The International Journal of Advanced Manufacturing Technology (2020) 110:511–522
https://doi.org/10.1007/s00170-020-05890-x

ORIGINAL ARTICLE

Tool wear state recognition based on gradient boosting decision tree and hybrid classification RBM

Guofa Li[1,2] · Yanbo Wang[1,2] · Jialong He[1,2] · Qingbo Hao[3] · Haiji Yang[1,2] · Jingfeng Wei[1,2]

Received: 1 May 2020 / Accepted: 5 August 2020 / Published online: 13 August 2020

Abstract
Machined surface quality and dimensional accuracy are significantly affected by tool wear in machining process. Tool wear state (TWS) recognition is highly desirable to realize automated machining process. In order to improve the accuracy of TWS recognition, this research develops a TWS recognition scheme using an indirect measurement method which selects signal features that are strongly correlated with tool wear to recognize TWS. Firstly, three time domain features are proposed, including dynamic time warping feature and two entropy features. The time, frequency, and time-frequency domain features of the vibration and force signals are extracted to form a feature set. Secondly, gradient boosting decision tree (GBDT) is adopted to select the optimal feature subset. Lastly, contrastive divergence (CD) and RMSspectral are used to train hybrid classification RBM (H-ClassRBM). The trained H-ClassRBM is used for TWS recognition. The PHM challenge 2010 data set is used to validate the proposed scheme. Experimental results show that the proposed features have better monotonicity and correlation than the classical features. Compared with CD and Adadelta, CD and Adagrad, and CD and stochastic gradient descent with momentum, the H-ClassRBM trained by CD and RMSspectral improves recognition accuracy by 1%, 2%, and 2%, respectively. Compared with feedforward neural network, probabilistic neural network, Gaussian kernel support vector machine, and H-ClassRBM, the proposed TWS recognition scheme improves recognition accuracy by 37%, 51%, 9%, and 8%, respectively. Therefore, the proposed TWS recognition scheme is beneficial in improving the recognition accuracy of TWS, and provides an effective guide for decision-making in the machining process.

Keywords Tool wear state recognition · Hybrid classification RBM · Contrastive divergence · RMSspectral · Gradient boosting decision tree

1 Introduction

In the era of intelligent manufacturing, high productivity and high machining accuracy have become important benchmarks for measuring the manufacturing levels of countries. Severe tool wear and breakage can cause interruptions in the machining process, useless products, and damage to machine tools. Approximately 20% of machine downtime is caused by excessive tool wear in actual industrial processes [1]. Note that real-time tool wear state recognition can effectively increase machine tool utilization by 50%, increase productivity by 35%, and reduce production costs by 30% [2, 3]. Therefore, accurate tool wear state recognition is significant in ensuring the stability of manufacturing processes and improvement of product quality.

Real-time tool wear measurement is difficult to put in practice as the tool is continuously in contact with the workpiece during machining. For this reason, a plethora of indirect approaches for tool wear state recognition have been proposed utilizing sensor signals [4]. A complete tool wear state recognition scheme using an indirect measurement method mainly consists of signal acquisition, feature extraction, feature selection, tool wear state recognition model, and training of tool wear state recognition model [5].

The raw signal during milling process has a lot of useless information including noise. For reliably and effectively

✉ Jialong He
hejl@jlu.edu.cn

[1] Key Laboratory of CNC Equipment Reliability, Ministry of Education, Jilin University, Changchun, Jilin, China

[2] School of Mechanical and Aerospace Engineering, Jilin University, Changchun 130022, China

[3] Office of Academic Affairs, Aviation University of Air Force, Changchun 130000, China

monitoring tool wear state, effective features are necessary to depict the dynamic relationship between raw signal and tool wear. Features extracted from the original signal include the time domain, frequency domain, and time-frequency domain features [6]. The commonly used frequency domain features include statistical features of frequency spectrum [7]. The commonly used time-frequency domain features include the energy ratios of different frequency domains to total energy [8], wavelet energy features [9], and average amplitude of each intrinsic mode function [10]. As for time domain features, Liu et al. [11] extracted the root mean square, skewness, and kurtosis of force signals. Guo et al. [12] proposed related-similarity features to mine beneficial tool degradation information. Que et al. [13] used the dynamic time warping (DTW) algorithm to calculate the similarity coefficient of vibration signals between degraded and normal condition to evaluate the health status of steam turbines. He et al. [14] used the DTW algorithm to calculate the similarity coefficient of vibration signals between degraded and normal bearing to monitor the performance degradation of bearings. Javed et al. [15] proposed trigonometric features to mine beneficial tool degradation information. However, the above studies still contain defects in feature extraction. Firstly, sensor signals often have unequal lengths because of truncation and outlier removal. In addition, signals often have such deformations as amplitude translation, amplitude expansion and contraction, discontinuity, time axis expansion and bending, and linear drift. Related-similarity features do not consider deformations of signals and can only calculate the correlation of equal-length signals. The dynamic time warping algorithm can be used to calculate the correlation of unequal-length signals. However, so far, the dynamic time warping algorithm has not been used for feature extraction of tool wear. Secondly, non-linear and non-steady force and vibration signals often present a multimodal distribution. Trigonometric features have low ability to describe the uncertainty of multimodal distributions.

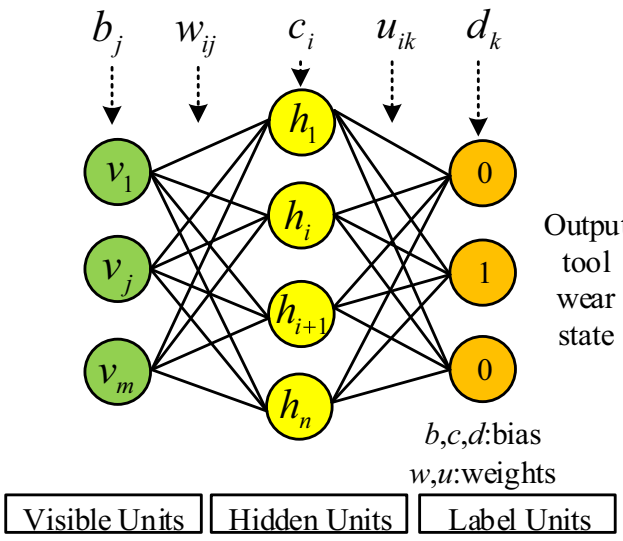

Fig. 1 Illustration of ClassRBM

However, each feature is not always correlated with tool wear. A subset of features that best recognizes tool wear state must be selected instead. The existing feature selection methods can be divided into three categories: filter methods, wrapper methods, embedded methods. Filter methods use a certain metric to estimate the validity of a feature subset, which is independent of the subsequent tool wear state recognition model. The commonly used metrics include correlation (Cor), monotonicity (Mon), and robustness [16, 17]. Wrapper methods generally employ optimization algorithms (such as genetic algorithm) to search for optimal features with high correlation with tool wear state [6]. Although filter methods are computationally fast, the feature subset selected by it does not match the subsequent tool wear state recognition model well, and its performance is low. Wrapper methods are computationally complex, and the feature subset selected by it is biased. Compared with filter methods and wrapper methods, the embedded methods combine the feature selection process and the model training process into one, which can find the optimal feature subset faster and more efficient. In the present study, gradient boosting decision tree (GBDT) is a commonly used feature selection method [18, 19].

