Journal of Intelligent Manufacturing (2023) 34:3365–3387
https://doi.org/10.1007/s10845-022-02005-z

Cross-domain adaptation network based on attention mechanism for tool wear prediction

Jianliang He[1] · Yuxin Sun[1] · Chen Yin[1] · Yan He[2] · Yulin Wang[1]

Received: 8 March 2022 / Accepted: 1 August 2022 / Published online: 28 August 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

On-machine monitoring of tool wear is of great significance in machining processes to enhance machining efficiencies and reduce cost. For the data-driven tool wear prediction, the consistent feature distribution is critical to the machining learning model. In CNC machining of complex parts, the data distribution is complex due to the time-varying operational parameters. Even the considered as perform well of transfer learning methods may not mine consistent feature distribution representations and may not explain the transferable components to break through the bottleneck of prediction performance. The attention mechanism has shown its unique advantages with the adaptive weighted structure to capture the specific feature in tool wear state prediction. In order to further capture the specific characteristic frequencies and determine the specific transferable components for different working conditions and enhance the performance of cross-domain tool wear prediction, we propose a new cross-domain adaptation network based on attention mechanism(CDATT). Firstly, a data filtering method based on the singular value decomposition and mixed Gaussian model is proposed for removing the idling signals to avoid the effects of invalid samples adaptively. Then, the attention mechanism is introduced to build a prediction network for highlighting the specific tool wear frequency and making the wear characteristic more significant. Next, the joint distribution adaptation regularization term is utilized to construct a network loss function to build the CDATT model for different working conditions. Finally, experiments are carried out on a machine tool to verify the effectiveness of the proposed method. The result shows that the proposed CDATT model can capture the specific tool wear frequency and identify the tool wear state accurately under different machining parameters.

Keywords Tool wear prediction · Cross-domain adaptation · Gaussian mixture model · Attention mechanism · Joint distributed adaptation

Introduction

In the machining process of machine tools, cutting tools play an irreplaceable role that contact with metal workpieces directly and the cutting edge will wear out gradually under the combined effects of machinery, heat, chemistry, and abrasive particles (Colantonio et al. 2021). Tool wear will impact the surface quality and precision of the machining process and even damage the workpiece and machine tools when heavy damage to tools occurs. Therefore, the research on the tool wear prediction technology has been a significant research topic in the academic and engineering fields.

Thanks to the quick development of artificial intelligence techniques, the data-driven of machine learning and signal processing methods have shown unique advantages and great potential in mechanical equipment fault monitoring and status recognition. For example, the wavelet analysis and

✉ Yulin Wang
wyl_sjtu@126.com

Jianliang He
hjlhdh@njust.edu.cn

Yuxin Sun
sunyuxin@njust.edu.cn

Chen Yin
yinchen@njust.edu.cn

Yan He
heyan@cqu.edu.cn

1 School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

2 State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing 400030, China

empirical mode decomposition (Madhusudana et al. 2018; Yin 2022), hidden Markov model(HMM) Boutros and Liang (2011), neuro-fuzzy inference system (ANFIS) Soualhi et al. (2020), kernel extreme learning machine (KELM) Tobon-Mejia et al. (2012), dynamic Bayesian networks Lei et al. (2020), support vector machine(SVM) Li et al. (2021). Compared with general tools wear prediction methods which are composed of feature extracting and model classification, deep learning has an excellent feature extraction capability and pattern recognition effect (Hinton and Salakhutdinov 2006; LeCun et al. 2015). The models based on deep learning theory have already entered in machinery health monitoring field and received great attention in the field of industrial data analysis (Lei 2020, 2018), such as artificial neural networks (ANN) (SK and Shankar 2020; Chang et al. 2020), convolutional neural network(CNN) (Brili et al. 2021; Yin 2021), long-short term memory (LSTM) Cai et al. (2020), sparse auto-encoder (SAE) (Dou 2020; Ou et al. 2021), deep stack auto-encoder (DSAE) He et al. (2021). Data-driven methods can process mechanical signals efficiently, and provide accurate diagnostic results that do not need much prior knowledge Yin (2021). Therefore, they are becoming more and more popular in tool wear prediction research.

The tool wear prediction methods based on deep learning not only extract features adaptively, but also construct network models with significant application performance in the machining processing. However, the existing deep learning-based tool wear prediction methods still have some limitations. (1) The existing methods cannot filter and select useful data adaptively. Different from the rolling bearings which are working in a stable working condition relatively, the milling process of the cutter includes multiple stages such as forward milling, reverse milling, and idling process. These complex working conditions make the monitoring data mixed with multiple data distributions. (2) The parameters of the machining process are varying in the milling operations and the traditional tool wear prediction methods cannot realize the cross-working conditions prediction. The data collected from the different parameters of spindle speed, cutting depth, and feed speed can reflect the tool wear status, but there are heavy differences among each other. From a statistical point of view, the distribution of collected data from different working conditions varies greatly. How to use the machine learning methods to build models among them with different data distributions is a cross-domain adaptation problem. The general assumption of machine learning datasets is independent and identical distributed(i.i.d). However, the existing method cannot build models and analyze the tool wear stages in these variable working conditions simultaneously. How to filter the useless information and capture the meaningful information adaptively in different working conditions are the keys to constructing a good tool wear prediction model.

One solution for the aforementioned problem is the transfer learning technique. Transfer learning is a new paradigm of machine learning, which aims to solve challenging learning problems with a small number of labeled samples in the target field. In computer vision Wang and Deng (2018), natural language processing Niehues and Cho (2017), recommendation systems Kanagawa et al. (2019), bioinformatics Sevakula et al. (2018) and prognostic and health management Mao et al. (2022) have achieved great application. Through the structure of network model parameter sharing, pre-training network transferring, and deep network joint distribution adaptation, the effect of information transferring from data with different data distributions is realized. The existing modeling based on transfer learning has also achieved good results in the tool wear prediction. Marei et al. (2021) utilized the deep convolutional network of ResNet-18 to analyze the tool wear image, and use the transfer method to predict the tool wear width. Mamledesai et al. (2020) adopted the pre-trained model deep convolutional network to realize the recognition of the tool wear status in the target cutting tools utilizing network model parameter transfer. Yao et al. (2022) built models from continuous sensor data based on long and short-term memory networks and deep transfer reinforcement learning, completed tool wear verification experiments and achieved good results. However, the above-mentioned tool wear prediction method just only relies on the prediction accuracy to judge the improvement of the effect. And these methods cannot specifically explain which components play a key part in the neural network model and which information from raw data can improve the performance of the tool wear prediction.

The attention mechanism is a new network construction method in recent years, which already realizes the selection of sensitive features by adding an attention layer to the network and it can obtain a weight vector to select samples adaptively (Vaswani 2017). The attention mechanism has achieved good application effects nowadays in natural language processing (Bahdanau et al. 2014), image classification (Wang 2017), computer vision (Fu et al. 2017; Mnih et al. 2014) and other fields. Related models have been introduced into bearing fault diagnosis (Li et al. 2019; Ding et al. 2022) and remaining life prediction Zhou et al. (2020), Shifat and Jang-Wook (2020) and have achieved good results. Existing studies have shown that the attention mechanism can help to capture the specific frequencies (Li et al. 2019), so this method can explain the changes in the frequency of fault features more effectively when a fault occurs. We are very interested in this adaptive capture function of the attention mechanism on specific fault frequencies, and our research also focuses on the problem of milling tool wear prediction in complex machining working conditions. Different from the rolling bearings worked in a stable working condition relatively, the milling process of the cutter includes multiple stages such as forward milling, reverse milling, and idling,

and it is also faced with different combinations of machining parameters, such as varying spindle speed, feed and depth of cut. These complex working conditions make the monitoring data mixed with multiple data distributions. Improving the performance of cross-domain modeling and the applicability of the method in complex working conditions is the key to our research. Therefore, combining the advantages of the adaptive feature extraction capability of the attention mechanism and the cross-domain modeling of transfer learning, we propose a tool wear prediction method based on the cross-domain and attention mechanism.

The deep transfer learning methods mainly include two categories: maximum mean discrepancy(MMD)-based domain adaptation and domain-adversarial training networks (Ganin 2016). The MMD-based domain adaptation method has achieved good results by its good feature adaptation ability. The domain-adversarial training networks also realize the knowledge transfer process by extracting common features across domains. However, the domain adversarial training method (Mao et al. 2022) is similar to the generative adversarial network(GAN) (Creswell 2018) method, and there are some problems in compute convergence and model stability. Training domain-adversarial networks also needs to achieve Nash equilibrium (Goodfellow 2014), but sometimes the method based on gradient descent may not be able to achieve. The MMD-based domain adaptation method is derived from the transfer component analysis(TCA) (Pan et al. 2010), and the mathematical principle of the theory is complete relatively. The methods such as joint distribution adaptation(JDA) (Long et al. 2013) developed on the basis of TCA have also achieved good applications in many fields. Furthermore, MMD-based domain adaptation methods contain only one loss function, which is easy to adjust and optimize. To solve the problem of tool wear prediction under complex working conditions, this paper introduces the attention mechanism to the MMD-based domain adaptation tool wear prediction problem. We believe that the attention mechanism can capture the sensitive feature of the tool wear state with the help of the cross-domain adaptation platform under different working conditions. Specifically, the contribution of this paper can be summarized as follows:

1. Compared with general samples filter methods, the proposed data filtering method based on singular value decomposition and Gaussian mixture model can filter the signals of different machining stages adaptively, such as milling and idling processes. And it can provide effective and clean data for further modeling.
2. Compared with the general domain adaptation method, the proposed tool wear prediction method can extract the common feature of the milling operation under different machining parameters. By integrating joint distribution adaptation and model complexity regularization terms, tool wear characteristics can be extracted from different working conditions with better discrimination.
3. Compared with the general tool wear prediction method, the proposed attention mechanism-based neural network can weigh the sparse feature adaptively. By mining the relationship between the milling frequency spectrum and the tool wear state, the proposed model can capture the sensitive characteristic frequency and it is helpful to explain and analyze the relationship between the wear state and frequency spectrum under different working conditions.

The rest part of this paper is organized as follows. Section "Background" is provided some related background and algorithms. In Sect. "Proposed Method", the specific implementation and steps of the algorithm proposed in this paper are provided. Section "Experiment Results and Analysis" is introduced the experiment and results analysis, and Sect. "Conclusion" provides the summary of this paper.

Background

Gaussian mixture model

In this paper, the Gaussian mixture model (GMM) cluster is used to monitor milling tools data to divide the different machining stages. Gaussian models are including a single Gaussian model and Gaussian mixture model according to the different data distribution clusters.

The Gaussian mixture model is established by using the Gaussian probability density function to accurately quantify data, and then decompose it into several parts by normal distribution probability density function. The whole GMM is composed of K Gaussian distributions and each Gaussian distribution can be regarded as a component. These probability density function of GMM can be composed by linear addition, as shown in Eq. (1).

$$p(x) = \sum_{k=1}^{K} p(k)\, p(x \mid k) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, Cov_k) \quad (1)$$

where $k = 1 \cdots K$, and the K is the clusters number of GMM, π_k is the influence factor of each Gaussian distribution on the data. μ_k and Cov_k are the mean value and covariance matrix, respectively.

Considering that a data set with N sample points and assuming that the data set obeys a certain distribution, the probability density function of its Gaussian distribution is recorded as $p(x)$. It is necessary to obtain the probability distribution model determined by the parameters π_k, μ_k and

Cov_k to ensure that the current data set is generated with the highest probability. And the probability density function can be expressed by Eq. (2).

$$p(x) = \prod_{i=1}^{N} p(x_i) \tag{2}$$

The above equation is called the likelihood function. Usually, the probability of a single point is very small. The multiplication of many decimals in the computer can easily cause floating-point overflow, so we take the logarithm to convert the product into a summation, as shown in Eq. (3).

$$p(x) = \sum_{k=1}^{K} \log p(x_i) = \sum_{k=1}^{K} \log \left[\sum_{k=1}^{K} \pi_k N(x_i \mid \mu_k, Cov_k) \right] \tag{3}$$

After obtaining the log-likelihood function, the parameters π_k, μ_k and Cov_k are solved by maximizing the likelihood function.

Because of the summation in the logarithmic function, it is impossible to find the maximum value by directly using the derivative to solve the equation. To solve this problem, the method of randomly selecting points from the GMM is generally used to solve the problem, which is divided into E-step and M-step, namely EM algorithm (Dempster et al. 1977). The EM algorithm is a maximum likelihood estimation method used to solve probabilistic model parameters in data sets with incomplete data or hidden variables. The E-step uses the existing estimates of hidden variables to calculate its maximum likelihood estimate, and M-step maximizes the maximum likelihood value obtained in E-step to calculate the value of the parameter. By iteratively using E-step and M-step until the function converges, the best parameter value is obtained.

Joint distribution adaptation

Joint distribution adaptation (JDA) is a classic method in transfer learning (Long et al. 2013). It is constructed formula for adaptation data distributions by considering both the marginal distribution probability adaptation and the conditional probability distribution adaptation from the source data and the target data. Generally, the specific implementation is expressed by maximum mean discrepancy (MMD). The MMD method is a measure of the difference between two distributions. Given two data sets which named as source data set and target data set, X_s and X_t, and there is a common relationship $P_s(X_s) \neq P_t(X_t)$. These two data sets are mapped to the low-dimensional reproducing kernel Hilbert space (RKHS) through a nonlinear mapping function ϕ, as shown in Equation (4).

$$MMD_H(X_s, X_t) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi\left(x_i^s\right) - \frac{1}{n_t} \sum_{i=1}^{n_t} \phi\left(x_i^t\right) \right\|_H \tag{4}$$

where n_s and n_t are the samples number of two data sets. In Eq. (4), The empirical estimate of the difference between the two distributions is considered the distance between the two data distributions in RKHS and the value of MMD close to zero means that the two distributions are similar with each other. In transfer learning, MMD is generally used to construct regularization terms for constraints in feature learning, so that the learned feature distribution is more similar between two different domains.

