MECHANICAL ENGINEERING SCIENCE

ISSN:2661-4448(online)
2661-443X(print)
Volume 4 No.2 2022

VISER
www.viserdata.com

COMPANY INTRODUCTION

Viser Technology Pte. Ltd. was founded in Singapore with branch offices in both Hebei and Chongqing, China. Viser focuses on publishing scientific and technological journals and books that promote the exchange of scientific and technological findings among the research community and around the globe. Despite being a young company, Viser is actively connecting with well-known universities, research institutes, and indexation database, and has already established a stable collaborative relationship with them. We also have a group of experienced editors and publishing experts who are dedicated to publishing high-quality journal and book contents. We offer the scholars various academic journals covering a variety of subjects and we are committed to reducing the hassles of scholarly publishing. To achieve this goal, we provide scholars with an all-in-one platform that offers solutions to every publishing process that a scholar needs to go through in order to show their latest finding to the world.

Mechanical Engineering Science

Publisher: Viser Technology Pte. Ltd.

ISSN: 2661-4448(online)
2661-443X(print)

Frequency: Semi-annual

Add.: 21 Woodlands Close, #08-18
Primz Bizhub SINGAPORE (737854)

https://www.viserdata.com/

Mechanical Engineering Science

Volume 4 No.2 (2022)

CONTENTS

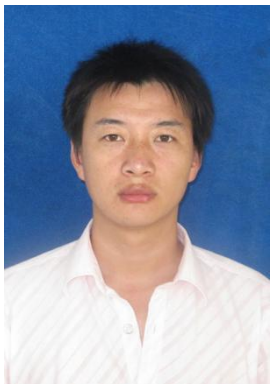
VISER

Special issue message

With the development of advanced equipment manufacturing and the promotion of informationization and industrialization integration strategy, the traditional equipment fault diagnosis technology is difficult to meet the demand. The application of fault prediction and health management (PHM) technology can provide support for the intelligent improvement of industrial equipment condition monitoring and fault diagnosis. With the development of deep learning theory, its prediction effect is significantly higher than that of machine learning technology, and it has been widely used in condition monitoring and fault diagnosis in recent years. However, due to the defects in the network structure, it is still challenging to fully reveal the effective features present in the monitoring signal. Therefore, this journal makes comprehensive use of machine vision, data-driven, machine learning, deep learning and other methods, and combines expert experience and knowledge to transform sample characteristics into valuable information, so as to complete condition monitoring and fault diagnosis.

To reflect the latest research results of condition monitoring and fault diagnosis in time and provide opportunities for academic exchange, we specially organized a special issue of condition monitoring and fault diagnosis. Six papers will be published in this issue, discussing the advantages and disadvantages of machine learning algorithms and deep learning theory, reconstructing and optimizing the algorithm, so as to improve the efficiency and accuracy of the condition monitoring and fault diagnosis model. Experimental verification and reliability analysis are carried out in the aspects of tool wear state monitoring, brake device life prediction, machine tool spindle fault diagnosis and gear reliability analysis. The method proposed in this journal can be widely used in various fields, which lays a theoretical foundation and scientific basis for improving the development of intelligent operation and health management in equipment manufacturing industry, conforms to the development trend of intelligent control and network interaction in the future, and has certain practical significance.

Planner Author



Zhenliang YU received his bachelor's degree and master's degree from Liaoning University of Technology in 2007 and 2011 respectively, and his PhD degree from Northeastern University in 2020. From 2011 to 2014, I worked in the Safety Equipment Research Institute of Shanxi Fenxi Electromechanical Co., LTD. (China Shipbuilding Industry Corporation) as the project leader. Design and develop a number of projects and equipment, won many awards. During my PhD, I participated in the National Natural Science Foundation project "Research on Efficient static/dynamic reliability Analysis and Reliability Optimization Design Method of Complex mechanical structures" and was responsible for the research on efficient static and dynamic reliability analysis method of mechanical structures. Technical basic scientific research project "Reliability simulation analysis technology of high-speed flexible mechanism of weapon equipment", responsible for reliability analysis method research; Thousands of domestic CNC lathe reliability improvement project, "high-grade CNC machine tools and basic manufacturing equipment" science and technology major special project, responsible for CNC machine tool quality reliability improvement; The national Natural Science Foundation of China Key project "Research on Dynamic and Gradual Reliability Robust Design Theory of key parts of machinery" is responsible for the research on dynamic reliability method of key parts of machinery.

The research proposes a reliability analysis method based on PC-Kriging model and Isomap-Clustering update strategy. Firstly, in order to determine the optimal basis function of the Kriging model, a truncated set of sparse polynomials is used as the candidate set of the optimal basis function. Minimum Angle regression (LARS) was used to calculate the number of possible polynomial basis function sets and rank the basis function candidates, and Akaike Information Criterion (AIC) was used to determine the optimal polynomial form. Secondly, the dimension reduction method of Isomap and the k-means Clustering analysis algorithm are used in the new Isomap-clustering strategy to determine the representative points among hundreds or even thousands of candidate points. The Isomap-Clustering strategy can update the PC-Kriging model with several representative points near the limit state in each iteration. For implicit functional functions or time-consuming finite element computations, a new method is proposed to determine the optimal MC candidate sample pool size. Firstly, the confidence interval of the actual failure probability is estimated according to the relative error of the failure probability at the confidence level of 0.95. That is, based on the lower limit of the confidence interval, a new method is proposed to estimate the optimal number of MC samples. Subsequently, based on the upper limit of the confidence interval, an adaptive sampling region strategy similar to radial centralized sampling is

proposed to concentrate the candidate sampling points in the important high probability density region. Secondly, the k-means ++ clustering technology and the learning function LIF were used to complete the adaptive experimental design to realize parallel computing, that is, k sampling points were added in each iteration to update the PC-Kriging model, so that the accuracy of the limit state could be improved in different regions at the same time. In order to solve the reliability problem with small failure probability and long simulation time, a new reliability analysis method based on PC-Kriging model and radial concentrated importance sampling strategy (RCA-PCK) is proposed. Because the objective function needs to be calculated a lot of time in the optimization process, the research proposed a reliability analysis optimization method based on PC-Kriging model and PSO optimization algorithm (PCK-PSO). He is currently a lecturer at the School of Mechanical and Power Engineering, Yingkou Institute of Technology. He has published more than 20 papers, including more than 10 SCI and EI papers, and is the reviewer of many journal papers, such as Structures (JCR Area 1). Reliability Engineering & System Safety (JCR Zone 1, Top); Applied Mathematical Modelling (JCR Area 1, Top), etc.

He is currently a lecturer in the School of Mechanical and Power Engineering of Yingkou Institute of Technology, and has been awarded the "Double teacher and Double Energy Type" teacher for three consecutive years from 2021 to 2023. He has presided over 2 projects at the provincial level and above, 1 project at the municipal level, and 3 projects at the university level. His current research interests include mechanical reliability analysis, intelligent operation and maintenance, and life prediction.

About the Author

Shuo WANG received the M.Sc. Degree from College of Mechanical Engineering, Shenyang University of Science and Technology,in 2014. After graduation, he engaged in the research of automobile engine related technologies and obtained the engineer title in 2018. During his work, he participated in 5 major technology research projects and 3 general projects of the enterprise. At the end of 2020, he began to teach in Yingkou Institute of Technology, mainly teaching machine tool numerical control technology and metal cutting principles and tools. In 2021, he was awarded as a "double-qualified and double-capable" teacher. In 2022, he won the "Special Contribution Award" of Yingkou Institute of Technology and the title of lecturer, and in 2023, he won the certificate of International Career Coach，Guide students to participate in 1 provincial college student innovation and entrepreneurship project and 2 university level projects. During the university, he participated in 1 project of the provincial Department of Education, 1 project of the provincial Department of Science and Technology, 1 project of the doctoral double innovation project, presided over 1 project of the university level scientific research, 1 project of the university level educational reform, participated in 2 projects of the university level scientific research, and published more than 10 papers. His current research interests include data mining, deep learning and intelligent instrumentation, and he has made a series of achievements in the field of condition monitoring and fault diagnosis.

Guangchen Xu is an assistant professor in Yingkou Institute of Technology. He obtained his master degree from the Zhejiang University of Technology in 2010. He has successfully completed one research project supported by the National Natural Science Foundation of Liao ning and is currently working on another. He has published over 20 papers, including more than 10 papers indexed by SCI and EI. His primary research interests lie in the areas of machinery dynamics, machine design, and machine tool vibration analysis.

Peng LIU graduated from Northeastern University, Yingkou Institute of Technology, metalworking training teacher, Yingkou Institute of Technology TRIZ innovative method studio leader, TRIZ innovative method first level tutor. Mainly engaged in practical teaching work, good at mechanical processing equipment maintenance and debugging. Responsible for the selection of schools in TRIZ Innovation Method Competition, led the studio members to win the national second prize for many times, and won the title of excellent instructor. He has also won the provincial first prize for many times in other discipline competitions.

About the Author

Yongqi Guo is a lecturer in YingKou Institute of Technology. He obtained his Master degree from Jilin University in 2016. From Jul 2017-Sep 2020, he worked at Brilliance Auto Research Institute, mainly engaged in lightweight design of automotive structures and suspension dynamics analysis. His current primary research interests lie in the areas of automobile lightweight, including composite structural optimization design, Multi-objective optimization.

Jingbo WANG, Master, Professor. He is one of the first batch of famous teachers and a "double qualified and double capable" teacher in Yingkou Institute of Technology. Currently, he is an expert of science and technology project evaluation in Liaoning Province, a member of the Professional Title Evaluation Committee of Yingkou, a special expert of safety production in Yingkou and an expert of innovation and Entrepreneurship project evaluation in Yingkou Entrepreneurship Valley. He has published more than 20 papers such as SCI, EI and Chinese cores., presided over or participated in more than 10 in Liaoning Province and Yingkou Municipal Education and Scientific Research Projects, obtained more than 20 national, provincial and municipal innovation and entrepreneurship awards, and has 7 national vocational skills certificates, and is the Editor-in-chief or deputy editor-in-chief in 5 national teaching materials, and has obtained 6 national patents and 1 copyright.

Fengqin ZHAO received a Doctor's degree in engineering from the College of Agricultural Engineering, Shenyang Agricultural University in 2002. He is now a professor and master supervisor in the School of Mechanical and Power Engineering, Yingkou Institute of Technology. He is a famous undergraduate teaching teacher of ordinary colleges and universities in Liaoning Province. He has supported 1 project of National Natural Science Foundation, presided over and participated in more than 10 projects at provincial level and above. Her current research interests include mechanical control mechanism and system research, machine tool fault diagnosis.

Viser Technology Pte. Ltd.
Mechanical Engineering Science
DOI: 10.33142/mes.v4i2.9082

Research Article Open Access

A multi-source information fusion method for tool life prediction based on CNN-SVM

Shuo WANG, Zhenliang YU*, Peng LIU, Man Tong WANG

School of Mechanical and Power Engineering, Yingkou Institute of Technology , Yingkou, China

***Corresponding Author:** Zhenliang YU, email address:yuzhenliang_neu@163.com*

Abstract:
For milling tool life prediction and health management, accurate extraction and dimensionality reduction of its tool wear features are the key to reduce prediction errors. In this paper, we adopt multi-source information fusion technology to extract and fuse the features of cutting vibration signal, cutting force signal and acoustic emission signal in time domain, frequency domain and time-frequency domain, and downscale the sample features by Pearson correlation coefficient to construct a sample data set; then we propose a tool life prediction model based on CNN-SVM optimized by genetic algorithm (GA), which uses CNN convolutional neural network as the feature learner and SVM support vector machine as the trainer for regression prediction. The results show that the improved model in this paper can effectively predict the tool life with better generalization ability, faster network fitting, and 99.85% prediction accuracy. And compared with the BP model, CNN model, SVM model and CNN-SVM model, the performance of the coefficient of determination R2 metric improved by 4.88%, 2.96%, 2.53% and 1.34%, respectively.
Keywords: *CNN-SVM; tool wear; life prediction; multi-source information fusion*

1 Introduction

CNC machining center is a set of high-tech, high precision, high efficiency in one of the high precision end equipment, specifically for processing complex curved parts, its key technology to improve the level of equipment manufacturing industry is of great significance. CNC machining center due to the complexity of the processing object tool wear more serious, when the tool wear exceeds a given threshold will greatly affect the accuracy of the workpiece processing, resulting in the processing of product quality is not up to standard, not only waste processing input time and economic losses, and even lead to machine accidents[1] . For complex curved parts with high precision machining requirements, how to make the tool wear before the critical threshold for intelligent tool change will be an important research direction for the future high-end manufacturing industry.

Currently, data-driven methods combining sensor monitoring data with machine learning algorithms are widely used for tool life prediction [2]. Monitoring data refers to the extraction of tool wear features in the time domain, frequency domain, and time-frequency domain using sensor technology to collect raw signals. However, most experts and scholars predict tool wear for only one signal[3] , which often has a low prediction accuracy, so in this paper, three signals, cutting vibration signal, cutting force signal and acoustic emission signal, are collected in

real time, and a multi-source information fusion strategy is adopted to fuse the features extracted from each signal and construct a sample feature matrix, so as to improve the accuracy of tool life prediction. Machine learning algorithm is to use the extracted tool wear features as the input of the model, simulate the whole process of tool life degradation, and compare the current working state with the historical data to complete the prediction of the remaining tool life [4]. Common machine learning algorithms include BP neural network, RBF neural network, Support vector machine (SVM) etc.

Wei Weihua [5] et al. optimized the BP neural network by genetic algorithm, which improved the optimization and learning ability of the model and ensured the efficiency and accuracy of tool wear recognition. Weiqing Cao[6] et al. diagnosed the tool wear fault by fusing the information of RBF neural network and D-S evidence theory, and the experiment showed that the model could effectively diagnose the tool wear fault and its prediction accuracy was improved. The model can effectively diagnose the tool wear fault and its prediction accuracy is improved. Zhang Kun[7] et al. constructed DCM-SVR model to predict the tool wear value of

Received on October 23, 2022; Accepted on December 14, 2022

machining process, which can correct the systematic error online and compensate the predicted value, and the results of comparison with other methods show that the prediction performance of DCM-SVM is improved by 28.7% and the root mean square error is decreased by 64.7%. Although the traditional tool life prediction methods have achieved certain results, the prediction model can only tap the shallow features of the sample data, the generalization ability is insufficient, the network fitting speed is slow, and it relies more on signal processing techniques and expert experience.

In 2006, Hinton [8] et al. proposed the theory of deep learning, and convolutional neural network (CNN) is a typical representative of deep learning. Convolutional neural networks (CNNs) have powerful feature extraction ability, can adaptively mine the deep features of the input data, and get rid of the model's over-reliance on signal processing techniques and expert experience, so they have been widely used and researched by scholars in recent years. P. K. Ambadekar [9] et al. established a tool life prediction system using CNN convolutional neural networks, where an inverted microscope regularly takes The images of the tool were used as input and different categories of tool wear were used as output, and the results showed that the accuracy of the prediction using CNN model reached 87.26%, which can meet the practical needs of production. However, the output layer of the convolutional neural network (CNN) generally consists of a fully-connected layer and a Softmax layer. When dealing with data with a high degree of nonlinearity, the number of features in the output of the fully-connected layer increases proportionally, which can cause the overfitting phenomenon[10] ; moreover, the prediction performance of the Softmax layer is not as good as that of the support vector machine (SVM) in dealing with regression problems. Therefore, the combination of convolutional neural network (CNN) and support vector machine (SVM) can make up for the shortcomings of the above CNN model.

CNN-SVM is a tool life regression prediction model proposed by combining convolutional neural network (CNN) and support vector machine (SVM) methods, but if we want to continue to improve the model prediction performance we need to optimize the model hyperparameters, such as penalty parameterρ and kernel function width g, etc. Currently, the more common parameter optimization methods include manual parameter tuning, random optimization [11] , gradient-based optimization[12] , and genetic algorithm optimization[13]. The genetic algorithm is scalable and easy to combine with other algorithms, and it can achieve fast optimization with less computation time and high robustness when computational accuracy is required, so it has attracted a lot of attention from scholars in the field of hyperparametric optimization in recent years [14] .

Therefore, this paper uses a CNC machining center as a platform to collect cutting vibration signals, cutting force signals and acoustic emission signals of tools under different wear states in real time using sensor technology, and proposes a tool life prediction model based on CNN-SVM optimized by genetic algorithm (GA). The model uses CNN convolutional neural network as a feature learner and SVM support vector machine as a trainer for regression prediction. The powerful computational capabilities of the convolutional and pooling layers of the CNN convolutional neural network model are utilized to reduce the loss rate of tool wear features during translation and effectively control the fitting ability of the model; meanwhile, the powerful depth search and global search capability of the genetic algorithm is utilized to optimize two parameters, penalty factor c and kernel function radius g, in the SVM support vector machine to improve the tool life prediction accuracy.

2 Construction of CNN-SVM-GA prediction model

2.1 Convolutional Neural Network (CNN)

Convolutional neural network (CNN)[15] is a kind of neural network, a typical representative of deep learning, fundamentally it is a further extension of BP neural network, its main difference is the convolutional operation and pooling operation, which can realize local connection and weight sharing and greatly shorten the training time. CNN network structure contains not only the input layer, fully connected layer and output layer in BP network, but also its unique convolutional, pooling and RELU layers, the training model parameters still use gradient descent method to finally complete the regression prediction task. The principle is as follows:

(1) The sample feature matrix is input to the CNN convolutional neural network for convolutional operation. The sample information is indirectly characterized by the local features of the sample through the weight value of each layer derived from the convolutional operation, and the higher the layer is, the more detailed the local features are extracted, and also the spatial continuity of the sample is maintained:

$$X_i^k = \sum_{j=1}^{n} W_i^{kj} \otimes X_{i-1}^j + b_i^k \qquad (1)$$

Where X_i^k denotes the feature matrix of the kth neuron at the output of the ith layer, andW_i^{kj} denotes the weight value of the kth neuron in the ith layer, and $\otimes$ denotes the convolution operator, and X_{i-1}^j denotes the feature matrix of the jth neuron at the output of layer i-1, and b_i^k is the bias coefficient of the kth neuron in layer i.

(2) In order to improve the prediction accuracy of the tool wear life model, the CNN network uses ReLU function for nonlinear activation, which has good non-saturation characteristics to avoid the gradient

disappearance phenomenon. The activation function is shown in equation (2):

$$V_i^k = \mathrm{Relu}\left(X_i^k\right) = \begin{cases} 0, x_i^k < 0 \\ x_i^k, x_i^k > 0 \end{cases} \quad (2)$$

Where x_i^k is the X_i^k each eigenvalue in the feature matrix.

(3) Each tool wear feature data is input to the pooling layer after convolution operation, and the pooling type is chosen as maximum pooling, which can retain the original features and reduce the parameters of network training, and improve the robustness of the extracted features. The maximum pooling is shown in equation (3):

$$C_i^k(s,t) = \underset{\substack{1+(s-1)Q \le d \le sQ \\ 1+(t-1)P \le h \le tP}}{\mathrm{Max}} \left\{V_i^k(d,h)\right\} \quad (3)$$

where $V_i^k(d,h)$ is the eigenvalue of column h of row d of the ith feature matrix input to the pooling layer, and $C_i^k(s,t)$ is the eigenvalue of the sth row t column of the ith feature matrix obtained after pooling, and P and Q are the length and width of the pooled region, respectively.

(4) The n feature matrices of dimension S ×T, which are derived from each row of the sample feature matrix after two convolution and pooling operations, are input to the global average pooling layer. The dimensionality of the pooling kernel of the global average pooling layer is kept consistent with the dimensionality of the feature matrix, and the n feature matrices are dimensionality reduced to reduce the covariance of the sample features and avoid the influence of redundant features, thus reducing the training time of the LSTM long and short term memory network, so the whole CNN model finally outputs a feature vector $X_t = \{x_1, x_2, \ldots, x_i, \ldots, x_j, \}$ where xi is calculated as shown in equation (4):

$$x_i = \frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T} C_i^k(s,t) \quad (4)$$

According to the above, CNN networks also have shortcomings, such as overfitting when encountering data sets with a small number of features or high nonlinearity, which affects the accuracy of prediction. To address this problem, the SVM classifier needs to be used instead of the Softmax classifier in the CNN model to compensate for this disadvantage.

2.2 Support vector machine (SVM)

Support vector machine (SVM) [16] was proposed in 1995 by Cortes and Vapnik et al. Based on statistical theory, this learning model has a supervised mechanism that can perform tasks such as pattern recognition, classification, and regression analysis. In this paper, the feature vector output from the global average pooling layer is used as the input of the SVM support vector machine model. The biggest advantage of the SVM algorithm is that it can handle data with high nonlinearity, and the number of features in the data set has basically no effect on its model complexity, so it can accomplish regression prediction for data sets with relatively large number of features. The mathematical model of SVM is shown in equation (5):

$$\begin{cases} \min \frac{1}{2}\|w\| + \rho \sum_{r=1}^{L} \xi_r \\ s.t.\, y_r(wX_r + b) + \xi_r \ge 1, r = 1,2, \cdots, L \end{cases} \quad (5)$$

Where w is the normal vector of the hyperplane, andρ is the penalty parameter, the ξ_r is the relaxation factor, b is the offset coefficient, and X_r is the feature vector of the rth sample, andy_r is the tool wear value, L is the total number of feature samples, and the total number of samples in this paper is 315.

The model of Eq. (5) is mostly used to deal with linearly divisible sample feature data, but the tool life sample data is linearly indivisible, so it is necessary to introduce the kernel function to up-dimension each labeled sample data. In this paper, the Gaussian radial basis kernel function is used to transform the nonlinear data of each label state into linear data in high-dimensional space, so that the analysis is possible, and then the optimal classification hyperplane is constructed based on the principle of maximizing the classification interval to complete the prediction of tool life, and the Gaussian radial basis kernel function is shown in Eq:

$$K(X) = \mathrm{sgn}\left(\sum_{r=1}^{L} a_r^* y_r \exp\left(-\frac{\|X_r - X\|^2}{2g^2}\right) + \theta^*\right) \quad (6)$$

where sgn is the sign function, a_r^* is the Lagrangian multiplier, g is the kernel function width, and X is the sample label data, and θ^* is the configuration factor.

The width parameter g and the penalty coefficient c of the radial basis kernel function are the focus of the SVM algorithm tuning, which directly affect the training speed and prediction accuracy of the model, so how to find the optimal c and g parameter matching is the key of SVM model regression analysis.

2.3 Genetic Algorithm (GA)

Genetic Algorithm (GA) [17] is an intelligent algorithm that originates from the laws of nature and the mechanism of superiority and inferiority among living organisms. Using genetic algorithm, global search for superiority can be achieved, usually with three most important steps of selection, crossover and mutation, which are similar to the genetic laws of individual biological chromosomes. Therefore, this algorithm is widely used to solve search problems or to optimize some hyperparameters. Firstly, through coding, the set of strings of problem solutions is transformed into individuals that can be recognized by the genetic algorithm. Therefore, individuals with high adaptation values will survive and generate the next generation,

while individuals with low adaptation values will be eliminated; secondly, individuals with medium adaptation values will be "crossed over" to generate new individuals, which will form a new population with the original adaptive individuals; finally, the new population will be "mutated", i.e. Finally, the new population is subjected to "variation", i.e., the adaptation value of some individuals in the population is changed; so on and so forth, the whole population develops to a higher level and finally evolves the most adaptive individuals, i.e., the optimal solution, to complete the task of global optimization, etc. The optimization process of the genetic algorithm (GA) is shown in Figure 1. In this paper, it is the genetic algorithm (GA) that is used to complete the selection of hyperparameters in the SVM model, so as to improve the prediction accuracy and precision of the model.

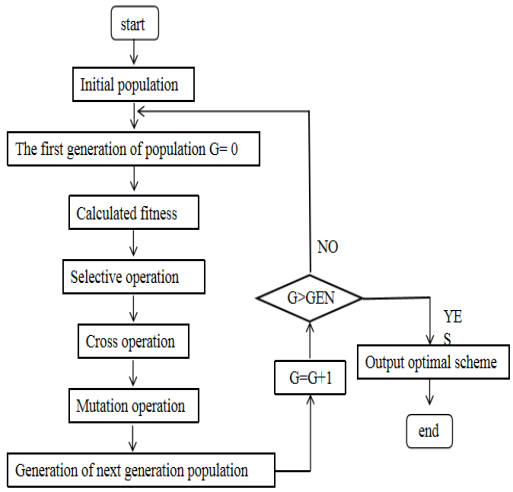


Figure 1 Flow chart of Genetic algorithm (GA)

2.4 CNN-SVM model

The essence of CNN-SVM convolutional support vector machine multi-input single-output regression prediction model is to use the CNN convolutional neural network model as a feature fuser and the SVM support vector machine as a trainer for regression prediction. The principle is firstly based on CNN convolutional neural network structure, using its convolutional layer in the network to obtain the weight parameters, pooling layer for dimensionality reduction, the sample set can be automatically feature mining and extraction from the input information without doing complex pre-processing, and fusion of features from shallow to deep as the network is continuously passed backwards. Its fusion pattern framework diagram is shown in Figure 2. Then the output feature vector (fusion value) is directly used as the input of SVM support vector machine for training, and the SVM model transforms these fusion values from low-dimensional space to high-dimensional space after CNN model processing, and then constructs an optimal decision function with the principle of maximizing classification interval to complete the regression prediction problem of data in low-dimensional space, which can realize the tool life prediction in the milling process using this method. Intelligent prediction of tool life in milling. The structure diagram of the CNN-SVM model is shown in Figure 3.

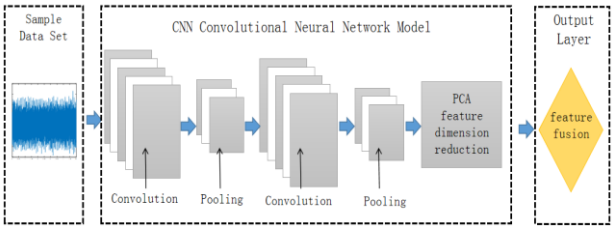


Figure 2 CNN model fusion model framework diagram

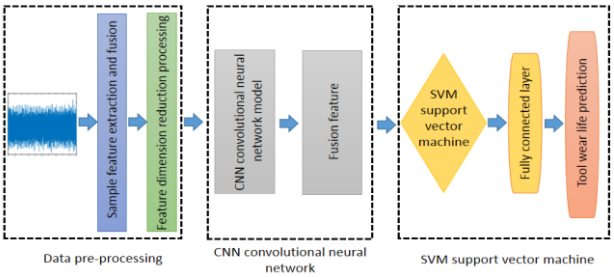


Figure 3 Structure of CNN-SVM model

2.5 CNN-SVM-GA hybrid model

The hybrid CNN-SVM model constructed in this paper uses genetic algorithm (GA) to optimize the two parameters of penalty factor c and kernel function radius g in the tool life prediction model of CNN-SVM described above. The resulting optimal solution is decoded as a parameter of the support vector machine to improve its generalization ability, speed up the network fitting, and make the tool wear prediction more accurate. The algorithmic flow of the tool life prediction technique based on CNN-SVM optimized by genetic algorithm is shown in Figure 4, and the specific steps are as follows:

Step 1: The original signal (7 channels) related to tool wear is processed for noise reduction and feature extraction and fusion in the time domain, frequency domain and time-frequency domain, respectively.

Step 2: Using Pearson's correlation coefficient formula for the above feature data to perform dimensionality reduction and random division of them to construct the training set and test set of the model.

Step 3: building a convolutional neural network, trained using the training and test sets from step 2, the output of which is a feature vector.

Step 4: Perform PCA feature dimensionality reduction on the extracted feature vectors to reduce the training time of the SVM and form a new training and test set.

Step 5: The SVM model is trained with the training set formed in step 4, and the g and c parameters of the support vector machine are optimized using a genetic algorithm.

Step 6: Input the test set formed in step 4 to the improved CNN-SVM model to test the model diagnostic effect.

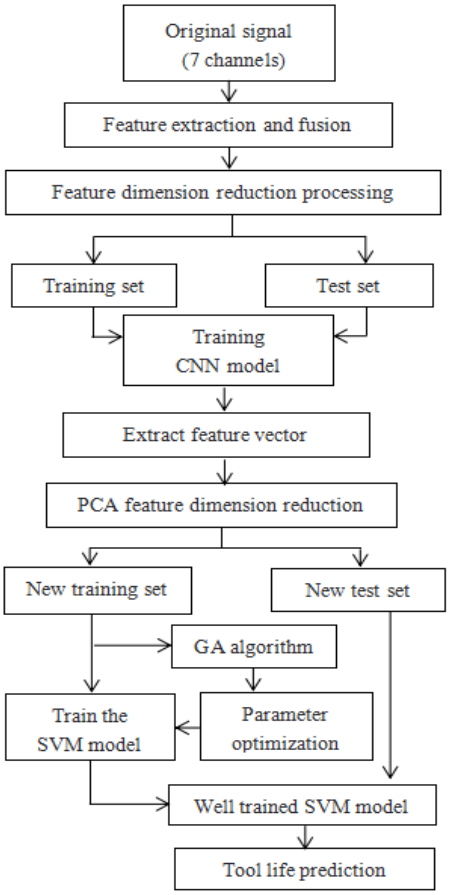


Figure 4 Improved CNN-SVM lifetime prediction model

3 Tool wear experiment process

The experimental data were obtained from the open data of the 2010 High Speed CNC Machine Tool Health Prediction Competition of the Prediction and Health Management Society (PHM), New York, USA[18] , whose tool wear experimental conditions are shown in Table 1.