Using selected signal characteristics as inputs, a tool wear state recognition model can be built. Counter propagation neural networks [11], multilayer perceptron (MLP) [20, 21], support vector machine (SVM) [22], and a deep coupled restricted Boltzmann machine [23] are commonly applied data-driven methods for tool wear state recognition. Larochelle et al. [24] proposed a hybrid classification RBM (H-ClassRBM). For character recognition and document classification, the classification accuracy of H-ClassRBM was superior to SVM, random forest, and RBM+Feedforward neural network (FNN) [24, 25]. For character recognition and document classification, the classification accuracy of H-ClassRBM trained by reconstruction error was superior to SVM, random forest, RBM+FNN, and deep belief network (DBN) [26]. In summary, the classification ability of H-ClassRBM is stronger than DBN, MLP, and SVM. In addition, H-ClassRBM has strong nonlinear fitting ability, strong robustness, low space-time complexity, clear mathematical model, and good interpretability, which is suitable for solving supervised classification problems. However, so far, H-ClassRBM has not been used to recognize tool wear state.

Two main approaches make H-ClassRBM learning substantially efficient [23, 27, 28]. The first approach is to design an efficient Markov chain Monte Carlo method to obtain good representative samples from the model distribution, thereby reducing the variance of the estimated gradient. The most commonly used approach is a contrastive divergence (CD) algorithm. The second approach is to design superior optimization strategies that are robust to the noise in estimated

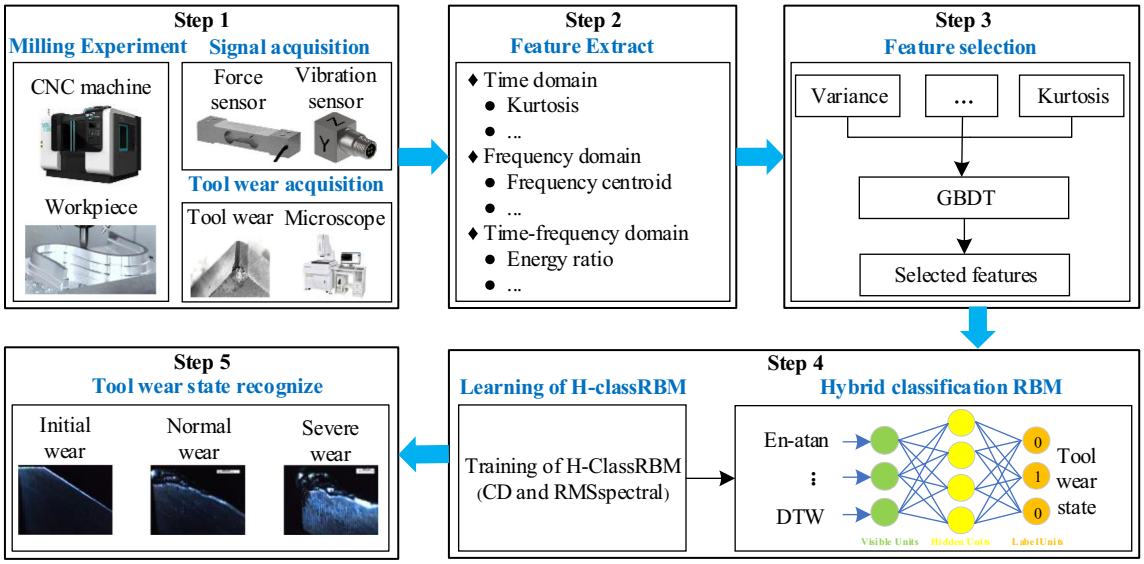

Fig. 2 Flowchart of the tool wear state recognition

gradient. The second class of approaches includes stochastic gradient descent with momentum (SGDM) [29], AdaGrad [30], AdaDelta [31], and RMSprop [32]. Carlson developed a stochastic spectral descent algorithm [33] and SSD-adapt algorithm [34] to train RBM. Subsequently, Carlson proposed RMSspectral that combined a non-Euclidean gradient method with preconditioning [33]. Compared with AdaGrad and AdaDelta, among others, RMSspectral could significantly improve the training efficiency and performance of RBM, FNN, and convolutional neural nets [33]. However, so far, RMSspectral has not been used to train H-ClassRBM.

In order to improve the recognition accuracy of tool wear state, a tool wear state recognition scheme based on GBDT and H-ClassRBM is developed. Firstly, the force and vibration signal of the machining process is collected. Secondly, in order to better depict the dynamic relationship between raw signal and tool wear, three time domain features including a dynamic time warping feature and two entropy features are proposed. The time, frequency, and time-frequency domain features of the vibration and force signals are extracted to form a feature set. Thirdly, feature selection is performed by GBDT, which avoids the processing complexity of high-dimensional nonlinear feature data and weakens the noise component characteristics of the signal. Fourthly, for the first time, CD and RMSspectral are used to train H-ClassRBM, which improves the training process of H-ClassRBM. Lastly, for the first time, the trained H-ClassRBM is employed to evaluate the dimensionality reduction features and obtain the tool wear state.

The remainder of this paper is organized as follows: Section 2 presents the proposed method in detail and gives a brief review of the relevant theories. Section 3 introduces the experimental setup and data set, and discusses the experimental results. Section 4 concludes this paper.

Table 1 Detailed description of sensor data

No.	Data description
1	F_x: force (N) in the x-direction
2	F_y: force (N) in the y-direction
3	F_z: force (N) in the z-direction
4	V_x: vibration (g) in the x-direction
5	V_y: vibration (g) in the y-direction
6	V_z: vibration (g) in the z-direction

2 Methodology

2.1 Gradient boosting decision tree

The basic idea of the GBDT is combining a series of weak base classifiers into a strong one. Different from the traditional boosting methods that weight positive and negative samples, GBDT makes global convergence of algorithm by following the direction of the negative gradient. If there are N sample points (x_i,y_i) and $i = 1,..,N$ (as in Eq. 1), then the goal of optimization is to find the optimal parameters α_m and β_m. SoftMax is the loss function. $F_m(x_i)$ is a prediction function obtained after the mth iteration, where $h(x_i; \alpha)$ ("base learner") are usually chosen to be simple functions of x_i with parameters α.

$$(\alpha_m, \beta_m) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} (L(y_i, F_m(x_i)) + \beta h(x_i; \alpha)) \quad (1)$$

Table 2 Extracted features of each type of signal

Cutting force (x, y, z dimensions) and vibration (x, y, z dimensions)					
Time domain				Frequency domain	Time-frequency domain
Mean value	Root mean square	Square root value	Absolute mean	Frequency centroid	Energy ratio of (3,0)~(3,7)
Skewness	Kurtosis	Maximum value	Minimum value	Frequency variance	Mean amplitude of IMF1~3
Peak-to-peak value	Variance	Waveform index	Peak index	Maximum energy frequency	-
Impulse factor	Tolerance index	Skewness index	Kurtosis index	-	-
En-asinh feature	*En-atan feature*	*DTW feature*	-	-	-

IMF1~IMF3 refer to the 1st to 3rd intrinsic mode functions

En-asinh feature and *En-atan feature* are the two proposed entropy features. *DTW feature* is the proposed dynamic time warping feature

The GBDT process often includes the following steps.

