Gated Recurrent Units Based Neural Network For Tool Condition Monitoring

Huan Xu[1], Chong Zhang[2], *Student Member, IEEE*, Geok Soon Hong[1], Junhong Zhou[3], *Member, IEEE*, Jihoon Hong[3], *Member, IEEE*, Keng Soon Woon[1]
[1] Department of Mechanical Engineering, [2] Department of Electrical and Computer Engineering
National University of Singapore, Singapore
[3] Manufacturing Execution and Control Group
Singapore Institute of Manufacturing Technology (SIMTech), A*STAR, Singapore
Emails: xuhuan@u.nus.edu, zhangchong@u.nus.edu, mpehgs@nus.edu.sg

***Abstract*—Tool condition monitoring (TCM) is a prerequisite to ensure high finishing quality of workpiece in manufacturing automation. One of the most important components in TCM system is tool wear estimation. How to achieve estimation with high accuracy is still an open question. In the past few decades, recurrent neural network (RNN) has shown a great success in learning long-term dependence of the sequential data. However, traditional RNNs (e.g., vanilla RNN, etc.) suffer gradient vanishing or exploding problem as well as long computational training time when the model is trained through back propagation through time (BPTT). To address these issues, we propose a gated recurrent units (GRU) based neural network to estimate the tool wear for tool condition monitoring. The GRU neural network can analyze time-series data on multiple time scales and can avoid gradient vanishing during training. A real-world gun drilling experimental dataset is used as a case study for tool condition monitoring in this paper. The performance of the proposed GRU based TCM approach is compared with other well-known models including support vector regression (SVR) and multi-layer perceptron (MLP). The experimental results show that the proposed GRU based TCM approach outperforms other competing models on this real-world gun drilling dataset.**

I. Introduction

Tool wear includes the gradual failure, chipping, broken of cutting tools due to regular operation. The tool wear increases as the cutting proceeds, which directly reduces the tool residual life. Subsequently, the friction between the worn tool and the workpiece increases, so as to the power consumption and vibration in the system. Eventually, the worn tool produces a rough surface of workpiece, distortions in dimension [1]. In numerous manufacturing processes, late replacement of worn tools may lead to unexpected machine downtimes and influence the product quality, which subsequently result in high costs. Early replacement of defective tools causes less efficient use of the tools. Tool condition monitoring (TCM) system is essential to monitor the tool condition based on direct (e.g., optical) or indirect sensing methods. TCM improves the quality and efficiency of manufacturing processes. Therefore, TCM systems have been widely used to pursue many industrial demands such as high-level competitiveness, high process productivity and standardized quality of the produced parts. Accurate tool wear estimation is a key enabler of TCM. In this paper, we mainly focus on the problem of tool wear estimation in TCM.

In the past years, researches in tool condition monitoring (TCM) [2]–[10] are mainly done by analyzing the features of force, torque, vibration and acoustic emission (AE) signals. Konig et al. [3] and Jemielniak et al. [2] used the Root Mean Square (RMS) of AE signal to detect the catastrophic tool failure. El-wardany et al. [4] showed the discriminant features of vibration signal were robust and effective in tool condition monitoring. Li et al. [11] modeled the functional equation of tool wear rate and feed cutting force estimated by feed motor current based on neuro-fuzzy techniques. As with the improving performance of machining learning techniques and neural networks in data analyzing, neural networks methods are also widely applied in tool wear diagnostics and estimation. Lin and Ting [5] acquired trust force and torque signal, and they found that neural network had better performance than other traditional regression algorithms. Issam et al. [6] utilized vibration signal features and multi-layer perceptron (MLP) model to classify tool wear state defined by the tool wear rate. Sun et al. [12] applied support vector machine (SVM) to do multi-classification to detect the tool flank wear state using features of AE signal. Various sensor signals were used in the past research, but there were few research on the comparison between the effectiveness of different signals and models in tool condition monitoring.

Although many variants of neural networks [13], [14] and other machine learning techniques can have a good performance on classification problems for tool condition diagnostics, signals are considered as independent discrete data points in these application. Zhang et al. [7], [13], [15] proposed a time window process to learn features from a sliding time window rather than a single data point. In order to effectively learn high-level features and optimize the ensemble model, Zhang et al. [15] proposed the multi-objective deep belief networks ensemble (MODBNE) to achieve more accurate and generalized predictions. Consider the multi-state nature of TCM, Zhang et al. [8] proposed a multi-state diagnosis and prognosis (MDP) framework with feature learning for TCM. However, the tool condition state depends on the previous tool state because tool wear condition grows to be worse

978-1-5090-6014-6/18/$31.00 ©2018 IEEE

and can never become better than the tool wear condition at previous time. Thus, other methods like Hidden Markov Model (HMM) utilizing the dependence on previous state have a better behavior. Fish et al. [16] made use of the dynamic characteristics of HMM to develop a multiclass prediction in milling process. Ozgur et al. [17] introduced multivariate coupled HMM to allow for both short- and long-term dynamics of AE signal. Omid et al. [18] applied multimodal HMM to improve the performance of traditional physically segmented HMM. HMM are basically recursively taking previous state into consideration when predicting current state, so it can obtain better result when dealing with sequential signal. In previous research, the machine learning techniques mainly used in tool wear monitoring are SVM, MLP and HMM. Besides HMM, recurrent neural network (RNN) can also utilize the dynamic characteristics in time sequence. Few research studies used RNN in tool condition monitoring (TCM). To the best of our knowledge, only Zhao et al. [19] applied bidirectional gated recurrent units (GRU) in machine health monitoring.