Table 1 Experimental conditions for tool wear

Hardware Conditions	Model and main parameters	Cutting Conditions	Parameters
CNC Milling Machine	CNC Milling Machine Roders Tech RFM760	Spindle speed	10400
Force Gauge	Three-way force gauge Kistler 9265B	Feeding speed	1555
Charge amplifier	Multi-channel charge amplifier Kistler 5019A	Back draft	0.2
Milling Material	cube Inconel 718	Side-draft amount	0.125
Tools	Ball end carbide milling cutter 3 teeth	Feed amount	0.001
Data Acquisition Cards	Data Acquisition Cards NI DAQ	Sampling frequency	50
Wear Gauge	Microscope LEICA MZ12	Cooling conditions	Dry cutting

In the process of machining, the spindle speed was 10400 RPM, the feed was 0.001 mm, the feed speed was set to 1555 mm/min, the tool side draft was 0.125 mm, and the tool back draft was 0.2 mm. The shape of the milled part was square, and the end face was milled by round-trip milling, and the length of the milled part was about 108 mm. The surface length is about 108 mm, and the machining process does not use cutting fluid.Δt The wear value of the rear face of the three teeth of the ball end mill was checked after each time. In this paper, the experimental data set of the first tool of C1 group is selected, and the data set collects and monitors the data of X, Y and Z axes cutting force signals, X, Y and Z axes vibration signals and acoustic emission signals, with a total of 7 channels, each channel walking 315 times, the acquisition frequency is 50 KHz per channel, and the number of sampling points is above 200000 each time, and its related specific data acquisition system is shown in Figure The specific data acquisition system is shown in Figure 5.

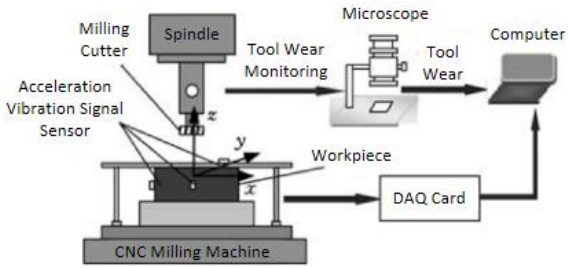


Figure 5 Tool wear data acquisition system

Since the milling cutter used in the experiment has three teeth, the wear of the three teeth was measured after everyΔt The wear of the three teeth was measured after each time. Figure 6 shows the wear curve of the first group of test tools, the purple curve is the wear of the first tooth, the blue curve is the wear of the second tooth and the yellow curve is the wear of the third tooth. In this paper, the average value of the wear of these three tool teeth is taken to represent the actual wear of the tool, and this average value is the sample target value of the improved CNN-SVM convolutional support vector machine model, i.e., the output data. From the figure, it can be seen that the tool wear is faster at the beginning of the tool wear period, flatter when it enters the middle period, and faster at the later period, which is consistent with the theory related to tool wear, which indirectly verifies the accuracy of the data set.

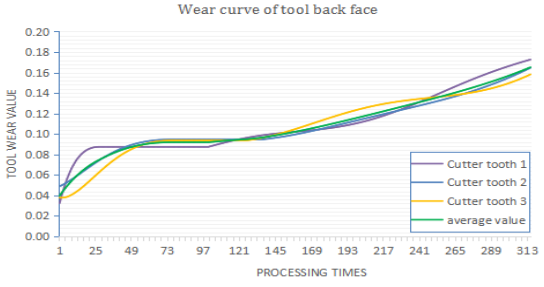


Figure 6 Test tool wear variation curve

4 Pre-processing of tool wear characteristics

4.1 Feature extraction and fusion

During CNC machining, sensor technology is used to collect real-time signals related to tool wear. In this paper, cutting vibration signals are collected using a Kistler 8636C piezoelectric accelerometer, cutting force signals are collected using a Kistler 8152 three-way platform dynamometer, and acoustic emission signals are collected using a Kistler 9265B acoustic transmitter. The number of signal data collected above is huge, and there is a lot of noise, which is often caused by the instability of the system at the moment of cutting in and out of the tool, so it is necessary to perform noise reduction processing on the various types of raw signals collected above. The number of times the tool is walked in this experiment is 315, and the number of acquisition points for each knife walk is about 200000 or more. In order to avoid adverse effects during model training, the sampling points with data labels of 50001 to 100000 in each acquisition signal are extracted for feature extraction and fusion in this paper.

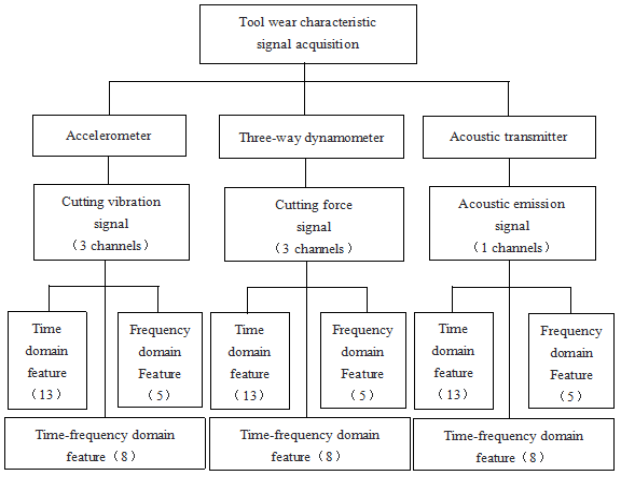


Figure 7 Wear feature extraction and fusion scheme

The feature quantities related to the tool wear state are extracted in the time domain, frequency domain, and time-frequency domain for the above three types of signals, respectively. In order to realize the intelligent tool wear prediction, the time domain features of the original signal are extracted, including 13 kinds, namely, mean value, standard deviation, skewness, cliffness, maximum value, minimum value, peak-to-peak value, root mean square, amplitude factor, waveform factor, impact factor, margin factor, and energy; the frequency domain features are extracted, including 5 kinds, namely, frequency domain amplitude mean, center of gravity frequency, mean square frequency, variance frequency, and frequency variance; the time The extraction of frequency domain features mainly uses wavelet packet analysis to subdivide the original signal into different frequency bands, when the tool wear state changes the energy parameters of different frequency bands will also change, so the energy of each frequency band is the extracted time-frequency domain features. The wavelet packet decomposition is performed on the original signal, and the number of decomposed layers is set to 3, all of which are completed by db5 wavelet base, and the frequency domain is divided into 8 frequency bands, so that 8 time-frequency domain features are extracted. In this experiment, the original signals of cutting vibration signal (3 channels), cutting force signal (3 channels) and acoustic emission signal (1 channel) are extracted and fused every Δt time, as shown in Figure 7, 13 time domain features, 5 frequency domain features and 8 time-frequency domain features are extracted from each channel signal, so 26 features can be extracted from each channel signal, for a total of 7 channels and 182 features in total. The total number of features is 182.

4.2 Feature dimensionality reduction processing

The speed of the tool wear prediction model fitting operation is closely related to the number of features, the more features the more complex the model is, the slower the operation is, so it is necessary to filter and optimize all the features. The best way is to find the correlation between the above mentioned 182 features and the tool wear, and to delete the uncorrelated or weakly correlated features, thus optimizing the extraction of the tool wear signal features and making the model operation speed increase. The Pearson correlation coefficient is the most widely used correlation coefficient analysis method, which can be used to measure the correlation between the extracted eigenvalues and the tool wear amount [19]. It is calculated as shown in equation (7):

$$P_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \quad (7)$$

whereP_{xy} denotes the Pearson correlation coefficient of the signal feature x and the tool wear value y. where n denotes that there are n sets of signal values andx_i denotes the ith value of the signal characteristic value, andy_i denotes the ith value of tool wear. The Pearson correlation coefficient formula is used to calculate the correlation between the above 182 features and the tool wear values. Figure 8 shows the correlation of Pearson coefficients for each feature, the red area is$|P_{xy}| < 0.5$ the features that are weakly correlated, with a total of 48 feature values; the yellow area is$0.5 \le |P_{xy}| < 0.9$ The yellow area is for the features that are moderately correlated, with a total of 87 feature values; the green area is for the features that are strongly correlated, with a total of 87 feature values.$|P_{xy}| \ge 0.9$ The green area is for the features that are strongly correlated, with a total of 47 eigenvalues. In this paper, the 47 strongly correlated features are used as the input data for the training and

prediction of CNN-SVM model, so as to improve the computational speed and accuracy of tool wear prediction.

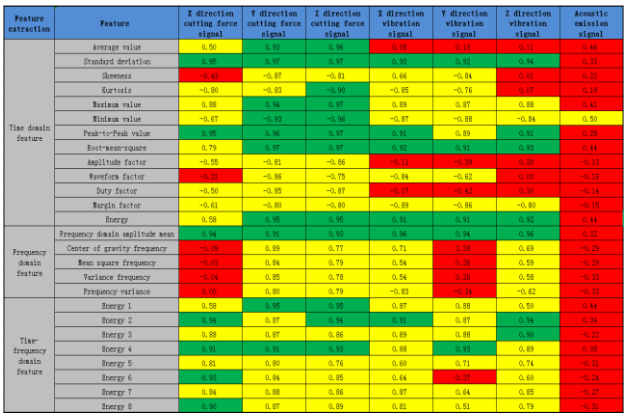


Figure 8 Correlation of Pearson coefficients for each feature

5 Tool wear experiment results and analysis

5.1 Construction of the sample set data

In this paper, firstly, sensor technology is used to collect the signals related to tool wear (7 channels), secondly, the original signals are subjected to noise reduction processing, feature extraction, feature fusion, dimensionality reduction processing and other operations respectively, and 47 strongly correlated features are derived to form a feature matrix as the input data for the training and prediction of the life prediction model, and its sample feature matrix dimension is 315×47; the wear of the three teeth of the milling tool is Finally, the above feature matrix is randomly sampled and feature coded, and then the training set and test set are divided, and the first 200 data are taken as the training set and the remaining data are taken as the test set.

5.2 Setting of prediction model parameters

In this paper, the sample set data are input to a CNN-SVM model based on genetic algorithm (GA) optimization for tool life prediction, where the initial learning rate parameter of CNN convolutional neural network is set to 0.001, the cross-entropy function is used as the loss function of the whole model, and the Adam optimizer is selected to optimize the hyperparameters, which is set to make the model generalization ability stronger. Second, the Softmax classifier on the fully connected layer is replaced with the SVM algorithm to better handle data with high nonlinearity, and an optimal decision function is constructed to complete the regression prediction of tool wear.

The CNN-SVM-GA model selects the penalty parameter in the SVM modelρ and the kernel function width g, which are both set between 0 and 3, as the 2 parameters for the optimization search process. The genetic algorithm (GA) adopts the strategy of superiority selection, crossover and variation to find the optimal hyperparameter pairing, with the crossover rate set to 0.35, the variation rate set to 0.1, the population size set to 20, and the evolutionary generation set to 3000. The specific parameters are shown in Table 2. Ten optimization operations were performed according to the parameters in Table 2, and the average value was taken as the final result, where the penalty parameter ρ The optimized kernel function g is 1.421.ρ The optimized parameters, g, are migrated to the CNN-SVM model to complete the tool life prediction.

Table 2 Genetic algorithm (GA) parameter settings

GA algorithm parameters	Parameter Value
Evolutionary Algebra	3000
Population size	20
Crossover Rate	0.5
Variation rate	0.1

In order to quantify the prediction performance of the tool life model, three objective evaluation indicators are selected, namely the mean absolute error MAE, the root mean square error RMSE and the coefficient of determination R2. Among them, the mean absolute error MAE can obtain an evaluation value, but the comparison between different models is necessary to reflect the model's merit; the mean square error RMSE can measure the deviation between the observed value and the true value, the smaller the RMSE value, the better our model is. The smaller the RMSE value is, the better the model is; the coefficient of determination R2 can directly characterize the merit of the model, and the closer the value of the coefficient of determination R2 is to 1, the higher the accuracy and precision of the prediction model is. The three evaluation indicators are calculated as shown in equations (8) to (10):

$$MAE = \frac{\sum_{t=1}^{m}|y_t - \hat{y}_t|}{m} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{m}(y_t - \hat{y}_t)^2}{m}} \tag{9}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{m}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{m}(y_t - \bar{y})^2} \tag{10}$$

where, m is the number of samples output from the fully connected layer, the number of samples in this paper is 315, and$\hat{y}_t$ is the predicted value of tool wear, andy_t is the actual value of tool wear.

5.3 Tool life prediction results

Based on the open data of the CNC machining center tool health prediction contest, the CNN-SVM algorithm optimized by genetic algorithm (GA) was used for tool wear regression prediction, and its test set prediction results are shown in Figure 9. The mean absolute error MAE value of the model was calculated to be 0.7231, the root mean square error RMSE value was 0.8292, and the coefficient of determination R2 value was 0.9985. The

results show that the regression prediction of tool life can be effectively performed using the CNN-SVM-GA-based model with good results.

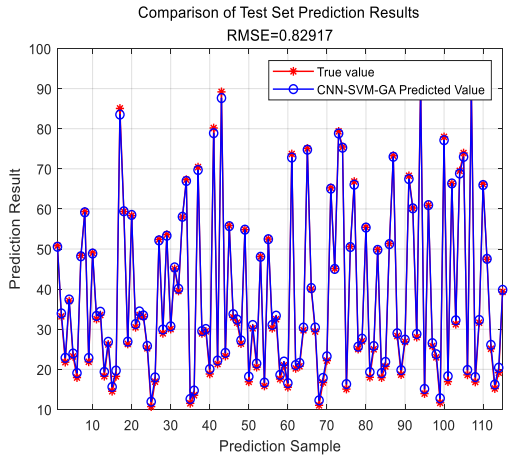


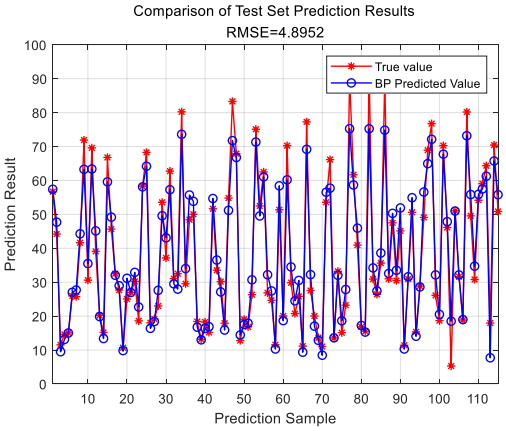
Figure 9 CNN-SVM-GA test set prediction results

Table 3 shows the effect of genetic algorithm (GA) on the tool life regression prediction model, where the penalty parameters of the CNN-SVM model ρ and hyperparameters such as kernel function width g are chosen randomly by relying on manual, it can be seen that the CNN-SVM model optimized using genetic algorithm (GA) has the best tool life prediction. Compared with the CNN-SVM model, its mean absolute error MAE and root mean square error RMSE are reduced and the coefficient of determination R2 is improved, and its performance index reaches 0.99, while the performance index of the CNN-SVM model with manually selected parameters is maintained at a maximum of about 0.98. This is mainly because the hyperparameter optimization of the CNN-SVM model by Genetic Algorithm (GA) has obtained more accurate hyperparameter pairings, found the most critical attributes affecting the accuracy of tool life prediction, and avoided the blindness of setting parameters, thus improving the prediction effect.

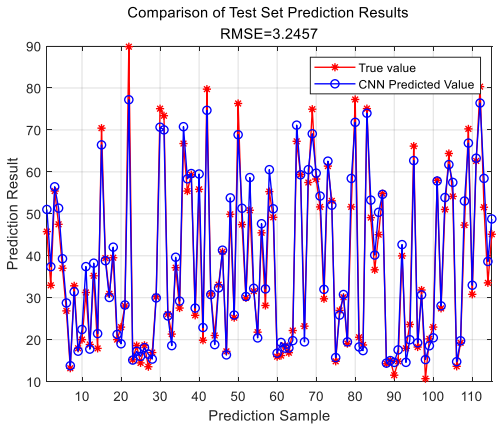
Table 3 Effect of genetic algorithm (GA) on the prediction model

Algorithm	Hyperparameters		Test set prediction results		
	Penalty Parameter	Kernel width	MAE	RMSE	R2
CNN-SVM	0.5	0.5	2.4859	2.8570	0.9817
	1	1	1.1671	1.4557	0.9851
	2	2	3.2250	4.1678	0.9628
	3	3	4.1927	5.7604	0.9296
CNN-SVM-GA	0.511	1.421	0.7231	0.8292	0.9985

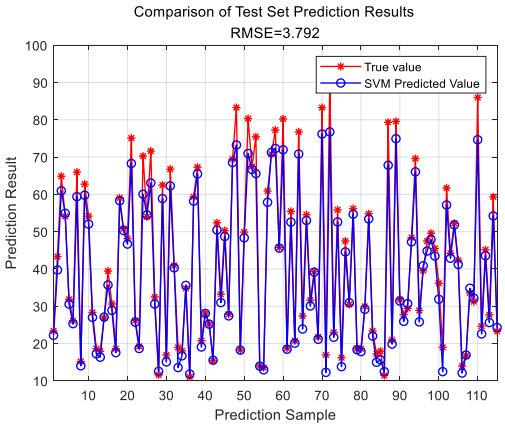
To further validate the prediction performance of CNN-SVM-GA based tool life, a comparative analysis was performed with other traditional prediction models in the past, such as BP neural network, CNN convolutional neural network, SVM support vector machine, and CNN-SVM model. Figure 10 shows the comparison results of four traditional tool life prediction models, and it can be seen from Figure 9 and Figure 10 that the root mean square error RMSE performance of the five tool life prediction models is ranked as CNN-SVM-GA < CNN-SVM < CNN < SVM < BP, and their root mean square error is reduced by 83.06%, 78.13%, 74.45%, and 43.04%, respectively. It can be seen that the CNN-SVM model based on genetic algorithm (GA) optimization proposed in this paper has obvious advantages in tool life prediction, which is because the CNN-SVM-GA model can deeply mine the hidden layer features of the data with high nonlinearity, the feature extraction is comprehensive, and the selection of hyperparameters does not have any dependence on expert experience.



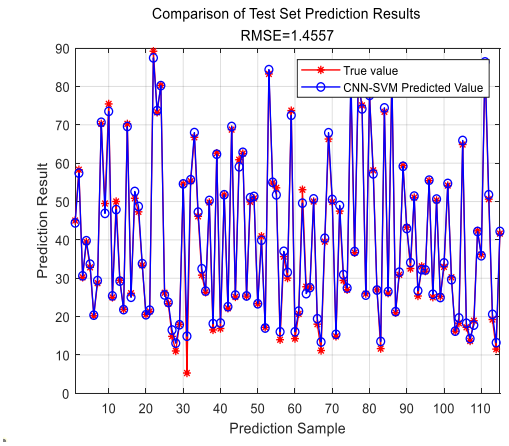
(a)



(b)



(c)



(d)

Figure 10 Prediction results of the four traditional models

(a) BP model (b) CNN model (c) SVM model (d) CNN-SVM model

Table 4 Comparison of prediction performance results of five models

Algorithm	Test set prediction results		
	MAE	RMSE	R2
BP Neural Network	3.6862	4.8952	0.9498
CNN Algorithms	2.5274	3.2457	0.9732
SVM Algorithms	2.5671	3.7920	0.9689
CNN-SVM Algorithm	1.1671	1.4557	0.9851
CNN-SVM-GA Algorithm	0.7231	0.8292	0.9985

Table 4 shows the comparison results of the prediction performance of the five models, and it is found that the CNN-SVM-GA model using multi-channel feature fusion for tool life prediction has the smallest mean absolute error MAE, and the index performance ranking is CNN-SVM-GA < CNN-SVM < CNN < SVM < BP, which is reduced by 80.38%, 71.83%, 71.39%, and 38.04%; the coefficient of determination R2 of the CNN-SVM-GA model proposed in this paper is 0.9985, which is closest to 1. The index performance is ranked as CNN-SVM-GA ＞ CNN-SVM ＞ CNN ＞ SVM ＞ BP, which is improved by 1.34%, 2.53%, 2.96%, and 4.88%, respectively. These two results once again prove that using the CNN-SVM-GA model proposed in this paper for tool life prediction is more effective and can achieve more effective tool life prediction and health management in the milling process.

6 Conclusion

This paper completes the construction of a tool life sample dataset based on machine vision, feature extraction, and information fusion, and also proposes a CNN-SVM tool life prediction model based on genetic algorithm (GA) optimization. The model uses convolutional neural network (CNN) model as the feature fusion and support vector machine as (SVM) as the trainer for tool life regression prediction. And the prediction accuracy of the model is improved by using genetic algorithm (GA) to find the superiority of hyperparameters in the model. The results show that:

(1) The mean absolute error MAE value of 0.7231, root mean square error RMSE value of 0.8292, and coefficient of determination R2 value of 0.9985 were obtained for tool life regression prediction using CNN-SVM-GA model. This indicates that the model can effectively predict the remaining life of the tool with good results.

(2) The tool life prediction model is parameter-seeking by genetic algorithm (GA), and its decision coefficient R2 performance index reaches 0.99, which reduces the subjective influence of manual selection of parameters and avoids the blindness of setting parameters, thus improving the model prediction accuracy.

(3) Compared with the BP model, CNN model, SVM model and CNN-SVM model, the mean absolute error MAE and root mean square error RMSE values of the CNN-SVM-GA model proposed in this paper are reduced, and the value of the coefficient of determination R2 is improved to be closest to 1. This indicates that the constructed tool life prediction model has stronger generalization ability, faster network fitting and tool wear prediction is more accurate.

In the future, this CNN-SVM-GA tool wear prediction model can be widely used in various factories for CNC machining tool life management and other fields. By making real-time prediction of tool life, it can realize predictive maintenance of CNC machining tools and can perform intelligent tool change before tool wear is at a critical threshold, which is in line with the future development trend of intelligent control and network interactive production.

Author Contributions: For Conceptualization, methodology, analysis, and writing original draft preparation, Wang Shuo; writing review and full-text editing, Yu Zhenliang; writing—original draft preparation, Liu Peng; writing — original draft preparation,Wang Mantong.

Conflicts of interest: The authors declare no conflict of interest.All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The research work financed with the means of Basic Scientific Research Youth Program of Education Department of Liaoning Province, No.LJKQZ2021185; Yingkou Enterprise and Doctor Innovation Program (QB-2021-05).

References

[1] Danil Yu Pimenov, Andres Bustillo, Szymon Wojciechowski, et al.. Artificial intelligence systems for tool condition monitoring in machining: analysis and critical review [J]. Journal of Intelligent Manufacturing, 2022(1):67-68.

[2] Dong Liang, Wang Chensheng, Yang Guang, Huang Zeyuan, Zhang Zhiyue, Li Cen. An Improved ResNet-1d with Channel

Attention for Tool Wear Monitor in Smart Manufacturing [J]. Sensors, 2023, 23(3):89-90.
[3] Shi Yuen Wong, Joon Huang Chuah, Hwa Jen Yap. Technical data-driven tool condition monitoring challenges for CNC milling: a review [J]. The International Journal of Advanced Manufacturing Technology, 2020,107 (prepublish).
[4] Pu Xiaobo, Jia Lingxu, Shang Kedong, Chen Lei, Yang Tingting, Chen Liangwu, Gao Libin, Qian Linmao. A New Strategy for Disc Cutter Wear Status Perception Using Vibration Detection and Machine Learning [J]. Sensors, 2022,22(17).
[5] Wei Weihua, Cong Rui, Li Yuantong, Abraham Ayodele Daniel, Yang Changyong, Chen Zengtao. Prediction of tool wear based on GA-BP neural network [J]. Proceedings of the Institution of Mechanical Engineers,2022, 236(12).
[6] Cao W. The Diagnosis of Tool Wear Based on RBF Neural Networks and D-S Evidence Theory. IEEE China Council. IEEE Beijing Section, Sichuan Computer Federation. Proceedings of 2010 3rd IEEE International Conference on Computer Science and Information Technology VOL.7 [J]. Institute of Electrical and Electronics Engineers, 2010:429-431.
[7] Zhang Kun , Zhu Hongtao, Liu Dun, Wang Guoning, Huang Chuanzhen,Yao Peng. A dual compensation strategy based on multi-model support vector regression for tool wear monitoring [J]. Measurement Science and Technology, 2022,33(10).
[8] G. E. Hinton, R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks [J]. Science, 2006, 313(5786).
[9] P. K. Ambadekar, C. M. Choudhari. CNN based tool monitoring system to predict the life of cutting tool [J]. SN Applied Sciences, 2020, 2(4).
[10] Caesarendra Wahyu, Triwiyanto Triwiyanto, Pandiyan Vigneashwara, Glowacz Adam, Permana Silvester Dian Handy, Tjahjowidodo Tegoeh. A CNN Prediction Method for Belt Grinding Tool Wear in a Polishing Process Utilizing 3-Axes Force and Vibration Data [J]. Electronics, 2021, 10(12).
[11] Stefan Droste, Thomas Jansen, Ingo Wegener. Upper and Lower Bounds for Randomized Search Heuristics in Black-Box Optimization. Electron. Colloquium Comput [J]. Complex, 2003:48-48.
[12] Weifeng Lu, Bingyu Cai, Rui Gu. Improved Particle Swarm Optimization Based on Gradient Descent Method [J]. CSAE, 2020: 121-126.
[13] Salih Omran, Duffy Kevin Jan. Optimization Convolutional Neural Network for Automatic Skin Lesion Diagnosis Using a Genetic Algorithm [J]. Applied Sciences, 2023,13(5):87-88.
[14] Nagra Arfan Ali, Mubarik Iqra, Asif Muhammad Mugees, Masood Khalid, Ghamdi Mohammed A. Al, Almotiri Sultan H. Hybrid GA-SVM Approach for Postoperative Life Expectancy Prediction in Lung Cancer Patients [J]. Applied Sciences, 2022, 12(21):655-66.
[15] Gajera Himanshu K., Nayak Deepak Ranjan, Zaveri Mukesh A.. A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features [J]. Biomedical Signal Processing and Control, 2023,79(2):32-37.
[16] Fafa Chen, Chen Fafa, Cheng Mengteng, Tang Baoping, Chen Baojia, Xiao Wenrong. Pattern recognition of a sensitive feature set based on the orthogonal neighborhood preserving embedding and adaboost_SVM algorithm for rolling bearing early fault diagnosis [J]. Measurement Science and Technology,2020,31(10):89-91.
[17] Zheng Zhang, Liang Li, Wei Zhao. Tool Life Prediction Model Based on GA-BP Neural Network [J]. Materials Science Forum, 2016, 3901(836):956-957.
[18] Huimin Chen. A Multiple Model Prediction Algorithm for CNC Machine Wear PHM [J]. International Journal of Prognostics and Health Management, 2011,2(2);56-57.
[19] Li Yifan, Xiang Yongyong, Pan Baisong, Shi Luojie. A hybrid remaining useful life prediction method for cutting tool considering the wear state [J]. The International Journal of Advanced Manufacturing Technology, 2022,121(5-6):78-90.

Viser Technology Pte. Ltd.
Mechanical Engineering Science
DOI: 10.33142/mes.v4i2.9083

Research Article Open Access

Tool wear condition monitoring method of five-axis machining center based on PSO-CNN

Shuo WANG, Zhenliang YU*, Changguo LU, Jingbo WANG

Yingkou Institute of Technology , School of Mechanical and Power Engineering, Yingkou,China

***Corresponding Author:** Zhenliang YU, email address:yuzhenliang_neu@163.com*

Abstract:
The effective monitoring of tool wear status in the milling process of a five-axis machining center is important for improving product quality and efficiency, so this paper proposes a CNN convolutional neural network model based on the optimization of PSO algorithm to monitor the tool wear status. Firstly, the cutting vibration signals and spindle current signals during the milling process of the five-axis machining center are collected using sensor technology, and the features related to the tool wear status are extracted in the time domain, frequency domain and time-frequency domain to form a feature sample matrix; secondly, the tool wear values corresponding to the above features are measured using an electron microscope and classified into three types: slight wear, normal wear and sharp wear to construct a target Finally, the tool wear sample data set is constructed by using multi-source information fusion technology and input to PSO-CNN model to complete the prediction of tool wear status. The results show that the proposed method can effectively predict the tool wear state with an accuracy of 98.27%; and compared with BP model, CNN model and SVM model, the accuracy indexes are improved by 9.48%, 3.44% and 1.72% respectively, which indicates that the PSO-CNN model proposed in this paper has obvious advantages in the field of tool wear state identification.

Keywords: five-axis machining center; tool wear; PSO-CNN; intelligent monitoring

1 Introduction

Five-axis machining center is a set of high-tech, high precision, high efficiency in one of the high precision end equipment, specifically for processing complex curved parts, its key technology to improve the level of equipment manufacturing industry is of great significance. And five-axis machining center. Due to its flexibility, versatility and high throughput, the machining environment is more complex and tool wear is more severe. Tool wear beyond a given threshold can greatly affect the machining accuracy of the workpiece, resulting in poor quality of the machined product [1]. On the other hand, in order to ensure the machining accuracy, if the tool has a long remaining life, it will affect the economy of its use and increase the production cost, especially in the process of batch processing will also cause interruptions in the production beat, lower production efficiency and other problems. For complex curved parts with high precision machining requirements, how to make the tool wear before the critical threshold for intelligent tool change will be an important research direction for the future high-end manufacturing industry.