Step 1: The initial weak classifier is defined as a constant β, and $F_0(x)$ represents the initial weak classifier. The constant β makes the initial prediction loss function reach the minimum value.

$$F_0(x) = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, \beta) \tag{2}$$

Step 2: A classifier based on a decision tree is constructed in each iteration. The prediction function obtained after the mth iteration is assumed to be $F_m(x)$. The corresponding prediction loss function is $L(y, F_m(x))$. To immediately reduce the prediction loss function, the mth weak classifier ($\beta_m h(x; \alpha_m)$) should be built on the gradient descent direction of the prediction loss function generated by the first $m-1$ iterations. Among them, $-g_m(x_i)$ represents the building direction of the mth iteration of the weak classifier. $L(y_i, F_m(x_i))$ represents the prediction loss function generated by the first $m-1$ iterations.

$$\begin{aligned} -g_m(x_i) &= -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right], \\ F(x_i) &= F_{m-1}(x_i), i = 1, \ldots, N \end{aligned} \tag{3}$$

Step 3: The basic classifiers are used to fit sample data and get the initial model. According to the least square approach, parameter α_m of the model is obtained and the model $h(x; \alpha)$ is fitted.

$$\alpha_m = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} [-g_m(x_i) - \beta h(x_i; \alpha)]^2 \tag{4}$$

Step 4: Loss function is minimized. According to Eq. (4), a new step size of the model, namely the current model weight, is calculated.

$$\beta_m = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} L\left([y_i, F_{m-1}(x) + \beta h(x_i; \alpha_m)]^2\right) \tag{5}$$

Step 5: The prediction function obtained (Eq. 6) after each iteration is updated. If the corresponding prediction loss function satisfies the convergence condition, then the iteration is terminated.

$$F_m(x) = F_m(x) + \beta_m h(x; \alpha_m) \tag{6}$$

Table 3 Operating conditions

Parameter	Value
Spindle speed	10,400 rpm
Feed rate	1555 mm/min
Y depth (radial) of the cut	0.125 mm
Z depth (axial) of the cut	0.2 mm
Sampling rate	50 kHz/channel
Material	Stainless steel

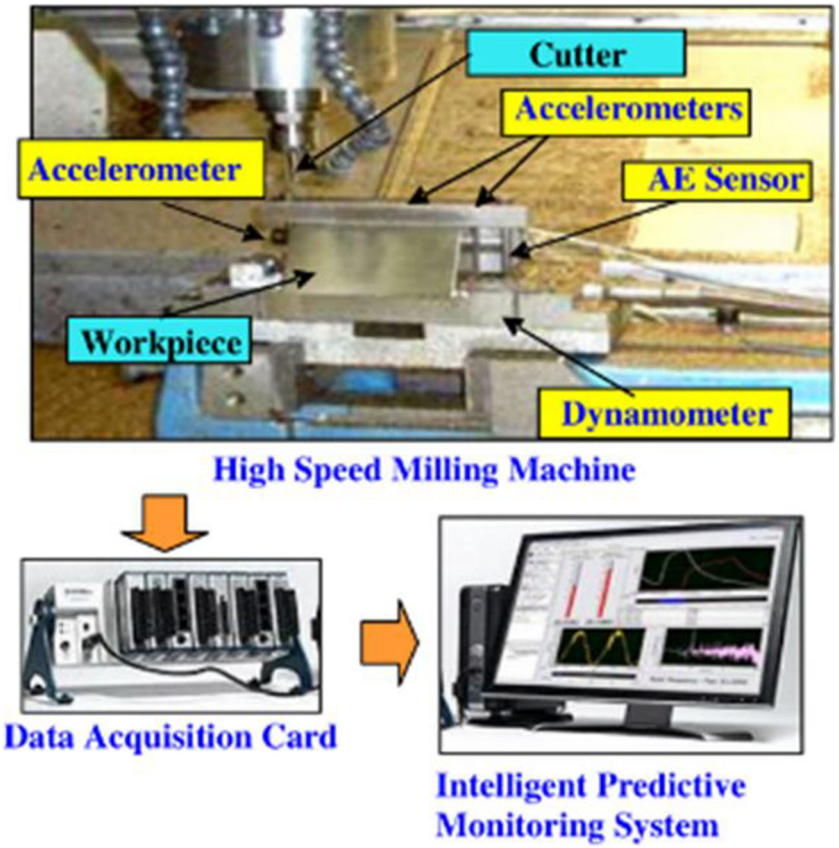

Fig. 3 Tool condition monitoring in high-speed milling process [35]

Compared with genetic algorithm, GBDT is more capable of processing non-linear, low-dimensional dense, and multi-feature types of data. Therefore, GBDT is used to select the optimal feature subset.

2.2 Hybrid classification restricted Boltzmann machine

Classification restricted Boltzmann machine (ClassRBM) models the joint distribution of an input x and target class $y \in \{1,\ldots,l\}$ using a hidden layer of binary stochastic units h. An illustration of the ClassRBM is provided in Fig. 1. In particular, $\theta = \{b,c,d,w,u\}$ are adjustable parameters.

The energy function is defined as follows:

$$E(y,x,h) = -h^T wx - b^T x - c^T h - d^T y - h^T uy \tag{7}$$

From the energy function, we assign probabilities to values of y, x, and h as follows:

$$p(y,x,h) = \frac{\exp(-E(y,x,h))}{Z} \tag{8}$$

where Z is a partition function, which ensures that Eq. 8 is a valid probability distribution.

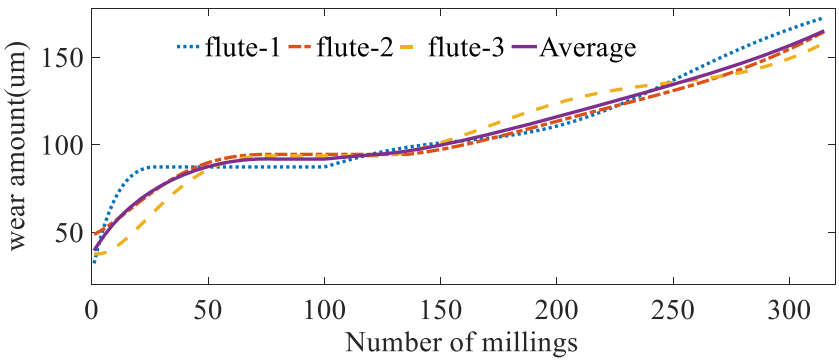

Fig. 4 Tool flank wear of the three flutes in the milling process

Table 4 Number of samples in each category for records c1, c4, and c6

Records	Category		
	Initial	Normal	Severe
c1	34	229	52
c4	85	176	54
c6	24	188	103

However, computing $p(y,x,h)$ or $p(y,x)$ is typically intractable. However, sampling may be performed from ClassRBM using Gibbs sampling (i.e., alternating between sampling a value for the hidden layer, given the current value of the visible layer, and vice versa). All required conditional distributions are simple. The following equation is provided when conditioning on the visible and label layers:

$$p(h_i = 1|y, v) = sigm\left(c_i + \sum_k y_k u_{ik} + \sum_j w_{ji} v_j\right) \tag{9}$$