The joint distribution adaptation method usually needs to consider two aspects of marginal distribution adaptation and conditional distribution adaptation, as shown in Eqs. (5) and (6).

$$\text{minD}\left(P_s(\phi(X_s)), P_t(\phi(X_t))\right) \tag{5}$$

$$\text{minD}\left(Q_s(Y_s \mid \phi(X_s)), Q_t(Y_t \mid \phi(X_t))\right) \tag{6}$$

where Y_s and Y_t are the labels of data set of X_s and X_t, correspondingly. The target function of Eq. (5) can be expressed by Eq. (7) by measuring the difference in the RKFS space after nonlinear mapping.

$$MMD_H^2(P_s, P_t) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi\left(x_i^s\right) - \frac{1}{n_t} \sum_{i=1}^{n_t} \phi\left(x_i^t\right) \right\|_H^2 \tag{7}$$

The target function in the conditional probability distribution in Eq. (6) is called conditional probability distribution adaptation. This step is essential to improve the accurate and robust distribution adaptation of the sample data in each category. However, since the label Y_t of the target field is unknown, previous related researches usually deal with this situation by using the pseudo-label of the target data in unsupervised field adaptation. With the help of pre-trained model by labeling source field data X_s, the predict labels of target data X_s is provided as pseudo-labels for the initial transfer learning model training.

Suppose there are C categories, where $c \in \{1, \cdots, C\}$. The MMD distance can be defined by the conditional probability $Q_s(x_s \mid y_s = c)$ of the source data set X_s and the conditional probability $Q_t(x_t \mid y_t = c)$ of the target data set X_t to measure the matching degree, as shown in Eq. (8).

$$MMD_H^2\left(Q_s^{(C)}, Q_t^{(C)}\right) = \left\| \frac{1}{n_s^{(C)}} \sum\nolimits_{x_i^s \in D_s^{(C)}} \phi\left(x_i^s\right) - \frac{1}{n_t^{(C)}} \sum\nolimits_{x_i^t \in D_t^{(C)}} \phi\left(x_i^t\right) \right\|_H^2 \tag{8}$$

where $D_s^{(C)} = \{x_i : x_i \in D_s \wedge y(x_i) = c\}$, and it represents the data of the corresponding category in the source data set. The $y(x_i)$ is the true label in the source data set. The $D_t^{(C)} = \left\{x_j : x_j \in D_t \wedge \hat{y}\left(x_j\right) = c\right\}$, and $\hat{y}\left(x_j\right)$ is the pseudo label in the target data set. Although the initial pseudo-label may not be completely accurate, the label can be iterated and updated during model optimization and gradient descent process. Therefore, the model can finally obtain the best prediction effect during the process of model optimization iteration.

Joint distribution adaptation is to combine two strategies of marginal distribution adaptation and conditional distribution adaptation to construct a unified regularization term, as shown in Eq. (9).

$$D_H^2(J_s, J_t) = MMD_H^2(P_s, P_t) + \sum_{c=1}^{C} MMD_H^2\left(Q_s^{(C)}, Q_t^{(C)}\right) \tag{9}$$

where J_s and J_t represent the joint distribution probability of source data set D_s and target data set D_t, respectively. By minimizing Eq. (9)., it can be ensured that the marginal probability distribution and the conditional probability distribution are approximated after data mapping.

Attention mechanism

In the research of this paper, the frequency domain data is mainly used to analyze and distinguish the wear and damage of the tool. The literature survey of the previous related research shows that not all frequency band data of spectrum have the same contribution to the failure characteristics of mechanical equipment (Bisu et al. 2012). Exploring the inherent relationship between specific vibration frequency bands and tool wear is the key to effective interpretation and construction of efficient tool wear status recognition.

The attention mechanism can realize adaptive input weighting and realize the goal of selection and positioning of specific frequency bands of tool wear. In this section, we will introduce the principle of attention mechanism in brief.

First, for the collected vibration data processed by the tool, the sample is divided and fast Fourier transform is performed to obtain the original input sample data, as shown in Equation (10).

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n], \mathbf{x}_i \in R^{N_{sub}} \tag{10}$$

where N_{sub} represents the length of the input sample after division. After feature extraction and network mapping by inputting data to the deep network model, the obtained feature representation vector $\mathbf{h}$ can be represented by Eq. (11).

$$\mathbf{h} = f_e(\mathbf{x}) = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n], \mathbf{h}_i \in R^{N_{att}} \tag{11}$$

where $f_e(\cdot)$ represents the deep feature extraction network, and $\mathbf{h}_i$ represents the hidden layer feature corresponding to sample $\mathbf{x}_i$. And N_{att} is the characteristic dimension of the attention layer.

Second, the attention mechanism generates a_i weight vector a for $\mathbf{h}_i$, which represents the important score of the corresponding data segment. The vector a_i can be calculated by a separately constructed attention network layer and normalized by the softmax function, as shown in Eqs. (12) and (13).

$$l_i = f_{att}(\mathbf{h}_i) \tag{12}$$

$$a_i = \frac{e^{l_i}}{\sum_{k=1}^{n} e^{l_k}} \tag{13}$$

In this paper, we use a simple attention model to realize the function of the attention mechanism through a single-layer neural network, as shown in Eq. (14).

$$l_i = f_{ac}(\mathbf{W}_{att}\mathbf{h}_i + b_{att}) \tag{14}$$

where $\mathbf{W}_{att}$ and b_{att} represent the weight and bias of the attention layer, respectively, and $f_{ac}(\cdot)$ is the network activation function. After the attention weight parameter is generated, the enhanced feature $\mathbf{v}_i$ is obtained by weighting with the original input hidden layer features, as shown in Eq. (15).

$$\mathbf{v}_i = a_i\mathbf{h}_i \tag{15}$$

The hidden layer features are enhanced by the attention importance parameter, and the weighted selection of the features in the deep network is realized.

Proposed method

To solve the problem of tool wear status recognition under complex working conditions, this paper introduces the GMM clustering, JDA and attention mechanism in Sect. "Background" to predict the wear states of cutting tools. This paper proposed a new cross-domain adaptation network based on attention mechanism for the tool wear prediction. Firstly, a data filtering method based on SVD and GMM is proposed for removing the idling signals adaptively. Then, the attention mechanism is introduced to build a prediction network for highlighting the specific tool wear frequency and making the wear characteristic more significant. Next, the joint distribution adaptation regularization term is utilized to construct a network loss function to build the CDATT model for different working conditions. Finally, experiments are carried out on a machine tool to verify the effectiveness of the proposed

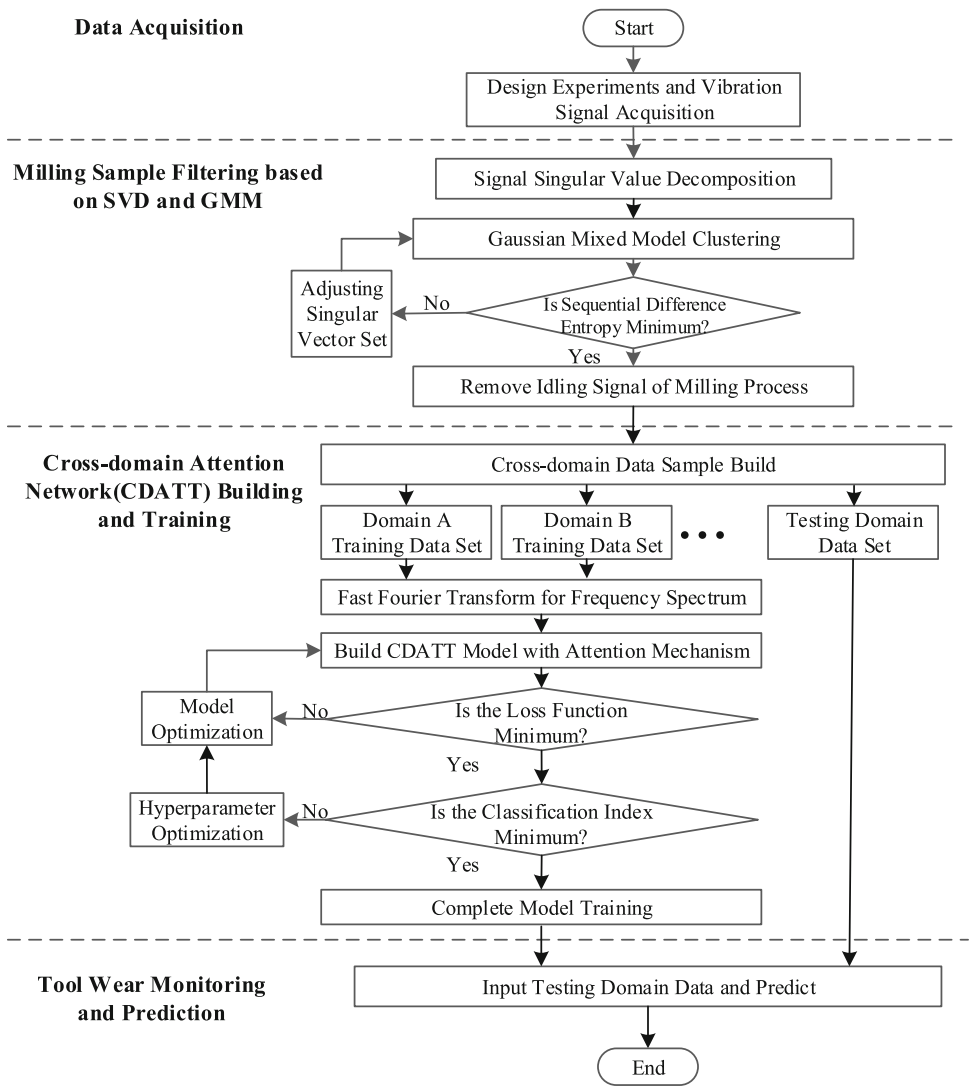

Fig. 1 The proposed tool wear prediction method flowchart

method. The flowchart of the proposed CDATT method is shown in Fig. 1.

The method proposed in this paper four parts are as follows.

Tool milling monitoring data acquisition

For the different work conditions, we collect the milling vibration signals data sets $\{S_i\}_{i=1}^{T}$ of cutting tools during machining process. The T is the working conditions number. In this paper, we set T=3 with the different spindle speeds of 2000rpm, 2600rpm and 3200rpm and the tool wear identification data under each working condition is expressed as $S_i = \left\{\mathbf{s}_j^i = (s_1, s_2, \cdots, s_l), y_j^i\right\}_{j=1}^{n}$. The $\mathbf{s}_j^i$ and y_j^i is the j-th vibration signal samples and the corresponding tool wear status label under working condition i, the l is the sample length of each signal respectively, and the n is the type of tool wear that needs to be identified. In this paper we set n=3 with normal, wear and breakage three different kinds of tools.

The data filter method based on SVD and GMM

For vibration signal sample $\mathbf{s}_j^i$ under each working condition, we convert signal sample vector to signal matrix $\mathbf{m}_j^i \in R^{a\times b}$, where $l = a \times b$. We use the SVD to obtain its singular value vector $v_j^i \in R^{a\times 1}$. Then the singular value matrix composed of n signal sets under the working condition i can be used $V^i \in R^{a\times n}$ to present. Take the first k columns of the singular value matrix V^i and use the principal component analysis method to reduce the dimension to a two-dimensional matrix $P_k^i \in R^{n\times 2}$ as the input of the GMM, and the clustering results of the sample labels $t^k \in \{0, 1\}^n$ can be obtained finally. Specially, we set the 0 as the idling samples, and set 1 as the milling samples.

Calculate the optimal differential sequence entropy of this vector e^i, the computing method is shown in Equation (16).

$$e^i = \arg\min_k \left(- \sum_{j=2\cdots n} p\left(\mid t_j^k - t_{j-1}^k \mid \right) \log p\left(\mid t_j^k - t_{j-1}^k \mid \right) \right) \tag{16}$$

Due to the large distribution difference between the idling signal and the milling signal, the singular value of the vibration signal can reflect the data distribution well in the feature space. When the sequence difference entropy is minimized, it indicates that the change times of cluster labels reach the minimum value between the milling and the idling processes. The clustering result at this time can effectively remove the influence of the signal's amplitude fluctuation.

Then we obtain the filtered data $S_i' = \left\{ \mathbf{s}_j^i, y_j^i \right\}_{j=1}^m$, where $m = \sum t^k$.

Cross-domain Attention Network(CDATT) building and training

After the data filter of the tool milling vibration signals, the fast Fourier transform is used to obtain frequency spectrum set $\{F_i\}_{i=1}^T$ under various working conditions, where $F_i = \left\{ \mathbf{f}_j^i = (f_1, f_2, \cdots, f_m), y_j^i \right\}_{j=1}^m$ represents the frequency spectrum data set and the corresponding labels. Using the frequency spectrum as the input of the deep network model can help the attention mechanism to capture the tool wear characteristic frequency, which is beneficial to explain the tool wear frequency band characteristics under different working conditions (Li et al. 2019). After the data transform, the training set and the test set are divided, and then a deep cross-domain adaptation tool wear prediction model based on the attention mechanism is constructed.