Tool Condition Monitoring (TCM) has been recognized as an important method for preventing excessive tool wear and maintaining part tolerances and surface quality during the milling process [2]. Its essence is the real-time acquisition of signals related to tool wear using sensor technology, as well as the capture of correlated features of tool wear using data-driven techniques to construct a reference model for feature monitoring. In the process of tool condition monitoring, it is usually necessary to pre-process the acquired raw signal, extract the effective features from the signal and construct a sample feature matrix as the input to the prediction model. The most commonly used feature extraction methods are: Empirical Mode Decomposition (EMD) [3], Fourier Transform [4] and Wavelet Packet Analysis [5] etc. Empirical mode decomposition (EMD) can effectively extract tool wear state features from the time and frequency domain, but it requires a high level of signal frequency processing and may suffer from severe endpoint effects and mode confounding in the process [6] . The Fourier transform is independently adaptive, allowing time domain features to be better revealed in the frequency domain, and is therefore widely used to extract frequency domain features of sample signals [7]. Wavelet packet

Received on October 14, 2022; Accepted on December 9, 2022

analysis is used to decompose the time domain features into different frequency bands by using different types of filters to refine the signal, so it is mostly used to extract the time-frequency domain features of the sample signal [8].

In the automated production process, a high-precision tool wear state prediction model can effectively predict the future tool wear degree, which is of great significance to improve the productivity and surface machining quality. Early scholars have made some achievements in constructing a tool wear state prediction model using mechanical learning techniques. Han Chengwen et al. identified two valuable features related to tool wear based on discrete wavelet transform (DWT) of thrust signal and artificial neural network (ANN), and then extracted them using DWT.This method can accurately estimate the CFRP drilling process tool wear[9]. Soufiane Laddada et al. used continuous wavelet transform for feature extraction and proposed an improved extreme learning machine (IELM) to map the input data by a nonlinear function in order to generate a degradation model to obtain health indicators to complete the prediction of the remaining tool life[10] . Liang Yu et al. used a combination of time domain, frequency domain and wavelet analysis to extract the force and vibration signals and constructed the IHDGWO-SVM model for tool wear prediction. The experimental results showed that the prediction accuracy of the model was 92 %, which was significantly higher than other models [11]. However, the machine learning method does not deeply mine the implicit information of the data, and its prediction accuracy and precision are not high.

In recent years, deep learning theory has been widely used in the field of tool condition monitoring, and Convolutional Neural Network (CNN) is a typical representative of deep learning. Convolutional neural networks (CNNs) have powerful feature extraction capabilities, and their convolution and pooling operations can adaptively mine the deep features of the input data, which can better approximate the objective function through a large number of nonlinear mappings and improved feature representations [12]. Therefore, a large number of researchers have started to use CNN network models for tool wear state recognition, such as Xin Cheng et al. conducted milling experiments on S45C steel under different machining parameters and used convolutional neural networks to mine potential features of multi-scale 2D signals to construct a wear state recognition model, and the results showed that the method can effectively recognize tool wear state [13]. Although CNN networks have achieved some achievements in tool wear status monitoring, how to avoid the overfitting phenomenon caused by gradient dispersion is an urgent problem to be solved.

To solve the above problem, the hyperparameters in the convolutional neural network (CNN) tool condition monitoring model need to be optimized, such as batch size and Epoch count and other key parameters. Currently, the more common hyperparameter optimization methods include random optimization [14], gradient-based optimization [15], genetic algorithm optimization [16], particle swarm algorithm optimization [17], etc. Particle swarm algorithm (PSO) has powerful search performance and individual optimization capability, and can choose adaptive weights according to the number of iterations, thus avoiding the phenomenon of global optimal solution omission due to too fast convergence, so it has been widely used and studied by scholars in recent years [18].

Therefore, this paper proposes a dynamic monitoring method for tool wear status based on machine vision, feature extraction, deep learning, and information fusion. The CNN convolutional neural network is used to mine the tool wear features, and the classifier is constructed in the output layer after a series of operations such as convolutional layer and pooling layer to output the tool wear status information; meanwhile, the particle swarm optimization algorithm (PSO) is used to optimize the hyperparameters in the CNN convolutional neural network to improve the accuracy and precision of the prediction model. It is verified that the PSO-CNN model proposed in this paper can accurately and efficiently predict the tool wear status, effectively ensure the machining quality of the part, improve the efficiency of tool use, and reduce the machining cost, which is an important step to realize the intelligence of CNC machining.

2 Tool wear condition monitoring method

2.1 PSO-CNN tool wear condition monitoring model

Convolutional Neural Network (CNN)[19] is a typical representative of deep learning, which is a locally connected and weight-sharing neural network structure consisting of input layer, convolutional layer, pooling layer, fully connected layer and output layer, and has obvious advantages for deep mining of data features. However, improper selection of hyperparameters in CNN networks can lead to slow convergence of the model and overfitting phenomenon. Therefore, this paper proposes a CNN convolutional neural network model based on the optimization of the PSO algorithm to classify and predict the tool wear state. The model firstly mines the features in the sample dataset deeply through a series of operations such as convolutional and pooling layers in the CNN network, The principle is as follows:

The sample feature matrix after batch normalization and dimensionality reduction is input to the CNN convolutional neural network for convolutional operation. The sample information is indirectly characterized by the local features of the sample through the weight value of each layer derived from the convolutional operation, and the higher the layer is, the more detailed the local features are extracted, and also the spatial continuity of the sample is maintained, and its convolutional operation is shown in equation (1):

$$X_i^k = \sum_{j=1}^{n} W_i^{kj} \otimes X_{i-1}^j + b_i^k \tag{1}$$

Where X_i^k denotes the feature matrix of the kth neuron at the output of the ith layer, and W_i^{kj} denotes the weight value of the kth neuron in the ith layer, and $\otimes$ denotes the convolution operator, and X_{i-1}^j denotes the feature matrix of the jth neuron at the output of layer i-1, and b_i^k is the bias coefficient of the kth neuron in layer i.

Each tool wear feature data is input to the pooling layer after convolution operation, and the pooling type is selected as maximum pooling, which can retain the original features and reduce the parameters of network training, and improve the robustness of the extracted features. The maximum pooling is shown in equation (2):

$$C_i^k(s,t) = \underset{\substack{1+(s-1)Q\le d\le sQ \\ 1+(t-1)P\le h\le tP}}{\text{Max}} \left\{V_i^k(d,h)\right\} \quad (2)$$

Where $V_i^k(d,h)$ is the eigenvalue of column h of row d of the ith feature matrix input to the pooling layer, and$C_i^k(s,t)$ is the eigenvalue of the sth row t column of the ith feature matrix obtained after pooling, and P and Q are the length and width of the pooled region, respectively.

For another，the PSO algorithm is introduced to optimize the hyperparameters of batch size and Epoch count in the CNN model, so as to Finally, a Softmax classifier is constructed in the output layer to predict the tool wear status and output the tool wear type, thus completing the prediction of the tool wear status of the 5-axis machining center. The 5-axis machining center will take different processing solutions according to different prediction results, such as the system will have a warning prompt when the tool enters into a sharp wear stage, and complete intelligent tool change and other operations, and its PSO-CNN tool wear state monitoring model is shown in Figure 1.

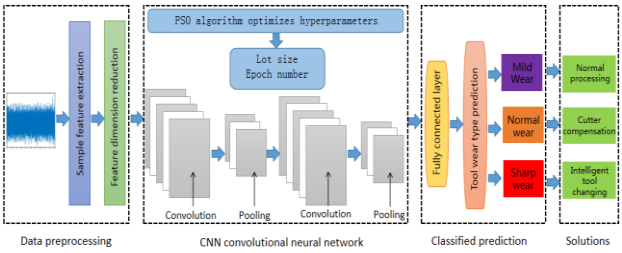


Figure 1 PSO-CNN tool condition monitoring model

2.2 Prediction process of PSO-CNN monitoring model

The tool condition monitoring process based on CNN convolutional neural network optimized by PSO algorithm proposed in this paper contains four main stages.

(1) The original signal is pre-processed to eliminate the noise effect, and then the feature quantities related to the tool wear state are extracted in the time domain, frequency domain, and time-frequency domain to construct the sample data set M.

(2) The sample data set M is randomly divided, and the first 200 samples are used as the training set and the remaining samples are used as the test set. The training set is input to the CNN network for model training, and the training process mainly includes two stages of forward propagation and backward propagation. Forward propagation is a series of operations such as convolution, pooling and full connection to obtain the output of the network, i.e., the probability distribution of the category of tool wear. Back propagation is to calculate the error between the probability value of the output of the CNN network and the standard answer, and then back propagate the calculated error to obtain the error of each layer, and finally fine-tune the whole network parameters by using gradient descent method to improve the whole CNN model.

(3)The PSO algorithm is introduced to optimize the two hyperparameters of batch size and Epoch count to derive the best combination of parameters, and the best parameters are used for forward propagation of the CNN network, and iterative operations are performed on the network connection weight matrix until the errors converge and then the operations are terminated to complete the optimal training of the final model.

(4)The test set is fed into the trained CNN model, and the three types of tool wear are output using the fully connected layer to complete the prediction of the type of tool wear state on a 5-axis machining center.

3 Construction of tool wear sample data set

3.1 Acquisition of cutting vibration signals and spindle current signals

This paper uses sensors to collect cutting vibration signals and spindle current signals during the milling process of a 5-axis machining center in real time, and uses an electron microscope to measure the corresponding tool wear values to provide data support for the realization of tool remaining life prediction, the Measuring system models is shown in Figure 2.

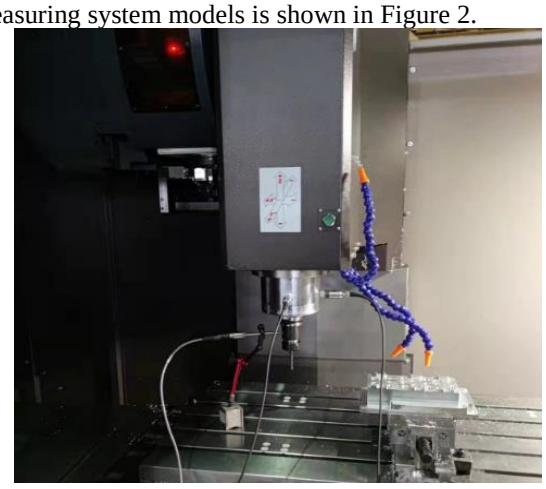


Figure 2 Measuring system models

3.1.1 Vibration signal acquisition scheme

The vibration signal is caused by the periodic vibration of the cutting system composed of machine

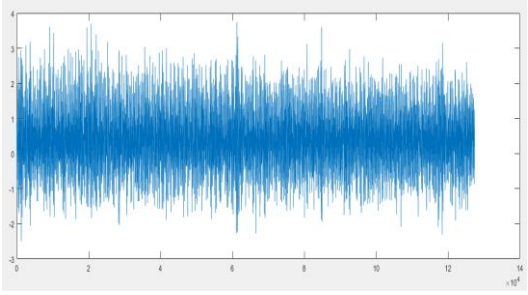
workpiece or tool, and the strength of the vibration between systems is closely related to the tool wear state. Acquisition of vibration signals generally choose acceleration sensors, according to the different measurement principles are broadly divided into three ways: piezo-resistive sensors, piezoelectric sensors and capacitive sensors, this paper uses BVM-YD-139 piezoelectric acceleration sensors to collect vibration signals, in the installation, you can use magnetic adsorption on the surface of the parts to be processed for detection, but the results of measuring the tool vibration signal by the location of the installation However, the result of measuring the tool vibration signal is affected by the location of the installation, and the strength of the machine tool system vibration and the interference of external environmental factors will also have an impact on the vibration signal acquisition, so the vibration signal collected needs to be processed for noise reduction.

3.1.2 Spindle current signal acquisition scheme

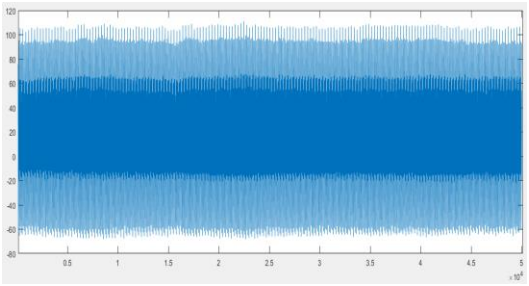
The spindle current during milling is the operating current generated by the spindle during the machining of the part. According to the relevant data, the more serious the tool wear, the higher the machine tool spindle current, which is almost linearly proportional, so the machine tool spindle current signal can indirectly reflect the tool wear status. This paper uses HC33C3 current sensor to acquire spindle current signal, which is characterized by simple installation, not restricted by machining environment and relatively wide application range. At the same time, the machine tool spindle current signal is easy to obtain and can be collected directly from the machine tool, but the spindle motor interferes with the collected data at the moment of starting and braking, so the collected current signal also needs to be processed for noise reduction.

3.1.3 Noise reduction processing of the original signal

In this paper, taking the cutting vibration signal as an example, the original vibration signal is collected every Δt, and its data volume is about 200000 or more, so the data labeled 50001～100000 in each collected signal is extracted for study to avoid the interference of the noise signal, and the comparison results of the original signal and the signal after noise reduction are shown in Figure 3. Then the signal data after noise reduction are extracted in the time domain, frequency domain and time-frequency domain respectively to extract the feature quantities related to the tool wear state in order to form the sample data set available for model training.



(a)



(b)

Figure 3 Comparison results between the original signal and the noise reduction signal

(a) Raw signal data (b) Signal data after noise reduction

3.2 Extraction of tool wear characteristics

3.2.1 Time domain feature extraction scheme

The time domain characteristics of the signal are for a certain time period of the milling process without limits of expansion, and discovering and analyzing the pattern of variables of interest as they change over time. Although the acquired signal possesses a continuously changing waveform, it is difficult to extract the features related to tool wear directly from the original signal due to the high sampling frequency and the limitations imposed by frequent noise interference, so time domain analysis is required. Time domain analysis is to process the original signal with relevant parameters calculation and data analysis, so that the extracted time domain features are more representative. In this paper, in order to realize the intelligent prediction and health management of tool wear, the time domain features of the original signal are mainly divided into dimensional and dimensionless features. The dimensional time domain features can directly reflect the various changes of the milling tool machining process, mainly including five kinds of time domain features, which are absolute mean, variance, rms, peak and peak-to-peak; the dimensionless parameters are obtained by dividing the same dimension, which can avoid the interference of signal The dimensionless parameters are obtained by dividing by the same magnitude, which can avoid the interference of signal amplitude and other factors, and also can reflect other information of tool wear. The dimensionless features mainly include five time-domain features, which are skewness indicator, cliffness indicator, peak factor, coefficient of variation, and waveform factor.

3.2.2 Frequency domain feature extraction scheme

The frequency domain characteristics of a signal describe the pattern between the variables associated with the observed signal in terms of frequency, which is more profound and convenient than the time domain analysis. Fourier Transform is the most commonly used method for frequency domain analysis, which essentially converts the signal in the time domain to the frequency domain and performs tool life prediction by extracting the spectral features of the sample signal. When the wear level of the

tool changes during the milling process, the frequency components of the signal spectrum will change, so by analyzing the frequency domain features, we can accurately characterize the signal spectrum information and learn whether the tool is in a healthy state or not. The frequency domain features extracted in this project mainly include four frequency domain features: frequency mean square, frequency center of gravity, frequency variance, and peak frequency.

3.2.3 Time-frequency domain feature extraction scheme

Due to the change of geometric features or process parameters of the machined part, the signal collected by the sensor during the tool wear signal monitoring process can change instantaneously and abruptly, so the signal on the time-frequency domain needs to be analyzed. In this paper, we use wavelet packet analysis to sample the high frequency signal and low frequency signal respectively during the layer-by-layer decomposition process. After the decomposition of high and low frequency signals, so that the low and high frequency parts have the same resolution , the signal is subdivided into different frequency bands, and the frequency band structure of the monitoring signal will change with the change of tool wear state, resulting in the change of energy parameters in different frequency bands, so the energy magnitude of each frequency band is The energy level of each frequency band can accurately characterize the degree of tool wear[20] , and the energy value of the frequency band is calculated as shown in equation (3):

$$E_n(x(t)) = \frac{1}{2^{-k}N - 1} \sum_{m-0}^{2^k-1} (x^{k,m}(i))^2 \qquad (3)$$

The time domain signal is decomposed into wavelet packets according to the above principle, and the number of decomposed layers is set to 3, all done by the db5 wavelet basis, and then the decomposed signal of each layer is reconstructed by wavelet coefficients for more accurate analysis. Because of the orthogonality of the wavelet packet basis, the energy of the frequency bands can be characterized by the wavelet packet coefficients of each frequency band. After 3 layers of decomposition, the frequency domain is divided into 8 frequency bands, and thus 8 time-frequency domain features are extracted.

3.3 Construction of the sample feature matrix

In this paper, the time domain, frequency domain and time-frequency domain features are extracted from the noise reduced data, while the noise reduction process is carried out every Δt time for the original data, i.e. the original cutting vibration signal and the original spindle current signal are extracted every Δt time. The above analysis extracts 10 time-domain features, 4 frequency-domain features and 8 time-frequency-domain features, making a total of 22 feature values, thus forming a sample matrix, i.e.: from t to t + Δt time, let the noise reduced cutting vibration data set as A = {A1，A2，⋯, AN} and the spindle current data set as B ={B1，B2，⋯，BN} and assuming that the above 22 features are calculated as Fi , where i = 1, 2, .⋯ , 22; then the extracted features for cutting vibration are Xi = Fi (A); and the features for spindle current are Yi = Fi (B), where i = 1, 2, and ⋯ , 22.

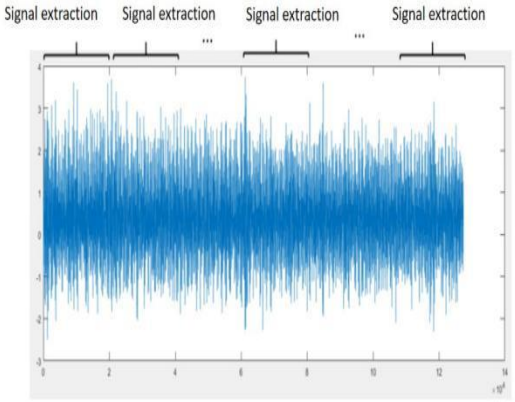


Figure 4 Extraction method of raw signal data

As shown in Figure 4, the above operation is repeated in the next Δt time, i.e. from t + Δt to t + 2Δt time, to calculate each sample feature value X, Y, until all features of all samples are extracted. However, the extraction of features in the time domain, frequency domain and time-frequency domain will have some data that are invalid and require corresponding dimensionality reduction, otherwise it will have a negative impact on the model training. For example, in the actual processing there is a spindle stall, similar to no load, or the spindle motor in the moment of starting and braking have a great impact on the collected data, which is a negative feature and should be identified and deleted. Therefore, for the sample set after feature extraction, the absolute average feature in each sample is thresholded, and if it is no-load data, stalled data or pulse data, the sample is deleted as a whole, and if it is not invalid data, the sample is retained. In this way, after screening all the samples, the remaining samples are the sample set after data processing, and each sample is the sample generated when the tool is cutting effectively. Based on this, a sample feature matrix is constructed for each signal with dimension N x 22 , the number of rows N of the matrix being the number of samples, and the structure of the cutting vibration sample eigenvalue X and the spindle current eigenvalue Y is shown in equations (4) and (5) as follows:

$$X = \begin{bmatrix} X_{1,1} & \cdots & X_{1,22} \\ \vdots & \ddots & \vdots \\ X_{N,1} & \cdots & X_{N,22} \end{bmatrix} \qquad (4)$$

$$Y = \begin{bmatrix} Y_{1,1} & \cdots & Y_{1,22} \\ \vdots & \ddots & \vdots \\ Y_{N,1} & \cdots & Y_{N,22} \end{bmatrix} \qquad (5)$$

3.4 Construction of the sample target matrix

Slight wear, normal wear, severe wear are the three major stages to characterize the degree of tool

wear during the milling process[21] .Table 1 gives the range of wear VB values of the back face of the tool in the three stages.

Table 1 Tool wear range at various stages of back tool face

Type	Wear phase	Rear tool face wear V_B values
1	Slight wear and tear	0-0.1mm
2	Normal wear and tear	0.1-0.5 mm
3	Rapid wear and tear	0.5mm or more

The high pressure and temperature between the rear face of the tool and the machined surface during the milling process of the 5-axis machining center causes its rear face to wear faster and reach the dullness standard before the front face, so this paper mainly uses the electron microscope to measure the wear value VB in the rear face area of the tool. The measurement is performed by sampling every Δt time and corresponds to the sample characteristics, and the magnitude of the rear face wear value VB at each moment is the sample target value of tool wear. Each sample target value can correspond to the wear stages in Table 1 to construct the sample target matrix Q. The dimension of the target matrix sample Q is N × 1 , which mainly contains three types of minor wear (set as label 1), normal wear (set as label 2) and sharp wear (set as label 3). The time domain, frequency domain, and time-frequency domain features of the cutting vibration signals and spindle current signals of the extracted tool under different wear states are fused with the target sample matrix Q using the multi-source information fusion technique, and finally a sample data set M is obtained, whose data set M is shown in equation (6):

$$M = \begin{bmatrix} X_{1,1} & \cdots & Y_{1,22} & \cdots & Q_1 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{N,1} & \cdots & Y_{N,22} & \cdots & Q_N \end{bmatrix} \tag{6}$$

Since the measured wear value of the back tool face is a few moments, and the tool wear value is a continuous curve, the coordinates of the actual wear value can be interpolated to generate a cubic polynomial fit curve, and the reliability of the sample data set can be verified by comparing it with the tool wear curve. In this paper, a cubic polynomial is used for the interpolation calculation, as shown in equation (7):

$$y(t,\omega) = \sum_{j=0}^{3} \omega_j t^j \tag{7}$$

Where ω_j ist^j the coefficient, y is the interpolated tool wear value, and t is the time. For the tool wear values y_i collected at time x_i, a total of N times were collected, the loss function of the cubic polynomial interpolation curve is shown in equation (8):

$$E_n(\omega) = \frac{1}{2}\sum_{i=0}^{N}[y(t_i,\omega) - y_i]^2 \tag{8}$$

And the difference curve coefficient ω_j can be found by calculation, which is shown in equation (9):

$$min\frac{1}{2}\sum_{i=0}^{N}[y(t_i,\omega) - y_i]^2 \tag{9}$$

The fitted curve of the cubic polynomial derived from the above calculation is shown in Figure 5. The fitted curves show that the tool wear is faster at the early stage, smoother when it enters the middle stage, and faster at the later stage, which is consistent with the curve situation of tool wear. The results show that this sample data set M can effectively characterize the tool wear state at each moment, and can be used as the input to the PSO-CNN model.

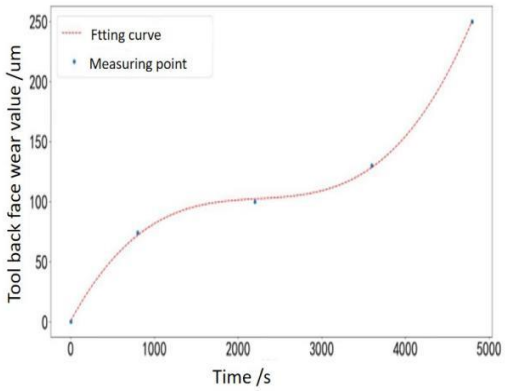


Figure 5 Cubic polynomial interpolation of tool wear curves

4 Experimental verification and analysis of tool wear

4.1 Structural parameters of CNN network model

In this experimental model, two hyperparameters, batch size and Epoch number , are selected as the object of the optimization process. To avoid the influence of external factors, the number of particle swarm individuals in the PSO algorithm is set to 10 and the maximum number of iterations is set to 50, as shown in Table 2. The optimized CNN model batch size parameter is set between 300 and 500, and the Epoch number is set between 5 and 15. The optimization was performed according to the parameter settings in Table 2, resulting in the best combination of hyperparameters with a batch size of 330 and an Epoch number of 10 iterations.

Table 2 Initial parameter settings for the PSO algorithm

PSO algorithm parameters	Parameter values
Number of individuals in the particle population	10
Maximum number of iterations	50
Cognitive factors c1, c2	2, 2
Inertia factor	0.5
Particle vector dimension	2

The optimized parameters of the PSO algorithm were input to the CNN model for tool wear prediction, and the specific parameters of the CNN network model were set as shown in Table 3. Table 3 shows that the CNN network structure contains two convolutional layers, two pooling layers and one fully connected layer. In order to improve the prediction performance of the model, the training process uses the RELU function for nonlinear activation, which has good non-saturation characteristics and can avoid the gradient disappearance phenomenon, and the activation function is shown in equation (10):

$$V_i^k = Relu\left(X_i^k\right) = \begin{cases} 0, x_i^k < 0 \\ x_i^k, x_i^k > 0 \end{cases} \tag{10}$$

where x_i^k is theX_i^k each eigenvalue in the feature matrix.

Table 3 CNN network structure parameters

Structural section	Network structure Name	Parameter settings
1	Convolutional layer 1 Batch standardisation layer 1 Pooling layer 1	Activation function: RELU Convolution kernel: 3*3 Maximum pooling
2	Convolutional layer 2 Batch standardisation layer 2 Pooling layer 2	Activation function: RELU Convolution kernel: 3*3 Maximum pooling
3	Dropout layer	25% discard
4	Output layer	Activation function: Softmax

In order to quantify the results of tool wear status monitoring, the precision, accuracy, recall, and F1-score values are selected as evaluation indexes in this paper, and the precision (Precision), accuracy (Accuracy), recall (Recall), and F1 values (F1-score) are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{14}$$

In the above equation, the values of TP, TN, FP, and FN can be found in the confusion matrix, and the confusion matrix is shown in Table 4 for the dichotomy example.

Table 4 Confusion Matrix

		True Value	
		1	0
Predicted value	1	TP	FP
	0	FN	TN

4.2 Prediction results of PSO-CNN model

In this paper, we use the convolutional neural network architecture based on particle swarm optimization for tool state recognition training, and it can be seen from the accuracy and loss function of the model in Figure 6: the accuracy of the model shows an increasing trend during the first 50 iterations, and then the accuracy gradually increases.

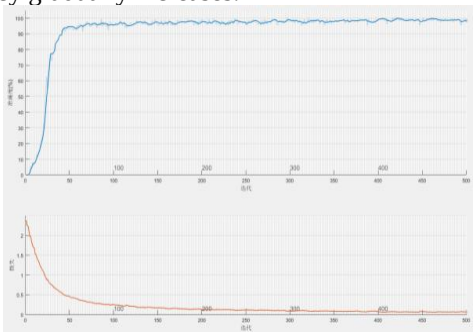


Figure 6 Accuracy and loss function graph

The sample data set M is randomly divided, and the first 200 samples are used as training sets to train the constructed PSO-CNN model. The predicted results of the training set are shown in Figure 7. which shows that only 2 out of 200 training samples were incorrectly identified, with an accuracy of 99.13%. The remaining samples are used as a test set to test the model, and the predicted results of the test set are shown in Figure 8. It can be found that only 2 out of 116 test samples were identified incorrectly, and the accuracy of the test set reached 98.28%; the results show that the PSO-CNN model constructed in this paper can effectively identify the tool wear status and achieve better results.

The confusion matrix of the PSO-CNN tool condition monitoring model test set is shown in Figure 9, which shows that the test set contains 42 samples of slight wear (label 1), 31 samples of normal wear (label 2) and 43 samples of sharp wear (label 3). The model proposed in this paper identifies all the slight wear samples correctly when testing them, and the test accuracy reaches 100%; when testing the normal wear samples, one sample is incorrectly identified as slight wear, and the accuracy is 96.8%; when testing the sharp wear samples, one sample is incorrectly identified as normal wear, and the accuracy is 97.7%.

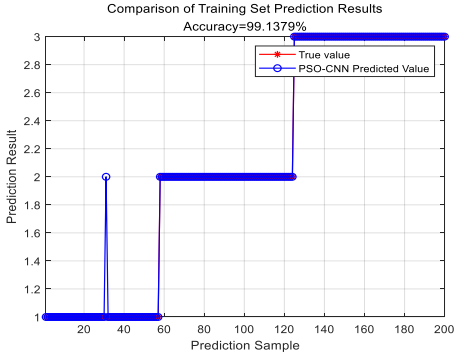


Figure 7 Training set prediction results

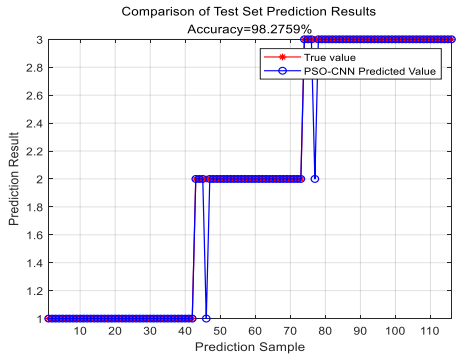


Figure 8 Test set prediction results

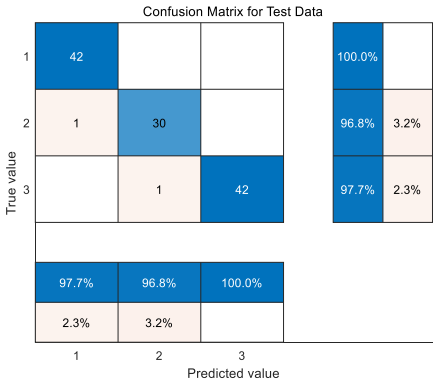


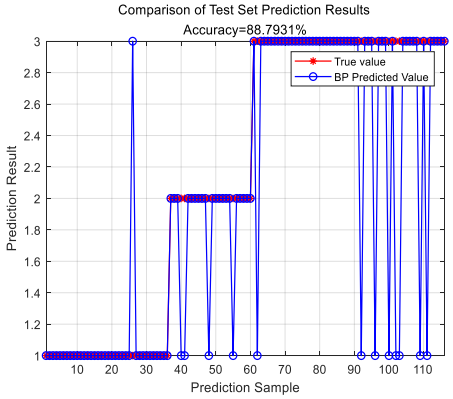
Figure 9 Confusion matrix

The evaluation indexes of tool condition monitoring can be calculated through the confusion matrix, and the results of the evaluation indexes of its three condition labels are shown in Table 5. For the accuracy rate index, it can be seen that the slight wear has the highest accuracy rate, and its performance is ranked as slight wear ＞ sharp wear ＞ normal wear; for the correct rate index, it can be seen that the accuracy rate of all three wear states is 98%, which is consistent with the previous analysis; for the recall rate index, it can be seen that the recall rate of slight wear and normal wear does not reach 100%, which indicates that there are other wear states incorrectly identified These four results further verify the effectiveness of the PSO-CNN tool state recognition model.