The following equations are provided when conditioning on the hidden layer:

$$p(v_j = 1|h) = sigm\left(b_j + \sum_i w_{ji} h_i\right) \tag{10}$$

$$p(y|h) = \frac{\exp\left(d_y + \sum_i \mu_{yi} h_i\right)\right)}{\sum_Y \exp(d_Y + \sum_i \mu_{iY} h_i)} \tag{11}$$

$$d_y = \sum_{k=1}^{l} d_k y_k, u_{yi} = \sum_{k=1}^{l} y_k \mu_{ki} \tag{12}$$

We have the following equation when conditioning on the visible layer:

$$p(y|v) = \frac{\exp(-F(y, v))}{\sum_{Y \in \{1,\ldots,y,\ldots,l\}} \exp(-F(y, v))} \tag{13}$$

Free energy $F(y,v)$, which marginalizes the hidden units, is deduced as follows:

$$F(y, v) = -\left(d_y + \sum_{i=1}^{n} softplus\left(c_i + u_{iy} + \sum_{j=1}^{m} w_{ji} v_j\right)\right) \tag{14}$$

$$softplus(x) = \log(1 + \exp(x)) \tag{15}$$

The training objectives of ClassRBM include the generative training, discriminative training, and hybrid training objectives. Similar to Eq. 16, generative training objective defines a value for the joint probability $p(y,v)$ and trains ClassRBM thereafter. Similar to Eq. 17, discriminative training objective defines a value for the conditional probability $p(y|v)$ and trains ClassRBM thereafter.

$$L(D_{\text{train}}; \theta) = -\sum_{(v,y) \in D_{\text{train}}} \log P_c(y, v; \theta) \tag{16}$$

$$L(D_{\text{train}}; \theta) = -\sum_{(v,y) \in D_{\text{train}}} \log P_c(y|v; \theta) \tag{17}$$

H-ClassRBM refers to a ClassRBM, the objective function of which is the hybrid objective function. The hybrid objective function combines the generative and discriminative training objectives, which are provided as follows:

$$\begin{aligned} L(D_{\text{train}}; \theta) = &-(1 + \alpha) \sum_{(v,y) \in D_{\text{train}}} \log P_c(y|v; \theta) \\ &-\alpha \sum_{(v,y) \in D_{\text{train}}} \log P_c(v, y; \theta) \end{aligned} \tag{18}$$

where D_{train} is the training set. The learning problem is to find optimal solutions for parameter θ that minimize the objective function $L(D_{\text{train}};\theta)$.

Compared with DBN, MLP, SVM, etc., H-ClassRBM has stronger nonlinear fitting ability, stronger robustness, lower space-time complexity, clearer mathematical model, and better interpretability, which is suitable for solving supervised learning problems. Therefore, H-ClassRBM is used to recognize tool wear state.

2.3 Proposed methodology

In order to improve the accuracy of tool wear state recognition, tool wear state recognition based on GBDT and H-ClassRBM is developed. The main steps of tool wear state recognition are shown in Fig. 2. The specific steps are as follows:

Step 1: As shown in Table 1, the force and vibration signals in the three axial directions of X, Y, and Z, and tool wear values are collected during the milling process.

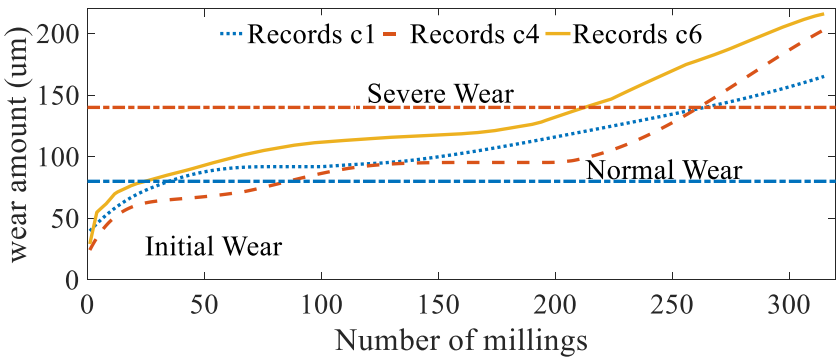

Fig. 5 Tool wear curve

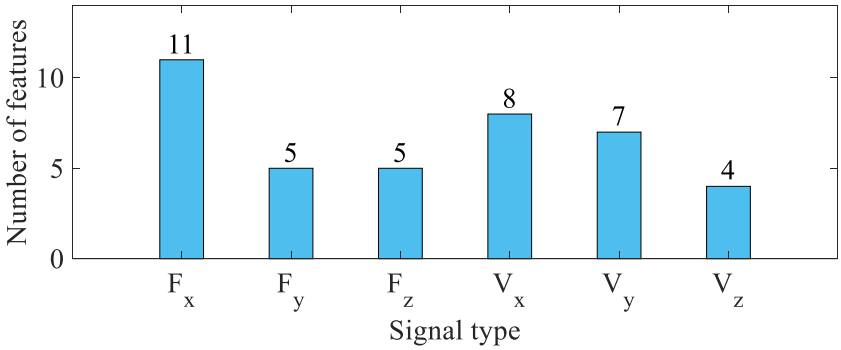

Fig. 6 Number of features selected from each type of signal

Step 2: As shown in Table 2, the time, frequency, and time–frequency domain features of each type of signal are extracted. There are 19 features in time domain, including three newly proposed features and 16 classic features. The features shown in italic in Table 2 are the newly proposed features. There are 3 features in frequency domain. There are 11 features in time-frequency domain, including the energy ratios of eight frequency sub-bands generated by performing Haar wavelet package transform with a three-level decomposition on an original signal, and the mean values of the three intrinsic mode functions obtained after three-level empirical mode decomposition of an original signal. Given that the sampling frequency of an original signal is as high as 50 Hz, the original signal must be resampled before calculating its DTW feature. As shown in Table 1, the six types of signals collected include F_x, F_y, F_z, V_x, V_y, and V_z. Thirty-three features of each type of signal are extracted. Therefore, 198 features (F_1, …, F_{198}) are extracted to form a feature set.

Step 3: All features in the feature set are normalized. Thereafter, GBDT is used to select a feature subset from the feature set. GBDT calculates the importance of each feature. The top few features with a sum of importance above 90% of the sum of the importance of all features are selected to form a feature subset.

Step 4: The feature subset and corresponding tool wear values are used as training set. CD and RMSspectral are used to optimize the H-ClassRBM parameters.

Step 5: The trained H-ClassRBM is used for tool wear state recognition.

Suppose $\{X_1,\ldots,X_m\}$ is a time series. Entropy features include the En-asinh ($H(asinh(x_1,\ldots,x_n))$) and En-atan ($H(atan(x_1,\ldots,x_n))$) features. The entropy feature first performs inverse trigonometric transformation on the original sequence, and solves the entropy of the transformed sequence thereafter. Meanwhile, *asinh*, *atan*, and $H(x_i)$ represent the inverse hyperbolic sine, arctangent, and information entropy functions.

Assume that a data sequence at time t is denoted as x^t and a data sequence at an initial operation time is denoted as x^0. The DTW algorithm is used to calculate the DTW distance between sequences x^t and x^0. The DTW feature of sequence (x^0) is defined as 0. The DTW feature of sequence (x^t) is defined as the DTW distance of sequences x^t and x^0.

