only C1, C4, and C6 milling cutters with wear results are used.

B. DATASET PREPROCESS

The datasets of C1, C4, and C6 are collected, and each dataset is composed of 315 cutting events, so it is called 315 data samples. Each data sample has a corresponding tool wear value. Different data samples have different time steps. Each data sample is an extremely long time series consisting of more than 200,000 time steps. Through the analysis of CNC milling scenarios, the first 224×224 time steps represent the normal operation state of milling. Therefore, the first 224×224 time steps of each data sample are retained as the length of the new sequence. Three experimental data sets D1, D4, and D6 are constructed from the milling cutters data of C1, C4, and C6 without outliers or missing values.

C. TRAINING AND TEST

The experimental evaluation process is shown in Figure 8. After data downsampling, the raw data is divided into train dataset and test dataset. The model parameters converged during the training process, and the test validation was completed on the test dataset.

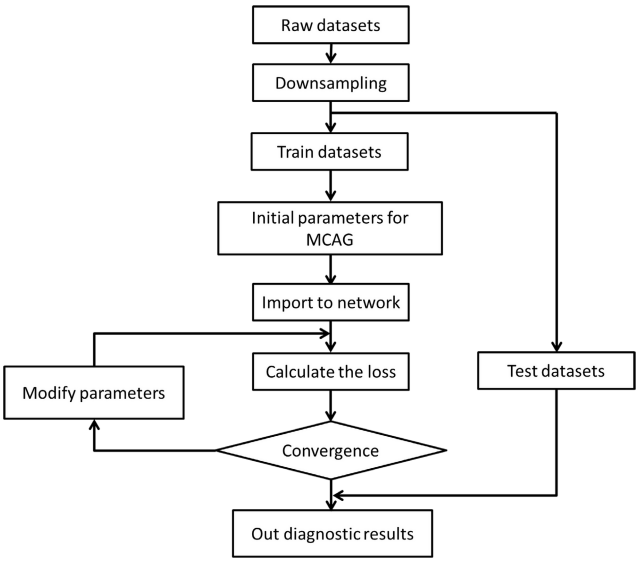

FIGURE 8. Flowchart of the steps for the training and test.

Milling cutter wear prediction evaluation experiments are conducted using the D1, D4, and D6 in PHM2010. Each milling cutter data is composed of 315 data samples, and each data sample measures the corresponding wear values. The average wear values of the three surface wear values of the milling cutter are calculated as the final wear value. To achieve the purpose of a full verification experiment, two groups of D1, D4, and D6 after dimensional-reduction resampling are divided into three different data sets, and three groups of verification experiments are carried out. In the first set of experiments, data from D1 and D4 (a total of 630 data samples) are used as a training set to produce a model P1 + 4, which is tested using data from D6; In the second set of experiments, data from D1 and D6 (a total of 630 data samples) are used as a training set to produce a model P1 + 6, which is tested using data from D4; The third set of experiments, using the data of D4 and D6 (a total of 630 data samples) is used as a training set to obtain the model P4 + 6, and the data of D1 is used to test the model; In each group of experiments, 20% of the data samples of the training set are used as validation sets to observe the training of the analysis model.

The model training network is set to 200 epochs using the Adam optimizer, the batch size is set to 32, the learning rate is set to 0.001, and the Dropout is set to 0.5 for overfitting mitigation. The hyperparameters of all models are tuned on the validation data and the best model is performed on the test data. The above process is independently repeated 5 times on each data set using different random seeds to get an average result.

D. EVALUATION METRICS

To quantitatively evaluate the verification results of the proposed model, two evaluation indexes are used in this evaluation experiment, namely mean square error (MSE) and mean absolute error (MAE). The calculation formula for MSE and MAE is as follows:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2 \tag{14}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right| \tag{15}$$

where $\hat{y}_i$ represents the predicted value and y_i represents the true value, N represents the number of samples.

TABLE 1. Performance of different algorithms for three cutters model.

Method	MSE(P1+4)	MSE(P4+6)	MSE(P1+6)
LSTM	950.13	96.40	634.77
GRU	1103.45	92.56	1101.97
Bi-GRU	864.365	131.30	994.05
SMAML	1057.83	160.69	339.70
M1	941.47	135.60	467.53
MCAG	521.74	79.60	230.62

TABLE 2. Performance of different algorithms for three cutters model.