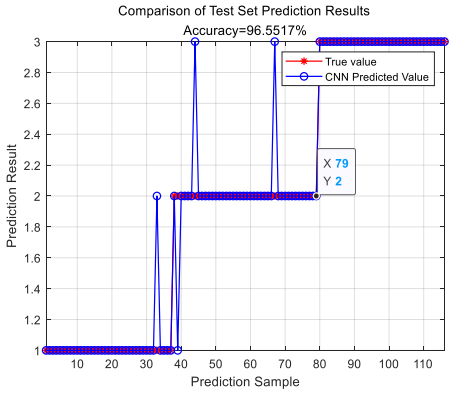
Table 5 Results of three tool condition evaluation indexes

Label classification	Accuracy rate	Accuracy	Recall Rate	F1 value	Test samples	Sample error
Slight wear and tear	1	0.98	0.98	0.99	42	0
Normal wear and tear	0.97	0.98	0.97	0.97	31	1
Rapid wear and tear	0.98	0.98	1	0.99	43	1

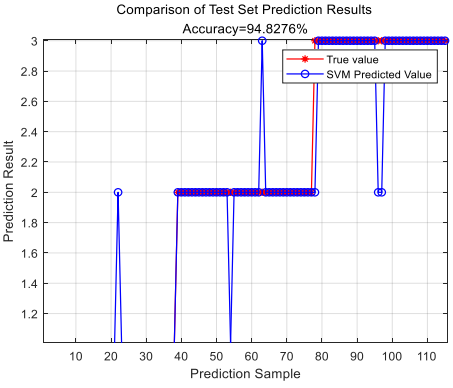
In order to further verify the recognition performance of PSO-CNN model tool wear status, a comparative analysis was performed with other traditional recognition models in the past, such as BP neural network, CNN convolutional neural network, and SVM support vector machine, and the prediction results of these three traditional tool wear status recognition models are shown in Figure 10. From Fig. 8 and Fig. 10, it can be seen that the prediction effects of the four tool wear state recognition models are ranked as PSO-CNN model ＞ CNN model ＞ SVM model ＞ BP model. It can be seen that the CNN model optimized based on the PSO algorithm proposed in this paper has obvious advantages in tool wear state recognition because the CNN network in the PSO-CNN model can perform deep mining of the hidden layer features using convolution and pooling operations, and the PSO algorithm is able to match the two hyperparameters of batch size and Epoch count in the CNN network for seeking the best, thus avoiding the blindness of setting parameters, thus improving the accuracy of the prediction model.



(a)



(b)



(c)

Figure 10 Prediction results of the three traditional models

BP model. (b) CNN model.(c) SVM mode

Table 6 shows the performance comparison results of the four tool wear state recognition models. The number of error samples identified by the BP model, SVM model and CNN model are 13, 6 and 4, respectively, and their accuracy rates are 88.79%, 94.83% and 96.55%, respectively. In contrast, the PSO-CNN model proposed in this paper identifies only 2 incorrect samples, and the accuracy rate is as high as 98.27%, which is 9.48%, 3.44%, and 1.72% higher than the above three traditional models respectively. This shows that the prediction accuracy of the tool wear status of the CNN model optimized based on the PSO algorithm is significantly higher than other models under the conditions of the same number of samples, and its generalization ability is stronger and the network fitting speed is faster, which indicates that the prediction of the tool wear status using the PSO-CNN model is more accurate and can more effectively realize the tool status monitoring and intelligent tool change during the milling process of the 5-axis machining center. This shows that using PSO-CNN model to predict the tool wear status will be more accurate and can more effectively realize the tool condition monitoring and intelligent tool change in the five-axis machining center.

Table 6 Performance comparison results of four prediction models

Network Model	Number of misidentified samples			Accuracy
	Minor Wear and tear	Normal wear and tear	Rapid Wear and tear	
BP Neural Networks	1	4	8	88.79%
SVM Support vector machines	1	2	3	94.83%
CNN Convolutional Neural Networks	1	3	0	96.55%
PSO-CNN Hybrid model	0	1	1	98.27%

5 Conclusion

In this paper, firstly, cutting vibration signals and spindle current signals are collected, and data features characterizing tool wear are extracted in the time domain, frequency domain and time-frequency domain; secondly, the tool wear values corresponding to the above features are measured by electron microscopy, and they are divided into three categories according to wear values: slight wear, normal wear and sharp wear, and the construction of sample data sets is completed; finally, the PSO-CNN model proposed in this paper is used to complete classification and prediction of tool wear status and compare and analyze with other models, the results show that:

(1) Parameter search optimization of CNN convolutional neural network by PSO algorithm yields the best combination of hyperparameters with a batch size of 330 and an Epoch count of 10. The blindness of setting parameters is avoided, thus improving the model prediction accuracy and precision.

(2) The prediction accuracy of the PSO-CNN model constructed in this paper reaches 98.27%, which can meet the requirements of monitoring the tool wear status and can realize the predictive maintenance of CNC machining tools, that is, intelligent tool change before the tool wear is in sharp wear .

(3) Comparing the prediction performance of the PSO-CNN model constructed in this paper with BP neural network, CNN convolutional neural network and SVM support vector machine, the results show that the PSO-CNN prediction model constructed in this paper has obvious advantages in the field of tool wear condition identification, and its accuracy indexes are improved by 9.48%, 3.44% and 1.72% respectively compared with other models.

In the future, this PSO-CNN tool wear state prediction model can be widely used in the fields of tool life prediction and intelligent operation and maintenance of CNC machine tools in various factories. By monitoring the cutting vibration signal and spindle current signal of the tool system in real time, the prediction of different tool wear states can be realized, and based on the prediction results the machine tool can make intelligent judgments and make corresponding processing, so as to improve the product machining quality and reduce the scrap rate, which has certain practical significance.

Author Contributions: For Conceptualization, methodology, analysis, and writing original draft preparation, Wang Shuo; writing review and full-text editing, Yu Zhenliang; writing — original draft preparation,Lu Changguo; writing — original draft preparation, Wang Jingbo.

Conflicts of interest: The authors declare no conflict of interest.All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The research work financed with the means of Basic Scientific Research Youth Program of Education Department of Liaoning Province, No.LJKQZ2021185; Yingkou Enterprise and Doctor Innovation Program (QB-2021-05).

References

[1] Danil Yu Pimenov. Andres Bustillo, Szymon Wojciechowski, et al. Artificial intelligence systems for tool condition monitoring in machining: analysis and critical review [J]. Journal of Intelligent Manufacturing,2022(3):34.

[2] Bagri Sumant, Manwar Ashish, Varghese Alwin, et al. Tool wear and remaining useful life prediction in micro-milling along complex tool paths using neural networks [J]. Journal

of Manufacturing Processes,2021(1):71.
[3] Olalere Isaac Opeyemi, Olanrewaju Oludolapo Akanni. Tool and Workpiece Condition Classification Using Empirical Mode Decomposition with Hilbert–Huang Transform of Vibration Signals and Machine Learning Models [J]. Applied Sciences,2023,13(4):78-79.
[4] Wang X, Zheng Y, Zhao Z, et al. Bearing fault diagnosis based on statistical locally linear embedding [J]. Sensors, 2015,15(7):16225-16247.
[5] Guan Shan Kang, Zhenxing Peng Chang. Analysis on cloud characteristics of wear acoustic emission signal for vehicle cutting tool [J]. Editorial Office of Transactions of the Chinese Society of Agricultural Engineering,2016,32(20):97-99.
[6] Jeon J.U, Kim S.W. Optical flank. wear monitor ing of cutting tools by image processing [J].wear, 1988,127(2): 207-217.
[7] Pyatykh A. S., Savilov A. V., Timofeev S. A.. Method of Tool Wear Control during Stainless Steel End Milling [J]. Journal of Friction and Wear,2022, 42(4):9-11.
[8] Wang Zhan, Leng Sheng, Min Tao, et al. Analysis of AE characteristics of tool wear in drilling CFRP/Ti stacked material [J]. MATEC Web of Conferences,2018(4):211.
[9] Han, Chengwen, Kim, Kyeong Bin, et al. Thrust Force-Based Tool Wear Estimation Using Discrete Wavelet Transformation and Artificial Neural Network in CFRP Drilling [J]. International Journal of Precision Engineering and Manufacturing, 2021 (1):898-899.
[10] Soufiane Laddada, Med. Ouali Si-Chaib, Tarak Benkedjouh, et al. Tool wear condition monitoring based on wavelet transform and improved extreme learning machine [J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2020, 234(5):1467-1468.
[11] Liang Yu, Hu Shanshan, Guo Wensen, et al. Abrasive tool wear prediction based on an improved hybrid difference grey wolf algorithm for optimizing SVM [J]. Measurement, 2022(1):187.
[12] Caesarendra Wahyu, Triwiyanto Triwiyanto, Pandiyan Vigneashwara, et al. A CNN Prediction Method for Belt Grinding Tool Wear in a Polishing Process Utilizing 3-Axes Force and Vibration Data [J]. Electronics, 2021, 10(12):76-78.
[13] Xin Cheng Cao, Bin Qiang Chen, Bin Yao,, et al. Combining translation-invariant wavelet frames and convolutional neural network for intelligent tool wear state identification [J]. Computers in Industry,2019(1):106.
[14] Stefan Droste, Thomas Jansen, Ingo Wegener. Upper and Lower Bounds for Randomized Search Heuristics in Black-Box Optimization. Electron [J]. Colloquium Comput. Complex, 2003(2):48-48.
[15] Weifeng Lu, Bingyu Cai, Rui Gu. Improved Particle Swarm Optimization Based on Gradient Descent Method [J]. CSAE, 2020(1): 121-126.
[16] Salih Omran, Duffy Kevin Jan. Optimization Convolutional Neural Network for Automatic Skin Lesion Diagnosis Using a Genetic Algorithm [J]. Applied Sciences, 2023,13(5):56-57.
[17] Zhang Xin, Jiang Yueqiu, Zhong Wei. Prediction Research on Irregularly Cavitied Components Volume Based on Gray Correlation and PSO-SVM [J]. Applied Sciences, 2023,13(3):79-87.
[18] Shi Jun, Zhang Yanyan, Sun Yahui, et al. Tool life prediction of dicing saw based on PSO-BP neural network [J]. The International Journal of Advanced Manufacturing Technology, 2022(123):11-12.
[19] Gajera Himanshu K., Nayak Deepak Ranjan, Zaveri Mukesh A. A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features [J]. Biomedical Signal Processing and Control,2023,79(P2).
[20] Wu Shun Xing, Li Peng Nan, Yan Zhi Hui, Zhang Li Na, Qiu Xin Yi, Yang Jin. Wavelet Packet analyses of Acoustic Emission Signal for Tool Wear in High Speed Milling [J]. Key Engineering Materials,2013(1):589-590.
[21] Liang Junhua, Gao Hongli, Xiang Shoubing, et al. research on tool wear morphology and mechanism during turning nickel- based alloy GH4169 with PVD-TiAlN coated carbide tool [J]. Wear, 2022(1):508-509.

Viser Technology Pte. Ltd.
Mechanical Engineering Science
DOI: 10.33142/mes.v4i2.9084

Research Article Open Access

Fault monitoring and diagnosis of motorized spindle in five-axis Machining Center based on CNN-SVM-PSO

Shuo WANG[1], Zhenliang YU[1*], Xu LIU[2], Zhipeng LYU[3]

1. School of Mechanical and Power Engineering, Yingkou Institute of Technology , Yingkou, China
2. Yingkou Dingsheng Heavy Industry Machinery Co. LTD, Yingkou, China
3. School of Mechanical and Power Engineering, Shenyang University of Chemical Technology , Shenyang, China

***Corresponding Author:** yuzhenliang, email address: yuzhenliang_neu@163.com*

Abstract:
A spindle fault diagnosis method based on CNN-SVM optimized by particle swarm algorithm (PSO) is proposed to address the problems of high failure rate of electric spindles of high precision CNC machine tools, while manual fault diagnosis is a tedious task and low efficiency. The model uses a convolutional neural network (CNN) model as a deep feature miner and a support vector machine (SVM) as a fault state classifier. Taking the electric spindle of a five-axis machining centre as the experimental research object, the model classifies and predicts four labelled states: normal state of the electric spindle, loose state of the rotating shaft and coupling, eccentric state of the motor air gap and damaged state of the bearing and rolling body, while introducing a particle swarm algorithm (PSO) is introduced to optimize the hyperparameters in the model to improve the prediction effect. The results show that the proposed hybrid PSO-CNN-SVM model is able to monitor and diagnose the electric spindle failure of a 5-axis machining centre with an accuracy of 99.33%. In comparison with the BP model, SVM model, CNN model and CNN-SVM model, the accuracy of the model increased by 10%, 6%, 4% and 2% respectively, which shows that the fault diagnosis model proposed in the paper can monitor the operation status of the electric spindle more effectively and diagnose the type of electric spindle fault, so as to improve the maintenance strategy.

Keywords: five-axis machining centres; CNN-SVM; spindle vibration; fault diagnosis

1 Introduction

Five-axis machining centre is a high technology, high efficiency, low energy consumption in one of the high-precision machine tools, widely used in the complex space surface processing, its core key components failure of intelligent identification to enhance the overall level of equipment maintenance technology is of great significance. The electric spindle is directly driven by an electric motor instead of a pulley drive and gear drive, which can achieve high-speed and steady-state operation of the machine tool spindle, and is a key functional component of the five-axis machining centre, whose working condition directly affects the spindle rotation accuracy and product processing quality[1]. It is a key functional component of a five-axis machining centre. Therefore, effective monitoring and accurate diagnosis of spindle faults is essential. Monitoring means timely warning when a spindle fault occurs, and diagnosis means intelligent identification of the type of fault for accurate maintenance at a later stage. Fault detection and diagnosis models are used to monitor and mine the vibration signals of each fault in the spindle and to construct a non-linear correlation with the actual fault. In the early days, a large number of scholars used machine learning methods to build prediction models for intelligent maintenance of motorized spindle, such as BP neural networks [2],RBF neural networks[3], Support vector machines (SVM)[4] etc.

Li Zhaolong [2] et al. collected temperature and axial thermal drift data of electric spindles at different rotational speeds, used fuzzy clustering and grey correlation analysis for feature extraction, and constructed a BAS-BP model to predict and compensate for the thermal errors of electric spindles, achieving better results. Shan Wentao [3] et al. proposed a block adaptive backstepping control method based on global RBF neural network. The backstepping control law and parameter update law were derived using Lyapunov theory to ensure the stability of the whole spindle system. C.K. Madhusudana [4] et al. collected vibration signals in the feed direction of the spindle in the healthy and faulty states of the milling cutter and used SVM models with different kernel functions to investigate and classify

Received on October 2, 2022; Accepted on December 16, 2022

selected features based on the discrete wavelet transform, and the results showed that spindle faults could be effectively diagnosed using this C- SVC model. Although the above-mentioned scholars have made some achievements using mechanical learning algorithms, the slow model fitting speed and low prediction accuracy have become urgent problems at this stage.

With the application of sensor technology and the rise of deep learning algorithms, has become a new trend to use piezoelectric acceleration sensors to acquire electric spindle vibration signals and construct fault diagnosis models for monitoring by deep learning algorithms, such as recurrent neural networks (RNN) [5], long and short term memory networks (LSTM) [6] and Convolutional Neural Networks (CNN) [7] etc. These predictive models have more powerful feature learning and mapping capabilities and can automatically mine deeper features for prediction without a priori knowledge or the help of human experts. However, recurrent neural networks (RNN) are prone to gradient disappearance or gradient explosion when diagnosing spindle faults, and researchers have used Long and Short Term Memory Networks (LSTM) to predict spindle faults [8]. Convolutional neural networks (CNNs) have been used for spindle fault monitoring and diagnosis in recent years because their convolution and pooling operations can improve the extraction of potential features in the hidden layer of the prediction model compared to LSTMs.

Wen Long[9] et al. proposed a CNN convolutional neural network for electric spindle bearing fault diagnosis, which can effectively perform fault monitoring, but there is still room to improve the accuracy of diagnosing specific fault types. This is due to the fact that when using a CNN diagnostic model to deal with functions with a high degree of non-linearity, the number of features output by the fully connected layer increases proportionally, reducing the generalisation capability of the model, which is not conducive to fault diagnosis of electric spindles. Support vector machines (SVMs), on the other hand, have an absolute advantage in dealing with non-linear data by using some kernel function to transform the input sample data from a low-dimensional space into a high-dimensional space, so that the originally non-linear data becomes linearly separable in the high-dimensional space[10] It uses a kernel function to transform the input sample data from a low-dimensional space to a high-dimensional space, so that the originally non-linear data becomes linearly separable in the high-dimensional space. Therefore, the combination of SVM and CNN can make up for the shortcomings of the above CNN model. The essence is that CNN is used as a feature learner to explore the deep features of the input data, and SVM is used as a trainer to construct the optimal classification hyperplane for fault classification prediction.

CNN-SVM is a multi-category diagnostic model proposed by combining convolutional neural network (CNN) and support vector machine (SVM) methods. Its model performance depends on the selection of model parameters, which include penalty parameters ρ and kernel function width g, etc., and it is crucial to select the optimal parameter pairing to further improve the model performance. The current more common hyperparameter optimisation methods are random optimisation search [11] , gradient-based optimisation[12], genetic algorithm optimisation[13], Particle swarm optimization[14]et al. The PSO algorithm can perform global optimization with fewer parameters, and its powerful search performance and individual optimization capability can accelerate the convergence speed of the model, so it has been widely used and studied by scholars in recent years [15] .This is why it has been widely studied in recent years.

In this paper, a hybrid CNN-SVM model based on particle swarm algorithm (PSO) optimisation is proposed. Firstly, the fully connected layer of the CNN model is replaced by a global average pooling layer to reduce the dimensionality of the output features and improve the generalisation capability of the model; secondly, the Softmax function of the CNN model is replaced by a support vector machine SVM classifier to complete the fault diagnosis of the electric spindle; finally, the hyperparameters in the SVM model are optimised using the PSO algorithm to derive the optimal solution to further improve the Finally, the PSO algorithm is used to optimise the hyperparameters in the SVM model and derive the optimal solution to further improve the fault diagnosis accuracy of electric spindles.

2 Construction of a CNN-SVM-PSO fault diagnosis method

To address the shortcomings of the CNN diagnosis model, this paper proposes a fault diagnosis model based on a CNN-SVM optimised by a particle swarm algorithm to identify the types of faults in the electric spindle system of a 5-axis machining centre. The improvements are:

(1) The sample feature matrix is pre-processed using batch normalisation techniques and then input into the CNN model, which reduces the complexity of the model and improves the convergence speed of the network with its unique structure of local connectivity and weight sharing.

(2) The fully connected layer of the CNN model is replaced by a global average pooling layer, and the features output after the convolution and pooling operations are reduced in dimensionality, which reduces the model parameters and lowers the training time of the SVM model.

(3) The SVM model is suitable for classification tasks dealing with problems with high non-linearity and makes up for the shortcomings of the CNN model, so the SVM model is used instead of the Softmax classifier in the CNN to classify and predict the electric spindle fault types, thus improving the model generalisation capability.

(4) Using the powerful search and global optimization-seeking capabilities of the PSO algorithm, the penalty parameter in the SVM modelρ and the two

parameters of kernel function width g are iteratively optimized to improve the accuracy of electric spindle fault diagnosis. The CNN-SVM-PSO fault diagnosis model is shown in Figure 1.

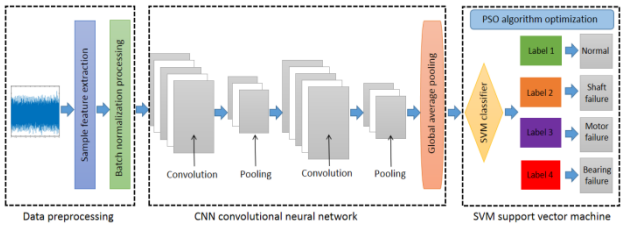


Figure 1 CNN-SVM-PSO fault diagnosis model

2.1 Acquisition of electric spindle vibration signals

During the operation of a five-axis machining centre, the electric spindle system will generate violent vibrations when problems occur in the core components such as the rotating shaft, motor and bearings, which are manifested by the loose and unbalanced phenomenon of the rotating shaft and coupling, the eccentric phenomenon of the air gap of the motor, as well as the damage failure of the bearings and rolling bodies. By monitoring the vibration of the machine tool spindle system when the above core components are abnormal, it is found that the frequency range of the vibration signals of various faults are slightly different, as shown in Table 1, so the spindle fault can be diagnosed by extracting the features of each fault vibration signal and finding the correlation between the sample features and the actual fault[16]. The sample features can then be correlated with the actual fault to diagnose the spindle fault.

Table 1 Frequency range of core component failures

Electric spindles Type of fault	Frequency range	Type of vibration
Unbalanced and loose rotating shafts and couplings	5 times Within working frequency	Low frequency vibration
Motor air gap eccentricity failure	2x Power frequency	Medium Frequency Vibration
Bearings and rolling elements Injuries	＞1KHz	High frequency vibration

The vibration information generated by the electric spindle system of the five-axis machining centre due to the above faults will be reflected in different ways, such as irregular fluctuations of the spindle motor current, the vibration of the outer casing of the spindle and the noise generated by the electric spindle system. The experiment is to use the 356A15 three-axis vibration acceleration sensor manufactured by PCB to monitor and collect the vibration signal generated by the outer casing of the spindle in real time under the high-speed rotating state of the electric spindle, the Measuring system models is shown in Figure 2, the Experimental equipment model parameters is shown in Table 2.

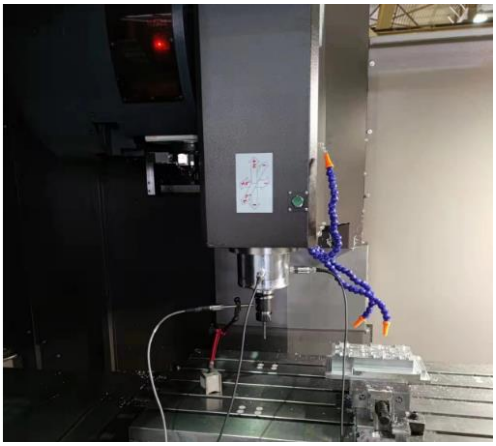


Figure 2 Measuring system models

Table 2 Experimental equipment model parameters

Serial number	Experimental equipment	Model parameters
1	Five-axis machining centres	SK5L-70100 i5M8
2	Acceleration sensors	PCB, Type 356A15
3	Data Acquisition Cards	NI-DAQ, 50HZ
4	Output Connector	BNC interface

In this paper, the raw vibration signals of the electric spindle system are collected in real time according to the above scheme. A total of four tag states are collected: normal (set as tag 1), spindle fault (set as tag 2), motor fault (set as tag 3) and bearing fault (set as tag 4). The number of samples collected for each of the four tag states is 100, giving a total of 400 data. As the raw signal data set collected contains 3 channels of X-axis, Y-axis and Z-axis vibration signals, a raw signal matrix of 400 x 3 is formed.

2.2 Electric spindle fault feature extraction

The instability of the five-axis machining centre electric spindle system at the moment of start/stop can interfere with the signal feature extraction, so the original signal needs to be processed for noise reduction. This experiment each acquisition signal data volume is about 200000 or more, so extract each acquisition signal in the label for 50001 ～ 100000 data for research, in order to avoid the interference of the noise signal, to the normal state of the data set as an example, its noise reduction signal results are shown in Figure 3.

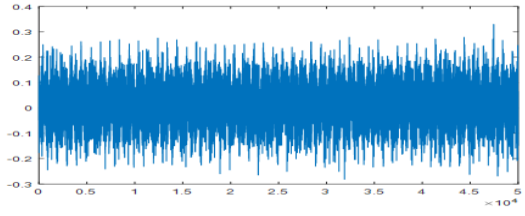


Figure 3 Spindle vibration signal data after noise reduction

After noise reduction, the original vibration signal is extracted in the time domain, frequency domain and time-frequency domain. 13 time-domain features are extracted in total, including mean value, variance, cliff index, peak factor, etc.; 5 frequency-domain features are extracted, including frequency-domain amplitude mean value, mean square frequency, variance frequency, etc.; the time-frequency domain features are extracted mainly by using wavelet packet analysis to subdivide the original signal into different frequency bands, and the energy value of each frequency band is the extracted time-frequency domain features. The energy value of each frequency band corresponds to the type of electric spindle fault, so the energy value of the frequency band is the extracted time-frequency domain features, and the energy value of the frequency band is calculated by the formula:

$$E_n\left(x(t)\right) = \frac{1}{2^{-k}N - 1}\sum_{m-0}^{2^k-1}\left(x^{k,m}(i)\right)^2 \qquad (1)$$

where En denotes the total energy of the original signal, j denotes the number of layers of wavelet packet decomposition, and$x^{k,m}(i)$ denotes the number of layers in the subspace $U_{j-k}^{2^k+m}$ of the signal x_{2^k+m} of the decomposed signal. In this experiment, the number of layers of wavelet packet decomposition of the original signal is set to 3, which are all done by the db5 wavelet base. The frequency domain is divided into 8 frequency bands, as shown in Figure 4, so that 8 time-frequency domain features are extracted. Therefore, 26 features can be extracted for each channel signal. The features of all channels are fused to produce 78 eigenvalues and the matrix is reorganised to produce a 400 x 78 eigenmatrix, which is the input to the electric spindle fault diagnosis model.

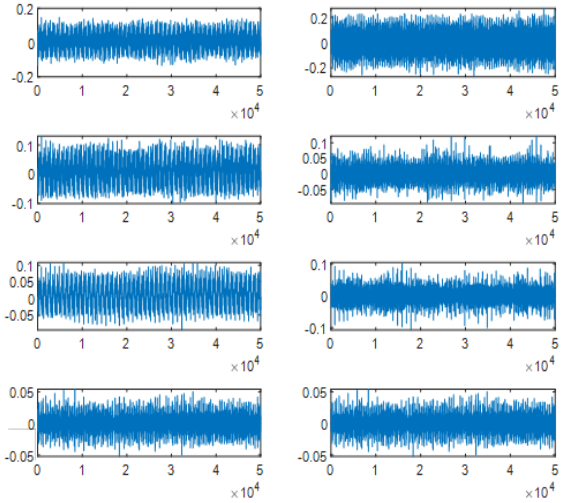


Figure 4 Frequency bands for wavelet packet decomposition

2.3 Fault diagnosis principle of CNN-SVM-PSO model

Firstly, the 400 x 78 sample feature matrix is reorganised using the batch normalisation technique; secondly, the sample data is input into the CNN model and passed into the global average pooling layer for feature dimensionality reduction after two successive convolution and pooling operations; finally, the reduced dimensional feature vector is passed into the SVM model optimised by the PSO algorithm for electric spindle fault diagnosis. The specific fault diagnosis principle is as follows:

According to the above, the 400×78 sample feature matrix was derived from the feature extraction of the original vibration signals of the four labels in the time domain, frequency domain and time-frequency domain, and the above feature matrix was batch normalized to avoid the occurrence of overfitting due to gradient dispersion, and the processing formula for batch normalization was

$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (2)$$

where X denotes the sample for each feature, andX_{min} denotes the minimum value of each feature, andX_{max} denotes the maximum value of each feature.

The sample information is indirectly characterized by the weight value of each layer derived from the convolution operation, the higher the layer, the more detailed the local features are extracted, and the spatial continuity of the sample is maintained[17] .The convolution operation is given by

$$X_i^k = \sum_{j=1}^{n} W_i^{kj} \otimes X_{i-1}^j + b_i^k \qquad (3)$$

where X_i^k denotes the feature matrix of the kth neuron at the output of the ith layer, andW_i^{kj} denotes the weight value of the kth neuron at layer i, and$\otimes$ denotes the convolution operator, andX_{i-1}^j denotes the feature matrix of the jth neuron at the output of layer i-1, andb_i^k is the bias coefficient of the kth neuron in layer i.

In order to improve the fault diagnosis performance of the prediction model, the CNN model uses ReLU function for non-linear activation, which has good non-saturation characteristics and avoids the gradient disappearance phenomenon. The activation function is as follows:

$$V_i^k = Relu\left(X_i^k\right) = \begin{cases} 0, x_i^k < 0 \\ x_i^k, x_i^k > 0 \end{cases} \qquad (4)$$

Where x_i^k is the value of theX_i^k the respective eigenvalues in the feature matrix.

The pooling type is chosen to be maximum pooling, which preserves the original features and reduces the parameters of network training, improving the robustness of the extracted features. The maximum pooling formula is:

$$C_i^k(s,t) = \underset{\substack{1+(s-1)Q \le d \le sQ \\ 1+(t-1)P \le h \le tP}}{Max}\left\{V_i^k(d,h)\right\} \qquad (5)$$

where$V_i^k(d,h)$ is the eigenvalue of column h of row d of the ith eigenmatrix input to the pooling layer, and$C_i^k(s,t)$ is the eigenvalue of the sth row t column of the ith feature matrix obtained after pooling, and P and Q are the length and width of the pooled region, respectively.