Table 5 Selected proposed features

Description of features	Symbol
En-asinh feature of F_x	F_{31}
En-asinh feature of V_z	F_{196}
En-atan feature of F_y	F_{32}
En-atan feature of V_y	F_{164}
En-atan feature of V_z	F_{197}
En-atan feature of V_x	F_{131}
DTW feature of F_x	F_{33}
DTW feature of F_z	F_{99}

3 Experimental verification

3.1 Tool condition monitoring system

The data used in this study are obtained from [35], which are part of the "prognostic data challenge 2010" data set from the Prognostics and Health Management (PHM) Society. The experiment is conducted on a high-speed CNC machine, and a ball-end milling cutter with three grooves are used to mill the workpiece of Inconel 718. The experiment device is shown in Fig. 3. During the experiment, the workpiece bevel is continuously milled to process the complete bevel. The machining parameters are as follows: spindle speed is 10,400 r/min, feed rate in the x-direction is 1555 mm/min, cutting depth in the y-direction (radial) is 0.125 mm, and cutting depth in the z-

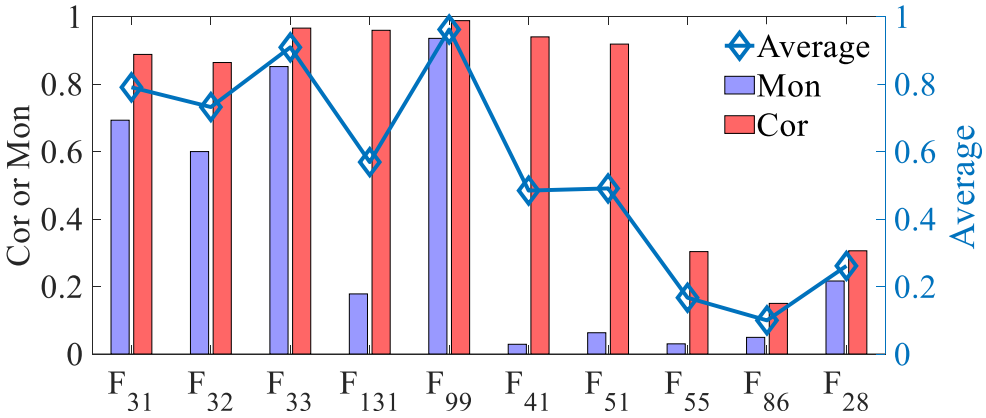

Fig. 7 Metrics comparison of features

direction (axial) is 0.2 mm. These values are shown in Table 3. During the milling process, the Kistler quartz three-component dynamometer is used to measure the cutting force, three Kistler piezoelectric accelerometers are used to measure the machine vibration signal, and the Kistler acoustic emission (AE) sensor is used to measure the acoustic emission signal generated during the cutting process. Seven signal channels (i.e., force_x, force_y, force_z, vibration_x, vibration_y, vibration_z, AE_RMS) are acquired using an NI PCI1200 data acquisition device with a sampling frequency of 50 kHz. According to ISO3685-1977, the tool wear is the flank wear width VB at a depth of half of the cutting depth. After a surface milling is performed, the flank wear of the three grooves is measured offline using a LEICA MZ12 microscope.

The tool wear data set includes six individual cutter records (i.e., c1, …, c6). Each cutter record corresponds to 315 millings of one milling cutter. The cutting tool is a 3-flute ball nose cutter. The wear value of each milling includes the wear values of the three grooves (i.e., flute_1, flute_2, and flute_3). However, the single wear value of flute_1, flute_2, or flute_3 has its limitation. Hence, the average of the wear values for flute_1, flute_2, and flute_3 are calculated as the final wear value of the cutting tool. Cutter records (records c1, c4, and c6) with real wear values are used. The corresponding wear value of the three flutes is shown in Fig. 4.

As shown in Fig. 5, the tool wear state is divided into three categories according to the change in tool wear value: initial, normal, and severe wear [36, 37]. The wear amounts of the initial, severe, and normal wear are (0, 80 μm), (80 μm, 140 μm), and (140 μm, ∞), respectively. Table 4 shows the number of samples in each category for each record.

3.2 Feature selection

A total of 198 statistical features (F_1, …, F_{198}) are extracted to form a feature set. GBDT is used for feature selection. The number of trees is set to 10, the maximum depth is set to 3, and the learning rate is set to 0.1. GBDT is used to calculate the importance of each feature. The top 40 features with the sum of importance above 90% of the sum of the importance of all features are selected.

As shown in Fig. 6, there are more features related to milling force signals than features related to vibration signals. Milling force signals can substantially characterize tool wear. The number of features in the X-axis is significantly greater than the number of features in the Y- and Z-axes. The X-axis signals can considerably characterize tool wear.

Table 5 shows the proposed features among the features selected. To gain deeper insight into the proposed features, the Mon and Cor of these features are shown in Fig. 7. The Mon and Cor metrics are normalized to the interval (0, 1).

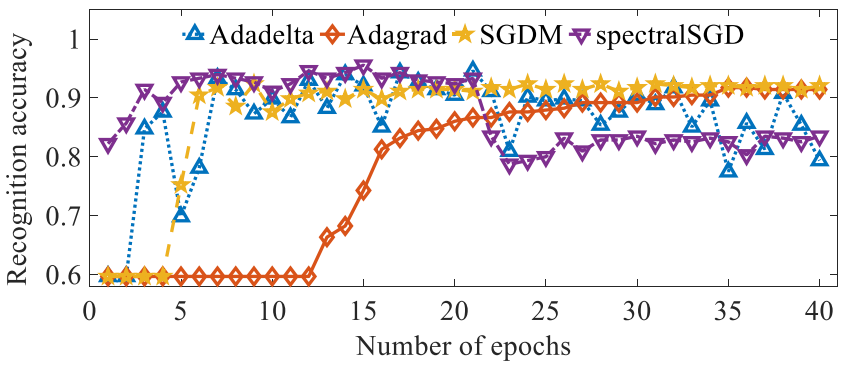

Fig. 8 Comparison of the recognition accuracy of the different algorithms

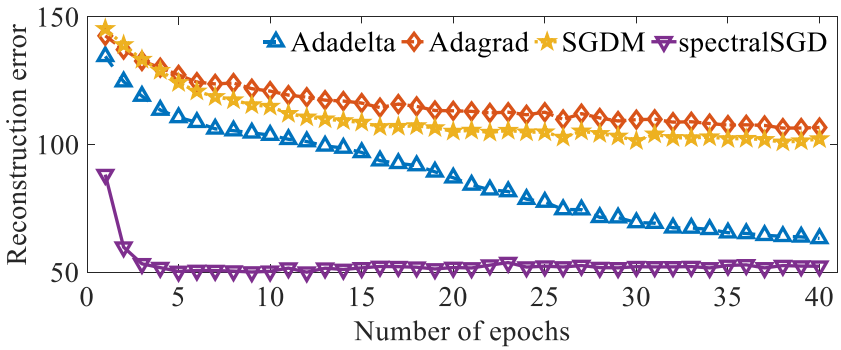

Fig. 9 Comparison of the reconstruction error of the different algorithms

F_{41}, F_{51}, F_{55}, F_{86}, and F_{28} are classical features. Although F_{131} has low Mon values, it has high Cor values as well. With the exception of F_{131}, the proposed features have high Mon and Cor values. Compared with the classic feature, the proposed features have high average values. Therefore, the proposed features can substantially characterize tool wear.