In the past few decades, RNN has shown a great success on sequential temporal data [**?**], [20]–[22]. RNN introduced a directed cycle to form connections between units with present and past data which allows to exhibit dynamic temporal behavior. However, RNN often suffers exploding and vanishing gradient problems, which results in failing to learn long-term dependencies. To avoid these problems, Hochreiter and Schmidhuber [**?**] proposed long short-term memory (LSTM) by introducing gating approach. Typical LSTM consists of an input gate, an output gate and a forget gate. However, the complex cell structure of LSTM makes it require longer time to train. To overcome this problem, Cho et al. [20] proposed a gated recurrent units (GRU) which only has two gates, i.e. update and reset. The simpler cell structure and fewer parameters make GRU to converge faster. Some empirical studies [21], [22] have shown that GRU has smaller computational time while performing comparably performance than LSTM.

Based on aforementioned literature review, a GRU based TCM approach is proposed in this paper to achieve accurate tool wear estimation based on temporal characteristics of historical sensory data. In this paper, we estimate the flank tool wear by using force signals. And we also compare the performance of support vector regression (SVR), MLP and GRU. The force signal is found to be good for tool wear estimation and it has good generalization ability in tool wear estimation in new cutting condition. And the experimental results show that GRU is better than SVR and MLP in tool wear regression problem.

The rest of the paper is organized as follows. Section II introduces the background of gated recurrent units based neural network. Section III presents the proposed GRU based TCM approach. Section IV shows the experimental results and discussions. Section V concludes this study and reveals the limitations of this paper.

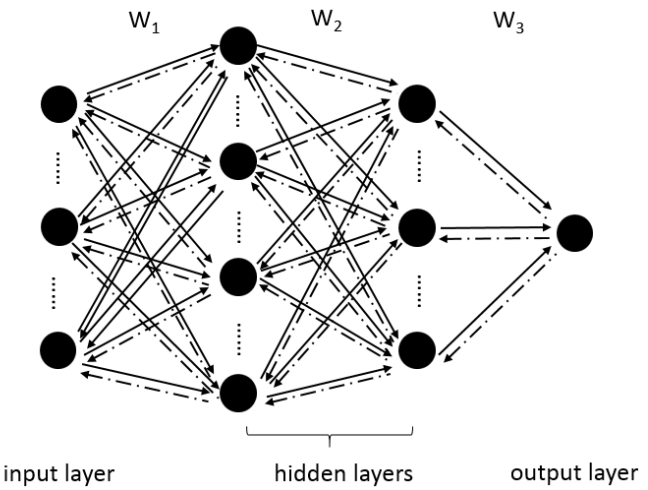


Fig. 1. The structure of MLP used in this paper.

II. Background

In this paper, the performances of MLP and GRU based neural networks are compared on tool condition monitoring problem. The brief introductions of these two techniques are subsequently presented in the following subsection.

A. Multi-Layer Perceptron (MLP)

MLP is multi-layer perceptron network which can learn the nonlinear mapping relationships between inputs and outputs. The output of each hidden layer is considered as the input of the next hidden layer. We use MLP with 2 hidden layers as an example. The structure is shown in Fig. 1. Suppose the dimension of input data is n. The solid line denotes the feedforward algorithm to map the feature vector $\mathbf{x}(t) = (x_1, x_2, ..., x_n)$ at time t to the desired output $y(t)$. The output at j^{th} hidden layer $h_j(t)$ is defined as Eq. 1, and the activation function used in MLP is rectified linear unit (ReLU) function which is described in Eq. 2.