The main goal of this paper is to verify that the attention mechanism can mine the mapping relationship between characteristics frequency and tool wear state adaptively. We set up three kinds of network layers for experimental verification, and the three network layers include:

(1) Feature mapping layers: Convert input frequency spectrum to hidden layer for feature mapping.
(2) Attention layers: Based on attention mechanism by Eqs. (10)–(15), build attention layer to weight the frequency spectrum adaptively.
(3) Fully connected layers: Build JDA regularization and model complexity regularization term for the features mapped by attention mechanism with Eqs. (7), (9) to share the different working conditions information Collobert and Weston (2008), Liu et al. (2012).

The proposed tool wear prediction network model is shown in Fig.2.

In Fig. 2, the network hidden layer units and regularization hyperparameters are selected and optimized by Bayesian optimization Snoek et al. 2012. In the fully connected layer of the network, the regularization terms of the network parameters are constructed, including model complexity constraints term $\lambda_2 \| W \|_F^2$ and JDA term $\lambda_1 D_H^2 (J_s, J_t)$. The loss function is constructed by using the cross-entropy loss for tool wear status prediction, as shown in Eq. (17).

$$\min_W \sum_{i=1}^T \sum_{j=1}^n -y_j^i \log \left(h\left(\mathbf{x}_j^i \right) \right) + \sum_{i=1}^T \left(\lambda_1 D_H^2 (J_s, J_t) + \lambda_2 \| W_t \|_F^2 \right) \tag{17}$$

where $h\left(\mathbf{x}_j^i\right) = g\left(W_t^{out} f\left(W_t^{in} \mathbf{x}_j^i + b^{in}\right) + b^{out}\right)$, it represents the mapping of the input layer to the data. $f(\cdot)$ and $g(\cdot)$ are the activation functions, W_t^{in} and W_t^{out} represent the input and output weights of the network, respectively. b^{in} and b^{out} are the biases of input and output. It should be noted that these parameters need to be solved by model training. The regularization term $D_H^2 (J_s, J_t)$ is shown in Eq. (9). λ_1 and λ_2 are the JDA term and the regularization parameter of the model complexity. By adjusting these two parameters, the model has the capability to extract features adaptively for different working conditions.

Tool wear monitoring and prediction

The data set is divided into the training set and test set, and select 70% of the milling samples under three different spindle rotation speeds are selected as the training set and select the rest 30% samples as the test set. The model is solved by neural network backpropagation and stochastic gradient descent (Kingma and Ba 2014). After training the deep network model, then input the test set to obtain the output labels $\{Y_i\}_{i=1}^T$ to obtain the tool wear states with healthy, wear, breakage label.

The procedures are shown in Algorithm 1.

Experiment results and analysis

To verify the performance of the CDATT model for tool wear status prediction method based on the attention mechanism in this paper, we use three different types of wearied tools with healthy, wear, breakage to conduct milling experiments on actual CNC machine tools. The deep learning experimental environment used in this paper is Python3.7, and the framework of PyTorch is 1.4.0. The computer configuration

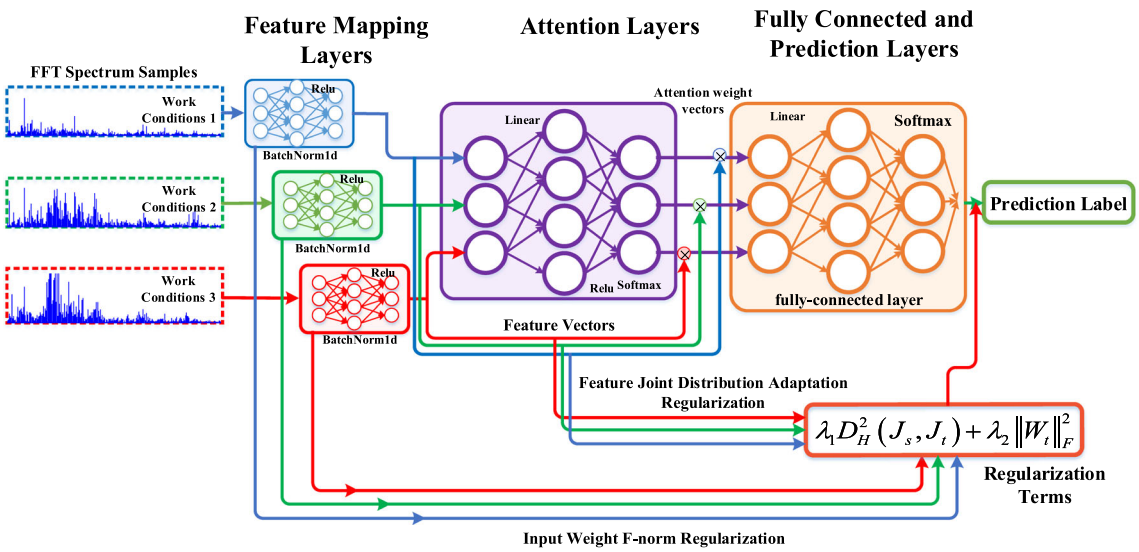

Fig. 2 The proposed CDATT tool wear prediction method

Algorithm 1 The Training Algorithm of CDATT Model

Require: The data sets under T conditions $\left\{S_i = \left\{\mathbf{s}_j^i, y_j^i\right\}_{j=1}^{n}\right\}_{i=1}^{T}$

Ensure: The labels $\{Y_i\}_{i=1}^{T}$ of testing set

1: Using SVD and GMM cluster to filter samples $\{S_i\}_{i=1}^{T}$ and Obtain the cluster results $\left\{S'_i = \left\{\mathbf{s}_j^i, y_j^i\right\}_{j=1}^{m}\right\}_{i=1}^{T}$ by minimum difference sequence entropy.

2: Using FFT to obtain the signal spectrum $\left\{F_i = \left\{\mathbf{f}_j^i, y_j^i\right\}_{j=1}^{m}\right\}_{i=1}^{T}$, then select 70% samples randomly as training set $\left\{F_i^{tr}\right\}_{i=1}^{T}$ and 30% as testing set $\left\{F_i^{ts}\right\}_{i=1}^{T}$ under each working conditions T.

3: Using the Adam algorithm to update weights and biases to train the CDATT model. $\nabla_{w,d} \sum_{i=1}^{T} \left(L\left(\mathbf{x}^i, Y^i\right) + \lambda_1 D_H^2\left(J_s, J_t\right) + \lambda_2 \|W_t\|_F^2\right)$, where the hyper-parameters of λ_1 and λ_2 are selected and optimized by Bayesian optimization.

4: Input the tool wear testing set $\left\{F_i^{ts}\right\}_{i=1}^{T}$ into the trained CDATT model and obtain the tool wear status labels $\{Y_i\}_{i=1}^{T}$.

is AMD-3700x processor, 16 Gb memory, and NVIDIA RTX 2060 GPU.

Data preparation

We mainly evaluate our CDATT model on two datasets: The UCB milling dataset Goebel 1996 and our own experiments dataset.

The Milling experiment dataset

In order to verify the performance of the proposed CDATT tool wear prediction method, a series of milling experiments were carried out. Experiments are conducted on KMC600SUMT five-axis machining center (produced by KEDE Company (Dalian, China)), as shown in Fig. 3. The workpiece used in experiments is #45 steel brick, and all cutting process was completed on the same workpiece. As the experiment progresses, the workpiece size didn't change significantly. Experiments were conducted with three different conditions of the flat-end cutting tool as shown in Fig. 3 which the one is health and two cutters are wear conditions with wear cutter and breakage cutter. Each cutting tool was under three different working conditions, where the rotation speed of spindle is 2000 RPM, 2600 RPM and 3200RPM. The vibration signal of each working conditions was collected by DongHua 1A314E acceleration sensor which was mounted on the spindle housing. The sampling frequency is 12 kHz and vibration signals include X, Y, Z three directions.

After collecting vibration signals of the cutting tool with healthy, wear and breakage cutters, we select some well-characterized signals to visualize in Fig. 4. The signal on the X direction in the time domain and frequency domain are visualized under 2600RPM working conditions.

The UCB milling dataset

The datasets are run on a milling machine under various operating conditions from the University of California at Berkeley (Goebel 1996; Agogino and Goebel 2007). This dataset was sampled by three different types of sensors (acoustic emission sensor, vibration sensor, and current sensor) at several locations. The total number of data was 16 different cases with 167 operating runs. To evaluate tool wear status, the flank wear was selected as the generally accepted parameter. At the beginning and the end of each run, the milling was

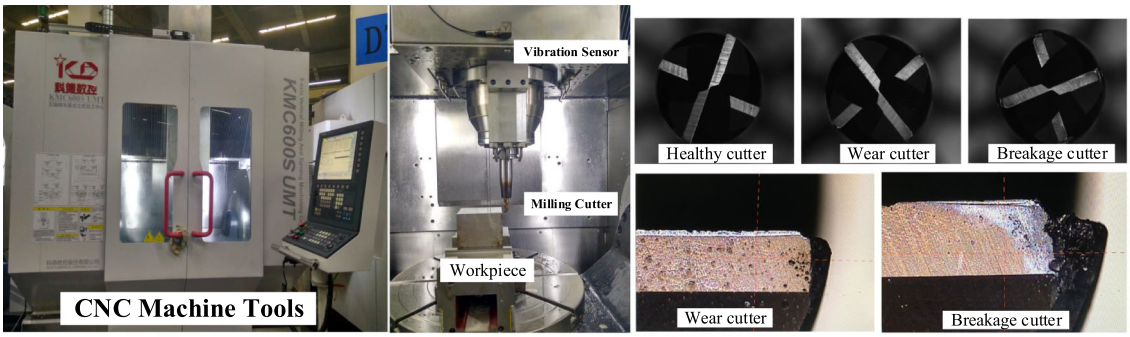

Fig. 3 KMC600SUMT five-axis machining center and the milling tools with different conditions of the heathy cutter, wear cutter and breakage cutter

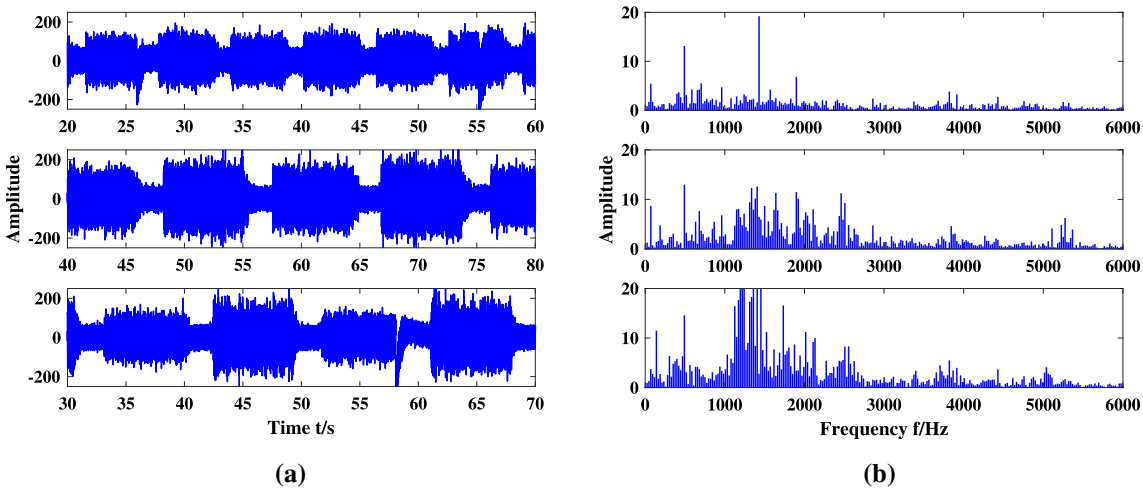

Fig. 4 The collected signals in **a** the time domain and **b** frequency domain accordingly by health cutter, wear cutter and breakage cutter

stopped, and the flank wear was measured. For each milling experiment, the number of runs depends on the degree of flank wear from zero up to a wear limit under various operating conditions. The basic experiment setup includes the spindle and the table of the machining center. An acoustic emission sensor and a vibration sensor are each mounted to the table and the spindle of the machining center. The signals from all sensors are amplified and filtered and then fed through two root mean square (RMS) converters before they enter the computer for data acquisition (Goebel 1996).

Take Case 1 for an example; the measured vibration data with 0.11mm and 0.29mm VBmax are illustrated in Fig. 5, respectively.

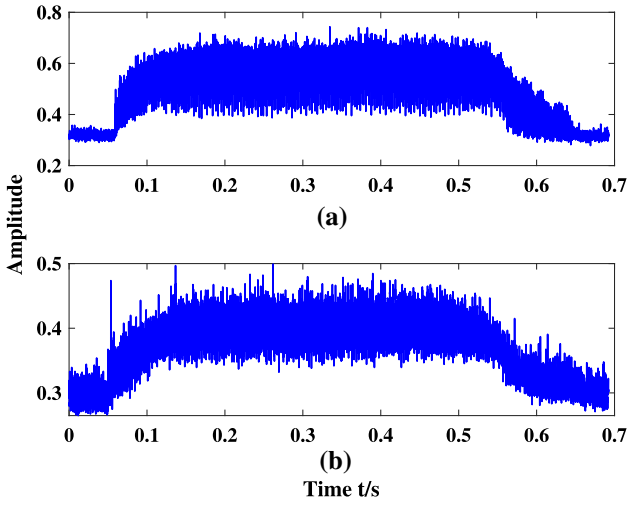

Fig. 5 The vibration signals' RMS of spindle and table in sequence of UCB milling dataset at Case 1 with VBmax of **a** 0.11mm and **b** 0.29mm

Results on datasets

Comparison methods

In order to verify the effectiveness of CDATT model, the proposed method is compared with several tool wear identification and domain adaptation methods.