The feature matrices of dimension S × T, which are derived from each row of the 400 × 78 sample feature matrix after two convolution and pooling operations, are fed into the global average pooling layer. The dimensionality of the pooling kernel of the global average pooling layer is kept consistent with the dimensionality of the feature matrix, and the n feature matrices are dimensionalized to output a feature vector Xr = {x1 , x2 , ... , xi , ... , xn , }, where xi is given by the formula

$$x_i = \frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T} C_i^k(s,t) \tag{6}$$

The feature vector output from the global average pooling layer is used as input to the SVM support vector machine model. The greatest advantage of the SVM algorithm is that the number of features in a dataset has essentially no effect on its model complexity, making it particularly suitable for classification tasks with relatively large datasets of features, and the mathematical model of the SVM is

$$\begin{cases} min\ \frac{1}{2}\|w\| + \rho\sum_{r=1}^{L}\xi_r \\ s.t.\ y_r(wX_r + b) + \xi_r \geq 1, r = 1,2, \cdots, L \end{cases} \tag{7}$$

where w is the normal vector to the hyperplane, andρ is the penalty parameter, the ξ_r is the relaxation factor, b is the offset coefficient, andX_r is the feature vector of the rth sample, the y_r is the fault class, L is the total number of feature samples, and the total number of samples in this paper is 400.

The model in Eq. (7) is mostly used to deal with linearly divisible sample characteristics data, but the electric spindle fault sample data is linearly indivisible, so it is necessary to introduce the kernel function to up-dimension each labeled sample data. In this paper, the Gaussian radial basis kernel function is used to transform the non-linear data of each labeled state into linear data in high dimensional space to make the analysis possible, and then the optimal classification hyperplane is constructed based on the principle of maximizing the classification interval to complete the fault diagnosis task, and its Gaussian radial basis kernel function formula is

$$K(X) = sgn\left(\sum_{r=1}^{L} a_r^* y_r exp\left(-\frac{\|X_r - X\|^2}{2g^2}\right) + \theta^*\right) \tag{8}$$

where sgn is the sign function, a_r^* is the Lagrangian multiplier, g is the kernel function width, and X is the sample label data, and θ^* is the configuration factor.

The five-axis machining centre spindle fault diagnosis has a total of four label states, in essence a multi-classification problem. In the fault diagnosis of the sample, each classifier scores the four label states and the label with the highest score is the final result of the fault diagnosis. The penalty parameterρ and kernel function width g directly affect the training speed and prediction accuracy of the model, so how to find the optimal ρ , g parameter pairing is the key to SVM model classification prediction [18]. This paper uses the PSO algorithm to perform the SVM classification prediction. In this paper, the PSO algorithm is used to optimise the hyperparameters in the SVM model to derive the optimal solution, and its PSO algorithm optimisation search process is shown in Figure 5.

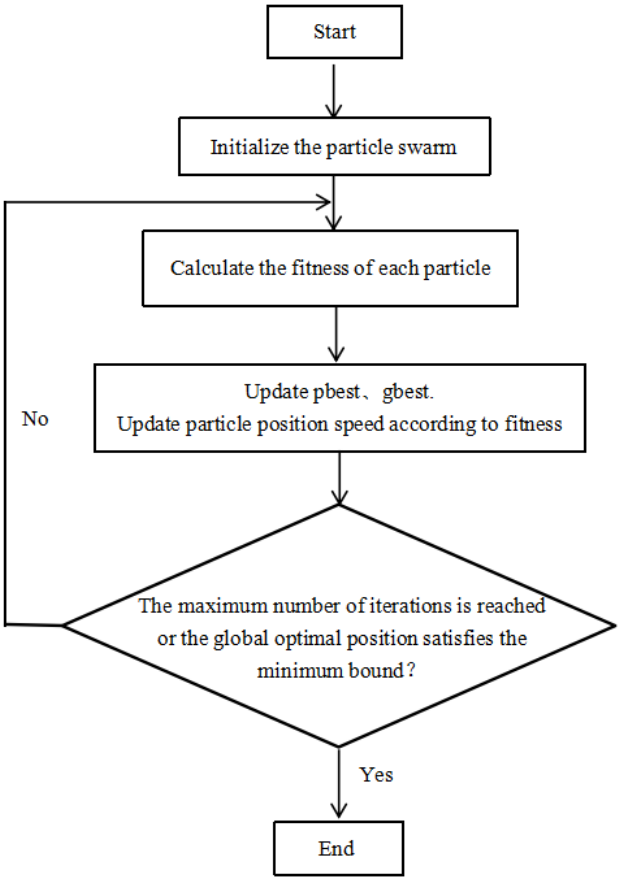


Figure 5 PSO algorithm optimisation process

2.4 Fault diagnosis process with CNN-SVM-PSO model

The process of electric spindle fault diagnosis based on CNN-SVM-PSO model mainly includes the following six stages: sample feature extraction, division of data set, training CNN model, training SVM model, optimization of model parameters and fault type diagnosis. The basic process is shown in Figure 6:

(1) Sample feature extraction: The original signals of the 3 channels related to the electric spindle vibration are extracted in the time domain, frequency domain and time-frequency domain respectively to form a sample feature matrix.

(2) Division of data set: The above sample matrix is normalized, the processed feature parameters are the model input, the four label states of the electric spindle are the model output, and the training data set and the test data set are randomly divided, with the ratio of training data set to test data set being 5:3.

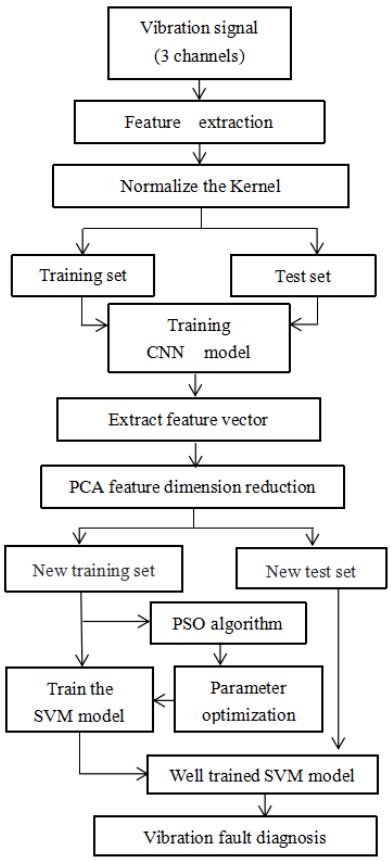


Figure 6 CNN-SVM-PSO fault diagnosis flowchart

(3) Training the CNN model: build a convolutional neural network and train it using the training and test sets from step 2. After two convolutional and pooling operations reduce the training time of the SVM by globally averaging one feature vector output from the pooling layer to form a new training and test set;

(4) Training the SVM model: train the SVM model with the training set formed in step 3, select the Gaussian radial basis kernel function as the basis function of the SVM classifier, initialize the penalty parameters ρ and kernel function width g.

(5) Model parameter optimization: Iterative optimization of the hyperparameters of the SVM model based on the training data set using the PSO algorithm to find the optimal c and g parameter pairing to improve the training speed and prediction accuracy of the model

(6) Fault type diagnosis: The test set formed with step 3 is input to the trained SVM model to identify the data fault type and provide a reference for electric spindle fault repair and troubleshooting.

3 Experimental analysis of electric spindle fault diagnosis

3.1 Setting of diagnostic model parameters

In this experiment, a 400×78 sample feature matrix was generated after feature extraction, corresponding to four labeled states, namely normal state (label 1), spindle fault (label 2), motor fault (label 3) and bearing fault (label 4). The CNN-SVM-PSO model was constructed by randomly disrupting the feature matrix and then batch normalising it to construct a training set and a test set, of which the number of training sets was 250 and the number of test sets was 150.ρ and the kernel function width g, both of which were set between 0 and 5, were selected as the target of the optimization process. To avoid interference from other factors, the number of particle swarm individuals in the PSO algorithm was set to 15 and the maximum number of iterations was set to 150, with the specific parameters shown in Table 3. Fifteen optimisation operations were carried out according to the parameters in Table 3, and the average value was taken as the final result, where the penalty parameterρ was 0.401 and the kernel function width g was 1.215. The optimizedρ , g parameters were migrated to the CNN-SVM model to complete the four label fault diagnosis.

Table 3 Initial parameter settings for the PSO algorithm

PSO algorithm parameters	Parameter values
Number of individuals in the particle population	15
Maximum number of iterations	150
Acceleration factors c1, c2	1.3, 1.5
Inertia factor	0.5
Particle vector dimension	2

3.2 Selection of diagnostic model evaluation indicators

In order to quantify the results of electric spindle fault diagnosis, this paper selects Precision, Accuracy, Recall and F1-score values as the evaluation indexes[19]The formulae for the calculation of Precision, Accuracy, Recall and F1-score are as follows

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{12}$$

In the above equation, the values of TP, TN, FP and FN can all be found in the confusion matrix, which is shown in Table 4 for the dichotomy example.

Table 4 Confusion matrix

		True value	
		Normal	Fault
Predicted value	Normal	TP	FP
	Fault	FN	TN

3.3 Electric spindle fault diagnosis results

This experiment takes the electric spindle of a five-axis machining centre as the research object, and uses the acceleration sensor to detect the vibration signals of four label states in real time, and forms the sample data after feature extraction, normalised and input to the CNN-SVM-PSO fault diagnosis model for fault identification, and the fault identification results of the training set obtained are shown in Figure 7, and it can be found that only 1 sample out of 250 training samples, The fault identification results of the test samples are shown in Figure 8, and it can be found that only 1 sample out of 150 test samples was diagnosed incorrectly, with an accuracy rate of 99.33%. The results show that the CNN-SVM-PSO model has a good effect in the diagnosis of electric spindle faults.

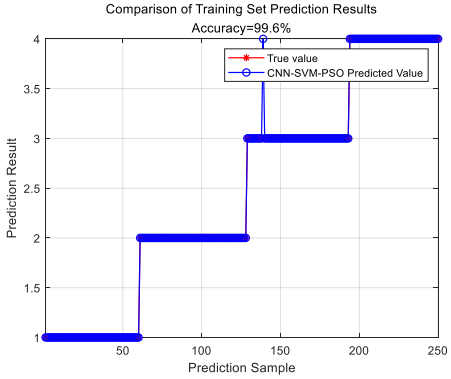


Figure 7 Training set spindle fault prediction results

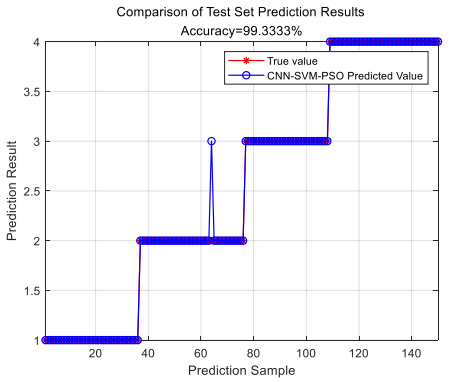


Figure 8 Test set spindle failure prediction results

The confusion matrix of the CNN-SVM-PSO model electric spindle fault diagnosis test set is shown in Figure 9. It can be seen that the test set contains 36 samples of normal state (label 1), spindle fault (label 2) 40, motor fault (label 3) 32 and bearing fault (label 4) 42, total 150 samples. In the diagnosis of the spindle fault (tag 2), one sample was incorrectly classified as a motor fault (tag 3), with an accuracy rate of 97.5%; no errors were found in the diagnosis of normal condition (tag 1), motor fault (tag 3) and bearing fault (tag 4), with an accuracy rate of 100%.

The evaluation index of electric spindle fault diagnosis can be calculated through the confusion matrix, and the results of the evaluation index of its four state labels are shown in Table 5. For the accuracy rate index, it can be seen that the accuracy rate of the spindle fault is the lowest, but it also reaches 97.5%, and all other states can reach 100%, which achieves a better result; for the correct rate index, it can be seen that the accuracy rate of all three wear states is 99.33%, which is consistent with the previous analysis; for the recall rate index, it can be seen that only the recall rate of the motor fault (label 3) does not reach For the F1 value metric, it can be seen that the minimum value of F1 for the four fault types is 0.985, which is close to 1. These four results further validate the superiority of the CNN-SVM-PSO model in the diagnosis of electric spindle faults.

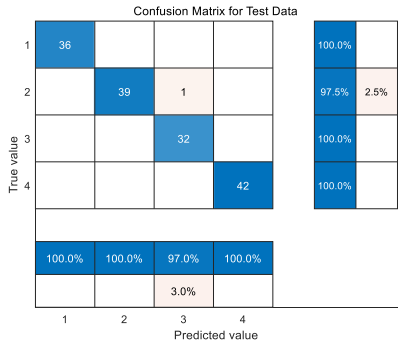


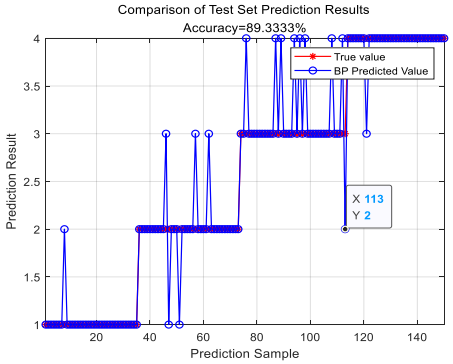
Figure 9 Fault diagnosis confusion matrix

Table 5 Results of the four fault diagnosis evaluations

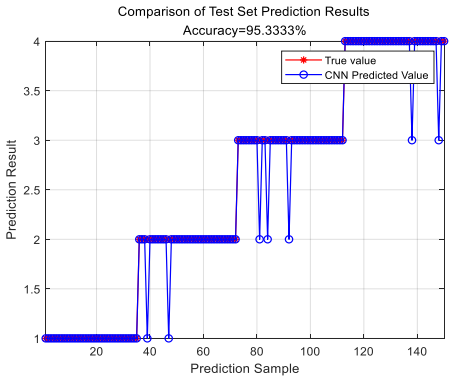
Label Classification	Precision	Accuracy	Recall rate	F1 value
1	100%	99.33%	100%	1
2	97.5%	99.33%	100%	0.987
3	100%	99.33%	97%	0.985
4	100%	99.33%	100%	1

In order to further verify the identification effect of the CNN-SVM-PSO electric spindle fault diagnosis model, the prediction effect was compared with other traditional fault diagnosis models in the past, such as BP neural network, CNN model, SVM model and CNN-SVM model, and the prediction results of these four traditional electric spindle fault diagnosis models are shown in Figure 10. From Fig. 8 and Fig. 10, it can be seen that the prediction effects of the five electric spindle fault diagnosis models are ranked as CNN-SVM-PSO ＞ CNN-SVM ＞ CNN ＞ SVM ＞ BP. It can thus be seen that the hybrid CNN-SVM model based on the optimization of PSO algorithm proposed in this paper has obvious advantages in electric spindle fault diagnosis, which is due to the ability in the CNN-SVM-PSO model to deep mining of data hidden layer features with high nonlinearity and comprehensive feature extraction, and the PSO algorithm is able to perform a deep mining of the penalty parameter in the SVM support vector machineρ The PSO algorithm is able to find the optimal pairing of two hyperparameters in the SVM support vector machine and the kernel function width g, which avoids the blindness of setting parameters and thus improves the accuracy of the prediction model. It is calculated that the CNN-SVM model optimized based on

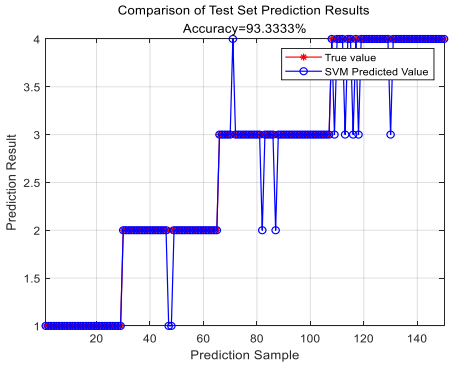
the PSO algorithm improves the accuracy by 2% over the traditional CNN-SVM model.



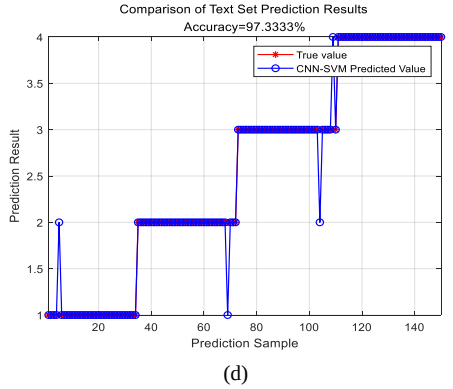
(a)



(b)



(c)



(d)

Figure 10 Prediction results of the four traditional models

(a) BP model. (b) CNN model. (c) SVM model. (d) CNN-SVM model

Table 6 shows the performance comparison results of the five electric spindle fault diagnosis models. The number of diagnostic error samples of BP model, SVM model, CNN model and CNN-SVM model are 16, 10, 7 and 4 respectively, and their accuracy rates are 89.33%, 93.33%, 95.33% and 97.33% respectively. In contrast, the CNN-SVM-PSO model proposed in this paper diagnosed only one wrong sample and the accuracy rate was as high as 99.33%, which improved the accuracy index by 10%, 6%, 4% and 2% respectively compared with the above four traditional models. This shows that under the conditions of consistent samples and the same number of samples, the prediction accuracy of the hybrid CNN-SVM model based on the optimised PSO algorithm for electric spindle fault diagnosis is significantly higher than the other models, and its generalisation ability is stronger and the network fitting speed is faster, which indirectly indicates that using the CNN-SVM-PSO model for electric spindle fault diagnosis is more accurate and can provide a reference for electric spindle fault repair and troubleshooting. This indirectly indicates that the CNN-SVM-PSO model is more accurate for electric spindle fault diagnosis and can provide a reference for electric spindle fault repair and troubleshooting.

Table 6 Performance comparison results of the five diagnostic models

Algorithm	Number of misidentified samples				Accuracy
	Normal Status	Bearing failures	Spindle failure	Motor failure	
BP Neural Network	1	5	9	1	89%
SVM Algorithms	0	2	3	5	93%
CNN Algorithms	0	2	3	2	95%
CNN-SVM algorithm	1	1	2	0	97%
CNN-SVM-PSO algorithm	0	1	0	0	99%

4 Conclusion

In this paper, a CNN-SVM fault diagnosis model based on PSO algorithm optimisation is proposed to classify and predict four labeled states: normal state, spindle fault, motor fault and bearing fault of an electric spindle, taking the electric spindle of a five-axis machining centre as the experimental object. The model uses a convolutional neural network (CNN) model as a deep feature miner and a support vector machine (SVM) as a fault state classifier to complete the diagnosis of electric spindle fault types. In order to improve the prediction accuracy of the model, the powerful search capability of the particle swarm algorithm (PSO) is used to search for the superparameters in the model. The results show that:

(1) The best hyperparameter pairing for the CNN-SVM electric spindle fault diagnosis model was found by the PSO algorithm, where the penalty parameterρ is 0.401 and the kernel function width g is

1.215, which reduces the subjective influence of manual parameter selection and avoids the blindness of setting parameters, thus improving the diagnostic accuracy.

(2) The CNN-SVM-PSO model can effectively monitor and diagnose the common types of faults in electric spindle systems, and its diagnostic accuracy reaches 99.33%.

(3) Under the same conditions, the diagnostic performance of the CNN-SVM-PSO model proposed in this paper was compared with the BP model, CNN model, SVM model and CNN-SVM model, and the results showed that the model constructed in the paper has obvious advantages in electric spindle fault diagnosis, and its accuracy indexes were improved by 10%, 6%, 4% and 2% respectively.

In the future, this CNN-SVM-PSO electric spindle fault diagnosis model can be widely used in the fields of spindle fault diagnosis and intelligent operation and maintenance of CNC machine tools in various factories. By monitoring the vibration signal of the electric spindle in real time, it is of practical significance to achieve early warning and display the type of fault when the vibration signal is abnormal, providing reference advice to maintenance personnel and improving maintenance efficiency.

Author Contributions: For Conceptualization, methodology, analysis, and writing original draft preparation, Wang Shuo; writing review and full-text editing, Yu Zhenliang; writing—original draft preparation, Liu Xu; writing—original draft preparation, Lv Zhipeng.

Conflicts of interest: The authors declare no conflict of interest.All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The research work financed with the means of Basic Scientific Research Youth Program of Education Department of Liaoning Province, No.LJKQZ2021185; Yingkou Enterprise and Doctor Innovation Program (QB-2021-05).

References

[1] Zhaolong Li, Wenming Zhu, Bo Zhu, et al.. Simulation analysis model of high-speed motorized spindle structure based on thermal load optimization [J]. Case Studies in Thermal Engineering, 2023(44):76-78.

[2] Li Zhaolong, Zhu Bo, Dai Ye, et al. Research on Thermal Error Modeling of Motorized Spindle Based on BP Neural Network Optimized by Beetle Antennae Search Algorithm [J]. Machines,2021,9(11):91-93.

[3] Wen-tao Shan, Xiao-an Chen. Block adaptive backstepping control for high-speed motorized spindle based on global RBF neural network [J]. Journal of Residuals Science & Technology,2016, 13(8):27-28.

[4] C.K.Madhusudana, N. Gangadhar, Hemantha KumarKumar, et al. Use of Discrete Wavelet Features and Support Vector Machine for Fault Diagnosis of Face Milling Tool [J]. Structural Durability & Health Monitoring,2018,12(2):34-35.

[5] Lee Hojin, Jeong Hyeyun, Koo Gyogwon, et al. Attention RNN Based Severity Estimation Method for Interturn Short-Circuit Fault in PMSMs [J]. IEEE Transactions on Industrial Electronics, 2020(1):29-30.

[6] Han Sung-Ryeol, Kim Yun-Su. A fault identification method using LSTM for a closed-loop distribution system protective relay [J]. International Journal of Electrical Power and Energy Systems, 2023(1),148.

[7] Zhou Yuankai, Wang Zhiyong, Zuo Xue, et al. Identification of wear mechanisms of main bearings of marine diesel engine using recurrence plot based on CNN model [J]. Wear, 2023(1),520-521.

[8] Chen Yu, Zhou Huicheng, Chen Jihong, et al. Spindle thermal error modeling method considering the operating condition based on Long Short-Term Memory [J]. Engineering Research Express, 2021,3(3):73-74.

[9] Wen Long, Li Xinyu, Gao Liang, et al. A New Convolutional Neural Network-Based Data-Driven Fault Diagnosis Method [J]. IEEE Transactions on Industrial Electronics, 2018,65(7):1256-1267.

[10] Tian-Luu Wu, Ji-Hwei Horng. Semantic Space Segmentation for Content-Based Image Retrieval using SVM Decision Boundary and Principal Axis Analysis [J]. Lecture Notes in Engineering and Computer Science, 2008,2168(1):451-455.

[11] Stefan Droste, Thomas Jansen, Ingo Wegener. Upper and Lower Bounds for Randomized Search Heuristics in Black-Box Optimization. Electron. Colloquium Comput [J]. Complex, 2003(4): 48-48.

[12] Weifeng Lu, Bingyu Cai, Rui Gu. Improved Particle Swarm Optimization Based on Gradient Descent Method [J]. CSAE, 2020(1): 121-126.

[13] Salih Omran, Duffy Kevin Jan. Optimization Convolutional Neural Network for Automatic Skin Lesion Diagnosis Using a Genetic Algorithm [J]. Applied Sciences,2023,13(5):1233-1235.

[14] Zhang Xin, Jiang Yueqiu, Zhong Wei. Prediction Research on Irregularly Cavitied Components Volume Based on Gray Correlation and PSO-SVM [J]. Applied Sciences,2023,13(3):445-446.

[15] Yu Sun, Dongpo He, Jun Li. The PSO optimisation SVM prediction model for the asphalt pavement environment and service fatigue life [J]. International Journal of Information and Communication Technology, 2022, 20(4):342-344.

[16] Jin Xin Zhang, Min Wang, Tao Zan, et al. Fault Detection of a High-Speed Electric Spindle. Advanced Materials Research, 2012, 1671(472-475):1568-15669.

[17] Gajera Himanshu K., Nayak Deepak Ranjan, Zaveri Mukesh A. A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features [J]. Biomedical Signal Processing and Control,2023,79(P2):423-433.

[18] Fafa Chen, Chen Fafa, Cheng Mengteng, et al. Pattern recognition of a sensitive feature set based on the orthogonal neighborhood preserving embedding and adaboost_SVM algorithm for rolling bearing early fault diagnosis [J]. Measurement Science and Technology,2020, 31(10):348-351.

[19] Wang Erhua, Yan Peng, Liu Jie. A Hybrid Chatter Detection Method Based on WPD, SSA, and SVM-PSO [J]. Shock and Vibration, 2020(77):45.

Viser Technology Pte. Ltd.
Mechanical Engineering Science
DOI: 10.33142/mes.v4i2.9085

Research Article Open Access

CNN-LSTM based on attention mechanism for brake pad remaining life prediction

Shuo WANG, Zhenliang YU*, Guangchen XU, Sisi CHEN

Yingkou Institute of Technology, School of Mechanical and Power Engineering Yingkou, China

***Corresponding Author:** Zhenliang YU, email address:yuzhenliang_neu@163.com*

Abstract:
In order to predict the remaining service life of brake pads accurately and efficiently, and to achieve intelligent warning, this paper proposes a CNN-LSTM brake pad remaining life prediction model based on an attention mechanism. The model constructs a non-linear relationship between brake pad features such as brake temperature, brake oil pressure and brake speed and brake pad wear data through convolutional neural network (CNN) and long and short term memory network (LSTM), as well as capturing the time dependence that exists in the brake pad wear sequence. The attention mechanism is also introduced to assign different weight values to the features output from multiple historical moments, highlighting the features with high saliency and avoiding the influence of invalid features, so as to improve the prediction effect of the remaining brake pad life. The results show that the proposed CNN-LSTM-Attention model can effectively predict the remaining life of brake pads, with the mean absolute error MAE value of 0.0048, root mean square error RMSE value of 0.0059 and coefficient of determination R2 value of 0.9636; and compared with the BP model, CNN model, LSTM model and CNN-LSTM model, the coefficient of determination R2 values are closest to 1, with an improvement of 8.26%, 5.25%, 3.99% and 1.85% respectively, enabling more effective monitoring and intelligent warning of the remaining brake pad life.

Keywords: *attention mechanism; CNN-LSTM; brake pads; life prediction*

1 Introduction

As people's living standards improve, the number of cars owned increases, and so does the probability of traffic accidents. As one of the important protection devices for safe driving, car brakes are of great concern, and their performance directly affects the personal safety of people driving cars. During the braking process, the brake pads and the brake discs produce relative motion, which instantly generates great temperature and friction, and the surface of the brake pads is prone to wear due to chemical reactions under high temperature and pressure. Therefore, it is necessary to make accurate life prediction and health management of the brake pads, so that the management system can make intelligent alarm according to the prediction result and remind the driver to replace the brake pads in time, thus avoiding major traffic accidents.

At the same time people's requirements for the reliability and safety of cars are getting higher and higher, and new requirements for the failure mechanisms, and diagnostic techniques of vehicle braking systems have been put forward, and the research literature on vehicle braking system fault diagnosis is becoming increasingly rich. Deng Fengman et al. based on fuzzy theory for hydraulic brake system fault diagnosis, the accuracy of the constructed ARX-RBQ diagnosis model is 92%, indicating that the use of the model can basically complete the fault diagnosis of hydraulic brake system [1]. However, the research on the remaining life prediction and intelligent warning of brake pads in vehicle braking system is very limited, and the early prediction is mainly for the design life of brake pads using theoretical or experimental methods to verify. Hao Mingshu et al. used the Manson-Coffin equation to predict the thermal fatigue life of disc brakes by studying the temperature and stress fields of disc brakes and deriving the average equivalent force at the hazardous parts of the disc [2].

The mid-to-late stage prediction mainly uses machine learning methods to extract and train features from the collected raw data, simulate the whole process of system degradation, and compare the current working state with historical data to complete the prediction of remaining life. The most commonly used machine learning methods mainly include BP neural networks [3], artificial neural networks (ANN) [4], Support vector machines (SVM) [5] etc. However, machine learning methods do not dig deep into the hidden information of

Received on October 9, 2022; Accepted on December 13, 2022

the data and do not consider the intrinsic correlation of the time series, which still needs to be improved.

In recent years, deep learning theory has emerged in the field of residual life prediction, which is able to extract deep features from complex data and combine them with time series information to predict residual life compared to traditional mechanical learning techniques [6]. The most commonly used deep learning algorithms include recurrent learning. The most commonly used deep learning algorithms include recurrent neural networks (RNN), long and short-term memory networks (LSTM) and convolutional neural networks (CNN).

Recurrent neural networks (RNNs) can handle time series data and can remember the intrinsic connections between systems in time steps, but are prone to gradient explosion or gradient disappearance and can only handle short-term memory problems [7]. Long Short Term Memory Networks (LSTM) can solve these problems by not only handling long term memory, but also by linking past and future time series [8]. Recently, work on the prediction of the remaining life of brake pads based on LSTM has been gradually carried out by Xu Meng The results show that the VMD-Bi LSTM model can meet the requirements of brake pad life prediction[9]. However, the prediction accuracy and precision of the LSTM network is not high for the time series with stronger non-linearity and more prominent non-smoothness [10].

In order to obtain better prediction results in the field of time-series data prediction, Riemer et al. proposed a neural network based on an attention mechanism for multi-source time-series data [11]. The input attention mechanism is introduced in the encoder stage to filter more relevant features for prediction, and the temporal attention is introduced in the decoder stage to extract the long-term time dependence of time series, thus avoiding the influence of invalid features and improving the model accuracy [12].