$$h_j(t) = relu(W_i h_{j-1}(t) + b) \quad (1)$$

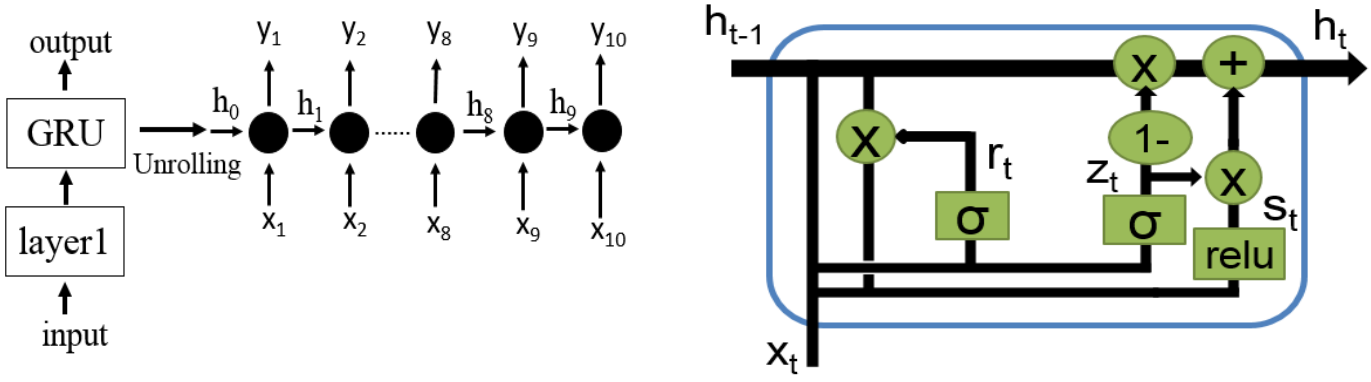
$$relu(x) = \begin{cases} x & \text{x>0} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where W_i is the weights at i^{th} iteration, b denotes the bias, $h_{j-1}(t)$ represents the output of $(j-1)^{th}$ hidden layer.

The dotted line represents the backpropagation updating algorithm [23]. It updates the parameter of networks based on the gradients of loss function. In this paper, we use first order gradient-based optimization algorithm, i.e. Adam [24], to optimize the weights W_1, W_2, W_3 and bias in MLP. The loss function E is defined as follows.

$$E(\theta) = \frac{1}{N}\Sigma(\hat{y} - y)^2 + \frac{1}{2}\lambda W W^T \quad (3)$$

where λ is the generalization parameter, y is the true output and $\hat{y}$ is the predicted output.



(a) (b)

Fig. 2. (a) Network structure with one fully connected layer and one GRU layer. (b) GRU cell structure.

B. Gated Recurrent Units (GRU) based Neural Network

Rather than traditional RNNs such as vanilla RNN, the state-of-the-art gated recurrent units (GRU) based neural network is applied to TCM for tool wear estimation. To solve the gradient exploding and or vanishing problems of vanilla RNN, LSTM was proposed [25] to learn long-term dependencies. In contrast with LSTM, GRU cell has a slight change to make cell structure simpler and faster to train. The network structure of GRU based neural network is shown as Fig. 2 (a). The first hidden layer is a fully connected layer with the activation function of $f(x) = tanh(x)$ to scale the input into the range of $[-1, 1]$. The second GRU hidden layer can be unrolled like the structure on the right in Fig. 2 (a).

The structure of GRU cell is shown in Fig. 2 (b). The forward algorithm is as follows, where U_z, U_r, U_h, W_z, W_r, W_h, W_Y are weights, h is the hidden state, r is the reset gate that decides the ratio of the memory of previous information, x is the input, subscript t is the current time step, $t-1$ is the previous time step. The hidden state h_t is estimated based on the weighted previous state h_{t-1} and current information s_t by the ratio z:

$$z = \sigma(x_t \mathbf{U_z} + h_{t-1}\mathbf{W_z}) \quad (4)$$

$$r = \sigma(x_t \mathbf{U_r} + h_{t-1}\mathbf{W_r}) \quad (5)$$

$$s_t = relu(x_t \mathbf{U_h} + (h_{t-1} * r)\mathbf{W_h}) \quad (6)$$

$$h_t = (1-z) * h_{t-1} + z * s_t \quad (7)$$

$$y_t = \mathbf{W_Y} h_t \quad (8)$$

III. GRU BASED TCM APPROACH

The proposed GRU based TCM approach includes four parts: 1) feature extraction; 2) feature selection; 3) GRU regression modeling; 4) evaluation. The flowchart of the proposed approach is shown in Fig. 3. Firstly, time-domain and frequency-domain features are extracted from the sensory data. Then the normalized features are selected via recursive feature elimination with cross validation (RFECV). Thirdly, the selected features are used as the inputs of the GRU based neural network for tool wear estimation. Finally, the regression

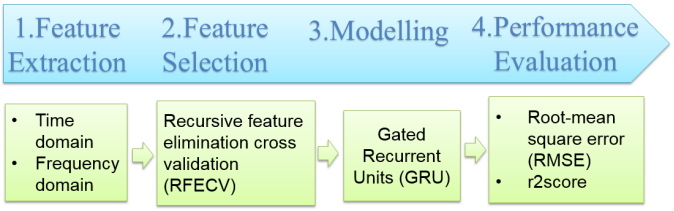


Fig. 3. The flowchart of the proposed GRU based TCM framework.

performance is evaluated with two evaluation metrics (i.e. RMSE, r2score). The following subsections explain each part in details. A real-world gun drilling experimental dataset is used as a case study.