Method	MAE(P1+4)	MAE(P4+6)	MAE(P1+6)
LSTM	22.07	10.43	19.27
GRU	25.83	9.87	25.57
Bi-GRU	24.01	8.27	22.87
SMAML	22.05	10.11	13.40
M1	27.20	7.78	15.29
MCAG	20.42	6.15	11.03

E. RESULTS AND DISCUSSION

To further evaluate the performance of MCAG, some other popular algorithms are selected for comparative experiments. These popular algorithms, including LSTM, GRU, Bi-GRU [35], and SMAML [26], proved the rationality and validity of the prediction model. The M1 ablation experiment is completed to prove the validity of MNT loss function.

The prediction results of the above wear prediction models are calculated respectively. Table 1 shows the MSE evaluation experimental results of different algorithms, and Table 2 shows the MAE evaluation experimental results of different algorithms. According to the result in the table, the MSE and MAE of the MCAG are the smallest through the evaluation of three different datasets, which is far smaller than the results of LSTM, GRU, Bi-GRU, and SMAML, showing better prediction performance. The main reason may be that the proposed MCAG takes advantage of the network architecture of the autoencoder and the superiority of the new monotonicity loss function to improve the accuracy of the prediction results. Through ablation experiments, the performance of MCAG is better than that of M1, which verifies the validity of the MNT loss function proposed. Therefore, the MCAG algorithm has a wide application prospect in the prediction of cutter wear values in milling machines.

The comparison curves of the predicted wear value and the real wear value of the three groups of evaluation experiments

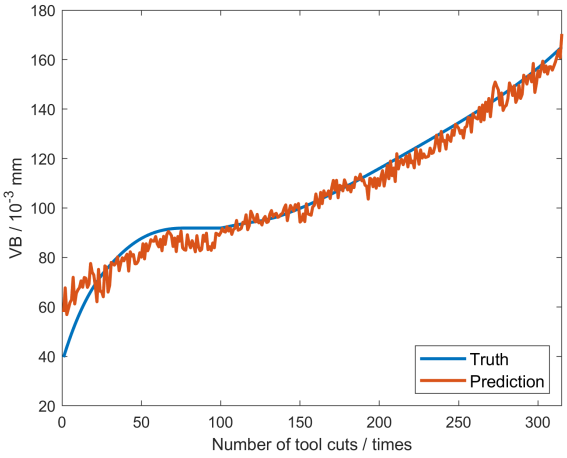

FIGURE 9. C1 prediction wear value.

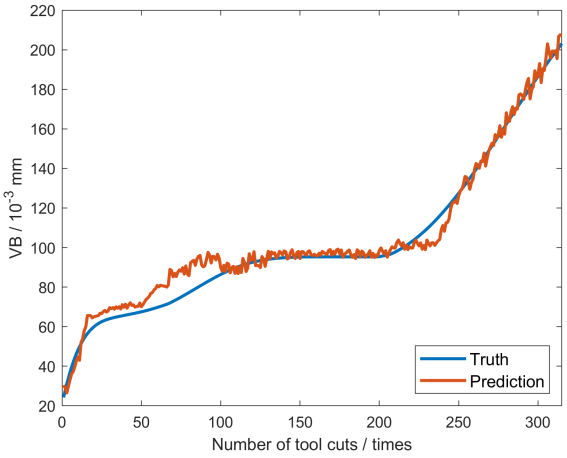

FIGURE 10. C4 prediction wear value.

are shown in the figure 9, 10, 11, where the yellow dotted line is the predicted wear value of the milling cutter, and the blue solid line is the real wear value. By observing the comparison curve, the predicted curve of the C1 milling cutter fluctuates greatly in the early stage, and the overall forecast curve of the C6 and C4 milling cutter is smooth and close to the real wear curve. The changing trend of C6 and C4 is almost consistent with the real curve. Analysis of the wear prediction curve of the three milling cutters, it is found that the prediction effect of the early wear value of the milling cutter is not good, the prediction curve of the middle wear is most in line with the real value curve, and the late wear of the milling cutter wear is not good, but can reflect the wear trend of the tool. The possible reasons for this result are: In the early wear stage of the milling cutter, the working condition of the milling machine is unstable, so there is an error in the collected data signal, and there is a large error between the predicted value and the real value. In the middle and late wear stage of the milling cutter, the working condition of the milling machine is relatively stable, and the collected signal data is more real, so the predicted curve of the wear value is more in line with the real curve of the wear value.