Convolutional neural networks (CNNs) are also widely used in various models for lifetime prediction because their convolution and pooling operations can improve the ability to mine potential features of complex data in prediction models compared to long and short-term memory networks (LSTMs) [13]. However, CNN networks are only able to extract the most important features of the data. However, CNN networks can only extract spatial features of brake pad wear and avoid temporal information, which leads to incomplete extraction of brake pad wear prediction features and reduced accuracy and efficiency of prediction[14]. Therefore, it has become an inevitable trend to combine CNN models with LSTM models.

The wear of automotive brake pads is a process of gradual degradation over time, which is by nature an asymptotic, non-linear and non-stationary time series with a severe dependence on time. Therefore, based on machine vision, feature extraction, deep learning, attention mechanism and other techniques, this paper proposes a CNN-LSTM brake pad remaining life dynamic evaluation method based on attention mechanism improvement, using CNN model to mention mining potential deep features in space and capturing time series information in time through LSTM model, so that the temporal features and spatial features of the data can be fully utilized. thereby improving the accuracy of brake pad wear prediction. Finally, an attention mechanism is introduced to deal with the difference in importance of the CNN-LSTM output features to enhance the influence of important time-series features in the model, avoid memory loss and gradient dispersion caused by too long a step, and improve the model prediction effect. The research of this method will propose a new theory and method for the prediction of the remaining life of brake pad wear, laying a theoretical foundation and scientific basis for improving the development of China's automobile manufacturing industry and automobile maintenance industry.

2 Life estimation options for automotive brake pads

The braking principle of a car is to use the friction between the brake pads and the brake disc to convert the kinetic energy of the car moving forward into the heat energy after friction, thus stopping the car. As shown in Figure 1, when the car brakes, the caliper piston pushes the brake pad under the action of hydraulic fluid, and the brake pad and the brake disc come into contact with each other to produce sliding friction, which eventually holds the brake disc to stop the car. Most of the brake pads are made of polymer-based composite materials, so this paper uses the quantitative calculation of wear of composite materials as a reference to estimate the wear of brake pads and obtains the following equation:

$$\Delta H = \alpha P^a V^b t^c \quad (1)$$

whereΔH is the amount of wear generated during the braking process of the car brake pad, P is the oil pressure of the hydraulic oil pushing the piston, V is the relative velocity between the brake pad and the brake pad, t is the friction time during the braking process, andα is the compensation coefficient of brake pad wear, a, b and c are the indices of brake oil pressure, braking speed and braking time respectively.

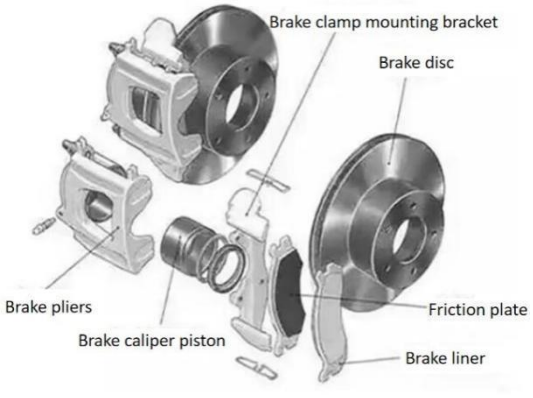


Figure 1 Structure of a car brake

During the braking process, the brake pads and brake discs generate a lot of heat in contact with each other, causing a chemical reaction on their surface resulting in wear. If we consider the frictional heat generated by the chemical reaction, the frictional heat coefficient is θ, theΔH_V In order to consider the amount of wear on the brake pads of the car after the chemical reaction, we are able to derive the following formula:

$$\Delta H_V = \Delta H * \theta = \alpha\theta P^a V^b t^c \quad (2)$$

Four parametric correction equation according to the Arrhenius formula:

$$\theta = \beta(T/T_0)^n e^{-E/RT} \quad (3)$$

where β, n is a constant; E is the activation energy generated by friction between the brake pad and the brake disc; R is the molar gas constant; T is the real-time temperature of the brake pad; T_0 is the initial temperature of the brake pad; and

It is therefore possible to derive an equation for the amount of brake pad wear after taking into account the chemical reaction:

$$\Delta H_V = \Delta H * \theta = \alpha\theta P^a V^b t^c = \alpha\beta P^a V^b t^c (T/T_0)^n e^{-E/RT} \quad (4)$$

The above equation shows that the real-time temperature of the brake pads, the oil pressure of the hydraulic fluid pushing the piston and the relative speed between the brake pads and the brake pads are all decisive factors in the wear of the car's brake pads.

3 Construction of a method for predicting the remaining life of brake pads

In order to improve the accuracy and precision of the brake pad remaining life prediction model, this paper proposes a life prediction model based on the improved CNN-LSTM with attention mechanism, which outputs the wear value of the brake pad by detecting the braking speed, braking pressure and braking temperature, so as to calculate the remaining thickness of the brake pad according to the initial amount of the brake pad, and will generate a failure alarm prompt when the remaining thickness exceeds the wear threshold, with The improvements are:

(1) The extracted braking speed, brake oil pressure and brake temperature features are batch normalised to improve the generalisation capability of the model, avoid over-fitting and improve the convergence speed of the model.

(2) The unique structure of the CNN model with local connectivity and weight sharing allows the complexity of the network to be reduced, and the spatial continuity of the sample features is maintained after convolution and pooling operations.

(3) The Long Short Term Memory Network (LSTM) is a further optimisation of the traditional RNN network, capable of handling longer time series data while avoiding gradient disappearance or gradient explosion phenomena.

(4) The introduction of the Attention mechanism can handle the importance variability of the CNN-LSTM output features, complete with the assignment of different weight values to avoid the influence of invalid features and improve the model accuracy. Its CNN-LSTM-Attention brake pad remaining life prediction model is shown in Figure 2.

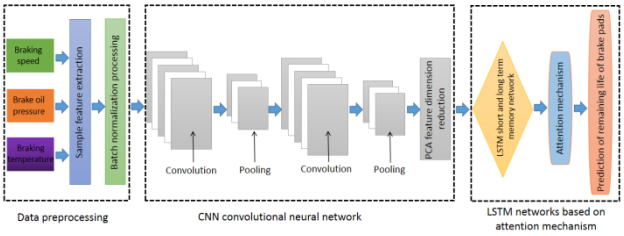


Figure 2 CNN-LSTM-Attention remaining life prediction model

3.1 Construction of the sample data set

Braking speed refers to the relative speed of sliding friction between brake pads and brake pads. The relative speed is measured by the speed sensor. In this paper, according to the national road safety regulations, the vehicle speed is controlled between 40km/h and 120km/h, so the extracted braking speed range is 354r/min to 1061r/min. Brake oil pressure refers to the hydraulic oil pressure that the piston pushes the brake pad to lock the brake disc. The pressure of the hydraulic oil to push the piston is extracted through the hydraulic pressure sensor. In this paper, according to the relevant requirements of the automobile brake performance, the brake pressure is controlled at 0.8Mpa to 1.6Mpa; Braking temperature refers to the instantaneous temperature generated by the friction between brake pads and brake discs. Real-time temperature of brake pads is extracted by temperature sensor, and the extracted temperature ranges from 47.4℃ to 84.7℃. The braking parameters are shown in Table 1.

Table 1 Selection range of braking parameters

Braking parameters	Sensors	Parameter range
Braking speed	Speed Sensors	354r/min to 1061r/min
Brake oil pressure	Oil pressure Sensors	0.8Mpa to 1.6Mpa
Braking temperature	Temperature Sensors	47.4° C to 84.7° C

In this paper, the raw data of the above three braking parameters and the wear of the brake pads after braking are extracted separately, but the wear of the brake pads after a single braking is small and difficult to measure, so the braking feature extraction experiments are conducted every Δt time. In this experiment, the number of braking cycles in Δt time was set to 300, and the braking parameters were kept constant, so each feature extraction experiment was able to obtain three time-domain features: braking speed, braking oil pressure and braking temperature. A total of 50 feature extraction experiments were carried out, so a 50 x 3 feature sample matrix can be

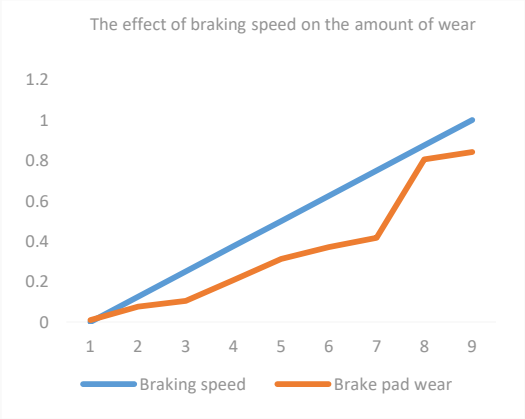
derived, which is the input to the brake pad residual life prediction model; at the same time, the thickness of the brake pad before and after braking in Δt time is measured, and the difference is divided by the number of braking times to determine the amount of brake pad wear after each braking, so a 50 x 1 target sample matrix can be derived, which is the output of the brake pad residual life prediction model. This matrix is the output of the brake pad residual life prediction model.

In order to improve the generalization ability of the prediction model and to find out the degree of influence of the three braking parameters on brake pad wear, the 50 x 3 feature sample matrix and the brake pad wear values obtained above were normalized by the normalization process formula:

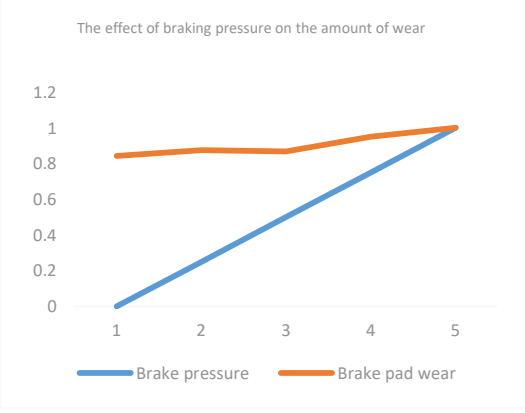
$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5}$$

Where X is the sample of each feature, and X_{min} is the minimum value of the sample feature, and X_{max} is the maximum value of the sample feature.

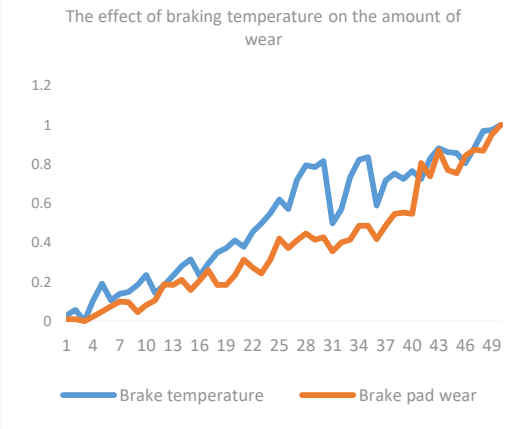
Figure 3 shows the effect of each braking parameter on brake pad wear after normalisation. From Figure 3, it can be seen that, according to the experiment conducted in accordance with the above requirements, with the increase of braking speed, braking pressure and braking temperature, the wear of automobile brake pads all course upwards; however, braking speed and braking temperature have a greater effect on brake pad wear, while braking pressure has no significant effect on brake pad wear.



(a)



(b)



(C)

Figure 3 Effect of braking parameters on brake pad wear.(a) Effect of braking speed on the amount of wear.(b) Effect of braking pressure on the amount of wear.(c) Effect of braking temperature on the amount of wear

3.2 Prediction principles of CNN-LSTM-Attention models

3.2.1 Convolutional Neural Networks (CNN)

The CNN convolutional neural network proposed by LeCun Y et al. is a typical representative of deep learning and is widely used for processing spatial features [15]. In this paper, CNN convolutional neural networks are used to extract the local correlation features between the sample data of the braking system and the wear and tear values in the target samples, and remove the unstable information and noise while maintaining the spatial continuity of the samples, resulting in a high-dimensional feature matrix as the input to the LSTM network, which is based on the following principles:

(1) Convolution operations are performed on the batch normalised sample matrix by using a convolution kernel of suitable dimensionality to abstractly represent the brake pad wear features in space. Let the jth feature data output from layer i-1 be X_{i-1}^{j} , and in order to improve the prediction accuracy of the model, this paper chooses the Relu function as the activation function, and its convolution operation can be represented by equation (6):

$$V_i^k = Relu\left(W_i^{kj} \circledast X_{i-1}^j + b_i^k\right) \tag{6}$$

where V_i^k is the kth feature data output from the next layer after the convolution operation, and W_i^{kj} is the weight value of the convolution kernel, and $\circledast$ is the convolution operator, and b_i^k is the bias value of the feature data in the next layer.

(2) The purpose of pooling is to reduce the dimensionality of the feature samples while keeping the number of features unchanged to avoid overfitting. Take the ith feature matrix as an example, let the input feature matrix of the pooling layer be $V_i^k(s,t)$ and its matrix dimension is s × t. The i-th feature matrix obtained after pooling is $C_i^k(m,n)$, whose dimension is m × n, then the maximum pooling operation can be expressed by equation (7):

$$C_i^k(m,n) = \underset{\substack{1+(m-1)Q \le s \le mQ \\ 1+(n-1)P \le t \le nP}}{Max} \{V_i^k(s,t)\} \quad (7)$$

where P is the length of the pooling window and Q is the width of the pooling window.

(3) Each sample in the 50×3 feature matrix is subjected to two convolution and pooling operations, and is able to produce j feature matrices of dimension m×n. The above feature matrix is subjected to PCA dimensionality reduction to reduce the covariance of the sample features and avoid the influence of redundant features, thus reducing the training time of the LSTM long and short-term memory network, so that the whole CNN model finally outputs a feature vector Xt = {x1 , x2 , ... , xi , ... , xj , }

3.2.2 Long Short Term Memory (LSTM) Network

CNN convolutional neural networks are able to mine local spatial features related to brake pad wear, but it is difficult to extract longer time series data, so this paper uses LSTM long and short term memory networks to further process the feature vectors output from CNN models to construct the link between sample features and time series. lstm networks were proposed by Hochreiter and Schmidhuber in 1997 proposed in 1997[16] , the principle of which is as follows:

(1) Through the forgetting doorf_t The state of the previous level of units C_{t-1} Performing forgetting or memory processing;

(2) By input gatei_t The input sample features areX_t A logical calculation is performed to update the memory of the whole system, which is transmitted according to the path set by the system to generate a new memory featureC_t The calculation of the new memory feature Ct constructed by the input gate and the forgetting gate can be expressed in equation (8):

$$C_t = f_t \otimes C_{t-1} + i_t \otimes tanh(H_{t-1}) \quad (8)$$

(3) Output gateso_t Memory featuresC_t The timing features are output by a control operationH_t and transfer to the next layer of cells, the timing characteristicsH_t The calculation can be expressed in equation (9):

$$H_t = o_t \otimes tanh(C_t) \quad (9)$$

Following the above principle is able to extract the temporal features of the samples, and its LSTM network structure is shown in Figure 4.

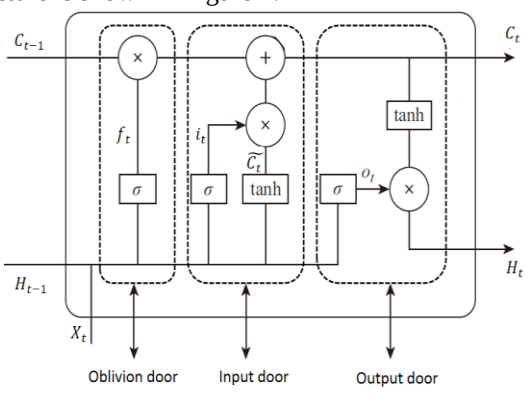


Figure 4 LSTM network gate cell structure

3.2.3 LSTM model based on attention mechanism

The CNN-LSTM model proposed above can achieve deep mining of brake pad wear features in space and time, and has obtained strong generalization ability and faster network fitting speed, but there is still room to improve the accuracy of the model. The attention mechanism can give different weight values to each feature according to the significance of the sample features, thus avoiding the interference of invalid features and improving the prediction accuracy of the model [17]. The attention mechanism based LSTM network model is shown in Figure 5.

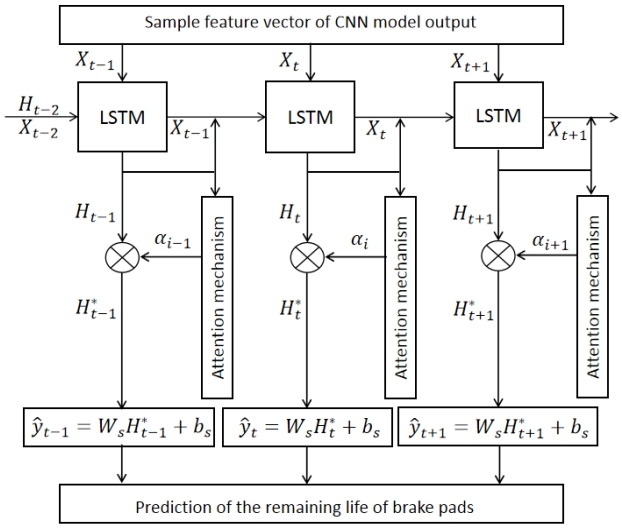


Figure 5 LSTM model based on attention mechanism

The specific principle of the LSTM network model based on the attention mechanism is as follows:

(1) Decode the hidden layer state of the input brake pad wear feature Xt for each Δt time $H_{t,i}$ and then apply the$Attention_Score$ function to compare the hidden layer states $H_{t,i}$ with the output of the LSTM network H_t correlation of the LSTM network, and the score of each sample feature at each Δt time is calculated$E_{t,i}$ which is calculated as shown in equation (10):

$$E_{t,i} = Attention\_Score(H_{t,i}, H_t) \quad (10)$$

(2) Based on the scores of each sample feature at each Δt time, a softmax function was used to value the attention weights of the input brake pad wear featuresα_i were calculated as shown in equation (11);

$$\alpha_i = \frac{exp[Attention\_Score(H_{t,i}, H_t)]}{\sum_{j=1}^{m} exp[Attention\_Score(H_{t,j}, H_t)]} \quad (11)$$

(3) The attention weights of the brake pad wear features areα_i with the output of its LSTM networkH_t weighted aggregation operation, resulting in a new brake pad wear feature vector H_t^* which is calculated as shown in equation (12):

$$H_t^* = \sum_{i=1}^{m} H_t * \alpha_i \quad (12)$$

where m is the number of nodes in the output of the

fully connected layer.

In summary, the weight calculation process of the attention mechanism is implemented through the attention layer, the input of which is the temporal feature vector extracted by the LSTM networkH_t and the output is a feature vector of H_t^* The new brake pad wear feature vector H_t^* The new brake pad wear feature vector is input to the fully connected layer to predict the remaining brake pad life and obtain the brake pad wear value $\hat{y}_t$ The new brake pad wear feature vector is input to the full connection layer to complete the prediction of the remaining brake pad life. It is calculated as shown in equation (13):

$$\hat{y}_t = W_s H_t^* + b_s \tag{13}$$

where $\hat{y}_t$ is the predicted value of brake pad wear, the$W_s$andb_s are the weights and bias values of the fully connected layer, respectively.

3.3 CNN-LSTM-Attention Model prediction process

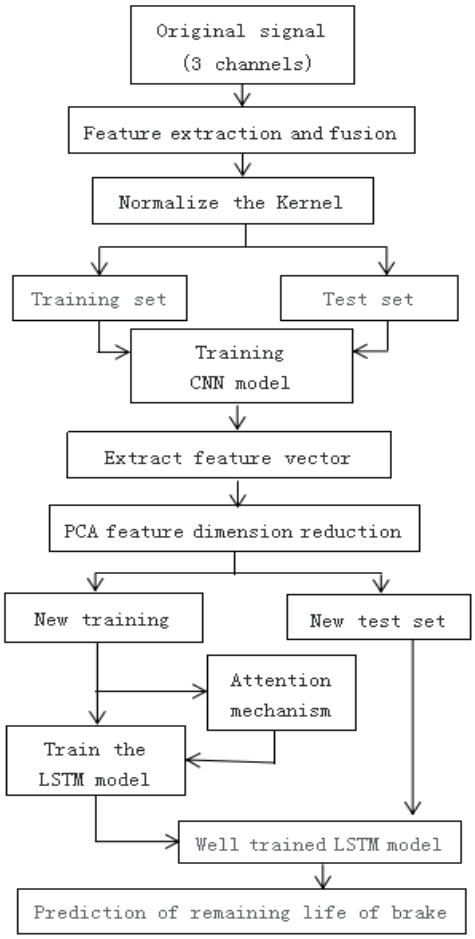


Figure 6 CNN-LSTM-Attention Model prediction process

The CNN-LSTM based Attention model for regression prediction of the remaining brake pad life contains the following six steps, feature extraction and processing, division of sample data, training of CNN model, training of LSTM model, weight assignment of attention mechanism and prediction of brake pad wear, the CNN-LSTM-Attention model prediction process is shown in Figure 6, and the specific steps are as follows:

(1) Feature extraction and fusion of the raw signals from the 3 channels of brake speed, brake pressure and brake temperature as well as brake pad wear values to form a 50 x 4 sample matrix.

(2) The above 50×4 sample matrix was batch normalised and its order was randomly disordered to divide the training and test sets of the CNN model in a ratio of 3:2.

(3) The CNN network is constructed, and the training and test sets from step 2 are used to perform convolution and pooling operations to extract a spatial feature vector Xr, which is dimensionally reduced by PCA to form a new training and test set.

(4) Construct an LSTM network and apply forgetting gates, input gates and output gates to the training set output from step 3 to extract a temporal feature vector Ht.

The attention mechanism is introduced to assign weights to each wear feature, eliminate invalid features, output a new wear feature vectorH_t^* and complete the training of the LSTM model.

The test set output from step 3 is fed into the LSTM model trained in step 5 to complete the regression prediction of the remaining brake pad life.

4 Experimental analysis of brake pad life models

4.1 Setting of structural parameters of the prediction model

Based on the advantages of Convolutional Neural Network (CNN) in mining spatial features and the characteristics of Long Short Term Memory Network (LSTM) in processing temporal features, this paper proposes an Attention-CNN-LSTM based brake pad life prediction model. After comparing the experimental prediction effects, the optimal model structure and parameter configuration selected in this paper is shown in Table 2, which mainly includes an input layer, CNN layer, LSTM layer, Attention layer, Dropout layer and output layer. The model first passes a 50×4 sample dataset through the input layer to the CNN layer, which mines the deep features of the brake pad wear data and uses them as input to the LSTM layer after two convolutions and pooling. The LSTM layer then learns the non-linear relationship between brake pad wear and the input features as well as the time dependence present in the brake pad wear sequence. Finally the attention mechanism uses a scoring function to assign greater weight values to the brake pad wear features at important moments, and the output layer is used to obtain the brake pad wear prediction values. Thus, the essence of the model is in obtaining a mapping between the current moment's brake pad state and the brake pad wear values at multiple historical moments.

Table 2 Structural parameters of CNN-LSTM-PSO model

1	Input layer	Sample data set Matrix dimension: 50 x 4
2	Convolutional layer 1	Activation function: RELU Convolution kernel: 3 x 3 Maximum pooling
	Batch standardisation layer 1	
	Pooling layer 1	
3	Convolutional layer 2	Activation function: RELU Convolution kernel: 3 x 3 Maximum pooling
	Batch standardisation layer 2	
	Pooling layer 2	
4	LSTM layer	Learning rate: 0.004 Number of hidden layer units: 50 Activation function: Sigmoid
5	Attention layer	Attention weighting values: α_i Scoring function: Attention_Score
6	Dropout layer	25% discard
7	Output layer	Activation function: Softmax

4.2 Comparison of predictive model evaluation indicators

In order to quantify the predictive performance of the brake pad residual life model, three objective evaluation metrics are selected, namely the mean absolute error MAE, root mean square error RMSE and coefficient of determination R2. The mean absolute error MAE can be used to obtain an evaluation value, but a comparison between different models is required to reflect the model's merit. The smaller the RMSE and the closer the coefficient of determination R2 is to 1, the higher the accuracy and precision of the prediction model. The three evaluation indicators are calculated as follows:

$$MAE = \frac{\sum_{t=1}^{m}|y_t - \hat{y}_t|}{m} \tag{14}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{m}(y_t - \hat{y}_t)^2}{m}} \tag{15}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{m}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{m}(y_t - \bar{y})^2} \tag{16}$$

where, m is the number of samples output from the fully connected layer, the number of samples in this paper is 50, and $\hat{y}_t$ is the predicted value of brake pad wear, and y_t is the actual value of brake pad wear.

4.3 CNN-LSTM-Attention Model prediction results

The CNN-LSTM-Attention model proposed in this paper is verified by using the data published by the Disc Brake Simulation Experimental Research Group of China University of Mining and Technology [18]. The data set was experimented using a disc brake simulated braking test bed, where information from three channels of braking speed, braking pressure and braking temperature were collected using each sensor at Δt time intervals, while the wear thickness of the brake pads was measured. The data set was feature extracted and fused to form a final 50 x 4 sample matrix, which was fed into a CNN-LSTM model based on an attention mechanism to perform regression prediction of the remaining life of the brake pads, the prediction results of which are shown in Figure 7. The average absolute error MAE value of the model is 0.0048, the root mean square error RMSE value is 0.0059 and the coefficient of determination R2 value is 0.9636. The results show that the CNN-LSTM model based on the attention mechanism can effectively predict the remaining life of brake pads and achieve better results.

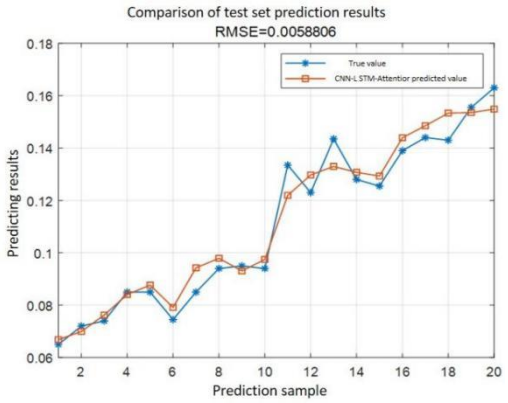
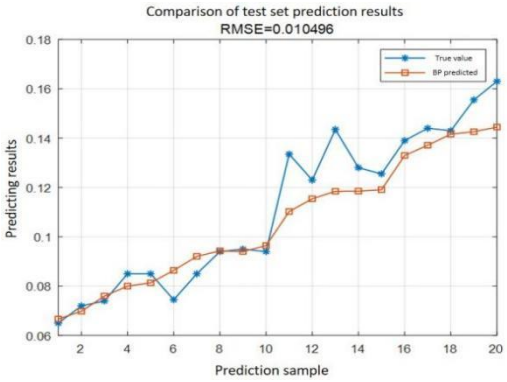


Figure 7 CNN-LSTM-Attention model prediction results

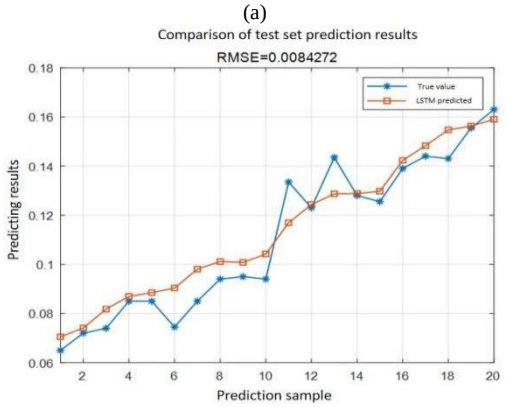
In order to further validate the prediction performance of the CNN-LSTM brake pad residual life model based on the attention mechanism, a comparative analysis with other traditional prediction models in the past, such as BP neural network, CNN model, LSTM model and CNN-LSTM model, was carried out. Figure 8 shows the comparison results of the four traditional life prediction models. From Figure 8, it can be seen that the CNN-LSTM-Attention model proposed in this paper has 43.8%, 35.2%, 29.8% and 16.9% lower RMSE values compared to the BP, CNN, LSTM and CNN-LSTM models respectively; and the CNN-LSTM-Attention model predicts a brake pad wear curve that is closer to the real brake pad wear curve than the other four prediction models, and the error curve has the smallest fluctuation range.

It can be seen that the prediction performance of the CNN-LSTM brake pad remaining life model based on the attention mechanism proposed in this paper has certain superiority. This is because other traditional prediction models have a single algorithm and incomplete feature extraction, while the CNN-LSTM model is not only capable of mining deep spatial features, but also better able to handle temporal features; at the same time, different weight values are given to the brake pad wear data at different moments in the input sample under the action of the attention mechanism, which strengthens the attention to the wear data at key moments in order to more accurately represent the brake pad This results in a more accurate representation of the brake pad wear

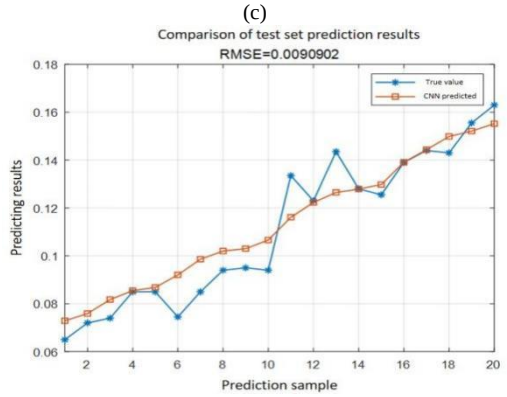
feature information, thus making the generalization ability of the whole model stronger and further improving the accuracy of brake pad wear prediction.