A. Feature Extraction

Force signals are sampled at a high frequency, so the data size will be a huge number. To save computational time in industry automation systems, it is necessary to segment the raw data and extract features from each segmentation to form a feature vector $\mathbf{x}(\tau) = (x_1, x_2, ..., x_n)$. In this process, raw signal can provide the meaningful information directly and the influence of raw signal's randomness on the estimation performance can be minimized [26]. The extracted features and their equation are summarized in Table I. The time-series signals are segmented using a sliding window and the data in each segmentation will be represented by a new extracted feature vector. Since the gun drilling is a periodic rotational process, we use the number of sampling points per revolution as the window size defined as Eq. 9, and the sliding width is the half of the window size.

$$Window = f_s \frac{60}{SP} \quad (9)$$

where f_s is the sampling frequency (Hz) and SP is the spindle speed (RPM).

Force signals have three channels, each channel represents one direction of the cutting force. In this paper, there are mainly features adopted in [27], [28] and [29]. The force features represent the amplitude and variance of force signal

TABLE I
DETAILS OF EXTRACTED FEATURES FOR FORCE SENSOR SIGNALS.

Features	Force features	Equation
1	Standard deviation	$x_\sigma = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(x_j - x_\mu)^2}$
2	Peak	$Peak = \max_{j=1,\cdots,N} x_j$
3	Root Mean Square (RMS)	$x_{rms} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(x_j)^2}$
4	Mean	$x_\mu = \frac{1}{N}\sum_{j=1}^{N} x_j$
5	Dynamic component	$\hat{x} = \frac{1}{N}\sum_{j=1}^{N}(x_j - x_\mu)$
6	Skewness	$x_{sk} = \frac{\sum_{j=1}^{N}(x_j - x_\mu)^3}{(N-1)\sigma^3}$
7	Kurtosis	$x_{ku} = \frac{\sum_{j=1}^{N}(x_j - x_\mu)^4}{(N-1)\sigma^4}$
8	Variable	$x_v = x_\mu - median(x_{1,\cdots,N})$
9	First order differential	$\frac{\partial x}{\partial t} = x_\mu\|_t - x_\mu\|_{t-1}$
10	Second order differential	$\frac{\partial^2 x}{\partial t^2} = \frac{\partial x}{\partial t}\|_t - \frac{\partial x}{\partial t}\|_{t-1}$

during the drilling process. Totally, there are 10 features of each channel and detailed force features are shown in Table. I.

B. Feature Selection

After features extraction, there are many features representing the useful information from the raw signals. However, redundant features not only increase analysis complexity and computational time, but also bring noise into training process which may worsen the generalization ability of the model, so meaningful features have to be selected before used in the model. Therefore, selecting the informative feature subsets is an important step to get an accurate estimation. Before feature selection, all the features are normalized into the same range of $[-1,1]$ by Eq. 10. In this way, all the features contributes equally to the output.

$$Norm(x_i) = \frac{x_i - \bar{x}}{x_i^{max} - x_i^{min}} \tag{10}$$

where x_i denotes the i^{th} feature, $\bar{x}$ is the averaged value of all the features.

In this paper, recursive feature elimination with cross validation (RFECV) method is used for feature selection. The flowchart of RFECV is shown in Fig. 4. There are two steps and the details of each step of REFCV are presented as follows:

step 1: Determine the number of features to select. First, all features are employed to train an estimation model and scored their performance with cross validation. Then recursively eliminate the unimportant features until the dimension of each remaining feature vector is 1. After that, the dimension of selected features is determined to be the dimension of the feature vector that has achieved the best average cross-validation score. The importance of each feature is represented by its corresponding weight obtained in the estimation model.

step 2: Select features. With the determined dimension of features to be selected, we train the estimation model SVR and eliminate the unimportant features. Then put the remaining features into the estimation model.

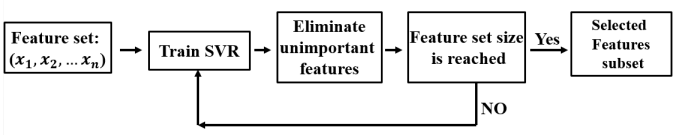


Fig. 4. The flowchart of RFECV for feature selection.

Repeat the elimination process until the dimension of the remaining feature vector is achieve.

IV. EXPERIMENTAL RESULTS

A. Dataset

In this paper, a real-world gun drilling experimental dataset is used as a case study for tool condition monitoring. The gun drilling experiment is conducted on a DMU 80P 5-axis Machining Center. The workpiece is in round shape with the length of 200mm and diameter of 110mm. The workpiece material is Inconel alloy 718. The type of gun drilling tools is N8 with the length of 350mm. The experiment is done under the same condition with the spindle speed of 1000 rpm and the feed rate of 16 mm/min.

Force signals are acquired during the process. The experimental setup is shown in Fig. 5. The dynamometer (model KISTLER 9171A) is mounted on the machine spindle, which is used to measure the force signals in three directions. Raw force signals are sampled by data acquisition system (model NI cDAQ9178) at a sampling frequency of 25600Hz.

The drilling process stops after drilling for every 10mm, then the tool is pulled out to record tool wear images and measure the tool wear using Keyence microscope. After recording images, the tool is installed back to continue drilling without resharpening. The tool wear is measured 20 times for each tool life. The signals are acquired using LABVIEW software.