3.3 Comparison of different learning algorithms for H-ClassRBM

The reconstruction error can accurately represent the log-likelihood of RBM to the training data, and is defined by taking $\|v-v^g\|^2$, where v are original samples and v^g are the samples obtained after Gibbs sampling. The recognition accuracy on the test set accurately reflects the classification ability and generalization ability of H-ClassRBM. Therefore, we use reconstruction error and recognition accuracy as an index to evaluate the trained H-ClassRBM.

Two main classes of approaches make the learning of H-ClassRBM considerably efficient. CD is a classical algorithm that belongs to the first class. The idea of CD-k is to approximate the second term in the log-likelihood gradient by a sample for a k-step Gibbs distribution, in which k is set to 1. The second category includes SGDM, Adagrad, Adadelta, and RMSspectral, among others. For SGDM, the learning rate η is set to 0.1. The momentum term γ is set to 0.9. For Adagrad, the learning rate η is set to 0.01. The smoothing term ε is set to 1e−8. For Adadelta, the learning rate η is set to 0.01. The smoothing term ε is set to 1e−8. The momentum term γ is set to 0.9. For RMSspectral, the smoothing term ε is set to 1e−8. The learning rate η is set to 0.01. The decay parameter α is set to 0.9. SGDM, Adagrad, Adadelta, and RMSspectral are compared.

Table 6 Description of the trained model datasets

Trained model label	Training datasets		Testing datasets
T1	Records c1	Records c4	Records c6
T2	Records c1	Records c6	Records c4
T3	Records c4	Records c6	Records c1

As shown in Fig. 8, the recognition accuracy of each learning algorithm increases with increasing number of iteration. RMSspectral converges to 0.94 at the epoch of 7. The Adadelta converges to 0.93 at the epoch of 7. SGDM converges to 0.92 at the epoch of 7. Adagrad converges to 0.92 at the epoch of 36. Adagrad has the slowest convergence speed and the lowest stable recognition accuracy. The learning rate of Adagrad is extremely low at 0.01. Adagrad's accumulation of the squared gradients in the denominator causes the learning rate to decrease and eventually become infinitesimally low, at which point the algorithm is no longer able to acquire additional knowledge. The convergence speed and stable recognition accuracy of SGDM and Adadelta are the same. The learning rate of SGDM is 0.1. Adadelta is an extension of Adagrad and seeks to reduce its aggressive, monotonically decreasing learning rate. Although the learning rate of RMSspectral is also 0.01, it has the fastest convergence speed and highest stable recognition accuracy. RMSspectral combines a non-Euclidean gradient method (stochastic spectral descent) with a preconditioned gradient descent method (RMSprop). RMSprop is identical to Adadelta. RMSspectral updates parameters in the normed space operates (l_∞ and *Shztten*_∞ norm). As the number of iterations increases, the recognition accuracy of RMSprop rapidly increases and fluctuates, decreases rapidly thereafter, and continues fluctuating. After 21 epochs, RMSprop falls into overfitting. Therefore, the recognition accuracy of RMSprop decreases rapidly, and continues fluctuating.

As shown in Fig. 9, the reconstruction error of each learning algorithm decreases with increasing number of iterations. The reconstruction error is negatively correlated with the recognition accuracy. RMSspectral has the fastest convergence and the smallest stationary reconstruction error. Furthermore, H-ClassRBM trained with RMSspectral has the least training time and highest accuracy. Therefore, RMSspectral and CD are used to train H-ClassRBM.

3.4 Comparison of different recognition algorithms

Cross-validation is a type of statistical analysis tool and is widely utilized to evaluate the performance of classification

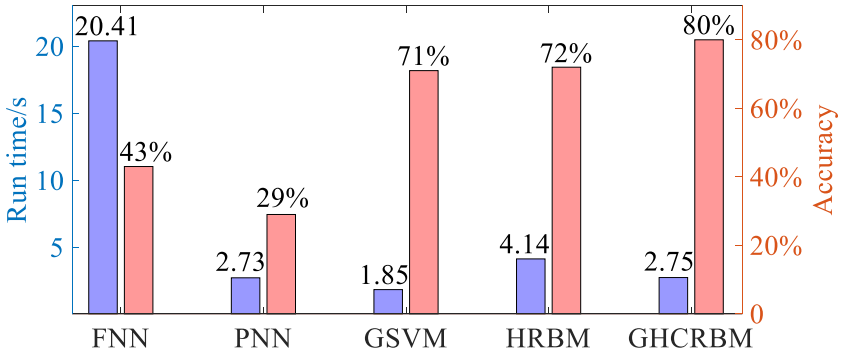

Fig. 10 Running time and accuracy of the recognition algorithms

accuracy. Three-fold cross-validation is applied to evaluate the performance of FNN, probabilistic neural network (PNN), Gaussian kernel support vector machine (GSVM), and H-ClassRBM. Table 6 details the data set used. For H-ClassRBM, the numbers of neurons in the visible, hidden, and label layers are 40, 30, and 3, respectively. The stopping strategy is early-stop strategy. The batch size comprises 30 samples. The training algorithms are CD and RMSspectral. The bias terms are initialized with zero values, and connection weights are generated randomly from a Gaussian distribution. Accuracy and running time are selected as evaluation indicators. The results are shown in Fig. 11.