1. One-dimensional convolutional neural network with joint distribution adaptation regularization term(1D-CNN-JDA)
2. Multi-scale Inception convolutional neural network with joint distribution adaptation regularization term(Inception-JDA)
3. Long-short term memory network with joint distribution adaptation regularization term(LSTM-JDA)
4. The basic network with the residual network with joint distribution adaptation regularization term(ResNet-JDA)
5. Support Vector Machines with transfer component adaptation (SVM+TCA)
6. Support Vector Machines with joint distribution adaptation (SVM+JDA)
7. Geodesic flow kernel (GFK)
8. Support Vector Machines with the time domain statistical features (SVM+TF)
9. Support Vector Machines with frequency domain statistical features (SVM+FF)
10. Support Vector Machines with empirical mode decomposition (SVM+EMD)
11. Support Vector Machines with wavelet packet decomposition (SVM+WPD)

The methods of No.1-7 are using cross-domain adaptation technique. And the methods 1-4 are deep learning-based methods with the joint distribution adaptation regularization term. The deep learning-based methods 1-4 are commonly used with good performance, and 1D-CNN (Eren et al. 2019), Inception Salam and El Hibaoui (2021), LSTM Cai et al. (2020), ResNet Zhou et al. (2019) models have good application effects in industrial data analysis, fault diagnosis and health management. The methods of No.8-11 are effective methods for tool wear state identification, including tool wear identification based on time domain and frequency domain features (Madhusudana et al. 2016; Liao et al. 2019, EMD features Nie et al. (2011) and WPD features Benkedjouh et al. (2015). These methods are not including transfer regularization terms because these performances can be used as a benchmark for traditional machine learning methods for comparing. The methods of No.5-7 are cross-domain methods, the TCA method is to achieve cross-domain transfer by remapping the feature space (Pan et al. 2010), the JDA method is an irreplaceable part of the method proposed in this paper by considering both marginal probability distribution and conditional probability distribution (Long et al. 2013). These two transfer learning methods use the SVM as a classifier to compare the performance. The GFK method is to measure the source domain manifold closest target domain by proposing the rank of domain metric (Gong et al. 2012). Here, the above method is used to compare the tool wear status recognition.

Implementation details

For the method of 1–4 based on the deep learning neural network, the frequency spectrum obtained by fast Fourier transform from the original vibration signal is used as the input to train the deep learning model to identify the test samples. For the method of 8-11, the tool wear state recognition process is to use the extracted features from the vibration signal and then input features into the classifier to identify directly. It should be noted that the source field and target field of methods 1-11 are vibration signals collected under different working conditions. The method 1-7 is to realize cross-domain adaptation for tool wears states identification by adapting different samples under different spindle rotation speeds in the first dataset.

In order to ensure the fairness of the comparison algorithm, each comparison method must be ensured that the model obtains the best results by model optimization and optimal parameter settings and choosing. For SVM, we choose the LIBSVM Chang and Lin 2011 as the compare method and grid search and cross-validation are conducted to select the best parameters. And we select the RBF kernel as the kernel function of SVM, the regularization parameter C and kernel parameter g are both searched in the interval $\{-6, 6\}$ to the power of 2. The EMD and WPD methods use the normalized power spectrum as a feature, and both extract the first 8-dimensional sub-signals energy spectrum as features and use LIBSVM as a classifier to predict the tool wear states. The input hidden number and learning rate are selected and optimized by Bayesian optimization in the interval $\{10, 200\}$ and $\{0.0001, 0.1\}$. The cross-domain comparing method 5-7 are selected RBF kernel and the kernel parameters are using grid search and cross-validation to determine in the interval $\{-6, 6\}$ to the power of 2. And the transformed feature dimension of GFK, TCA, JDA are also searching in the interval $\{10, 100\}$ for the best-comparing performance.

For the building of the compare deep network, our construction principle is that the network is the simplest when the network can work and converge (Eren et al. 2019; Salam and El Hibaoui 2021; Cai et al. 2020; Zhou et al. 2019). We use the basic units in Pytorch to build a deep network, and using the code sequence to represent the whole structure of the net-

Table 1 The network structure of method of CDATT network

No.	Layer type	Channel size	Output size	Parameters number
1	Linear-1	2	512*2	40
2	Linear-2	2	512*2	40
3	Linear-3	2	512*2	40
4	ReLU-4	20	512*20	–
5	BatchNorm1d-5	20	512*20	1024
6	Linear-6	512	20*512	262,656
7	Softmax-7	512	20*512	–
8	Flatten-8	512	20*512	–
9	Dropout-9	512	20*512	–
10	Linear-10	3	1*3	40,964
11	Softmax-11	3	1*3	–

work by setting its initial code for the basic unit. For example, the L is the linear mapping layer. The B represents batch-normalization layer. The C denotes the one-dimensional convolution layer. The M means the one-dimensional maximum pooling layer. And the L represents LSTM layer. The F is the flatten layer. The Re denotes the activation function of ReLU. The S is softmax layer. The R represents the ResNet block layer. the I means inception layer. For the 1D-CNN-JDA, the sequence of network structure parameters is as follows LLL-B-C-B-M-C-B-M-B-F-L-Re-S. For the LSTM-JDA network, the sequence of network structure layer is LLL-B-L-M-B-L-M-B-F-L-Re-S. For the ResNet-JDA network, the code sequence is LLL-B-C-B-M-R-R-M-B-F-L-Re-S, where the ResNet layer is donated as C-C. For the Inception-JDA network, the sequence code is LLL-B-C-B-I-M-I-M-B-F-L-Re-S. It is should be noted that the LLL means the shared feature mapping layer so that to add the JDA-norm term, just like the structure in Fig. 2. The network of the proposed CDATT is shown in Table 1.

The proposed method CDATT and the four compare deep learning methods in this paper uses Bayesian optimization (Snoek et al. 2012) to obtain the best network structure parameters and regularization hyper-parameters. The Bayesian optimization is an adaptive hyper-parameter search method that it can predict the next parameters combination with the most possibility based on the currently hyper-parameters combination. Assuming the function of hyper-parameter optimization $f(\lambda)$ obeys a Gaussian process, then $p(f(\lambda) \mid \lambda)$ is a normal distribution. The Bayesian optimization process is to build a Gaussian process model based on the existing N groups of experimental results $H = \{\lambda_n, c_n\}_{n=1}^{N}$, and calculate the posterior distribution $p(f(\lambda) \mid \lambda, H)$, where c_n is the observed value of $f(\lambda_n)$. In order to make $p(f(\lambda) \mid \lambda, H)$ close to its true distribution, the sample space needs to be sampled enough. However, the generation cost of each sample in hyper-parameter optimization is very high, and it is necessary to use as few samples as possible to make $p(f(\lambda) \mid \lambda, H)$ close to the true distribution. Therefore, it is necessary to define an acquisition function $a(\lambda, H)$ to determine whether a sample can provide more income for modeling $p_\theta(f(\lambda) \mid \lambda, H)$. The larger the gain of $a(\lambda, H)$, the closer its modified Gaussian process is to the true distribution of the objective function.

Therefore, we set the expected improvement(EI) function and assuming that $c^* = \min\{c_n, 1 \le n \le N\}$ is the optimal value in the current existing samples, the expected improvement function is:

$$EI(\lambda, H) = \int_{-\infty}^{\infty} \max\left(c^* - c, 0\right) p(c \mid \lambda, H)\, dc \tag{18}$$

The expected improvement is to define a sample λ under the current model $p(f(\lambda) \mid \lambda, H)$, the best expectation $f(\lambda)$ of the best result c^*. In addition to the expected improvement function, there are improvement probability, Gaussian process confidence upper bound (Bergstra et al. 2011). And the general steps of Bayesian optimization is shown in Algorithm 2.

In actual use, we use the skopt module of Scikit-Optimize in pytorch to optimize the loss function and find the optimal hyper-parameter combination of λ_1 and λ_2.

Results on tool wear prediction

Milling sample filtering

As shown in Fig. 6, the milling operations include three parts: forward milling, reverse milling and idling. However, the signal from forward milling and reverse milling can reflect the healthy states of tools since the tools don't contact the workpiece in the idling process. Therefore, filter the signal samples adaptively for the collected data is the key to identify the tool wear states.

The existing time-series-based cluster methods are still unable to select samples adaptively for the complex signals

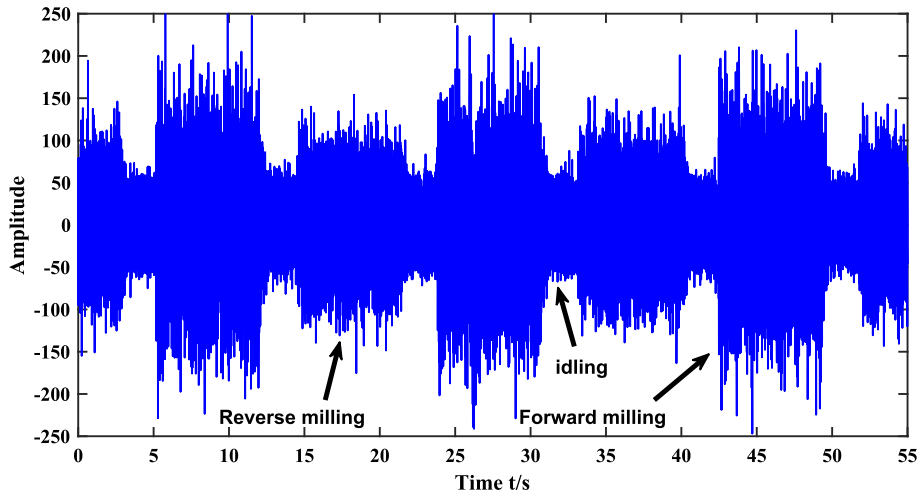

Fig. 6 The monitoring vibration signal of the cutting tool includes forward milling, reverse milling, and idling

Algorithm 2 The Bayesian Optimization for Network Hyper-parameter Search Algorithm

Require: Optimization objective function $f(\lambda)$, number of iterations T, profit function $a(\lambda, H)$.
Ensure: λ with minimum c in H
1: Initialization $H=\emptyset$.
2: Randomly initialize the Gaussian process and compute the $p_{gp}(f(\lambda) \mid \lambda, H)$.
3: for $t = 1$ to T do
$\lambda' \leftarrow \arg\max_{\lambda} a(\lambda, H)$
$c' = f(\lambda')$
$H \leftarrow H \cup (\lambda', c')$
fit new model according to the updated H and compute the $p_{gp}(f(\lambda) \mid \lambda, H)$.
4: end for
5: Return λ with minimum c in H.

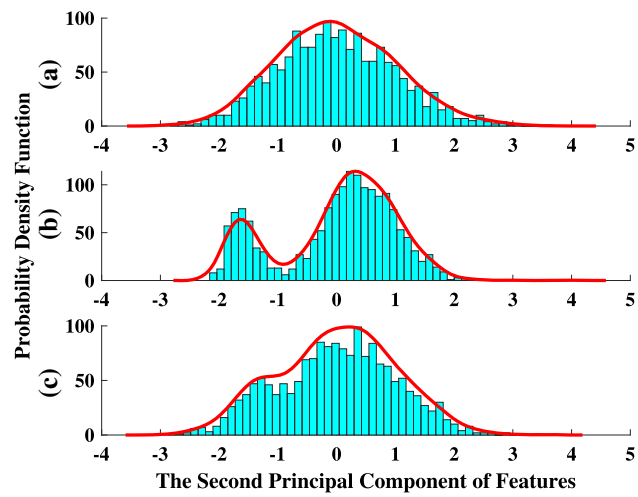

Fig. 7 Probabilistic density function of the SVD values **a** All SVD vector **b** SVD vector from the 5-th dimension to the last **c** SVD vector from the 14-th dimension to the last

under different working conditions. The amplitude and periodicity will change sharply under different rotational speeds. Therefore, the feature-based time-series clustering method is important to use for tool wear identification. In this paper, we proposed a new clustering method based on SVD and GMM for milling samples filtering. We will make a detail explain about the proposed method. The proposed method utilizes SVD vectors as input features, and the different SVD vector component selected will influence the performance of the clustering effect due to their different distribution characteristics. As is shown in Fig. 7.

The singular value analysis is based on a special matrix. Then we build a matrix by the one-dimensional signal and conduct singular value decomposition to extract the singular value vector as the feature. However, the number of singular values is determined by the small rank of the matrix. In order to ensure the number of singular values is maximized for a larger parameter search range, the rows and the columns of the matrix need to be closer as possible. Therefore, we convert the length of 512 signal vector into a matrix with 16x32 to compute its singular value vector and conduct the cluster algorithm. As shown in Fig. 7, we convert the vibration signal into a matrix with 16x32 dimensions, then we obtain 16-dimensional singular values. We make a visualization about the probabilistic density function of these singular values with all SVD vectors, the 5-th and the 14-th dimension to the last. From the SVD vector's probability density distribution we can obtain that the different sets of SVD vector selection will be shown different feature distributions. How to select the best clustering features set is important for clustering results for sample filtering.