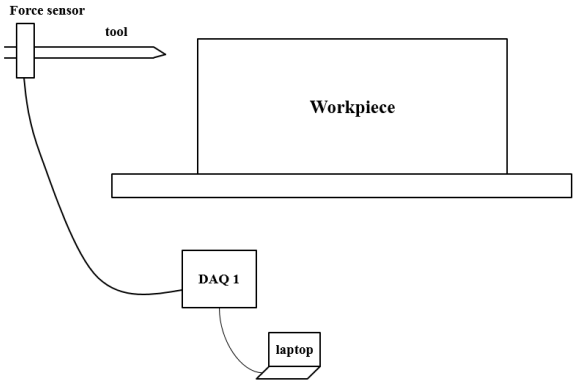


Fig. 5. Experimental setup of the gun drilling. One dynamometer is mounted on the machine spindle. The dynamometer is connected to data acquisition system (DAQ). The signal outputs of DAQ are sampled using the LABVIEW program in the laptop.

Fig. 6 shows the raw force signals acquired during drilling process. $force_x$ is radial force. $force_y$ is thrust force. And $force_z$ is tangential force.

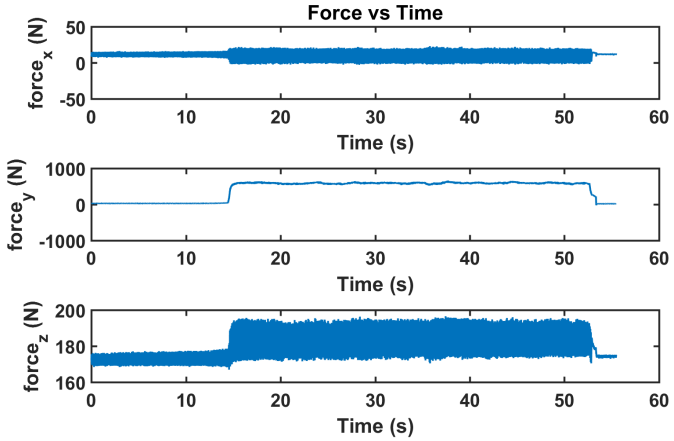


Fig. 6. Illustration of three raw force signals. $force_x$, $force_y$, $force_z$ represent the radial force, thrust force and tangential force, respectively. The beginning part of each signal represents the force signal when the machine starts and the spindle is rotating. Once starting to drill the workpiece, the amplitude of force signals increase dramatically. The middle part of each signal represents the force signal during drilling process. After machining is finished, the spindle is still rotating. The force signal looks like that at the end part.

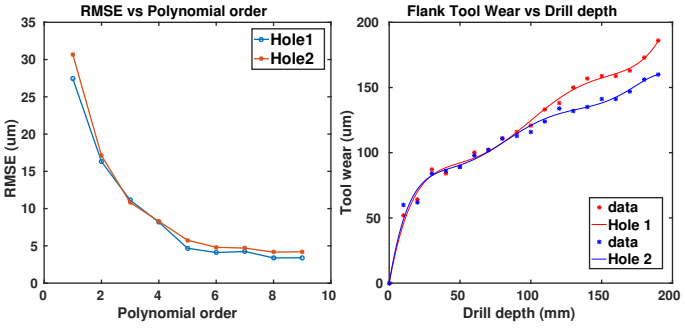
TABLE II
MEASURED FLANK TOOL WEAR

Depth (mm)	Hole1 (um)	Hole2 (um)
0	0	0
10	52	60
20	64	62
30	87	84
40	84	86
50	89	89
60	100	98
70	102	102
80	111	111
90	116	113
100	121	116
110	133	124
120	138	134
130	150	132
140	157	135
150	159	141
160	159	141
170	163	147
180	173	156
190	186	160

B. Analysis of Tool Wear Fitting

To estimate flank tool wear continuously during gun drilling process, GRU based neural network is implemented with the selected features. In the above-mentioned experiment, the flank tool wear is measured 20 times for each tool life. The measured flank tool wear is shown in the Table II. Two holes are drilled locating 46mm away from the center of the workpiece's flat surface. As can be seen in the table, even thought the two holes are drilled with the same feed rate and spindle speed, the flank progressive tool wear rate is different. One of the reasons is that the workpiece material is inhomogeneous.

Polynomial fitting method is used to fit the tool wear curve. To choose the polynomial fitting order, different fitting order from 1 to 10 have been tried and the root mean square error (RMSE) of the fitting results are calculated. The relation of



(a) (b)

Fig. 7. (a) The relation between polynomial orders and their corresponding RMSE values. The left plot shows that RMSE of polynomial fitting result decreasing with polynomial order. (b) The relation between flank tool wear and drill depth. The right plot shows the polynomial fitting result with the order of 6.