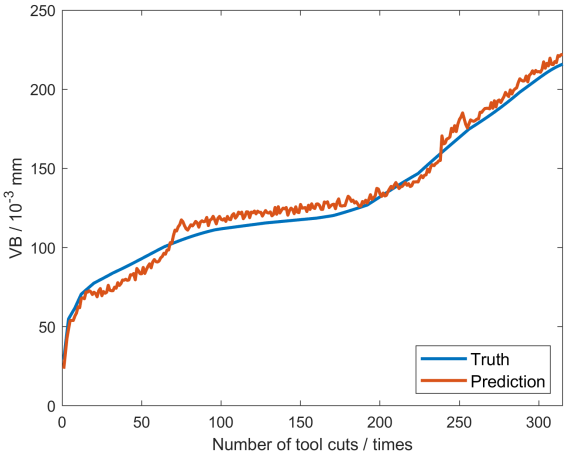

FIGURE 11. C6 prediction wear value.

In summary, it can be concluded that the proposed method can preliminarily predict the wear trend and wear value of tool wear.

The statistic P-value was calculated to compare the statistically significant between the different models. First, the null hypothesis (H0, there is no significant difference between the predictions of the two models) and the alternative hypothesis (H1, there is a significant difference between the predictions of the two models) are established. For the statistically significant between MCAG and each algorithm, the difference value and the corresponding standard deviation are calculated separately to obtain the t-value. Based on the t-value and the degrees of freedom (DF), use the t-distribution table to obtain the corresponding p-value. All calculated p-values are very close to 0 (less than 0.0001). This extremely small p-value indicates a significant difference between the predicted results of MCAG and each algorithm.

Through the analysis of the above evaluation result data and the comparison between the predicted curve and the real curve, it is found that there is still a significant gap between the wear value predicted by the MCAG model and the real wear value. The MCAG model needs further optimization and improvement in future work. For example, in the encoders, a deep convolutional structure can be used instead of the ordinary convolutional neural networks. In the decoder, the model proposed in this paper can be optimized and improved by learning more about the self-attention mechanism in the Transformer. These will continue in the future work.

V. CONCLUSION AND FUTURE WORK

A new tool wear prediction method named MCAG based on multi-sensor data is proposed. The original data are collected by multiple sensors in the machine milling process. The structural of the proposed method within encoder and decoder adaptively enhances the weight of relevant features and suppresses the irrelevant information. By training and evaluation on the PHM2010 dataset, the proposed method

achieves 6.15 MAE and 79.6 MSE, indicating that the method has an improved ability in result. The following conclusions can be obtained:

1) A novel MCAG network is proposed. MCAG is composed of an augmented AE with MHA. The encoder part of MCAG encompasses multiple CNNs, dynamically extracting signal features. On the other hand, the decoder segment of MCAG incorporates MHA and GRU, which enhance the relevant feature weights and extract diverse, long-term, deep features. To optimize the model training process, a monotonicity loss function is defined.
2) The original dataset is resampled and three different evaluation experiments are performed on the reconstructed dataset to prove the validity of the proposed model. The p-value statistical significance tests are conducted to prove that the methods performance differences are statistically significant.

This study has the potential for applications in predicting the remaining useful life of tools in diverse machine milling scenarios, but there are limitations. As an example, The metric index and generalization of the algorithm proposed in this paper can be further improved and validated, and the results have not been tested in the real-time system. In future research, efforts will be directed towards enhancing result accuracy by more advanced model, such as GPT. To prove the generalization of the proposed algorithm, validation experiments on other CNC milling datasets (like NASA milling dataset) are needed, and an extensive exploration of practical applications in Industry 4.0 will be embarked upon, aiming to bridge the gap between theoretical advancements and real-world implementation.