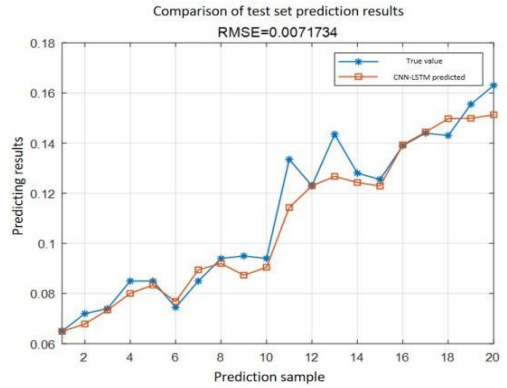
(a)



(c)



(b)



(d)

Figure 8 Prediction results of the four traditional models

(a) BP model (b) CNN model (c) LSTM model (d) CNN-LSTM model

Table 3 shows the calculation results of the five model evaluation metrics. Compared with the improved CNN-LSTM model based on the attention mechanism proposed in this paper and the CNN-LSTM model, the mean absolute error MAE and root mean square error RMSE are reduced and the coefficient of determination R2 was improved, This result demonstrates the role of the attention mechanism in predicting brake pad wear. In addition, compared with other traditional prediction models, the CNN-LSTM-Attention The mean absolute error MAE of the prediction model was the smallest, with a reduction of 37.7%, 31.4%, 28.4% and 2.04% compared to the BP, CNN, LSTM and CNN-LSTM models respectively; the value of the coefficient of determination R2 was closest to 1, with an improvement of 8.26%, 5.25%, 3.99% and 1.85% , respectively, these two results again prove that using the CNN-LSTM-Attention The two results again demonstrate that the CNN-LSTM-model proposed in this paper is more accurate in predicting the brake pad thickness wear value, and can be more effective in monitoring and intelligently warning the remaining brake pad life.

Table 3 Comparison results of the five model evaluation indicators

Lifetime prediction models	RMSE	MAE	R2 Value
BP Neural Networks	0.0105	0.0077	0.8840
CNN models	0.0091	0.0070	0.9130
LSTM model	0.0084	0.0067	0.9252
CNN-LSTM Models	0.0071	0.0049	0.9458
CNN-LSTM-Attention model	0.0059	0.0048	0.9636

5 Conclusion

In this paper, we use sensor technology to collect the raw signals from 3 channels of brake speed, brake pressure and brake temperature, and also measure the wear thickness of brake pads, and construct a sample matrix after feature extraction and fusion, then propose a CNN-LSTM prediction model based on attention mechanism to predict the remaining life of brake pads, and conduct a comparative study with other traditional prediction models, the results show that:

(1) With the increase of braking speed, braking pressure and braking temperature, the amount of brake pad wear of the car is on the rise; however, braking speed and braking temperature have a greater influence on the amount of brake pad wear, while braking pressure has no significant influence on the amount of brake pad wear.

(2) Using the CNN-LSTM-Attention model for regression prediction of brake pad wear values with a mean absolute error MAE value of 0.0048, a root mean square error RMSE value of 0.0059 and a coefficient of determination R2 value of 0.9636, which indicates that the model can effectively predict the remaining life of brake pads with good results.

(3) Compared with the BP model, CNN model, LSTM model and CNN-LSTM model, the CNN-LSTM-Attention model proposed in this paper mean absolute error MAE and root mean square error RMSE values were reduced, and the value of the coefficient of determination R2 was improved to be closest to 1. This indicates that the constructed brake pad life prediction model has less error, better accuracy and better results.

In the future, the CNN-LSTM -Attention brake pad residual life prediction model can be widely used in automotive manufacturing and car maintenance, etc. The model is of practical significance as it monitors the brake pad braking speed, braking pressure and braking temperature in real time, outputs the wear of the brake pad after each braking, and accumulates the wear after braking to calculate the remaining thickness of the brake pad, and generates a failure alarm indication when the remaining thickness of the brake pad exceeds the wear threshold to avoid accidents caused by brake failure.

Author Contributions: For Conceptualization, methodology, analysis, and writing original draft preparation, Wang Shuo; writing review and full-text editing, Yu Zhenliang; writing — original draft preparation, Xu Guangchen; writing — original draft preparation, Chen Sisi.

Conflicts of interest: The authors declare no conflict of interest.All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The research work financed with the means of Liaoning Provincial Science and Technology Department natural Science Regional Joint Fund project, No.2022-YKLH-03.

References

[1] Deng Fengman. Analysis of hydraulic brake system fault diagnosis based on fuzzy ARX-RBQ method [J]. Hydraulic Pneumatics and Seals, 2020, 40(6): 51-54.

[2] Hao Mingshu. Thermal-structural coupling analysis and life prediction of disc brakes [J]. Wuhan University of Technology, 2012(7):68-69.

[3] T.Liao, N.Zhang. Wear prediction of brake pads in EMUs using a BP neural network. Advanced Science and Industry Research Center.Proceedings of 2014 Proceedings of 2014 International Conference on Artificial Intelligence and Industrial Application (AIIA2014) [J]. WIT Press, 2014(6):47-56.

[4] Eker Muammer, Mutlu brahim, Aysal Faruk Emre, Atli Sinan, Yavuz brahim, Ergn Yelda Akin. The ANN Analysis and Taguchi Method Optimisation of the Brake Pad Composition [J]. Emerging Materials Research,2021(4):6-9.

[5] Eltayb N. S. M., Hamdy Abeer. LS-SVM Approach for Predicting Frictional Performance of Industrial Brake Pad Materials [J]. International Journal of Mechanical Engineering and Robotics Research, 2016,7(2):87-88.

[6] Changchang Che, Huawei Wang, Qiang Fu, Xiaomei Ni. Combining multiple deep learning algorithms for prognostic and health management of aircraft [J]. Aerospace Science and Technology, 2019,94(3):34-35.

[7] Zheng S, Ristovski K, Farahat A, et al. Long short-term memory network for remaining useful life estimation. Proceedings of the 2017 EEE International Conference on Prognostics and Health Management [J]. piscataway: EEE, 2017(7): 88-95.

[8] Zhang J, Wang P, Yan R, et al. Long short-term memory formachine remaining life predictiion [J]. Journal of Manufacturing Systems, 2018, 48(Pt C):78-86.

[9] Xu M., Wang Y. Kun. Remaining life prediction of DA40 aircraft carbon brake pads based on bidirectional long- and short-term memory networks [J]. computer Applications, 2021,41(05): 1527-1532.

[10] Liu Muyuan, He Junyu, Huang Yuzhou, Tang Tao, Hu Jing, Xiao Xi. Algal bloom forecasting with time-frequency analysis: A hybrid deep learning approach [J]. Water Research, 2022(2),219.

[11] Rremer M, Vempaty A, Calmonf P, et al. Correcting forecasts with multifactor neural attention [J]. International Conference on Machine Learning. 2016(3):3010-3019.

[12] Qin Y, Song D J, Chen H F, et al. A dual-stage attention-based recurrent neural network for time series prediction [J].AAAI Press, 2017(1):2627- 2633.

[13] P. K. Ambadekar, C. M. Choudhari. CNN based tool monitoring system to predict the life of cutting tool [J]. SN Applied Sciences, 2020,2(4):5-9.

[14] Xiaoyang Zhang, Xin Lu, Weidong Li, Sheng Wang. Prediction of the remaining useful life of cutting tool using the Hurst exponent and CNN-LSTM [J]. The International Journal of Advanced Manufacturing Technology, 2021(112):357-366.

[15] Lecun Y, Bottou L, Bengioy, et al. Gradient based learning applied to document recognition [J]. Proceedings of the IEEE, 1998,86(11) :2278-2324.

[16] Hochreiter S, Schmidhuber J. Long short- term memory [J]. Neural Computation, 1997,9 (8) :1735 - 1780.

[17] Chen HH, Wu G, Li JX et al. Advances in deep learning recommendation research based on attention mechanism [J]. Computer Engineering and Science, 2021, 43 (2) :370-380.

[18] Yao Wang. Research on the wear life prediction and failure warning method of automotive brake pads [J]. China University of Mining and Technology, 2018(7):68-69.

Viser Technology Pte. Ltd.
Mechanical Engineering Science
DOI: 10.33142/mes.v4i2.9086

Research Article Open Access

A CNN-LSTM-PSO tool wear prediction method based on multi-channel feature fusion

Shuo WANG[1], Zhenliang YU[1*], Yongqi GUO[1], Xu LIU[2]

1. School of Mechanical and Power Engineering, Yingkou Institute of Technology, Yingkou, China
2. Yingkou Dingsheng Heavy Industry Machinery Co. LTD, Yingkou, China

***Corresponding Author:** yuzhenliang, email address:yuzhenliang_neu@163.com*

Abstract:
In order to achieve predictive maintenance of CNC machining tools and to be able to change tools intelligently before tool wear is at a critical threshold, a CNN-LSTM tool wear prediction model based on particle swarm algorithm (PSO) optimization with multi-channel feature fusion is proposed. Firstly, the raw signals of seven channels of the machining process are collected using sensor technology and processed for noise reduction; secondly, the time-domain, frequency-domain and time-frequency-domain features of each channel signal are extracted, and a sample data set of spatio-temporal correlation of traffic flow is constructed by dimensionality reduction processing and information fusion of the above features; finally, the data set is input to the CNN-LSTM-PSO model for training and testing. The results show that the CNN-LSTM-PSO model can effectively predict tool wear with an average absolute error MAE value of 0.5848, a root mean square error RMSE value of 0.7281, and a coefficient of determination R2 value of 0.9964; and compared with the BP model, CNN model, LSTM model and CNN-LSTM model, its tool wear prediction accuracy improved by 7.56%, 2.60%, 2.98%, and 1.63%, respectively.

Keywords: *feature fusion; CNN-LSTM; tool wear; life prediction*

1 Introduction

The severity of tool wear during CNC machining plays a decisive role in the machining accuracy of products, and serious tool wear can reduce product quality, lead to increased scrap rate, and even lead to machine accidents. Therefore, in recent years, tool wear prediction has become a fundamental and prerequisite work in the field of tool life management and intelligent tool change. Early on, experts and scholars have made some progress by exploring the tool wear mechanism and combining Taylor's empirical formula for tool life prediction, the Andis Ābele et al. confirmed the validity of Taylor's empirical formula for predicting tool life and determined the coefficients of Taylor's formula, and finally obtained the formula for predicting the length of the cutting trajectory at the critical wear stage of the tool based on the cutting speed[1] . However, the Taylor's empirical formula only yields a fixed value of tool life, which does not correspond to the actual application of the tool, because the machining parameters are variable and the manufacturing environment is complex, which leads to the impossibility of the remaining tool life in the form of a fixed value.

Based on the above problems, researchers have started to use mechanical learning techniques to predict tool life. Commonly used mechanical learning prediction models are: random forest [2], BP neural network [3], support vector machine (SVM) [4], etc. Wei Weihua [5] et al. optimized BP neural network by genetic algorithm, so that the model's optimization and learning ability can be improved, which can effectively identify tool wear. Sarat Babu Mulpur [6] et al. used OGM-SVM model for real-time prediction of rear tool face wear based on extracted multi-sensor heterogeneous data features and also achieved good prediction results, but the prediction efficiency and accuracy were not high.

In the automated production process, a high accuracy life prediction model can be very effective in predicting the future tool wear level, which is important to study the tool wear at a critical threshold to enable intelligent tool change. Therefore, a large number of experts and scholars have applied deep learning theory in tool life prediction, such as recurrent neural networks (RNN) [7], long and short-term memory networks (LSTM) [8] and convolutional neural networks (CNN) [9], whose prediction effect is significantly higher than mechanical learning techniques. Recently, work on tool life prediction based on long and short term memory networks (LSTM)

Received on October 11, 2022; Accepted on December 3, 2022

has been carried out gradually. Ma Kaile [10]et al. analyzed the singularity of the original vibration signal to eliminate the effect of milling path and constructed a stacked LSTM model for tool wear prediction, and compared with models such as WOA-SVR, it was found that the method improved the accuracy of tool wear prediction. Although the LSTM network can perfectly process the timing information of tool wear, it is difficult to extract the deep features hidden in the samples, which leads to the incomplete extraction of tool wear prediction features and there is still room for improvement.

Convolutional neural networks (CNNs) have strong feature extraction capability and low computational complexity compared with long and short-term memory networks (LSTMs), and can tap deep features hidden in samples. Lim Meng Lip [11] et al. cropped the surface profile images of machined parts and input them into CNN networks for tool wear prediction, and the results showed that the CNN model can meet the tool wear prediction requirements with an accuracy of 98.9 % accuracy. Although these methods have been successful in predicting tool wear, it is still challenging to fully reveal the effective features present in the monitored signals due to the defects in the network structure [12].

As we all know, when the tool wear reaches the sharp wear stage, the system alerts for intelligent tool change, which can improve product machining accuracy and reduce tool management costs. The rule of tool wear is faster in the early stage, slower in the middle stage, and the fastest and most drastic in the late stage. It can be seen that using only one model for tool life prediction will lead to a single extracted feature, which is prone to overfitting. Therefore, combining convolutional neural network (CNN) and long and short-term memory network (LSTM) has become an inevitable trend, using CNN model to extract potential deep features in space and capturing time series information in time by LSTM model, so that the temporal and spatial features of the data can be fully utilized to make up for the shortcomings of the above single prediction model.

In order to further improve the prediction effect of the model, the hyperparameters in the prediction model must be optimized. The more common hyperparameter optimization methods include random optimization [13], gradient-based optimization [14], genetic algorithm optimization [15], particle swarm algorithm optimization [16], etc. The particle swarm algorithm (PSO) can perform global optimization with fewer parameters, and its powerful search performance and individual optimization capability can speed up the convergence of the model, so it has been widely used and studied by scholars in recent years [17].

Therefore, this paper proposes a CNN-LSTM tool wear prediction model with multi-channel feature fusion based on machine vision, feature extraction, deep learning and hyperparameter optimization, constructs a spatio-temporal correlation feature matrix of traffic flow so that the temporal and spatial features of the monitored signal can be fully utilized, and optimizes the hyperparameters in the prediction model using particle swarm algorithm (PSO), so as to improve the tool wear prediction accuracy. The research of this method will propose a new theory and method for tool wear remaining life prediction, and lay a theoretical foundation and scientific basis for improving the development of China's machine tool manufacturing industry and intelligent tool changing field.

2 Construction of CNN-LSTM-PSO prediction model

In order to improve the accuracy and accuracy of the prediction model of tool remaining life, a multi-channel feature fusion CNN-LSTM tool wear prediction model based on particle optimization was proposed in this paper. The output tool wear values were monitored by the vibration signals of three channels, the cutting force signals of three channels and the acoustic emission signals of one channel. Thus, predictive maintenance of NC machining tools can be realized, and tools can be changed intelligently before tool wear is in the critical threshold. The improvements are as follows:

(1) The characteristics of vibration signals, cutting force signals and acoustic emission signals were extracted by batch normalization and dimensionality reduction processing, which improved the generalization ability of the model, avoided overfitting phenomenon and improved the convergence speed of the model.

(2) CNN model reduces network complexity with its unique structure of local connection and weight sharing, and the spatial continuity of sample features is maintained after convolution and pooling operations.

(3) Long term memory network (LSTM) is a further optimization of the traditional RNN network, which can process longer time series data while avoiding the phenomenon of gradient vanishing or gradient explosion.

(4) Using the powerful search and global optimization ability of PSO algorithm, the two parameters of the initial learning rate parameter and the number of hidden layer units in LSTM network were iteratively optimized, which reduced the subjective influence of manual selection parameters, and thus improved the prediction accuracy of tool wear model. The CNN-LSTM-PSO tool wear prediction model is shown in Figure 1.

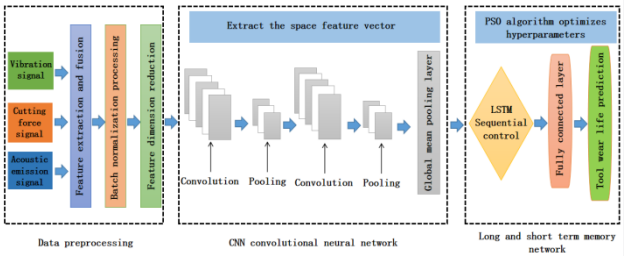


Figure 1 CNN-LSTM-PSO tool wear prediction model

2.1 Convolutional Neural Network (CNN)

Convolutional neural network (CNN) [18] is a kind of

neural network, which is a typical representative of deep learning and has obvious advantages for processing spatial data. The most important difference between CNN convolutional neural network and other traditional neural networks is the convolution operation and pooling operation, which can realize local connection and weight sharing. Therefore, the pre-processing part of this paper uses the CNN model to extract the spatial features of the 315×47 sample feature matrix, and its output is a one-dimensional spatial sequence matrix, which lays the foundation for the prediction of tool wear using the LSTM model. The principle is as follows:

(1) The sample feature matrix after batch normalization and dimensionality reduction is input to the CNN convolutional neural network for convolutional operation. The sample information is indirectly characterized by the local features of the sample through the weight value of each layer derived from the convolutional operation, and the higher the layer is, the more detailed the local features are extracted, and also the spatial continuity of the sample is maintained, and its convolutional operation is shown in equation (1):

$$X_i^k = \sum_{j=1}^{n} W_i^{kj} \otimes X_{i-1}^j + b_i^k \tag{1}$$

Where X_i^k denotes the feature matrix of the kth neuron at the output of the ith layer, and W_i^{kj} denotes the weight value of the kth neuron in the ith layer, and $\otimes$ denotes the convolution operator, and X_{i-1}^j denotes the feature matrix of the jth neuron at the output of layer i-1, and b_i^k is the bias coefficient of the kth neuron in layer i.

(2) In order to improve the prediction accuracy of the tool wear life model, the CNN network uses ReLU function for nonlinear activation, which has good non-saturation characteristics to avoid the gradient disappearance phenomenon. The activation function is shown in equation (2):

$$V_i^k = \mathrm{Relu}\left(X_i^k\right) = \begin{cases} 0, x_i^k < 0 \\ x_i^k, x_i^k > 0 \end{cases} \tag{2}$$

where x_i^k is theX_i^k each eigenvalue in the feature matrix.

(3) Each tool wear feature data is input to the pooling layer after convolution operation, and the pooling type is selected as maximum pooling, which can retain the original features and reduce the parameters of network training, and improve the robustness of the extracted features. The maximum pooling is shown in equation (3):

$$C_i^k(s,t) = \underset{\substack{1+(s-1)Q \le d \le sQ \\ 1+(t-1)P \le h \le tP}}{\mathrm{Max}} \left\{ V_i^k(d,h) \right\} \tag{3}$$

where $V_i^k(d,h)$ is the eigenvalue of column h of row d of the ith feature matrix input to the pooling layer, and $C_i^k(s,t)$ is the eigenvalue of the sth row t column of the ith feature matrix obtained after pooling, and P and Q are the length and width of the pooled region, respectively.

(4) The n feature matrices of dimension S ×T, which are derived from each row of the 315 ×47 sample feature matrix after two convolution and pooling operations, are input to the global average pooling layer. The dimensionality of the pooling kernel of the global average pooling layer is kept consistent with the dimensionality of the feature matrix, and the n feature matrices are dimensionality reduced to reduce the covariance of the sample features and avoid the influence of redundant features, thus reducing the training time of the LSTM long and short term memory network, so the whole CNN model finally outputs a feature vector $X_t = \{x_1, x_2, \ldots, x_i, \ldots, x_j, \}$ where xi is calculated as shown in equation (4):

$$x_i = \frac{1}{ST} \sum_{s=1}^{S} \sum_{t=1}^{T} C_i^k(s,t) \tag{4}$$

2.2 Long and short-term memory neural network (LSTM)

CNN convolutional neural networks are capable of mining local spatial features related to tool wear, but it is difficult to extract longer time series data. Recurrent neural networks (RNN) can perform temporal processing of tool wear data, but it is difficult to process for longer time series data, and gradient disappearance or gradient explosion occurs during operation. It is usually used to solve this phenomenon using long and short term memory networks (LSTM) or hierarchical RNNs [19]. Long Short-Term Memory Network (LSTM) is a further improvement of the traditional RNN network by introducing memory cells on the input, output, and forgetting past information to construct new cell statesC_t Realize the data transmission, and control the path of data transmission by logic operation through input gate, output gate, and forget gate, so as to complete the processing of longer time series data, and its LSTM network gate cell structure is shown in Figure 2. The new cell state Ct and the output state Ht of the LSTM core are constructed with the following equations:

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \tanh(H_{t-1}) \tag{5}$$

$$H_t = o_t \otimes \tanh(C_t) \tag{6}$$

where f_t is the forgetting gate, which serves to make the cell forget or remember the state of the previous cell C_{t-1} The input gate i_t is the input gate, which controls the input signal and thus updates the memory cell; the current cell state is obtained by reconstructing the cell through the forgetting gate and the input gate C_t The output gateo_t The output gates are used to control the state of the cell C_t The output gates are used to control the state of the cell so that it is transferred to the next cell.

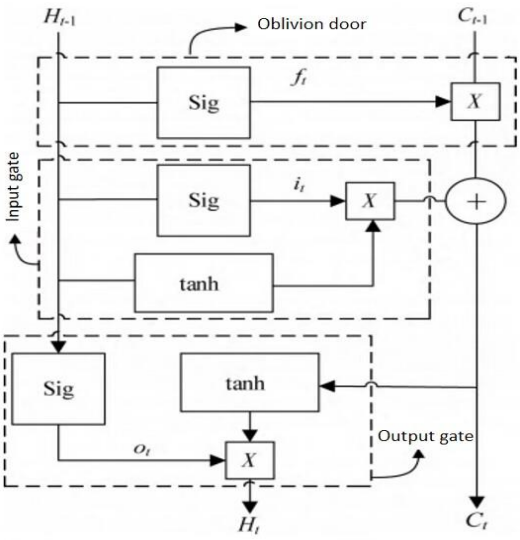


Figure 2 LSTM network gate cell structure

However, the LSTM model also has shortcomings, when dealing with data samples with a large number of features, overfitting is prone to occur, which requires the use of some optimization algorithms to find the optimal number of implied layers and initial learning rate and other parameters to increase the model nonlinear fitting performance and prediction accuracy [20].

2.3 Particle Swarm Optimization (PSO) algorithm

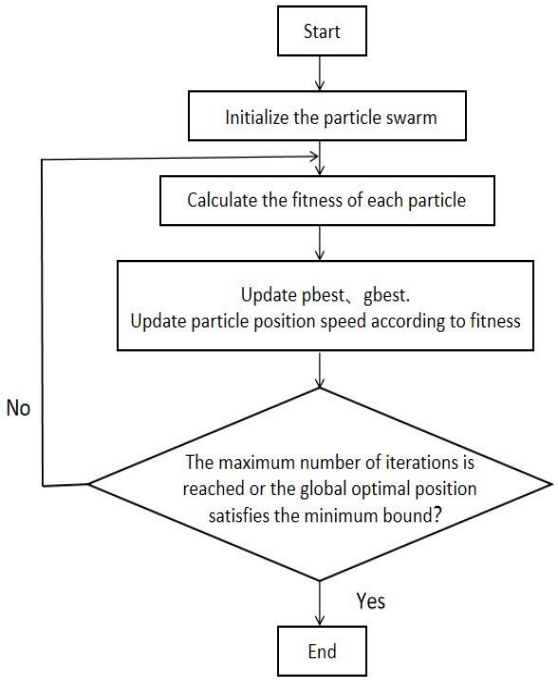


Figure 3 Particle swarm optimization algorithm optimization process

The particle swarm algorithm (PSO) is an intelligent algorithm developed by observing the social behavior of birds. The PSO algorithm is similar to the flock feeding process, and is widely used in the global optimization process of hyperparameters due to its simple principle and easy operation, which refers to the individuals in the population as a particle, and each particle is a possible solution of the optimized parameter in the global search space. Each particle is a possible solution of the optimized parameter in the global search space, and its characteristic index mainly includes three aspects: position, speed and fitness value. Firstly, the fitness value of each particle is calculated by the fitness function to memorize the optimal position and speed of all particles. In each iteration, the particle reaches a new position by adjusting the velocity component of any dimension and calculating it, and so on, until the particle finds the optimal position or reaches the number of iterations, so as to complete the optimization process of the particle in the multidimensional search space, the particle swarm optimization algorithm is shown in Figure 3. In this paper, we use the PSO algorithm to optimize the hyperparameters in the CNN-LSTM model and derive the optimal solution to avoid the overfitting phenomenon during model training.

2.4 CNN-LSTM-PSO hybrid model

In the regression prediction of tool wear, the convolutional layer in the CNN model is first used to obtain the weight parameters, and the pooling layer is used for dimensionality reduction to mine the local features related to tool wear, and its output is a one-dimensional spatial feature vector. The output feature vector is then trained as an LSTM model, which enables the two models to complement each other in time and space, thus improving the accuracy of prediction. Figure 4 shows the CNN-LSTM model prediction process based on particle swarm optimization for multi-channel feature fusion proposed in this paper. The essence is to use the CNN convolutional neural network model as a spatial feature extractor and the LSTM model as a trainer for regression prediction, based on which the superparameters such as initial learning rate and number of hidden layer units in the LSTM model are optimized by the PSO algorithm, so that the model nonlinear fitting performance is improved and the tool wear prediction effect is optimized, and the specific steps are as follows:

Step 1: The original signals of the 7 channels are processed for noise reduction and feature extraction and fusion in the time domain, frequency domain and time-frequency domain, respectively.

Step 2: Using Pearson's correlation coefficient formula to downscale the above feature data to construct the training and test sets of the model.

Step 3: Build a convolutional neural network, train it using the training set and test set from step 2, output a spatial feature vector, and form a new training set.

Step 4: The hyperparameters such as initial learning rate and number of hidden layer units in the LSTM model are used as optimization-seeking processing objects by the PSO algorithm, and the particle swarm optimization model is initialized.

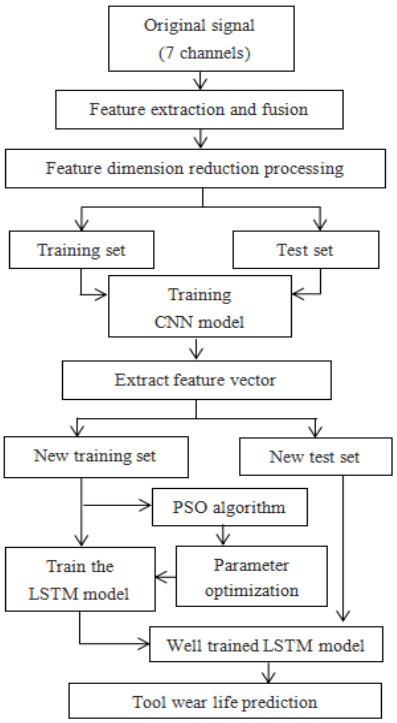


Figure 4 CNN-LSTM-PSO model prediction flow

Step 5: As shown in Figure 6, firstly, the particle's fitness value is calculated, secondly, pbest, gbest are updated according to the fitness, and finally, the position and velocity of the updated particle are recorded.

Step 6: When the maximum number of iterations is reached or the most suitable position is found, the whole loop is terminated and the optimal hyperparameters are derived. If the termination condition is not reached, then return to step 5 for the next iteration.

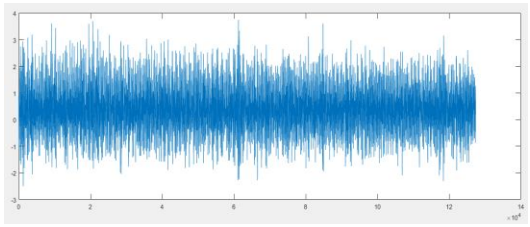
Step 7: Train the LSTM model with the training set formed in step 3 and the hyperparameters obtained in step 6, thus completing the regression prediction of tool wear.

3 Construction of tool wear sample dataset

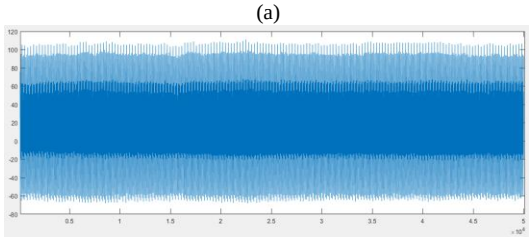
3.1 Multi-channel feature extraction and fusion

The experimental data were obtained from the open data of the 2010 High Speed CNC Machine Tool Health Prediction Contest of the Prediction and Health Management Society (PHM), New York, USA[21]. The dataset is the result of real-time tool wear monitoring experiments on six ball-ended milling tools. In this paper, the experimental dataset of group C1 is selected, where the experimental data of the first 200 tool walks of group C1 is used as the sample training set and the experimental data of the last 115 tool walks are used as the test set. The original signals in each dataset include X-axis, Y-axis and Z-axis cutting force signals, X-axis, Y-axis and Z-axis vibration signals and acoustic emission signals, among which cutting force signals and vibration signals contain 3 channels and acoustic emission signals are 1 channel signals, totaling 7 channels.

In this experiment, the tool is walked once every Δt time, and each time the tool is walked, the original signal of 7 channels can be collected, and the number of collected points of the original signal of each single walk is about 200000 or more, which shows that the number of signal data is huge and there is a lot of noise, and these noises are often caused by the instability of the system at the moment the tool is cut in and out. This requires noise reduction for all types of raw signals collected above to avoid adverse effects during model training. Therefore, the sampling points with data labels from 50001 to 100000 in the raw signal are collected respectively as the research object. The results of the comparison between the original signal and the noise reduction signal are shown in Figure 5, which shows that the signal fluctuation after noise reduction is uniform and noiseless. In this experiment, the number of tool walks in each channel is 315, and there are 7 channels in total, so the original signal can form a 315×7 tool wear signal matrix after noise reduction.