RMSE versus polynomial order is shown in Fig. 7(a). Based on the observation from Fig. 7(a), it is obvious that the order of 6 is the change point. When the order is more than 6, the change rate of RMSE becomes small, which is considered as overfitting. Thus, the 6^{th} order polynomial fitting is chosen as the option to fit the flank tool wear curve, as shown in Fig. 7(b). From the fitting result in Fig. 7(b), the experimental tool life includes two periods: initial tool wear stage, moderate tool wear stage. The fitting curve is consistent with the characteristic pattern of the tool wear curve as discussed in [30] and [31].

C. Analysis of Feature Selection

Support vector regression (SVR) [32] is adopted as the estimation model in feature selection with RFECV [33] method. In the step 1, the different average r2scores of RFECV with different numnber of features are obtained as shown in Fig. 8. It can be seen from the Fig. 8, when the number of remaining features is 17, the discarded features would not affect r2score accuracy too much.

Once the number of features selected is determined, eliminate the unimportant features recursively as described in step 2. After recursive feature elimination (RFE) process, the selected features for force signals are:

$$\begin{aligned} \mathbf{X} &= (X_1, X_3, X4, X_5, X_6) \\ \mathbf{Y} &= (Y_1, Y_2, Y_3, Y_4, Y_5, Y_7, Y_8) \\ \mathbf{Z} &= (Z_1, Z_4, Z_5, Z_6, Z_7) \end{aligned}$$

where X denotes the features of radial force, Y denotes the features trust force, Z denotes the features of tangential force, the subscript number is the corresponding feature in Table I.

Feature selection reduces the dimensionality of feature space. To test the result of feature selection, we train SVR with all and selected features from Hole1 and test the model on the features of Hole2. The results are shown in Table III. From the results, we can see that the test performance of model trained in selected features is better than that of model trained on all the features. We also apply MLP to estimate the tool

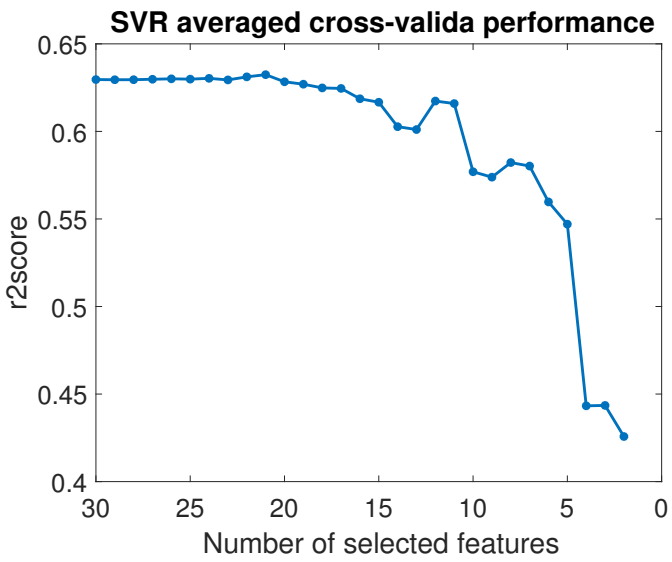


Fig. 8. Determination of the number of force features to be selected based on the average r2score via RFECV.

TABLE III
COMPARISON BETWEEN THE PERFORMANCES OF ALL FEATURES AND THE SELECTED FEATURES BASED ON RFECV WITH FORCE SIGNALS IN TERMS OF RMSE AND R2SCORE.

Model		Force	
		All Features	Selected Features
RFECV	RMSE	22.777	**19.224**
	r2score	0.630	**0.735**

wear. Fig. 9 shows the training loss of MLP decreases over iterations when using all and selected features. Although when the train process has converged, the train loss of MLP using all features is smaller than those of selected features, its test accuracy of a new hole is worse. Based on the aforementioned observation, the training process of using all the features is probably overfitted. This proves that feature selection process eliminates the noisy features and new feature subset keeps the most meaningful information of the original data.

D. Evaluations

The assessment criteria used in this paper are root mean square error (RMSE) and r2score. RMSE is defined as Eq. (11) and the r2score is defined as Eq. 12.

$$RMSE = \sqrt{\frac{1}{N}\sum(y_{pred} - y_{true})^2} \tag{11}$$

$$r2score = 1 - \frac{\sum(y_{pred} - y_{true})^2}{\sum(y_{pred} - \bar{y}_{true})^2} \tag{12}$$

The implementation details of GRU are described below. The number of hidden neuron is ranged in $[10, 20]$. The total number of layers of GRU is 4. Training parameters include learning rate $\gamma = 0.0001$, time steps $t = 10$, batch size of 40, the iteration number of 4000. All the experimental results are