In this paper, a Gaussian mixture model clustering method based on minimum differential entropy is proposed in Eq. (16). And the best clustering results are obtained by calculating the differential entropy of cluster labels as an evaluation index. The clustering results are shown in Fig. 8.

From Fig. 8. that the different SVD vector sets impact the GMM performance and clustering results. In Fig. 8c–d, the

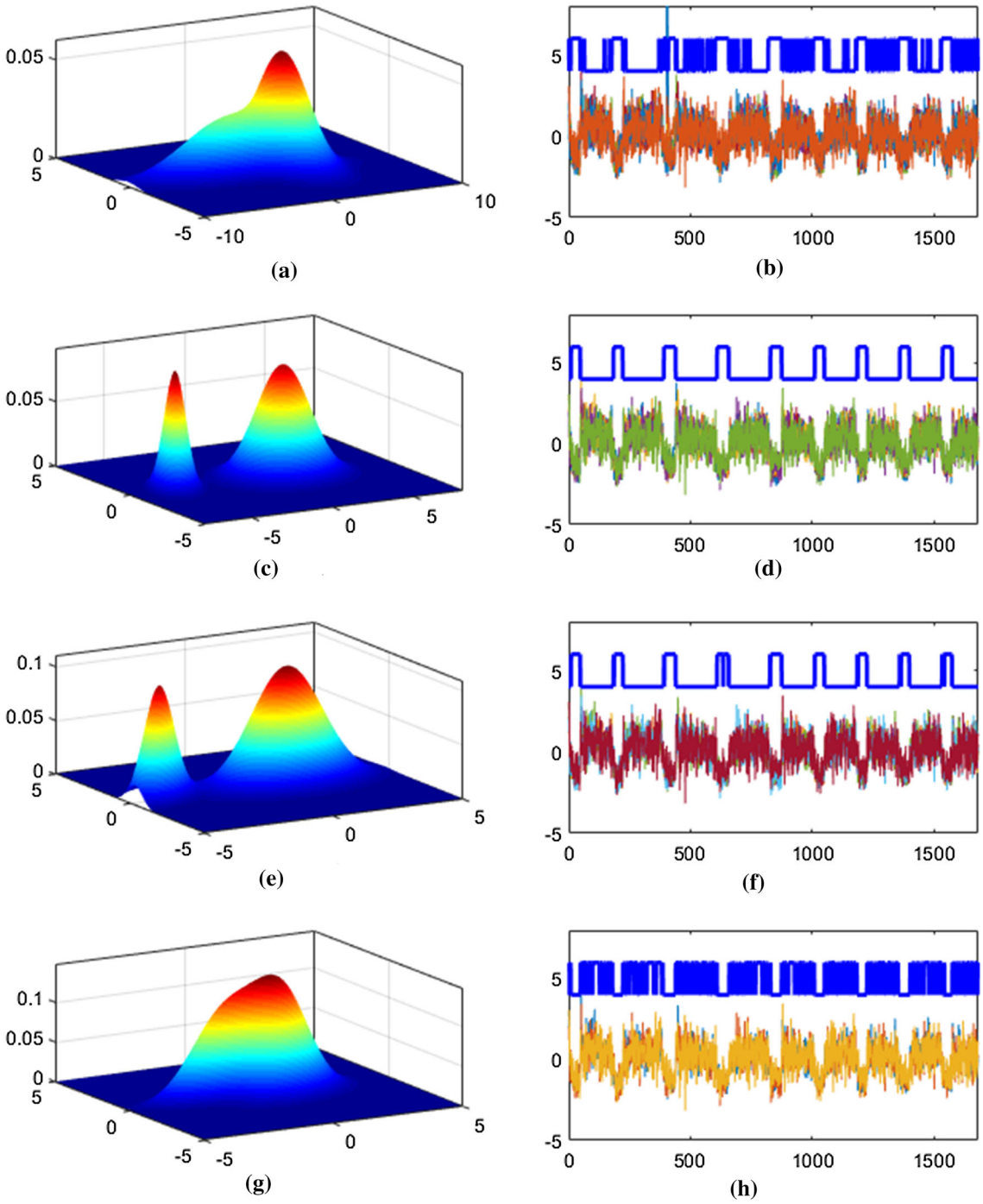

Fig. 8 GMM model clustering results of distribution figure and clustering results with **a**, **b**: All SVD vector.**c**, **d**: the 5-th dimension to the last.**e**, **f**: the 8-th dimension to the last. **g**, **h**: the 14-th dimension to the last

distance between the two Gaussian distributions is far away from each other, and the obtained labels are discriminative. In Fig. 8e, the two Gaussian distributions overlap slightly which indicates that the clustering label will be more aliasing in the clustering result of Fig. 8f. Too many features or too few features cannot effectively distinguish the vibration signal. As shown in Fig. 8b and h, it cannot effectively distinguish different categories.

In this paper, we can find the optimal SVD vector set to obtain the best cluster results by calculating the differential minimum differential entropy value of the cluster labels. The SVD vectors selection number and its corresponding GMM cluster labels' differential entropy is shown in Fig. 9.

As shown in Fig. 9, the minimum cluster label differential entropy is at the 5 which means the best SVD vector set is the 5-th dimension to the last. At this time, the milling sample

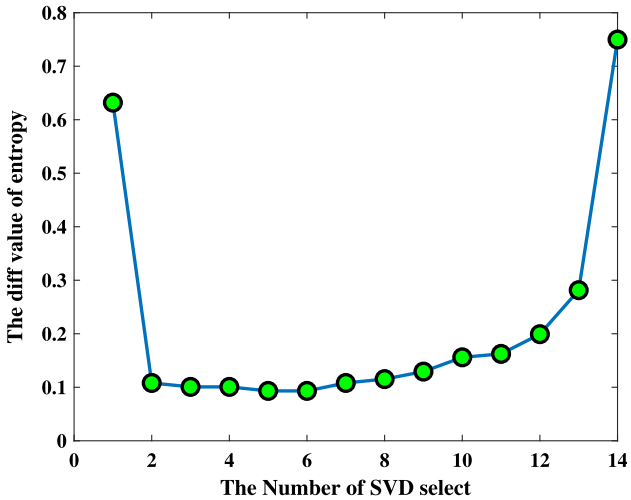

Fig. 9 The differential entropy changes figure of different SVD vector selection number

clustering results is shown in Fig. 10. From Fig. 10a that in the sample clustering result selected, the processing process and the gap can be well distinguished from the original signal. It can provide a good data quality for cross-domain tool wear monitoring.

This paper also compares the methods based on sequence similarity discrimination of time-series clustering methods, such as dynamic time warping(DTW) Keogh and Pazzani (2001), Hausdorff distance(HD) Maciel et al. (2016), longest common subsequence (LCSS) Banerjee and Ghosh (2001), k-means MacQueen (1967), density-based spatial clustering of applications with noise(DBSCAN) (Ester et al. 1996), the obtained clustering result is shown in Fig. 10b, and the obtained clustering result index is shown in Table 2. These include Rand index(RI) Hubert and Arabie (1985), Adjusted Rand index(ARI) Steinley (2004), Fowlkes and Mallows Index(FMI) Fowlkes and Mallows (1983), Normalized Mutual Information(NMI) (Vinh et al. 2010).

In this paper, there are two reasons for the combination of SVD and GMM in this paper:(1) The singular value vector was utilized for clustering due to its robustness of the feature representation. (2) The GMM method is a kind of soft clustering method to deal with the vibration signal changes drastically and discontinuously. Among the compare methods, the distance-based cluster methods are difficult to achieve good results on the dataset of tool wear. The reason is that the amplitude of the vibration signal on the CNC machine tools changes drastically, and it causes discontinuous changes in the feature. As a result, the sample boundary cannot be effectively determined based on distance and density-based methods, which makes it hard to accurately distinguish between milling and idling. The clustering method of GMM based on singular value features proposed in this paper utilizes the robustness of singular value vector as feature representation. In addition, GMMs are clustered by selecting components to maximize the posterior probability. The posterior probability of each data point represents the possibility of belonging to a class, rather than determining that it belongs to a certain class, so it is called soft clustering. It may be more suitable than k-means and DBSCAN clustering methods when the data fluctuates greatly and the distance of clusters center are closer. The clustering results and the corresponding metrics show that the proposed method is more suitable than other methods to filter the signals of different machining stages adaptively.

Cross-domain prediction for tool wear states

(1) Dataset One: The Milling Experiment with Different Rotational Speeds

After the clustering of milling signal samples, the proposed CDATT model is constructed to identify the cross-domain tool wear state. As the comparison of deep learning methods, this paper uses 1D-CNN-JDA model, Inception-JDA model, LSTM-JDA model, ResNet-JDA model to identify the tool wear state under different working conditions. For each vibration signal sample, we select the 512 as the length of the signal sample for extracting features or extract frequency spectrum. The number of samples for each tool wear state is constructed as 2000. The definition of different conditions for tool wear cross-domain state recognition proposed in this paper is shown in Table 3.

Firstly, we compare the convergence of the deep learning-based cross-domain tool wear prediction methods. The iterative optimization curve of these models and the results on the training set are shown in Fig. 11 and Table 4. Due to space limitations, we only show the prediction results of Domain A to B and Domain B to C. It should be noted that the shadow area of the curve represents the variance change of the prediction results after 20 iterations of the loop. As shown in Table 3, the Domain A, B, and C correspond to the main spindle speeds of 2000 RPM, 2600 RPM, and 3200 RPM respectively for the CNC machine tool.

From Fig. 11 and Table 4, the deep learning-based tool wear prediction methods are all got a good training accuracy. In terms of model training loss, the proposed CDATT method got the lowest loss values on the three domains, and the prediction accuracy obtained is relatively stable, with a minimum of 99.32% on 3200 RPM. The LSTM-JDA model has poor performance among the above methods. The accuracy on 2600 RPM and 3200RPM only reaches 73.99% and 73.19%, and the algorithm stability is insufficient. And the prediction accuracy of other deep learning methods is better relatively. In the Domain A to B task with relatively difficult prediction, the CDATT also achieved the highest prediction accuracy. Meanwhile, From Fig. 11 that among the three tasks, the prediction stability of the CDATT algorithm is also

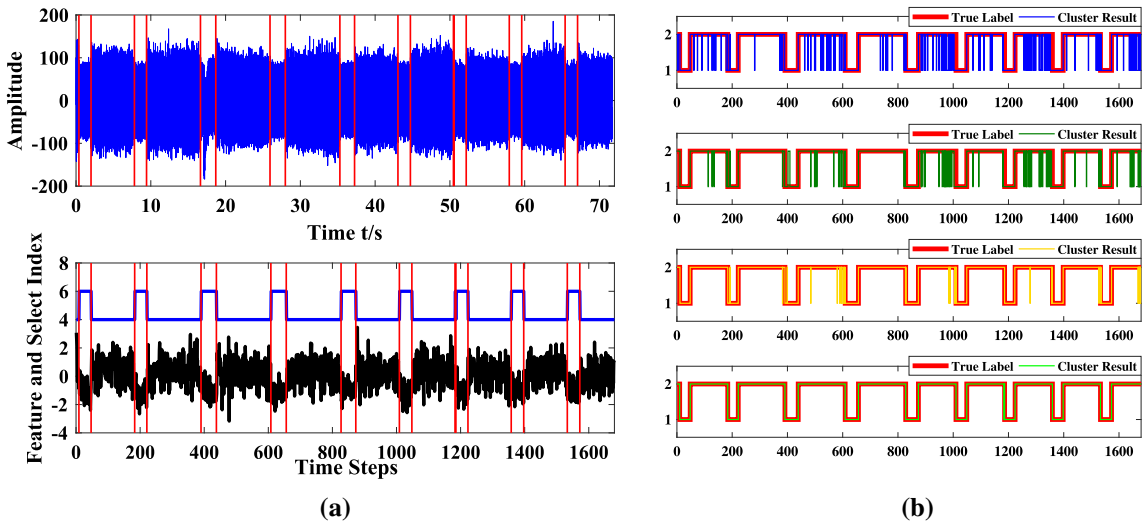

Fig. 10 **a** The clustering results of SVD features and vibration signal **b** The clustering results of the DTW, Hausdorff distance, LCSS and Our proposed method

Table 2 Comparison of clustering results of tool milling vibration signal samples

Metrics	DTW	HD	LCSS	K-means	DBSCAN	Proposed method
ARI	0.6038	0.7221	0.9205	0.7228	0.6045	**0.9974**
RI	0.8095	0.8693	0.9638	0.8807	0.8095	**0.9988**
FMI	0.6190	0.7386	0.9275	0.7613	0.6190	**0.9976**
NMI	0.5339	0.6116	0.8401	0.6141	0.5460	**0.9915**

Table 3 The definition of different working conditions setting

Domain name	Data type	Sample rate	Spindle speed	Tools healthy state
Domain A	Vibration signal	12kHz	2000 RPM	Healthy/wear/breakage
Domain B	Vibration signal	12kHz	2600 RPM	Healthy/wear/breakage
Domain C	Vibration signal	12kHz	3200 RPM	Healthy/wear/breakage