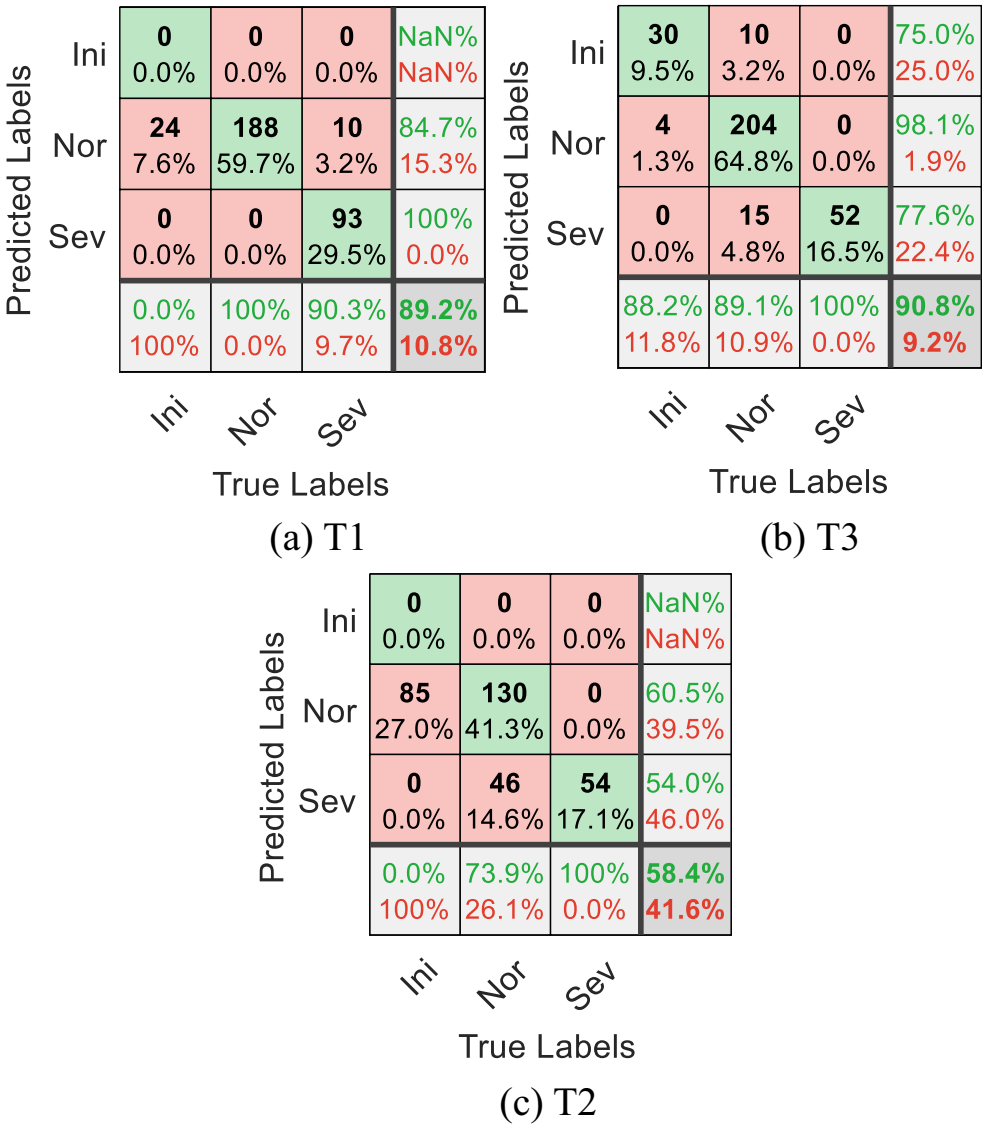

Fig. 11 Confusion matrices of **a** T1, **c** T2, and **b** T3

As shown in Fig. 10, HRBM refers to H-ClassRBM. GHCRBM refers to the combination of GBDT and H-ClassRBM. The combination of GBDT and H-ClassRBM has the highest accuracy. On the one hand, H-ClassRBM simultaneously models the joint distribution of the input and target and the relationships between the input variables. H-ClassRBM uses the generative training objective to regularize the discriminative training objective. Compared with FNN, PNN, and GSVM, H-ClassRBM has more powerful capabilities of feature extraction and classification for solving supervised learning problems. On the other hand, GBDT selects features that are highly correlated with tool wear from the original feature set, reduces overfitting, reduces the input dimension of the model, and enhances the generalization ability of the model. GBDT also reduces the number of features and the parameters of H-ClassRBM. Therefore, the running time of the combination of GBDT and H-ClassRBM is less than that of FNN and H-ClassRBM. Compared with PNN and GSVM, the combination of GBDT and H-ClassRBM has higher computational complexity and more running time. In summary, the combination of GBDT and H-ClassRBM slightly increases the running time by 0.9 s and improves the accuracy by 9% compared with GSVM.

As shown in Fig. 11, T3 has the highest accuracy, which is 90%, whereas T2 and T1 have lower accuracies of 50% and 70%, respectively. For T2 and T1, the accuracy of the initial wear is 0, thereby reducing the overall accuracy. On the one hand, the training data sets of T1 and T2 have serious class imbalance problems. On the other hand, the change ranges of the wear amount of c1, c4, and c6 are (39 μm, 165 μm), (24 μm, 203 μm), and (29 μm, 216 μm), respectively. The change range of the wear amount of c1 is smaller than that of c4 and c6. Therefore, the accuracy of the initial wear of T1 and T2 is the lowest (i.e., 0). T1 and T2 did not predict any samples as initial wear. Instead, all samples, the true category of which is initial wear, are predicted to be normal wear. Therefore, the recall of the initial wear of T1 and T2 is not a number (NaN). This prediction tends to be conservative and facilitates timely tool changes.

4 Conclusion

In this study, an automated tool wear condition recognition scheme for the milling process is developed using cutting force and vibration sensors. The PHM challenge 2010 data set is used to validate the proposed method. The major conclusions of this study are as follows.

(1) Compared with the classical time, frequency, and time-frequency domain features, the proposed features including a dynamic time warping feature and two entropy features have good monotonicity and correlation, which can well characterize tool wear.

(2) Compared with H-ClassRBM trained by CD and Adadelta, CD and Adagrad, CD and SGDM, the H-ClassRBM trained by CD and RMSspectral improves recognition accuracy by 1%, 2%, and 2%, respectively. The use of CD and RMSspectral to train H-ClassRBM can improve its generalization performance.

(3) Compared with H-ClassRBM, the combination of GBDT and H-ClassRBM improves recognition accuracy by 8%. Therefore, GBDT greatly reduces the input dimension of the H-ClassRBM model and improves the recognition accuracy of tool wear state.

(4) Compared with FNN, PNN, GSVM, and H-ClassRBM, the combination of GBDT and H-ClassRBM improves recognition accuracy by 37%, 51%, 9%, and 8%, respectively. Therefore, the proposed tool wear state recognition scheme based on GBDT and H-ClassRBM is beneficial in improving the recognition accuracy of tool wear state, and provides an effective guide for decision-making in the machining process.

Funding information This work is supported by the State Key Science & Technology Program of China (Grant No. 2019ZX04012-001); National Natural Science Foundation of China (51905209); Industrial Technology Research and Development Project of Jilin Province Development and Reform Commission, China (2019C040-2); Young and Middle-aged Scientific and Technological Innovation leaders and Team Projects in Jilin Province, China (20190101015JH); and Program for JLU Science and Technology Innovative Research Team (JLUSTIRT).

References

1. Zhu KP, Zhang Y (2019) A generic tool wear model and its application to force modeling and wear monitoring in high speed milling. Mech Syst Signal Process 115:147–161. https://doi.org/10.1016/j.ymssp.2018.05.045
2. Cao XC, Chen BQ, Yao B, Zhuang SQ (2019) An intelligent milling tool wear monitoring methodology based on convolutional neural network with derived wavelet frames coefficient. Appl Sci-Basel 9(18):26. https://doi.org/10.3390/app9183912
3. Chi YJ, Dai W, Lu ZY, Wang MQ, Zhao Y (2018) Real-time estimation for cutting tool wear based on modal analysis of monitored signals. Appl Sci-Basel 8(5):13. https://doi.org/10.3390/app8050708
4. Martinez-Arellano G, Terrazas G, Ratchev S (2019) Tool wear classification using time series imaging and deep learning. Int J Adv Manuf Technol 104(9-12):3647–3662. https://doi.org/10.1007/s00170-019-04090-6
5. Xie ZY, Li JG, Lu Y (2019) Feature selection and a method to improve the performance of tool condition monitoring. Int J Adv Manuf Technol 100(9-12):3197–3206. https://doi.org/10.1007/s00170-018-2926-5
6. Liao XP, Zhou G, Zhang ZK, Lu J, Ma JY (2019) Tool wear state recognition based on GWO-SVM with feature selection of genetic algorithm. Int J Adv Manuf Technol 104(1-4):1051–1063. https://doi.org/10.1007/s00170-019-03906-9
7. Wang GF, Xie QL, Zhang YC (2017) Tool condition monitoring system based on support vector machine and differential evolution optimization. Proc Inst Mech Eng B J Eng Manuf 231(5):805–813. https://doi.org/10.1177/0954405415619871