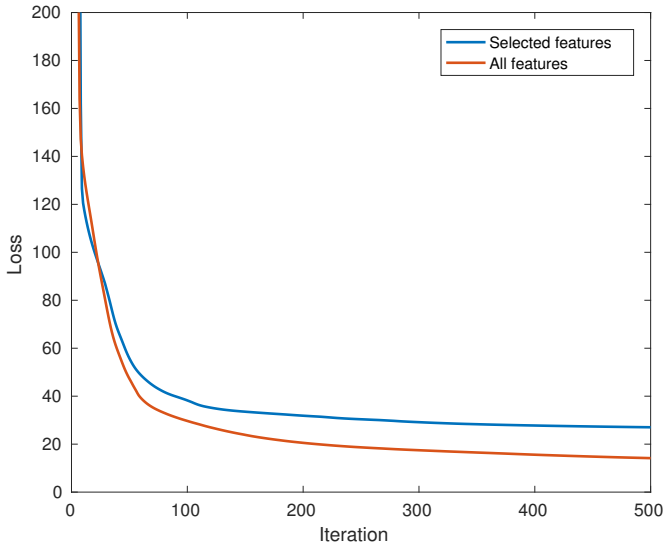


Fig. 9. The comparison between the training losses of MLP model with all features and selected features extracted from force signals.

TABLE IV
COMPARISON BETWEEN THE PERFORMANCE OF DIFFERENT MODELS (I.E. SVR, MLP AND GRU) WITH FORCE SIGNALS IN TERMS OF RMSE AND R2SCORE.

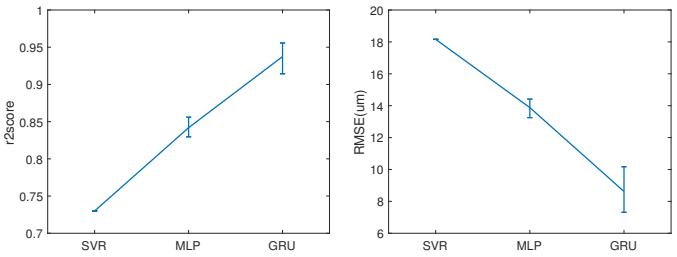
Model	Evaluation score		Force
SVR	RMSE	Train	14.372
		Test	18.159
	r2score	Train	0.870
		Test	0.730
MLP	RMSE	Train	6.272 ± 0.381
		Test	13.880 ± 0.535
	r2score	Train	0.975 ± 0.004
		Test	0.842 ± 0.014
GRU	RMSE	Train	3.952 ± 1.076
		Test	$\mathbf{8.622 \pm 1.550}$
	r2score	Train	0.990 ± 0.006
		Test	$\mathbf{0.937 \pm 0.019}$

trained based on dataset from Hole1 and tested on dataset from Hole2. The total number of training data is 22499 and the total number of test data is 23749 for force features set.

E. Comparison of Different Models

SVR, MLP and GRU models are trained on the training dataset of Hole1 and tested on the dataset of Hole2. We repeat the process for 5 times and average the assessment score. The training and testing performance of all models is shown in Table IV. The average evaluation performance is shown in Fig. 10.

From the Table IV, it can be observed that: By comparing the three different models (i.e. SVR, MLP, GRU), GRU has provided better performance than other competing models in this regression problem with temporal sequence data. This is because SVR and MLP treat the features as discrete points while GRU keeps the memory of previous states when estimates the current states.



(a) (b)

Fig. 10. (a) Comparison of the performance between SVR, MLP and GRU of force signals in terms of r2score. (b) Comparison of the performance between SVR, MLP and GRU of force signals in terms of RMSE.

V. Conclusion

In this paper, a GRU based TCM approach has been proposed for accurate tool wear estimation. A real-world gun drilling experimental dataset is used as a case study to evaluate the performance of the proposed approach. The proposed approach has been compared with SVR and MLP on this dataset. The experimental results have shown that the proposed approach outperformed other competing methods in terms of RMSE and r2score. Based on the experimental results, the effectiveness and generalization ability of force signals and GRU based TCM approach in progressive tool wear estimation is proven to be practical. The limitations of this study include the limited experimental conditions and small data size. In order to achieve a better generalization ability on new cutting conditions, further research needs to be done such as on other varying cutting conditions. In summary, GRU based TCM approach shows better performance in continuous tool condition monitoring system.

References

[1] N. Ambhore, D. Kamble, S. Chinchanikar, and V. Wayal, "Tool condition monitoring system: A review," *Materials Today: Proceedings*, vol. 2, no. 4-5, pp. 3419–3428, 2015.

[2] K. Jemielniak and O. Otman, "Tool failure detection based on analysis of acoustic emission signals," *J. Mater. Process. Tech.*, vol. 76, no. 1-3, pp. 192–197, 1998.

[3] W. König, K. Kutzner, and U. Schehl, "Tool monitoring of small drills with acoustic emission," *Inter. J. Machine Tools Manuf.*, vol. 32, no. 4, pp. 487–493, 1992.

[4] T. I. El-Wardany, D. Gao, and M. A. Elbestawi, "Tool condition monitoring in drilling using vibration signature analysis," *Inter. J. Machine Tools Manuf.*, vol. 36, no. 6, pp. 687–711, 1996.

[5] S. Lin and C. Ting, "Drill wear monitoring using neural networks," *Inter. J. Machine Tools Manuf.*, vol. 36, no. 4, pp. 465–475, 1996.