APPENDIX A
RESULTS OF THE PREDICTION WEAR VALUE

The results of the prediction wear value are in google-drive, please click DRIVE-URL.

REFERENCES

[1] G. Li, C. Wang, D. Zhang, and G. Yang, "An improved feature selection method based on random forest algorithm for wind turbine condition monitoring," *Sensors*, vol. 21, no. 16, p. 5654, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/16/5654
[2] K. Zhu and Y. Zhang, "A generic tool wear model and its application to force modeling and wear monitoring in high speed milling," *Mech. Syst. Signal Process.*, vol. 115, pp. 147–161, Jan. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327018303042
[3] L. Dong, C. Wang, G. Yang, Z. Huang, Z. Zhang, and C. Li, "An improved ResNet-1D with channel attention for tool wear monitor in smart manufacturing," *Sensors*, vol. 23, no. 3, p. 1240, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/3/1240
[4] T. Mohanraj, S. Shankar, R. Rajasekar, N. Sakthivel, and A. Pramanik, "Tool condition monitoring techniques in milling process—A review," *J. Mater. Res. Technol.*, vol. 9, no. 1, pp. 1032–1042, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2238785418313061
[5] Q. Wang, H. Wang, L. Hou, and S. Yi, "Overview of tool wear monitoring methods based on convolutional neural network," *Appl. Sci.*, vol. 11, no. 24, p. 12041, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/24/12041
[6] Y. Zhang, K. Zhu, X. Duan, and S. Li, "Tool wear estimation and life prognostics in milling: Model extension and generalization," *Mech. Syst. Signal Process.*, vol. 155, Jun. 2021, Art. no. 107617. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327021000121
[7] T. Mohanraj, J. Yerchuru, H. Krishnan, R. N. Aravind, and R. Yameni, "Development of tool condition monitoring system in end milling process using wavelet features and Hoelder's exponent with machine learning algorithms," *Measurement*, vol. 173, Mar. 2021, Art. no. 108671. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224120311830
[8] Z. Li, R. Liu, and D. Wu, "Data-driven smart manufacturing: Tool wear monitoring with audio signals and machine learning," *J. Manuf. Processes*, vol. 48, pp. 66–76, Dec. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1526612518315822
[9] M. Kuntoglu and H. Saglam, "Investigation of progressive tool wear for determining of optimized machining parameters in turning," *Measurement*, vol. 140, pp. 427–436, Jul. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224119303239
[10] M. Cheng, L. Jiao, P. Yan, H. Jiang, R. Wang, T. Qiu, and X. Wang, "Intelligent tool wear monitoring and multi-step prediction based on deep learning model," *J. Manuf. Syst.*, vol. 62, pp. 286–300, Jan. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S027861252100251X
[11] Y. Huang, Z. Lu, W. Dai, W. Zhang, and B. Wang, "Remaining useful life prediction of cutting tools using an inverse Gaussian process model," *Appl. Sci.*, vol. 11, no. 11, p. 5011, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/11/5011
[12] X. Xu, J. Wang, B. Zhong, W. Ming, and M. Chen, "Deep learning-based tool wear prediction and its application for machining process using multi-scale feature fusion and channel attention mechanism," *Measurement*, vol. 177, Jun. 2021, Art. no. 109254. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224121002645
[13] Y. Qin, X. Liu, C. Yue, M. Zhao, X. Wei, and L. Wang, "Tool wear identification and prediction method based on stack sparse self-coding network," *J. Manuf. Syst.*, vol. 68, pp. 72–84, Jul. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0278612523000304
[14] R. Li, X. Ye, F. Yang, and K.-L. Du, "ConvLSTM-Att: An attention-based composite deep neural network for tool wear prediction," *Machines*, vol. 11, no. 2, p. 297, 2023. [Online]. Available: https://www.mdpi.com/2075-1702/11/2/297
[15] A. K. W. Chee, R. F. Broom, C. J. Humphreys, and E. G. T. Bosch, "A quantitative model for doping contrast in the scanning electron microscope using calculated potential distributions and Monte Carlo simulations," *J. Appl. Phys.*, vol. 109, no. 1, Jan. 2011, Art. no. 013109, doi: 10.1063/1.3524186.
[16] A. K. Chee, "The mechanistic determination of doping contrast from Fermi level pinned surfaces in the scanning electron microscope using energy-filtered imaging and calculated potential distributions," *Microsc. Microanalysis*, vol. 28, no. 5, pp. 1538–1549, Oct. 2022, doi: 10.1017/S1431927622000642.
[17] Z. Kang, C. Catal, and B. Tekinerdogan, "Remaining useful life (RUL) prediction of equipment in production lines using artificial neural networks," *Sensors*, vol. 21, no. 3, p. 932, Jan. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/3/932
[18] X. Wu, Y. Liu, X. Zhou, and A. Mou, "Automatic identification of tool wear based on convolutional neural network in face milling process," *Sensors*, vol. 19, no. 18, p. 3817, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/18/3817
[19] H. Xu, C. Zhang, G. S. Hong, J. Zhou, J. Hong, and K. S. Woon, "Gated recurrent units based neural network for tool condition monitoring," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7.
[20] S. Vijay and B. Kuraichen, "Data driven prognostics of milling tool wear : A machine learning approach," in *Proc. Int. Conf. Comput. Perform. Eval. (ComPE)*, Dec. 2021, pp. 2–7.
[21] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
[22] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," Sep. 2014, *arXiv:1409.2329*.
[23] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, Mar. 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/036402139090002E