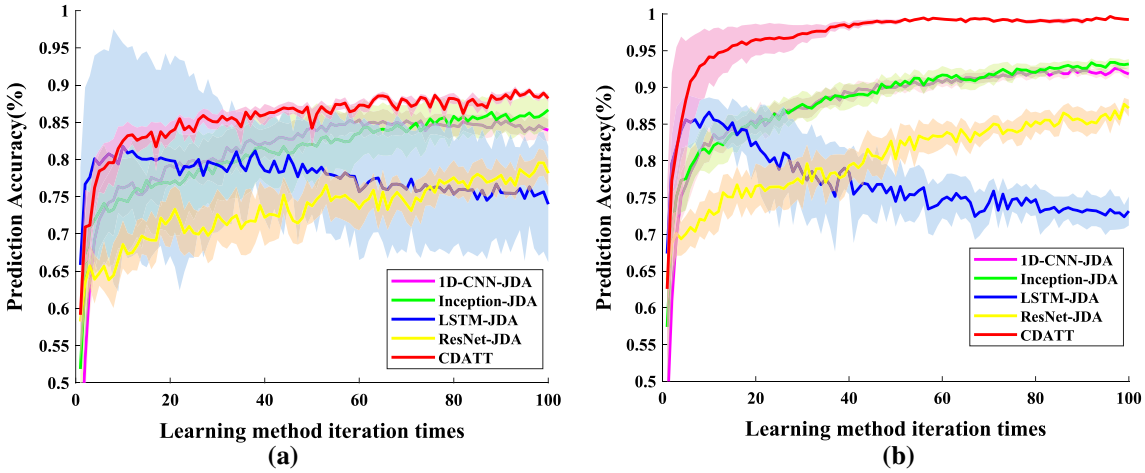

Fig. 11 The accuracy curve of deep learning methods for cross-domain tool wear prediction(**a**) From 2000RPM to 2600RPM (Domain A to B) (**b**) From 2600RPM to 3200RPM (Domain B to C)

Table 4 The training loss and prediction accuracy of comparing methods based on deep neural networks

Domain	Name	1D-CNN-JDA	Inception-JDA	LSTM-JDA	ResNet-JDA	CDATT
Domain A to B	Accuracy(%)	83.9	86.67	73.99	78.14	**88.21**
	Loss Value	1.0664	1.0568	1.065	1.0525	**0.8717**
Domain B to C	Accuracy(%)	91.83	93.19	73.19	87.18	**99.24**
	Loss Value	1.0503	1.0459	1.0554	1.0431	**0.7517**
Domain C to A	Accuracy(%)	90.98	91.09	92.78	87.2	**99.32**
	Loss Value	1.0503	1.0482	1.0473	1.0444	**0.7644**

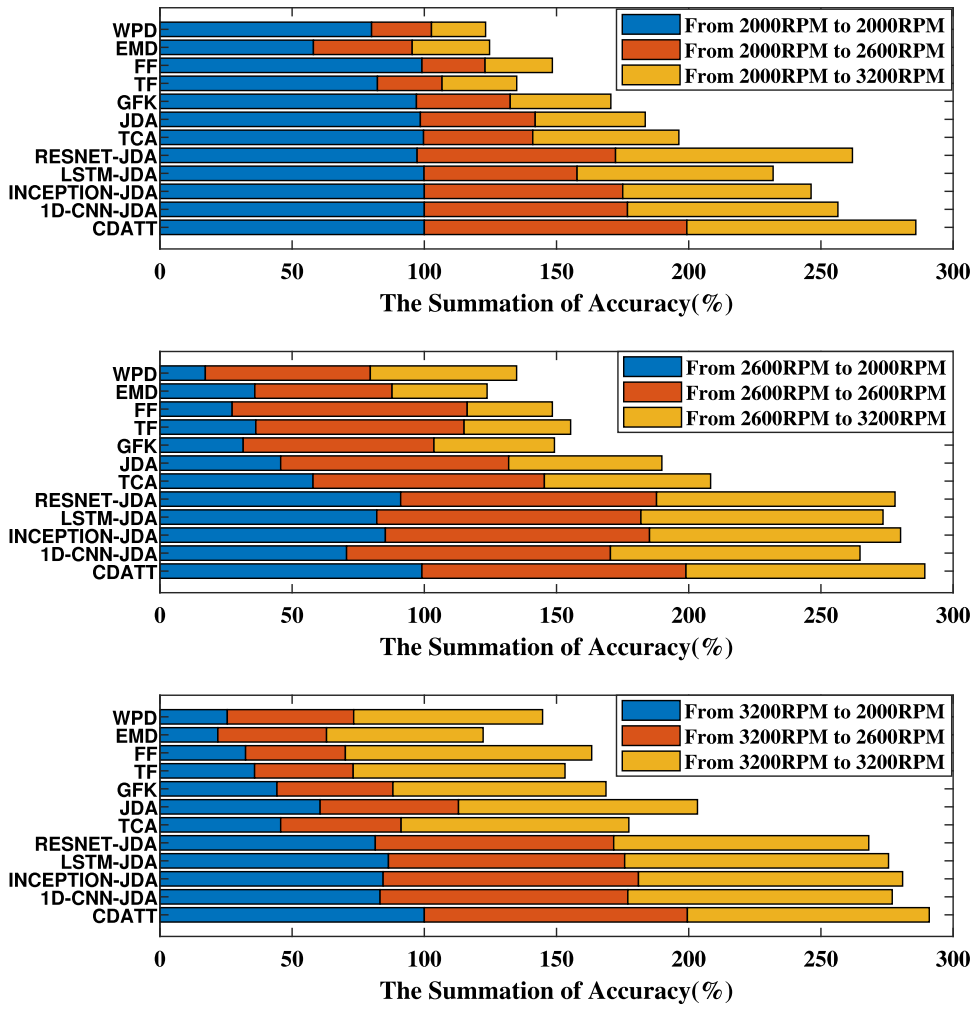

Fig. 12 Tool wear prediction results

the highest. In these cases, the CDATT cross-domain method can take into account both JDA and attention mechanism to make the prediction results stable.

Then considering the performance of cross-domain tool wear recognition, we compare non-deep learning methods of GFK, JDA, and TCA methods for comparison. The obtained prediction accuracy results are shown in Fig. 12–13.

From the Fig. 12, each prediction method obtained a good prediction performance with the same training and testing domain. The darker colors in the 2000 RPM to 2000 RPM, 2600 RPM to 2600 RPM and 3200 RPM to 3200 RPM columns in Fig. 13 represent better predictions. But considering the cross-domain identification of tool wear states in Fig. 12, it is clear that CDATT based on the attention mechanism have achieved better results than other methods, and the total accuracy length of the histogram is the highest. The prediction accuracy of CDATT on each task in Fig. 13 also verifies this results that the proposed method has the best performance. The prediction effects of TCA, JDA and GFK, which are also transfer learning, are got poor prediction accuracy

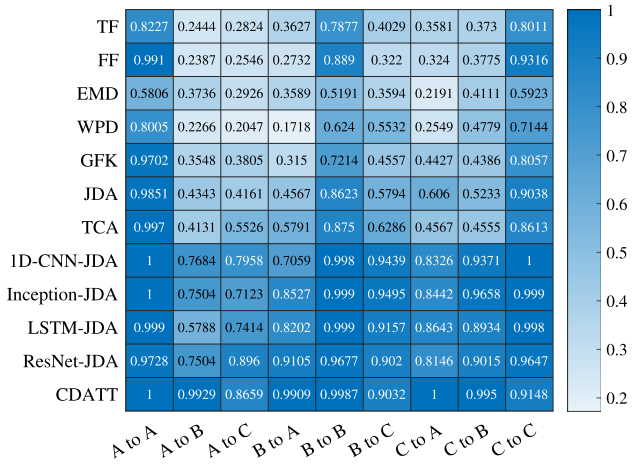

Fig. 13 Tool wear prediction accuracy results in detail on the Dataset One

in different domains which indicates that these algorithms cannot effectively classification.

Here we provide parameter search results based on SVM method, as shown in Table 5. The best hyper-parameters of λ_1 and λ_2 of the CDATT on the dataset one are 2.849e-1 and 58.

(2)Dataset Two: The UCB milling experiment with different flank wear VB.

In order to verify the effectiveness of the proposed CDATT method, we verify it on the mill dataset from UCB milling experiment (Goebel 1996; Agogino and Goebel 2007). We choose two different tool wear states with the different of maximum flank wear(VBmax), and set up domain D and domain E for experiments. The selected data table is shown in Table 6. For each vibration signal sample, we select the 512 as the length of the RMS signal sample for extracting features. The samples are constructed by sliding windows and the number of samples for each tool wear state is 130. To verify the necessity of the transfer model, we first visualize the feature distribution of the dataset, as shown in Fig. 14.

From the two domain distributions in Fig. 14, the second principal components of its statistical features are quite different in the data distribution, and the performance of data-driven machine learning method will be limited in this cross-domain prediction. We also compare the deep learning-based cross-domain methods and the proposed CDATT method to compare the tool wear prediction results. The obtained results are shown in Fig. 15.

From Fig. 15 we can observe that the numerical stability of the Inception-JDA with green color method is good, but the prediction accuracy is not so high. The prediction performance of 1D-CNN-JDA, LSTM-JDA and the ResNet-JDA methods get a stable tends in 300 iterations, but the proposed CDATT method reached the best prediction performance finally. The compare results of non-deep learning methods of GFK, JDA, and TCA and basic SVM method with TF, FF, EMD and WPD are shown in Fig. 16 and Table 7.

From Table 7, the cross-domain prediction task in Dataset two is relatively easy, and the basic non-deep learning methods TCA, JDA and GFK have achieved good results. In tasks of domain D to E, both JDA and TCA prediction accuracy exceeds 90%. And in the tasks of domain E to D, the accuracy of the above three methods has reached more than 90%, too. The CDATT method proposed in this paper achieves 100% and 98.9% prediction results in domain D to E and domain E to D prediction tasks, respectively. The above results in the prediction of Dataset two also show that the cross-domain prediction method based on JDA and attention mechanism has a better performance.

Here we provide parameter search results based on SVM method on Dataset two, as shown in Table 8. The best hyper-parameters of λ_1 and λ_2 of the CDATT on the dataset two are 2.13e-2 and 52.

Table 5 the regularization parameter(2^C) and RBF kernel parameter(2^g) of SVM-based methods

Domain	Name	TF	FF	EMD	WPD	TCA	JDA
Domain A	C	4.667	6	6	6	−0.28	−3.89
	g	2	3.33	0.67	2	5	2.27
Domain B	C	6	6	6	2	5	−0.55
	g	0.667	−0.67	−0.67	3.33	−4.1	4.09
Domain C	C	6	0.667	0.67	0.67	−0.17	5
	g	0.667	0.667	3.33	0.67	−2.27	−0.45

Table 6 The definition of different working conditions setting

Domain	Data type	VBmax	Tools healthy state
Domain D	Vibration signal	0.025/0.24/0.44	Healthy/wear/breakage
Domain E	Vibration signal	0.075/0.29/0.47	Healthy/wear/breakage

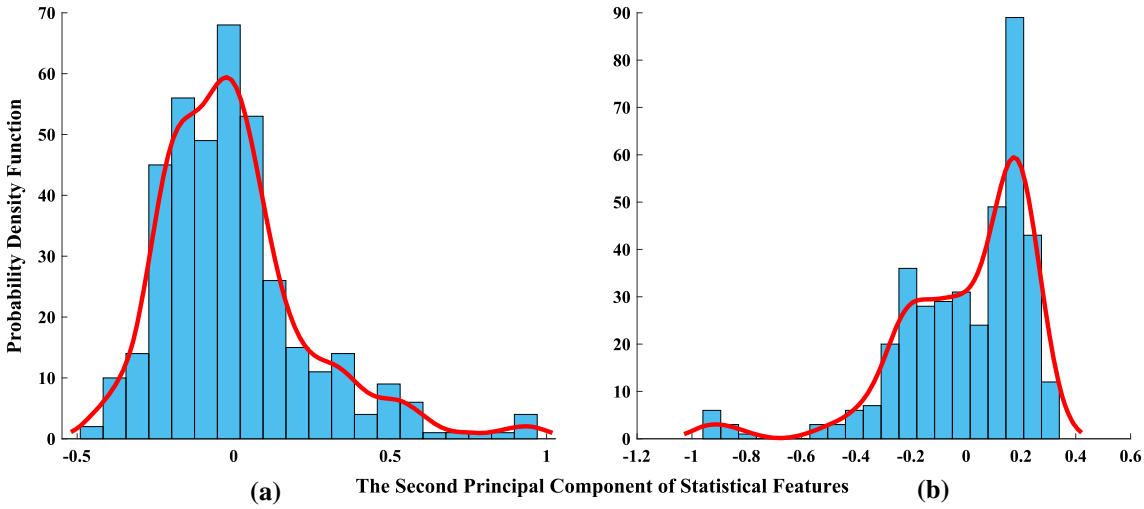

Fig. 14 The feature distribution of the selected dataset with **a** Domain D **b** Domain E

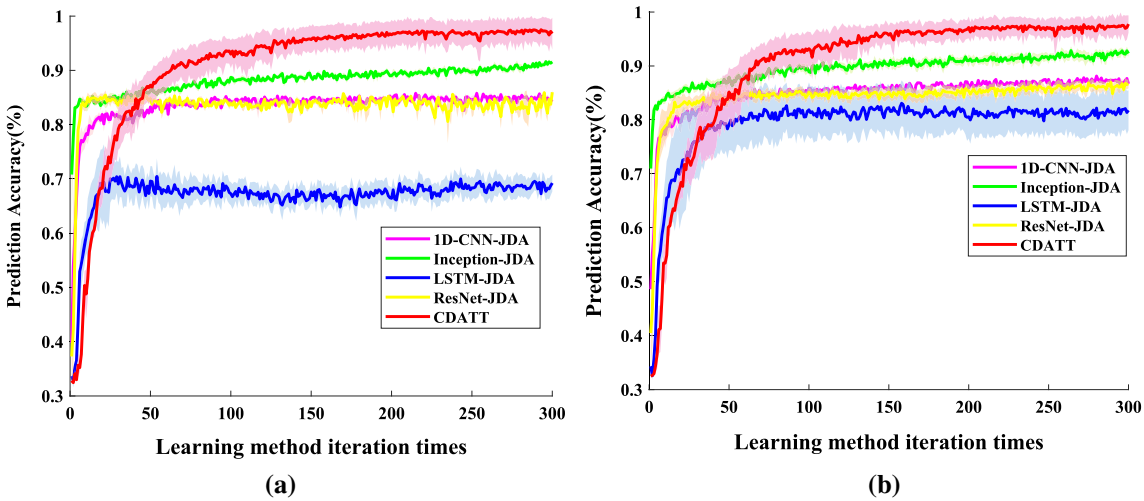

Fig. 15 The accuracy curve of deep learning methods for cross-domain tool wear prediction(**a**) Domain D to E **b** Domain E to D

Feature visualization and model parameter selection

To further analyze the reasons for the cross-domain recognition effect, we use the T-SNE (Van der Maaten and Hinton 2008) method to visualize each feature of the comparing methods, as shown in Fig. 17.