[6] I. Abu-Mahfouz, "Drilling wear detection and classification using vibration signals and artificial neural network," *Inter. J. Machine Tools Manuf.*, vol. 43, no. 7, pp. 707–720, 2003.

[7] C. Zhang, G. S. Hong, H. Xu, K. C. Tan, J. H. Zhou, H. L. Chan, and H. Li, "A data-driven prognostics framework for tool remaining useful life estimation in tool condition monitoring," in *2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp. 1–8, Sept 2017.

[8] C. Zhang, G. S. Hong, J. H. Zhou, K. C. Tan, H. Li, H. Xu, J. Hong, and H. L. Chan, "A multi-state diagnosis and prognosis framework with feature learning for tool condition monitoring," *IEEE T. Cybernetics, submitted*, 2017.

[9] C. Zhang, K. C. Tan, H. Li, and G. S. Hong, "A cost-sensitive deep belief network for imbalanced classification," *IEEE T. Neur. Net. Lear.*, 2017.

[10] J. Hong, J. H. Zhou, H. L. Chan, C. Zhang, H. Xu, and G. S. Hong, "Tool condition monitoring in deep hole gun drilling: A data-driven approach," in *Industrial Engineering and Engineering Management (IEEM), 2017 IEEE International Conference on*, pp. 2148–2152, IEEE, 2017.

[11] X. Li, A. Djordjevich, and P. K. Venuvinod, "Current-sensor-based feed cutting force intelligent estimation and tool wear condition monitoring," *IEEE T. Ind. Electron.*, vol. 47, no. 3, pp. 697–702, 2000.

[12] J. Sun, M. Rahman, Y. S. Wong, and G. S. Hong, "Multiclassification of tool wear with support vector machine by manufacturing loss consideration," *Inter. J. Machine Tools Manuf.*, vol. 44, no. 11, pp. 1179–1187, 2004.

[13] C. Zhang, J. H. Sun, and K. C. Tan, "Deep belief networks ensemble with multi-objective optimization for failure diagnosis," in *Syst. Man Cybernetics (SMC), 2015 IEEE Int. Conf.*, pp. 32–37, Oct 2015.

[14] C. Zhang, K. C. Tan, and R. Ren, "Training cost-sensitive deep belief networks on imbalance data problems," in *Neural Networks (IJCNN), 2016 International Joint Conference on*, pp. 4362–4367, IEEE, 2016.

[15] C. Zhang, P. Lim, A. Qin, and K. C. Tan, "Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2306–2318, 2017.

[16] R. K. Fish, M. Ostendorf, G. D. Bernard, and D. A. Castanon, "Multilevel classification of milling tool wear with confidence estimation," *IEEE T. Pattern Anal.*, vol. 25, no. 1, pp. 75–85, 2003.

[17] Ö. Çetin, M. Ostendorf, and G. D. Bernard, "Multirate coupled hidden Markov models and their application to machining tool-wear classification," *IEEE J. Sign. Proc.*, vol. 55, no. 6 II, pp. 2885–2896, 2007.

[18] O. Geramifard, J.-X. Xu, J.-H. Zhou, and X. Li, "Multimodal hidden markov model-based approach for tool wear monitoring," *IEEE T. Ind. Electr.*, vol. 61, no. 6, pp. 2900–2911, 2014.

[19] F. Yang, M. S. Habibullah, T. Zhang, Z. Xu, P. Lim, and S. Nadarajan, "Health index-based prognostics for remaining useful life predictions in electrical machines," *IEEE T. Ind. Electron.*, vol. 63, no. 4, pp. 2633–2644, 2016.

[20] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[21] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 2342–2350, 2015.

[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *Int. Conf. Mach. Learn.*, pp. 2067–2075, 2015.

[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] D. A. Dornfeld and M. F. DeVries, "Neural Network Sensor Fusion for Tool Condition Monitoring," *CIRP Annals - Manufacturing Technology*, vol. 39, no. 1, pp. 101–105, 1990.

[27] K. Zhu, Y. S. Wong, and G. S. Hong, "Multi-category micro-milling tool wear monitoring with continuous hidden Markov models," *Mech. Syst. Signal Process.*, vol. 23, no. 2, pp. 547–560, 2009.

[28] T. Liu and K. Anantharaman, "Intelligent classification and measurement of drill wear," *Journal of engineering for industry*, vol. 116, no. 3, pp. 392–397, 1994.

[29] J. H. Zhou, C. K. Pang, F. L. Lewis, and Z. W. Zhong, "Intelligent Diagnosis and Prognosis of Tool Wear Using Dominant Feature Identification," *IEEE T. Ind. Inform.*, vol. 5, no. 4, pp. 454–464, 2009.

[30] a. Thangaraj and P. K. Wright, "Computer-assisted Prediction of Drill-failure Using In-process Measurements of Thrust Force," *J. Eng. Ind.*, vol. 110, no. 2, p. 192, 1988.

[31] H. M. Ertunc and C. Oysu, "Drill wear monitoring using cutting force signals," *Mechatronics*, vol. 14, no. 5, pp. 533–548, 2004.

[32] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[33] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.