[24] Q. An, Z. Tao, X. Xu, M. El Mansori, and M. Chen, "A data-driven model for milling tool remaining useful life prediction with convolutional and stacked lstm network," *Measurement*, vol. 154, Mar. 2020, Art. no. 107461. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224119313284
[25] W. Xu, H. Miao, Z. Zhao, J. Liu, C. Sun, and R. Yan, "Multi-scale convolutional gated recurrent unit networks for tool wear prediction in smart manufacturing," *Chin. J. Mech. Eng.*, vol. 34, no. 1, p. 53, Dec. 2021, doi: 10.1186/s10033-021-00565-4.
[26] G. Wang and F. Zhang, "A sequence-to-sequence model with attention and monotonicity loss for tool wear monitoring and prediction," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.
[27] D. Hu and Z. Tang, "ResGRUA model for tool wear prediction based on encoder–decoder," in *Proc. 5th Int. Conf. Mech., Control Comput. Eng. (ICMCCE)*, Dec. 2020, pp. 1067–1070.
[28] J. Yao, B. Lu, and J. Zhang, "Multi-Step-Ahead tool state monitoring using clustering feature-based recurrent fuzzy neural networks," *IEEE Access*, vol. 9, pp. 113443–113453, 2021.
[29] M. Marei and W. Li, "Cutting tool prognostics enabled by hybrid CNN-LSTM with transfer learning," *Int. J. Adv. Manuf. Technol.*, vol. 118, nos. 3–4, pp. 817–836, Jan. 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253719191
[30] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:1658773
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*. Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.
[32] K. Cho, B. van Merrienboer, Ç. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1–15. [Online]. Available: https://api.semanticscholar.org/CorpusID:5590763
[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[34] Y. Yin, S. Wang, and J. Zhou, "Multisensor-based tool wear diagnosis using 1D-CNN and DGCCA," *Int. J. Speech Technol.*, vol. 53, no. 4, pp. 4448–4461, Jun. 2022, doi: 10.1007/s10489-022-03773-0.
[35] W. Zheng, P. Cheng, Z. Cai, and Y. Xiao, "Research on network attack detection model based on BiGRU-attention," in *Proc. 4th Int. Conf. Frontiers Technol. Inf. Comput. (ICFTIC)*, Dec. 2022, pp. 979–982.

CHUNPING DONG received the master's degree in computer application from Wuhan University of Technology, Hubei, China, in 2007. She is currently engaged in computer teaching in Weifang University. Her research interests include computer technology application based on deep learning and AI.

JIAQIANG ZHAO received the Ph.D. degree in optics from Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Anhui, China. He is currently a Professor with Weifang University. He has published many academic papers in Phys.Rev. A, Scientific Reports, and Phys.Lett.A. His research interests include opto-electronic information technology research and AI.