From the feature distribution of Fig. 17, the points with different colors but the same shape represent the same data from different training models. The higher the overlap degree of the same shape points are, the stronger the cross-domain ability of the model get. The farther the distance of the different shape points are, the stronger the classification ability of the model obtain. The feature distances from Fig. 17b, e and f are more farther with each other, but the rest three feature distributions are mixed with each other which indicate the LSTM-JDA model, the inception and the ResNet-JDA model may not suitable for cross-domain prediction of tool

	D to D	D to E	E to D	E to E
TF	0.9121	0.707	0.8205	0.9084
FF	0.9744	0.8938	0.9048	0.9707
EMD	0.7143	0.5311	0.6227	0.7839
WPD	0.8132	0.6081	0.7363	0.8278
GFK	0.9744	0.8513	0.9128	0.9795
JDA	0.9949	0.9282	0.9436	0.9795
TCA	0.9795	0.9385	0.9436	0.9744
1D-CNN-JDA	0.8971	0.8839	0.8661	0.9118
Inception-JDA	0.875	0.8482	0.8214	0.9154
LSTM-JDA	0.8125	0.7857	0.8839	0.9375
ResNet-JDA	0.9154	0.8214	0.8839	0.9154
CDATT	0.9926	1	1	0.989

Fig. 16 Tool wear prediction accuracy results in detail on the Dataset Two

Table 7 Tool wear prediction accuracy results in detail on the Dataset Two

Accuracy(%)	Domain D to D	Domain D to E	Domain E to D	Domain E to E
TF	91.21	70.7	82.05	90.84
FF	97.44	89.38	90.48	97.07
EMD	71.43	53.11	62.27	78.39
WPD	81.32	60.81	73.63	82.78
GFK	97.44	85.13	91.28	97.95
JDA	**99.49**	92.82	94.36	97.95
TCA	97.95	93.85	94.36	97.44
1D-CNN-JDA	89.71	88.39	86.61	91.18
Inception-JDA	87.5	84.82	82.14	91.54
LSTM-JDA	81.25	78.57	88.39	93.75
ResNet-JDA	91.54	82.14	88.39	91.54
CDATT	99.26	**100**	**100**	**98.9**

Table 8 the regularization parameter(2^C) and RBF kernel parameter(2^g) of SVM-based methods on Dataset two

Domain	Name	TF	FF	EMD	WPD	TCA	JDA
Domain D	C	4.67	6	3.33	6	−3.89	−1.67
	g	2	−0.67	−0.67	0.67	−3.18	−2.27
Domain E	C	0.67	6	6	0.67	0.55	−2.78
	g	0.67	0.67	−2	4.67	−1.36	−1.36

wear. In Fig. 17f, we can observe that the overlap degree of the same shape points are the largest and the distance of the different shape points are the furthest. It can prove the good performance of the proposed CDATT method. Considering the ATT method which doesn't include the JDA regulation term for cross-domain modeling of Fig. 17e, the distinction between 2000 RPM, 2600RPM and 3200 RPM of the ATT mdoel is closer then that of the CDATT.

This paper also considers the effect of different regularization parameters of CDATT on cross-domain tool wear prediction. By updating λ_1 and λ_2, the average cross-domain prediction accuracy is changed. From Fig. 18, under a specific parameter combination, CDATT can achieve the best tool wear recognition under different working conditions. From Fig. 18a, the three stages of CDATT variation trend are observed with the change of the parameter λ_2 in a specific parameter of $\lambda_1 \in (0.2, 0.4)$. When $\lambda_1 \in (0.23, 0.42)$ and $\lambda_2 \in (54, 60)$, the CDATT model obtains the best prediction accuracy. It is worth noting that there are also some extreme points in a rang $\lambda_1 \in (0.7, 1)$ and $\lambda_2 \in (26, 52)$, It shows that the deep learning-based model may not exist only one extreme point with the modify of loss function. Therefore, the hyper-parameter optimization method based on Bayesian optimization needs to increase the number of iterations, and increase the learning rate appropriately and add Dropout layer to improve performance. From Fig. 18b, as the $\lambda_1 \in (0, 0.1)$ parameter gradually increases, the CDATT prediction accuracy increases and then decreases. The best CDATT prediction area is relatively concentrated, distributed in the range $\lambda_1 \in (0.01, 0.03)$ and $\lambda_2 \in (45, 73)$. It is indicated that CDATT can achieve a stable and desirable performance in a wide range of parameter values. The best hyper-parameters λ_1 and λ_2 of the CDATT on the dataset one are 2.849e-1 and 58, and the best hyper-parameters on the dataset two are 2.13e-2 and 52.

To further verify the effectiveness of the proposed CDATT model, we constructed ablation experiments on the Dataset One and Dataset Two respectively. We consider the following four cases: (1) not add any regularization term, (2) only add the F-norm regularization term. (3) only add the JDA-norm regularization term. (4) add both JDA and F-norm term. For the JDA regularization term and F-norm term in Eq. (17), the corresponding terms are added and removed under the above optimal parameters to compare the network performance. From Fig. 19a, when only add the JDA term, the model has a good performance but it is not as good as adding both JDA and F-norm regularization term. Only add the F-norm of the network, the cross-domain performance of network relies on the public network structure in Fig. 2, and the performance is insufficient. From Fig. 19b, the performance of the network without any regularization term even exceeds the one with JDA term. We think the distribution of the data from Domain D and E is closer, relatively. And the public network part works well so that this model can achieve a better effect. However, on both dataset one and dataset two, the CDATT

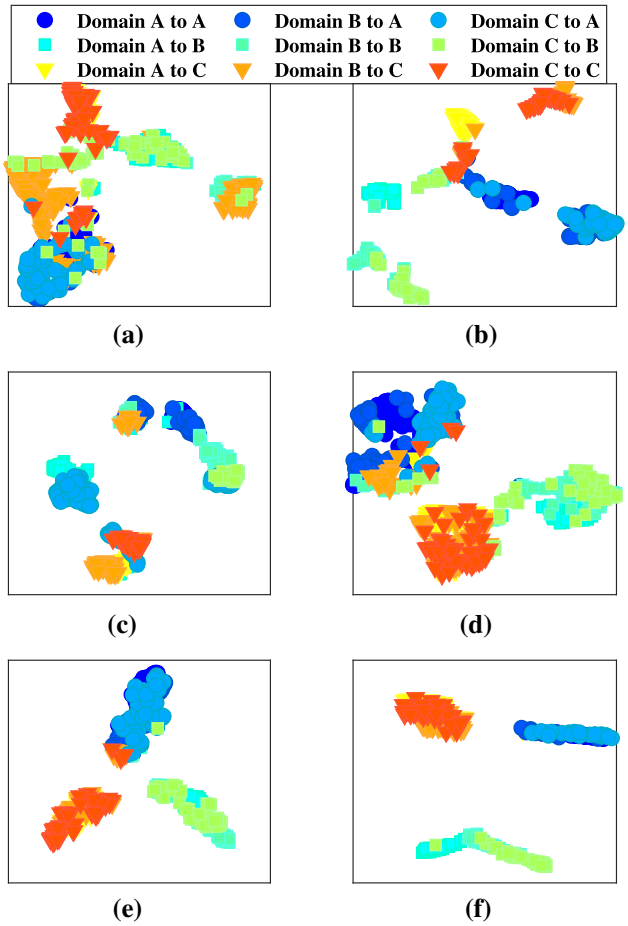

Fig. 17 Scatter plots of principal components features by T-SNE **a** 1D-CNN-JDA **b** LSTM-JDA **c** Inception-JDA **d** ResNet-JDA **e** ATT **f** CDATT

achieved the best prediction effect finally and it indicate the effectiveness of the loss function constructed in this paper.

Finally, to verify the effectiveness of the attention mechanism in this paper, the network weights of CDATT are visualized, and the attention weights and the corresponding specific locations of the spectrum are visualized, as shown in Fig. 20. It can be seen from the two dashed circles, the selected sensitive features are mainly concentrated in two different frequency regions. From reference (Bisu et al. 2012) we can obtained that the two regions are the workpiece frequency domain and milling tool frequency domain respectively. And our proposed method can further select wear features from this two domains. The CDATT tool wear states recognition based on the attention mechanism can realize the recognition of cross-domain tool wear states by selecting the frequency spectrum of a specific location. Compared with other sensitive feature selection or data distribution adaptation methods, this CDATT method can realize universal recognition of tool wear state under different working conditions. From Fig. 20b, most of the non-zero items of the weight are concentrated in the low-frequency part of the input spectrum, while the signal in the high-frequency part is 0 after model optimization. This performance indicates that the CDATT method can extract tool wear prediction features adaptively in the low-frequency part of the signal, and can guarantee the sparsity of the model.

Conclusion

In this paper, a new tool wear status prediction method based on cross-domain adaptation and attention mechanism is proposed. This method can not only filter milling signals, but also capture sparse features adaptively with good and stable wear representation. By analyzing the experimental results,

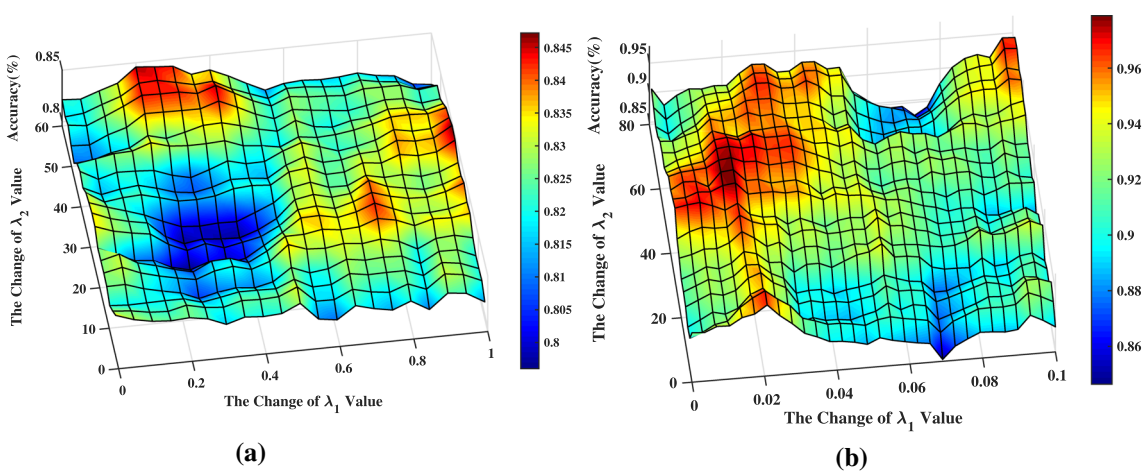

Fig. 18 The average prediction accuracy of CDATT with the regularization parameters changes of the λ_1 and λ_2 on **a** the Dataset One **b** the Dataset Two

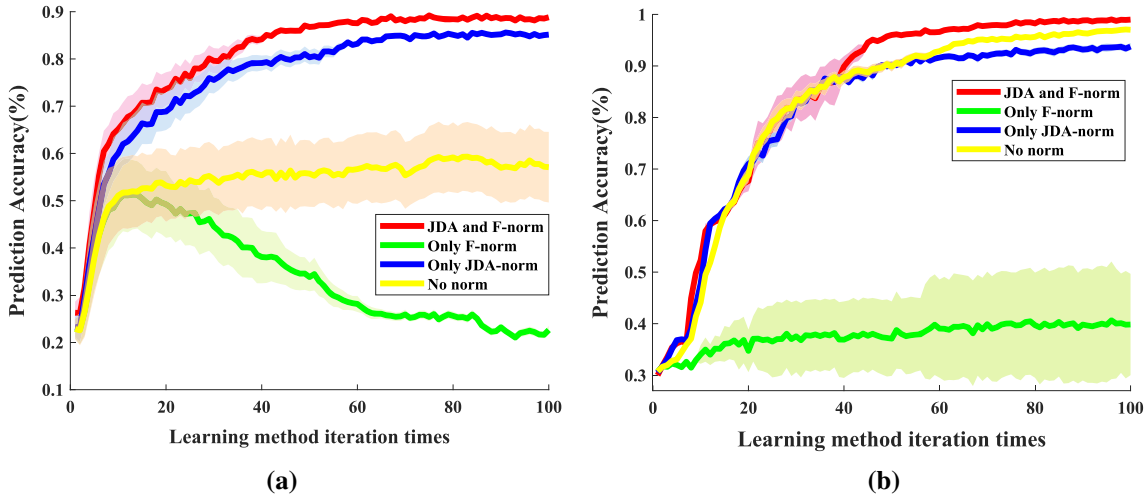

Fig. 19 The performance of ablation experiment on **a** Domain C of the Dataset One **b** Domain E of the Dataset Two