8. Benkedjouh T, Medjaher K, Zerhouni N, Rechak S (2015) Health assessment and life prediction of cutting tools based on support vector regression. J Intell Manuf 26(2):213–223. https://doi.org/10.1007/s10845-013-0774-6
9. Kong DD, Chen YJ, Li N (2018) Gaussian process regression for tool wear prediction. Mech Syst Signal Process 104:556–574. https://doi.org/10.1016/j.ymssp.2017.11.021
10. Hui-bin S, Wei-long N, Jun-yang W (2015) Tool wear feature extraction based on Hilbert-Huang transformation. J Vibr Shock 34(4):158–164. https://doi.org/10.13465/j.cnki.jvs.2015.04.027
11. Liu TI, Jolley B (2015) Tool condition monitoring (TCM) using neural networks. Int J Adv Manuf Technol 78(9-12):1999–2007. https://doi.org/10.1007/s00170-014-6738-y
12. Guo L, Li NP, Jia F, Lei YG, Lin J (2017) A recurrent neural network based health indicator for remaining useful life prediction of bearings. Neurocomputing 240:98–109. https://doi.org/10.1016/j.neucom.2017.02.045
13. Que ZJ, Xu ZG (2019) A data-driven health prognostics approach for steam turbines based on XGBoost and DTW. IEEE Access 7:93131–93138. https://doi.org/10.1109/access.2019.2927488
14. He ZY, Cheng WD, Wen WG (2020) A similarity comparison method of homologous fault response fragments under variable rotational speed. Shock Vib 2020:12–12. https://doi.org/10.1155/2020/8973678
15. Javed K, Gouriveau R, Zerhouni N, Nectoux P (2015) Enabling health monitoring approach based on vibration data for accurate prognostics. IEEE Trans Ind Electron 62(1):647–656. https://doi.org/10.1109/tie.2014.2327917
16. Zhang B, Zhang LJ, Xu JW (2016) Degradation feature selection for remaining useful life prediction of rolling element bearings. Qual Reliab Eng Int 32(2):547–554. https://doi.org/10.1002/qre.1771
17. Kong DD, Chen YJ, Li N (2017) Force-based tool wear estimation for milling process using Gaussian mixture hidden Markov models. Int J Adv Manuf Technol 92(5-8):2853–2865. https://doi.org/10.1007/s00170-017-0367-1
18. Rao H, Shi XZ, Rodrigue AK, Feng JJ, Xia YC, Elhoseny M, Yuan XH, Gu LC (2019) Feature selection based on artificial bee colony and gradient boosting decision tree. Appl Soft Comput 74:634–642. https://doi.org/10.1016/j.asoc.2018.10.036
19. Zhang Z, Li YB, Jin SS, Zhang ZY, Wang H, Qi L, Zhou RL (2018) Modulation signal recognition based on information entropy and ensemble learning. Entropy 20(3):18. https://doi.org/10.3390/e20030198
20. Twardowski P, Wiciak-Pikula M (2019) Prediction of tool wear using artificial neural networks during turning of hardened steel. Materials 12(19):15. https://doi.org/10.3390/ma12193091
21. Khajavi MN, Nasernia E, Rostaghi M (2016) Milling tool wear diagnosis by feed motor current signal using an artificial neural network. J Mech Sci Technol 30(11):4869–4875. https://doi.org/10.1007/s12206-016-1005-9
22. Jegorowa A, Gorski J, Kurek J, Kruk M (2019) Initial study on the use of support vector machine (SVM) in tool condition monitoring in chipboard drilling. Eur J Wood Wood Prod 77(5):957–959. https://doi.org/10.1007/s00107-019-01428-5
23. Ma M, Sun C, Chen XF, Zhang XW, Yan RQ (2019) A deep coupled network for health state assessment of cutting tools based on fusion of multisensory signals. IEEE Trans Ind Inform 15(12):6415–6424. https://doi.org/10.1109/tii.2019.2912428
24. Larochelle H, Mandel M, Pascanu R, Bengio Y (2012) Learning algorithms for the classification restricted Boltzmann machine. J Mach Learn Res 13:643–669
25. Luo LK, Zhang SF, Wang YD, Peng H (2018) An alternate method between generative objective and discriminative objective in training classification restricted Boltzmann machine. Knowl Based Syst 144:144–152. https://doi.org/10.1016/j.knosys.2017.12.032
26. Yin J, Lv JC, Sang YS, Guo JX (2018) Classification model of restricted Boltzmann machine based on reconstruction error. Neural Comput Applic 29(11):1171–1186. https://doi.org/10.1007/s00521-016-2628-6
27. Niu QM, Liu F, Tong QB, Cao JC, Zhang YH (2018) Health condition assessment of ball bearings using TOSELM. J Vibroeng 20(1):272–282. https://doi.org/10.21595/jve.2017.18978
28. Upadhya V, Sastry PS (2019) An overview of restricted Boltzmann machines. J Indian Inst Sci 99(2):225–236. https://doi.org/10.1007/s41745-019-0102-z
29. Chaudhuri A (2019) The minimization of empirical risk through stochastic gradient descent with momentum algorithms. In: Silhavy R (ed) Artificial intelligence methods in intelligent algorithms, Advances in Intelligent Systems and Computing, vol 985. Springer International Publishing Ag, Cham, pp 168–181. https://doi.org/10.1007/978-3-030-19810-7_17
30. Ruder S (2016) An overview of gradient descent optimization algorithms. arXiv
31. Zeiler MD (2012) ADADELTA: an adaptive learning rate method. Comput Sci
32. Dauphin YN, de Vries H, Bengio Y (2015) Equilibrated adaptive learning rates for non-convex optimization. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems 28, vol 28. Advances in Neural Information Processing Systems. Neural Information Processing Systems (Nips), La Jolla,
33. Carlson DE, Collins E, Hsieh YP, Carin L, Cevher V (2015) Preconditioned spectral descent for deep learning. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems 28, vol 28. Advances in Neural Information Processing Systems. Neural Information Processing Systems (Nips), La Jolla,
34. Carlson D, Hsieh YP, Collins E, Carin L, Cevher V (2016) Stochastic spectral descent for discrete graphical models. IEEE J Sel Top Signal Process 10(2):296–311. https://doi.org/10.1109/jstsp.2015.2505684
35. Li X, Lim BS, Zhou JH, Huang S, Phua SJ, Shaw KC, Er MJ (2009) Fuzzy neural network modelling for tool wear estimation in dry milling operation. In: Annual Conference of the Prognostics and Health Management Society, PHM 2009, September 27, 2009 - October 1, 2009, San Diego, CA, United States. Annual Conference of the Prognostics and Health Management Society, PHM 2009. Prognostics and Health Management Society, p et al.; General Atomics; IMPACT; Palo Alto Research Center (PARC); Ridgetop Group Inc.; Xerox
36. Zhu KP, Mei T, Ye DS (2015) Online condition monitoring in micromilling: a force waveform shape analysis approach. IEEE Trans Ind Electron 62(6):3806–3813. https://doi.org/10.1109/tie.2015.2392713
37. Lei Z, Zhou YQ, Sun BT, Sun WF (2020) An intrinsic timescale decomposition-based kernel extreme learning machine method to detect tool wear conditions in the milling process. Int J Adv Manuf Technol 106(3-4):1203–1212. https://doi.org/10.1007/s00170-019-04689-9