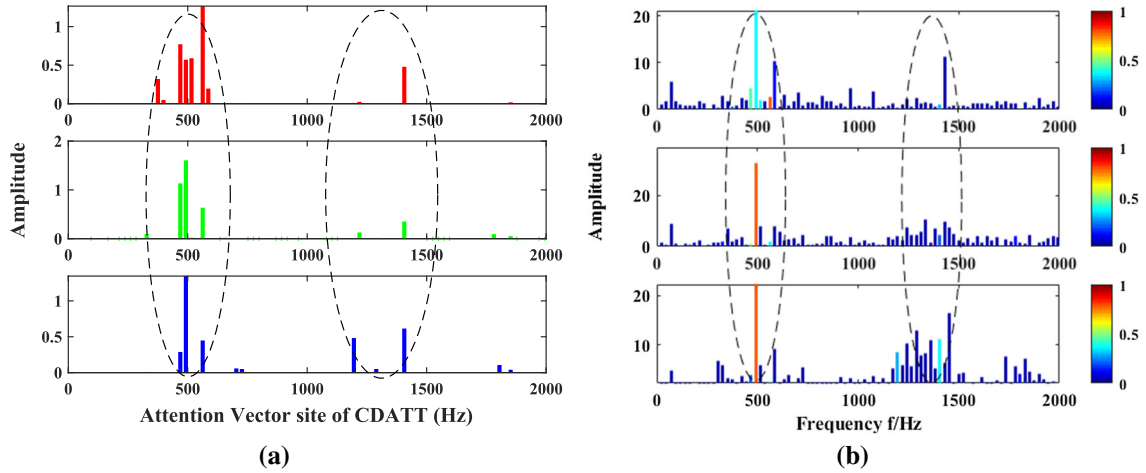

Fig. 20 The visualization of the attention mechanism on Dataset one with Domain A B and C with **a** attention weights **b** attention weights on corresponding spectrum

the main contributions of this paper are summarized as follows.

(1) The proposed time-series clusters method with SVD and GMM can filter the signals for different machining stages adaptively. And it can provide effective and clean data for further modeling.
(2) The proposed network makes a combine with attention mechanism and JDA regularization term, and extract the cross-domain features with similar data distribution to improve the transfer prediction performance.
(3) The proposed CDATT method can capture the sensitive parts of the frequency spectrum. It can help to explain the relationship between the tool wear state characteristic under different working conditions.

In the next work, our research interest will focus on the structure deep learning diagnosis model. Considering the multi-source monitoring data, the structure data can provide a better feature representation for improving tool wear prediction.

Acknowledgements This work was supported by the National Key Research and Development Program of China (2021YFB2012104), National Natural Science Foundation of China (52075267,52005332).

Declarations

Conflict of interest We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

Agogino, A. & Goebel, K. Best lab. in: Uc berkeley, milling data set. nasa ames prognostics data repository. http://ti.arc.nasa.gov/project/prognostic-data-repository (2007).

Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014) .

Banerjee, A. & Ghosh, J. Clickstream clustering using weighted longest common subsequences, Vol. 143, 144 (organizationCiteseer, 2001).

Benkedjouh, T., Medjaher, K., Zerhouni, N., & Rechak, S. (2015). Health assessment and life prediction of cutting tools based on support vector regression. *Journal of Intelligent Manufacturing, 26*(2), 213–223.

Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. Advances in neural information processing systems **24** (2011) .

Bisu, C. F., Zapciu, M., Cahuc, O., Gérard, A., & Anica, M. (2012). Envelope dynamic analysis: a new approach for milling process monitoring. *The International Journal of Advanced Manufacturing Technology, 62*(5), 471–486.

Boutros, T., & Liang, M. (2011). Detection and diagnosis of bearing and cutting tool faults using hidden markov models. *Mechanical Systems and Signal Processing, 25*(6), 2102–2124.

Brili, N., Ficko, M., & Klančnik, S. (2021). Automatic identification of tool wear based on thermography and a convolutional neural network during the turning process. *Sensors, 21*(5), 1917.

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, 31*(6), 1497–1510.

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(3), 1–27.

Chang, W.-Y., Wu, S.-J., & Hsu, J.-W. (2020). Investigated iterative convergences of neural network for prediction turning tool wear. *The International Journal of Advanced Manufacturing Technology, 106*(7), 2939–2948.

Colantonio, L., Equeter, L., Dehombreux, P., & Ducobu, F. (2021). A systematic literature review of cutting tool wear monitoring in turning by using artificial intelligence techniques. *Machines, 9*(12), 351.

Collobert, R. & Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning, 160–167 (2008).

Creswell, A., et al. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine, 35*(1), 53–65.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Ding, Y., Jia, M., Miao, Q., & Cao, Y. (2022). A novel time-frequency transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings. *Mechanical Systems and Signal Processing, 168,* 108616.

Dou, J., et al. (2020). An unsupervised online monitoring method for tool wear using a sparse auto-encoder. *The International Journal of Advanced Manufacturing Technology, 106*(5), 2493–2507.

Eren, L., Ince, T., & Kiranyaz, S. (2019). A generic intelligent bearing fault diagnosis system using compact adaptive 1d cnn classifier. *Journal of Signal Processing Systems, 91*(2), 179–189.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.,96*, 226–231.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association, 78*(383), 553–569.

Fu, J., Zheng, H. & Mei, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, 4438–4446 (2017).

Ganin, Y., et al. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research, 17*(1), 2030–2096.

Goebel, K. F. Management of uncertainty in sensor validation, sensor fusion, and diagnosis of mechanical systems using soft computing techniques (publisherUniversity of California, Berkeley, 1996).

Gong, B., Shi, Y., Sha, F. & Grauman, K. Geodesic flow kernel for unsupervised domain adaptation, 2066–2073 (organizationIEEE, 2012).

Goodfellow, I. et al. Generative adversarial nets. Advances in neural information processing systems **27** (2014) .

He, Z., Shi, T., Xuan, J., & Li, T. (2021). Research on tool wear prediction based on temperature signals and deep learning. *Wear, 478,* 203902.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.

Kanagawa, H., Kobayashi, H., Shimizu, N., Tagami, Y. & Suzuki, T. Cross-domain recommendation via deep domain adaptation, 20–29 (organizationSpringer, 2019).

Keogh, E. J. & Pazzani, M. J. Derivative dynamic time warping, 1–11 (organizationSIAM, 2001).

Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) .

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Lei, Y., et al. (2018). Machinery health prognostics: A systematic review from data acquisition to rul prediction. *Mechanical Systems and Signal Processing, 104,* 799–834.

Lei, Y., et al. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing, 138,* 106587.

Lei, Z., Zhou, Y., Sun, B., & Sun, W. (2020). An intrinsic timescale decomposition-based kernel extreme learning machine method to detect tool wear conditions in the milling process. *The International Journal of Advanced Manufacturing Technology, 106*(3), 1203–1212.

Li, X., Zhang, W., & Ding, Q. (2019). Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. *Signal Processing, 161,* 136–154.

Li, J., Lu, J., Chen, C., Ma, J., & Liao, X. (2021). Tool wear state prediction based on feature-based transfer learning. *The International Journal of Advanced Manufacturing Technology, 113*(11), 3283–3301.

Liao, X., Zhou, G., Zhang, Z., Lu, J., & Ma, J. (2019). Tool wear state recognition based on gwo-svm with feature selection of genetic algorithm. *The International Journal of Advanced Manufacturing Technology, 104*(1), 1051–1063.

Liu, J., Ji, S. & Ye, J. Multi-task feature learning via efficient l2, 1-norm minimization. arXiv preprint arXiv:1205.2631 (2012) .

Long, M., Wang, J., Ding, G., Sun, J. & Yu, P. S. Transfer feature learning with joint distribution adaptation, 2200–2207 (2013).

Maciel, L., Ballini, R., & Gomide, F. (2016). Evolving granular analytics for interval time series forecasting. *Granular Computing, 1*(4), 213–224.

MacQueen, J. et al. Some methods for classification and analysis of multivariate observations, Vol. 1, 281–297 (organizationOakland, CA, USA, 1967).

Madhusudana, C., Budati, S., Gangadhar, N., Kumar, H., & Narendranath, S. (2016). Fault diagnosis studies of face milling cutter

using machine learning approach. *Journal of Low Frequency Noise, Vibration and Active Control, 35*(2), 128–138.

Madhusudana, C., Gangadhar, N., Kumar, H., Narendranath, S., et al. (2018). Use of discrete wavelet features and support vector machine for fault diagnosis of face milling tool. *Structural Durability & Health Monitoring, 12*(2), 111.

Mamledesai, H., Soriano, M. A., & Ahmad, R. (2020). A qualitative tool condition monitoring framework using convolution neural network and transfer learning. *Applied Sciences, 10*(20), 7298.

Mao, W., Chen, J., Liu, J. & Liang, X. Self-supervised deep domain-adversarial regression adaptation for online remaining useful life prediction of rolling bearing under unknown working condition. IEEE Transactions on Industrial Informatics (2022) .

Mao, W., Liu, J., Chen, J., & Liang, X. (2022). An interpretable deep transfer learning-based remaining useful life prediction approach for bearings with selective degradation knowledge fusion. *IEEE Transactions on Instrumentation and Measurement, 71,* 1–16.

Marei, M., El Zaatari, S., & Li, W. (2021). Transfer learning enabled convolutional neural networks for estimating health state of cutting tools. *Robotics and Computer-Integrated Manufacturing, 71,* 102145.

Mnih, V., Heess, N., Graves, A. et al. Recurrent models of visual attention. Advances in neural information processing systems **27** (2014).

Nie, P., Xu, H., Liu, Y., Liu, X. & Li, Z. Aviation tool wear states identifying based on emd and svm, 246–249 (organizationIEEE, 2011).

Niehues, J. & Cho, E. Exploiting linguistic resources for neural machine translation using multi-task learning. arXiv preprint arXiv:1708.00993 (2017) .

Ou, J., Li, H., Huang, G., & Yang, G. (2021). Intelligent analysis of tool wear state using stacked denoising autoencoder with online sequential-extreme learning machine. *Measurement, 167,* 108153.

Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks, 22*(2), 199–210.

SK, T. & Shankar, S. Tool wear prediction in hard turning of en8 steel using cutting force and surface roughness with artificial neural network. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science **234** (1), 329–342 (2020) .

Salam, A., & El Hibaoui, A. (2021). Energy consumption prediction model with deep inception residual network inspiration and lstm. *Mathematics and Computers in Simulation, 190,* 97–109.

Sevakula, R. K., Singh, V., Verma, N. K., Kumar, C., & Cui, Y. (2018). Transfer learning for molecular cancer classification using deep neural networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16*(6), 2089–2100.

Shifat, T. A., & Jang-Wook, H. (2020). Remaining useful life estimation of bldc motor considering voltage degradation and attention-based neural network. *IEEE Access, 8,* 168414–168428.

Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems **25** (2012) .

Soualhi, M., Nguyen, K. T., & Medjaher, K. (2020). Pattern recognition method of fault diagnostics based on a new health indicator for smart manufacturing. *Mechanical Systems and Signal Processing, 142,* 106680.

Steinley, D. (2004). Properties of the hubert-arable adjusted rand index. *Psychological Methods, 9*(3), 386.

Tobon-Mejia, D. A., Medjaher, K., & Zerhouni, N. (2012). Cnc machine tool's wear diagnostic and prognostic by using dynamic bayesian networks. *Mechanical Systems and Signal Processing, 28,* 167–182.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research, 9*(11), 112.

Vaswani, A. et al. Attention is all you need **30** (2017). https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research, 11,* 2837–2854.

Wang, F. et al. Residual attention network for image classification, 3156–3164 (2017).

Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing, 312,* 135–153.

Yao, J., Lu, B., & Zhang, J. (2022). Tool remaining useful life prediction using deep transfer reinforcement learning based on long short-term memory networks. *The International Journal of Advanced Manufacturing Technology, 118*(3), 1077–1086.

Yin, C., et al. (2021). Early fault diagnosis of ball screws based on 1-d convolution neural network and orthogonal design. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 235*(5), 783–797.

Yin, C., et al. (2021). Robust wheel wear detection for solid carbide grinding under strong noise interference: A new approach based on improved ensemble noise-reconstructed empirical mode decomposition. *Wear, 486,* 204112.

Yin, C., et al. (2022). Weak fault feature extraction of rolling bearings based on improved ensemble noise-reconstructed emd and adaptive threshold denoising. *Mechanical Systems and Signal Processing, 171,* 108834.

Zhou, B., Cheng, C., Ma, G. & Zhang, Y. Remaining useful life prediction of lithium-ion battery based on attention mechanism with positional encoding, Vol. 895, 012006 (organizationIOP Publishing, 2020).

Zhou, G., Yang, X., Zhang, C., Li, Z., & Xiao, Z. (2019). Deep learning enabled cutting tool selection for special-shaped machining features of complex products. *Advances in Engineering Software, 133,* 1–11.

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

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