(a)



(b)

Figure 5 Comparison results between the original signal and the noise reduction signal.(a) Raw signal data.(b) Signal data after noise reduction

The 315×7 signal matrix after noise reduction is extracted in the time domain, frequency domain and time-frequency domain, and the time-domain information mainly includes mean, standard deviation, root mean square, etc., totaling 13 time-domain features; the frequency domain information mainly includes frequency domain amplitude mean, center of gravity frequency, mean square frequency, etc., totaling 5 frequency domain features; the wavelet packet decomposition is performed on the original signal, resulting in 8 frequency bands, and the energy of each frequency band is used as time-frequency domain information, totaling 8 time-frequency domain features. The energy of each frequency band is used as time-frequency domain information, and the total is 8 time-frequency domain features, so 26 features can be extracted from each channel signal. The features of all channels are fused to obtain 182 features, and the matrix is reorganized to obtain a 315×182 feature matrix.

3.2 Feature dimensionality reduction processing

The ball-head milling cutter used in this experiment has three teeth in the CNC machining process. In order to improve the accuracy and precision of tool wear prediction, the rear face wear of each tooth needs to be measured and its average value is taken to characterize the actual wear of the tool. In this experiment, there are 315 tool walks, and the average value of the measured wear after each tool walk is composed of a sample target matrix with a matrix dimension of 315×1. Each value in the sample target matrix is the output data of the CNN-LSTM-PSO wear prediction model. In this experiment, the LEICA MZ12 microscope was used to measure the tool rear face wear, and its C1 group tool wear variation curve is shown in Fig. 6, and its variation pattern is consistent with the temporal information mentioned in the previous section.

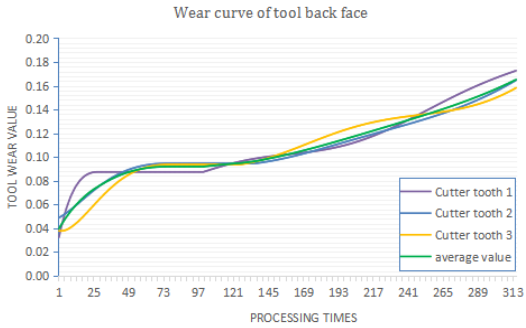


Figure 6 Test tool wear variation curve

According to the above, the extracted features yielded a feature matrix of 315 × 182 by multi-channel feature fusion, but not all the features can characterize the wear of the back tool face. In order to find the correlation between the feature matrix and the target matrix more clearly, the above multi-channel fused feature matrix and the tool wear value are normalized, and the normalized processing formula is

$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{7}$$

The correlation between the normalized sample data and the tool wear curve is shown in Figuer 7. It can be seen from the figure that there are many features that do not correlate with the tool wear values or have weak correlation that will interfere with the tool wear prediction model and should be given to be removed. And Pearson correlation coefficient is the most widely used correlation coefficient analysis method, which can be used to measure the correlation between the extracted feature values and tool wear [22]. Its calculation formula is:

$$P_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum {x_i}^2 - (\sum x_i)^2}\sqrt{n\sum {y_i}^2 - (\sum y_i)^2}} \tag{8}$$

where P_{xy} denotes the Pearson correlation coefficient of the signal feature x and the tool wear value y. The Pearson correlation coefficient formula is used to calculate the 315 × 182 feature matrix and filter out $|P_{xy}| \geq 0.9$ the strongly correlated features as the input of the prediction model. In total, 47 strongly correlated features are extracted through the calculation, taking the X-axis cutting force signal as an example, 7 strongly correlated features are obtained after dimensionality reduction, and the sample data of the signal after dimensionality reduction are shown in Figure 8, which shows that the noise signals with poor correlation are deleted, and the stripped out data with poor correlation are shown in Figure 9; in this paper, the 47 strongly correlated features are fused and reorganized, and the dimensionality of the sample feature matrix is 315×47. 47, and this sample feature matrix is the input data of CNN-LSTM-PSO wear prediction model.

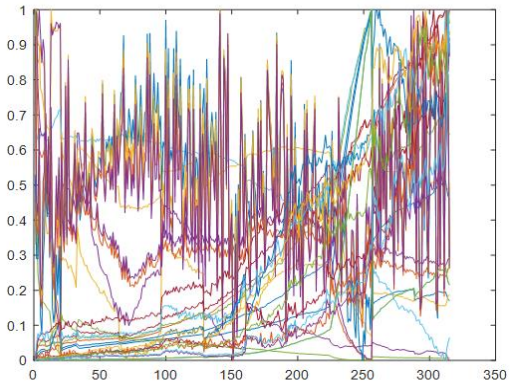


Figure 7 Normalised sample dataset

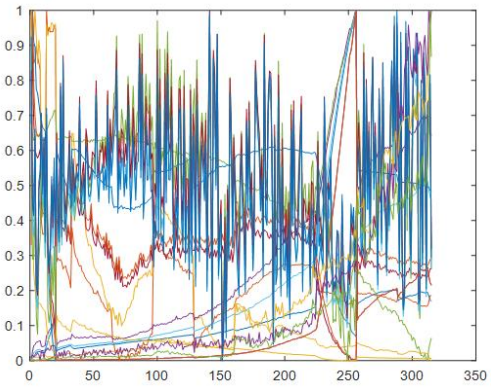


Figure 8 Sample data after dimensionality reduction

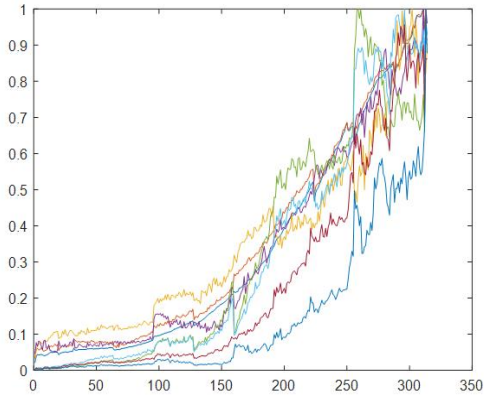


Figure 9 Deleted poor correlation data

4 Experimental verification and analysis of tool wear

4.1 Tool wear experimental conditions

The experimental conditions for tool wear are shown in Figure 10, whose cutting vibration signals were collected using a Kistler 8636C piezoelectric accelerometer, cutting force signals were collected using a Kistler 8152 three-way platform dynamometer, and acoustic emission signals were collected using a Kistler 9265B acoustic transmitter, whose relevant CNC machining cutting parameters are shown in Table1.

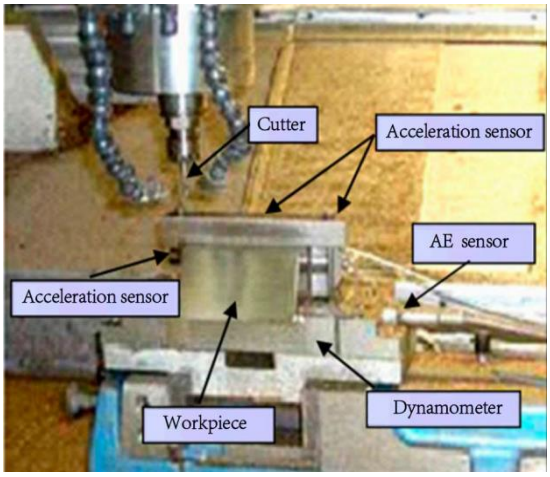


Figure 10 Experimental conditions for tool wear

Table 1 CNC machining cutting parameters

Main shaft Rotational Speed	Feeding Speed	Back draft	Side Eating Knife quantity	Feed amount	Cold cutting conditions
10400	1555	0.2	0.125	0.001	Dry cutting

In this paper, the raw signals related to tool wear are collected in real time according to the above experimental conditions, and each channel raw signal is processed by noise reduction, extraction, fusion and dimensionality reduction to obtain a 315×47 sample feature matrix, and a sample dataset with spatio-temporal correlation of traffic flow is jointly constructed with 315×1 sample target matrix with dimensionality of 315×48. CNN-LSTM-PSO The model first inputs the sample dataset into the multilayer CNN model to extract the spatial sequence features of the traffic flow data and outputs the spatial feature vector. Then the spatial feature vector is input to the multilayer LSTM model to extract the time series features of the data, thus combining the temporal features and spatial features. Finally, the PSO algorithm is used to optimize the hyperparameters in the CNN-LSTM model, so as to complete the prediction of tool wear.

4.2 Setting of prediction model parameters

In order to avoid the influence of external factors, the number of particle swarm individuals in the PSO algorithm is set to 15 and the maximum number of iterations is set to 60. The values of the initial learning rate parameter of the optimized CNN-LSTM model are set between 0.001 and 0.01, and the values of the number of hidden layer units are set between 1 and 100. The structural parameters of the tool wear prediction model after hyperparametric optimization based on the PSO algorithm are shown in Table 2.

Table 2 Structural parameters of CNN-LSTM-PSO model

Structural section	Network structure Name	Parameter settings
1	Convolutional layer 1 Batch standardisation layer 1 Pooling layer 1	Activation function: RELU Convolution kernel: 3*3 Maximum pooling
2	Convolutional layer 2 Batch standardisation layer 2 Pooling layer 2	Activation function: RELU Convolution kernel: 3*3 Maximum pooling
3	LSTM layer 1	Learning rate: 0.004 Number of hidden layer units: 50 Activation function: Sigmoid
4	LSTM layer 2	Learning rate: 0.004 Number of hidden layer units: 32 Activation function: Sigmoid
5	Dropout layer	25% discard
6	Output layer	Activation function: Softmax

In order to quantify the prediction performance of the tool life model, three objective evaluation indexes are selected, namely the mean absolute error MAE, the root mean square error RMSE and the coefficient of determination R2. Among them, the mean absolute error MAE can obtain an evaluation value, but the comparison between different models is required to reflect the model's superiority; the mean square error RMSE and the coefficient of determination R2 can directly characterize the model's superiority. The smaller the mean square error RMSE and the closer the coefficient of determination R2 is to 1, the higher the accuracy and precision of the prediction model. The three evaluation indicators are calculated as follows:

$$MAE = \frac{\sum_{t=1}^{m}|y_t - \hat{y}_t|}{m} \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{m}(y_t - \hat{y}_t)^2}{m}} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{m}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{m}(y_t - \bar{y})^2} \tag{11}$$

where, m is the number of samples output from the fully connected layer, the number of samples in this paper is 315, and $\hat{y}_t$ is the predicted value of tool wear, and y_t is the actual value of tool wear.

4.3 Tool life prediction results

In this paper, a CNN-LSTM model with multi-channel feature fusion using particle swarm optimization is used for tool wear regression prediction, and its test set prediction results are shown in Figure 11. The mean absolute error MAE value of the model was calculated to be 0.5848, the root mean square error RMSE value was 0.7281, and the coefficient of determination R2 value was 0.9964. The results show that the use of CNN-LSTM-PSO based model can effectively perform regression prediction of tool wear and achieve better results.

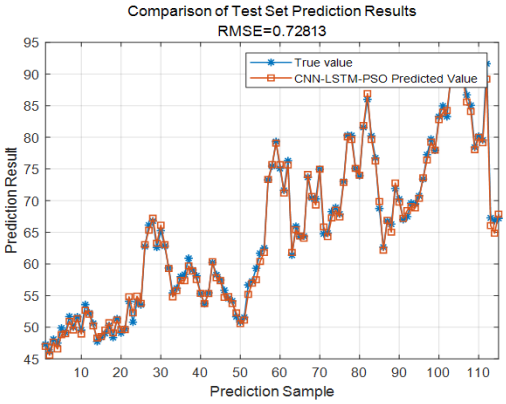


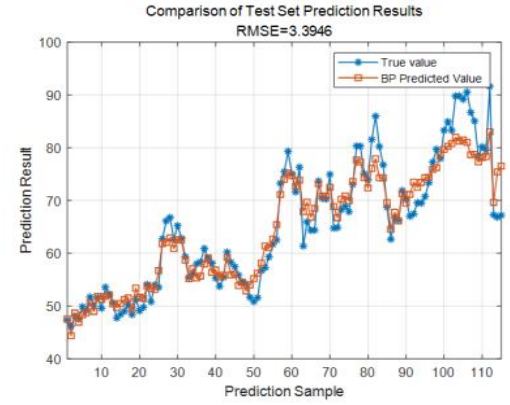
Figure 11 CNN-LSTM-PSO test set prediction results

Table 3 shows the effect of the PSO algorithm on the tool wear regression prediction model, where the hyperparameters such as the initial learning rate and the number of hidden layer units of the CNN-LSTM model rely on manual random selection, and it can be seen that the CNN-LSTM model optimized using the PSO algorithm has the best tool wear prediction. Compared with the CNN-LSTM model, its mean absolute error MAE and root mean square error RMSE are reduced and the coefficient of determination R2 is improved, and its performance index exceeds 0.99, while the performance index of the CNN-LSTM model with manually selected parameters is maintained at a maximum of about 0.98. This is mainly because the PSO algorithm obtained more accurate hyperparameter pairings after hyperparameter optimization of the CNN-LSTM model, which found the most critical attributes affecting the accuracy of tool wear prediction and avoided the blindness of setting parameters, thus improving the prediction results.

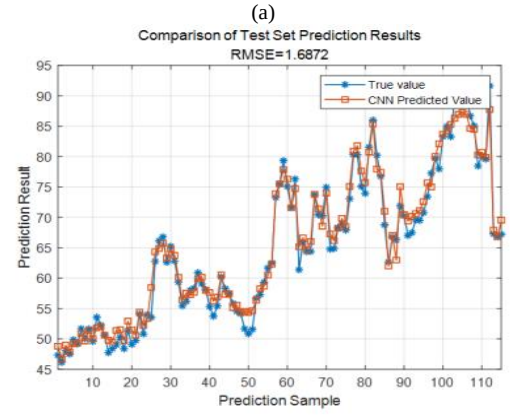
Table 3 Effect of PSO algorithm on prediction model

Algorithm	Initial learning rate	Number of hidden layer units		Test set prediction results		
		LSTM 1	LSTM 2	MAE	RMSE	R2
CNN-LSTM	0.01	100	50	2.9757	3.5829	0.9128
	0.01	60	20	2.2307	3.0005	0.9388
	0.001	100	50	2.0172	2.1781	0.9675
	0.001	60	20	0.9718	1.1914	0.9802
CNN-SVM-PSO	0.004	50	32	0.5848	0.7281	0.9964

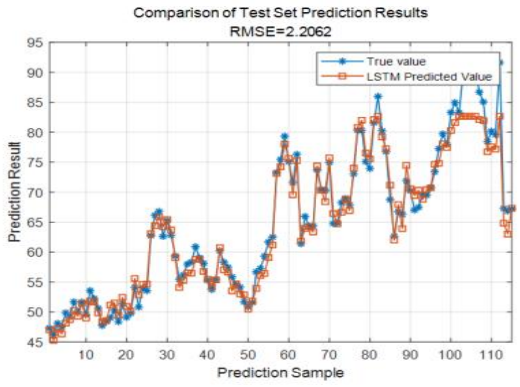
To further validate the prediction performance of CNN-LSTM-PSO based tool wear, a comparative analysis was performed with other traditional prediction models in the past, such as BP neural network, CNN model, LSTM model and CNN-LSTM model. Figure 12 shows the comparison results of the four traditional tool wear prediction models, and it can be seen from Figure 11 that the root mean square error RMSE values of the CNN-LSTM-PSO model proposed in this paper are reduced by 78.59%, 56.85%, 66.99%, and 38.89% compared to the BP model, CNN model, LSTM model, and CNN-LSTM model, respectively. This shows that the prediction performance of the CNN-LSTM tool wear prediction model optimized based on the PSO algorithm proposed in this paper is superior due to the single algorithm of other traditional prediction models, incomplete feature extraction, and over-reliance on signal processing techniques and expert experience.



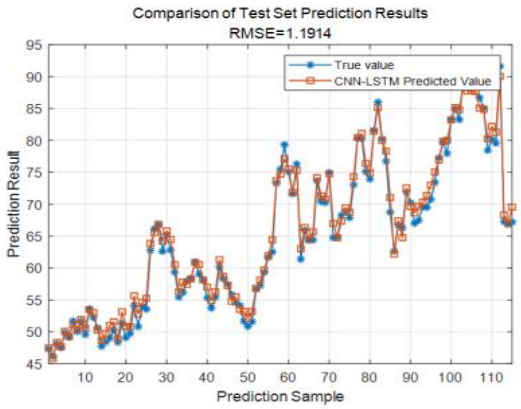
(a)



(b)



(c)



(d)

Figure 12 Prediction results of the four traditional models

(a) BP model. (b) CNN model. (c) LSTM model.(d) CNN-LSTM model.

Table 4 Comparison of prediction performance results of five models

Algorithm	Test set prediction results		
	MAE	RMSE	R2
BP Neural Network	2.5413	3.3946	0.9211
CNN Algorithms	1.3242	1.6872	0.9705
LSTM Algorithms	1.5425	2.2062	0.9667
CNN-LSTM Algorithm	0.9718	1.1914	0.9802
CNN-SVM-PSO Algorithm	0.5848	0.7281	0.9964

Table 4 shows the comparison results of the prediction performance of the five models. It is found that the CNN-LSTM-PSO model using multi-channel feature fusion for tool wear prediction has the smallest value of mean absolute error MAE, which is reduced by 76.98%, 55.84%, 62.09% and 39.82% compared to the BP, CNN, LSTM and CNN-LSTM models, respectively ; the value of the coefficient of determination R2 is closest to 1, which is 7.56%, 2.60%, 2.98%, and 1.63% higher compared to the BP, CNN, LSTM, and CNN-LSTM models, respectively. These two results again prove that the prediction of tool wear values using the CNN-LSTM-PSO model proposed in this paper is more accurate and can achieve more effective monitoring of remaining tool life and intelligent tool change.

5 Conclusion

In this paper, the open dataset of the tool health prediction competition is selected as the original data, and the original data is preprocessed using feature extraction and multi-channel fusion techniques, and then a CNN-LSTM model based on particle swarm optimization with multi-channel feature fusion is proposed to predict the tool wear values during milling machining, and compared with other single mechanical models and the traditional CNN-LSTM model analysis, and the results show that:

(1) In this paper, the CNN model is used to extract local features from the feature matrix after multi-channel fusion and dimensionality reduction to obtain important information of tool wear data and avoid the interference of tool wear data by other factors.

(2) The parameter search optimization of the tool wear prediction model by the particle swarm PSO algorithm reduces the subjective influence of manual parameter selection and avoids the blindness of setting parameters, thus improving the prediction accuracy.

(3) Tool wear regression prediction using the CNN-LSTM-PSO model has a mean absolute error MAE value of 0.5848, a root mean square error RMSE value of 0.7281, and a coefficient of determination R2 value of 0.9964. This indicates that the model can effectively predict the remaining life of the tool with good results.

(4) Compared with the BP model, CNN model, LSTM model and CNN-LSTM model, the mean absolute error MAE and root mean square error RMSE values of the CNN-LSTM-PSO model proposed in this paper have been reduced, and the value of the coefficient of determination R2 has been improved to be closest to 1. This indicates that the constructed tool life prediction model has less error, better accuracy and better.

In the future, the CNN-LSTM-PSO tool wear prediction model can be widely used in the fields of intelligent tool change and tool life management for CNC machining in various factories. By predicting the tool wear value in real time, it can realize intelligent tool change in advance when the tool wear is at the critical threshold, thus improving the machining accuracy of products.

Author Contributions: For Conceptualization, methodology, analysis, and writing original draft preparation, Wang Shuo; writing review and full-text editing, Yu Zhenliang; writing—original draft preparation, Guo Yongqi; writing—original draft preparation, Liu Xu.

Conflicts of interest: The authors declare no conflict of interest.All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The research work financed with the means of Basic Scientific Research Youth Program of Education Department of Liaoning Province, No.LJKQZ2021185; Yingkou Enterprise and Doctor Innovation Program (QB-2021-05).

References

[1] Andis Ābele, Henn Tuherm. Predictions of Cutting Tool Wear of Straight Milled Aspen Wood with Taylor's Equation [J]. Current Journal of Applied Science and Technology, 2016(5):7.

[2] Cynthia Deb, M. Ramesh Nachiappan, M. Elangovan, V. Sugumaran. Fault Diagnosis of a Single Point Cutting Tool using Statistical Features by Random Forest Classifier [J]. Indian Journal of Science and Technology, 2016, 9(33):45-46.

[3] Xu Yanwei, Gui Lin, Xie Tancheng. Intelligent Recognition

Method of Turning Tool Wear State Based on Information Fusion Technology and BP Neural Network [J]. Shock and Vibration, 2021(2):55-56.

[4] Alajmi Mahdi, Almeshal Abdullah. Estimation and Optimization of Tool Wear in Conventional Turning of 709M40 Alloy Steel Using Support Vector Machine (SVM) with Bayesian Optimization [J]. Materials,2021, 14(14):23.

[5] Wei Weihua, Cong Rui, Li Yuantong. Prediction of tool wear based on GA-BP neural network [J]. Proceedings of the Institution of Mechanical Engineers,2022, 236(12):8-9.

[6] Sarat Babu Mulpur, Babu Rao Thella. Multi-sensor heterogeneous data-based online tool health monitoring in milling of IN718 superalloy using OGM (1, N) model and SVM [J]. Measurement, 2022(199):723-724.

[7] Lee Hojin, Jeong Hyeyun, Koo Gyogwon. Attention RNN Based Severity Estimation Method for Interturn Short-Circuit Fault in PMSMs [J]. IEEE Transactions on Industrial Electronics, 2020(5):7.

[8] Han Sung-Ryeol, Kim Yun-Su. A fault identification method using LSTM for a closed-loop distribution system protective relay [J]. International Journal of Electrical Power and Energy Systems, 2023(148):5-8.

[9] Zhou Yuankai, Wang Zhiyong, Zuo Xue. Identification of wear mechanisms of main bearings of marine diesel engine using recurrence plot based on CNN model [J]. Wear, 2023(6): 520-521.

[10] Ma Kaile, Wang Guofeng, Yang Kai. tool wear monitoring for cavity milling based on vibration singularity analysis and stacked LSTM [J]. The International Journal of Advanced Manufacturing Technology, 2022(120):5-6.

[11] Lim Meng Lip, Derani Mohd Naqib, Ratnam Mani Maran, Yusoff Ahmad Razlan. tool wear prediction in turning using workpiece surface profile images and deep learning neural networks [J]. The International Journal of Advanced Manufacturing Technology, 2022(120):11-12.

[12] Jiahang L, Xu Z. Convolutional neural network based on attention mechanism and BiLSTM for bearing remaining life prediction [J]. Appl Intell. 2021(52):1076 -1091.

[13] Stefan Droste, Thomas Jansen, Ingo Wegener. Upper and Lower Bounds for Randomized Search Heuristics in Black-Box Optimization. Electron. Colloquium Comput [J]. Complex, 2003(77):48-48.

[14] Weifeng Lu, Bingyu Cai, Rui Gu. Improved Particle Swarm Optimization Based on Gradient Descent Method [J]. CSAE, 2020(2): 121-126.

[15] Salih Omran, Duffy Kevin Jan. Optimization Convolutional Neural Network for Automatic Skin Lesion Diagnosis Using a Genetic Algorithm [J]. Applied Sciences, 2023,13(5):7.

[16] Zhang Xin, Jiang Yueqiu, Zhong Wei. Prediction Research on Irregularly Cavitied Components Volume Based on Gray Correlation and PSO-SVM [J]. Applied Sciences,2023,13(3):87-89.

[17] Wang Ji, Zhou Jian, Mo Wen-An. Tool life prediction based on multi-source feature PSO-SVR neural network [J]. Journal of Physics: Conference Series, 2022,2366(1):754-756.

[18] Gajera Himanshu K., Nayak Deepak Ranjan, Zaveri Mukesh A.. A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features [J]. Biomedical Signal Processing and Control, 2023,79(2):46-50.

[19] Ning Zhang, Enping Chen, Yukang Wu, et al. A novel hybrid model integrating residual structure and bi- directional long short- term memory network for tool wear monitoring [J].The International Journal of Advanced Manufacturing Technology , 2022(120):6707-6722

[20] Li Xianwang, Qin Xuejing, Wu Jinxin, et al. tool wear prediction based on convolutional bidirectional LSTM model with improved particle swarm optimization [J]. The International Journal of Advanced Manufacturing Technology, 2022,123(11-12):89-92.

[21] Huimin Chen. A Multiple Model Prediction Algorithm for CNC Machine Wear PHM [J]. International Journal of Prognostics and Health Management, 2011,2(2):78-89.

[22] Li Yifan, Xiang Yongyong, Pan Baisong, et al. A hybrid remaining useful life prediction method for cutting tool considering the wear state [J]. The International Journal of Advanced Manufacturing Technology, 2022(121):5-6.

Viser Technology Pte. Ltd.
Mechanical Engineering Science
DOI: 10.33142/mes.v4i2.9087

Research Article Open Access

Multi-objective reliability optimization design of high-speed heavy-duty gears based on APCK-SORA model

Zhenliang YU*, Shuo WANG, Fengqin ZHAO, Chenyuan LI

School of Mechanical and Power Engineering,Yingkou Institute of Technology , Yingkou, China

***Corresponding Author:** Zhenliang YU, email address:yuzhenliang_neu@163.com*

Abstract:
For high-speed heavy-duty gears in operation is prone to high tooth surface temperature rise and thus produce tooth surface gluing leading to transmission failure and other adverse effects, but in the gear optimization design and little consideration of thermal transmission errors and thermal resonance and other factors, while the conventional multi-objective optimization design methods are difficult to achieve the optimum of each objective. Based on this, the paper proposes a gear multi-objective reliability optimisation design method based on the APCK-SORA model. The PC-Kriging model and the adaptive k-means clustering method are combined to construct an adaptive reliability analysis method (APCK for short), which is then integrated with the SORA optimisation algorithm. The objective function is the lightweight of gear pair, the maximum overlap degree and the maximum anti-glue strength; the basic parameters of the gear and the sensitivity parameters affecting the thermal deformation and thermal resonance of the gear are used as design variables; the amount of thermal deformation and thermal resonance, as well as the contact strength of the tooth face and the bending strength of the tooth root are used as constraints; the optimisation results show that: the mass of the gear is reduced by 0.13kg, the degree of overlap is increased by 0.016 and the coefficient of safety against galling Compared with other methods, the proposed method is more efficient than the other methods in meeting the multi-objective reliability design requirements of lightweighting, ensuring smoothness and anti-galling capability of high-speed heavy-duty gears.
***Keywords:** APCK-SORA model; high-speed heavy-duty gears; multi-objective reliability optimization design; k-means clustering method*

1 Introduction

High-speed and heavy-duty gears are widely used in aerospace and aviation, the marine industry and high-speed trains, and gear devices of various industries are also developing in the direction of high speed, heavy load and light weight. The surrogate model technique converts the actual complex structural problem into an approximate mathematical problem to be solved, which not only improves the computational efficiency of the optimised design model, but also allows the performance of the whole structure in the design space. Omar D. M et al [1] can change the contact pattern of tooth surfaces and proposed a structured optimisation method. Li et al [2] proposed a multi-objective ant colony optimisation model for improving the meshing performance and dynamic characteristics of gear transmission systems for high-speed heavy-duty herringbone gears used in the marine sector. Zhao et al [3] used the potential energy method to study the effect of tooth root cracking on gear meshing stiffness. Daniel et al [4] used a genetic algorithm to optimise the parameters of normal load, sliding speed and friction coefficient of a gear pair with multiple objectives and carried out an analytical calculation of the transmission efficiency of the gear. Dixit et al [5] used CRITIC (Criteria Importance through Intercriteria Correlation) method and Genetic Algorithm (GA) to obtain the optimal solution for multi-objective optimization considering the weight of the gear pair, power loss and gear heat treatment time. Maruti et al. [6] used an improved non-dominated sorting genetic algorithm (NSGA-II) to perform multi-objective optimization of three different gear profiles (unmodified profiles, smooth engagement profiles and high load-bearing energy profiles) and four ISO oil grades at two speeds. Edmund et al [7] carried out a multi-objective optimal design of a two-stage straight cylindrical gearbox with volume, power output and centre distance as objectives. Emna et al [8] considered both micro and macro parameters of gears, and used genetic algorithm to optimize gears for multi-objective optimization. Ekansh et al [9] carried out a multi-objective optimization of gears considering production cost, gear strength and noise impact parameters. Zhang [10] proposed a new collaborative strategy (C-RBMDO), which is a combination of a performance metric approach (PMA) and a parallel subspace optimization strategy (CSSO) and decouples the SORA for multidisciplinary optimal design of gears. At present, the sequence optimisation and reliability assessment methods are mainly based on the primary second order moment theory method for

Received on October 8, 2022; Accepted on November 31, 2022

reliability analysis calculation, which has limited computational efficiency and accuracy. Some scholars have used different reliability methods in combination with SORA to make up for their shortcomings. For example, Cho and Lee [11] used a convex programming approach to transform the reliability optimisation column into a series of subplanning columns in a convex design domain, and introduced the hybrid mean value method (Hybird Mean Value,) to improve the computational performance of the SORA method. Du et al. [12] proposed a new search format for MPTP points to improve the robustness of the SORA method.On this basis, Cheng et al. [13] used the change of Angle in the iteration process to determine the convergence performance of MPTP search, and proposed to reduce the calculation times of non-tight constraint reliability information and improve the calculation efficiency of SORA by using the feasibility determination method of probability constraints. Ilchi [14] used sequence optimization and reliability analysis (SORA) to reduce the RBDO based on an improved adaptive chaos control method computational cost, a two-step improved adaptive chaos control method (DS-MACC) was proposed to speed up the cycle of reliability analysis. Subsequently, Ilchi [15] proposed a sequential optimisation and reliability analysis (SORA) method based on a PMA method for selecting adaptive step sizes at normalised locations and negative gradient vectors at two consecutive iteration points. Kaveh A [16] used sequential optimization with reliability assessment (SORA) as a decoupling method and proposed a reliability design optimization (RBDO) framework based on a meta-heuristic algorithm for decoupling methods. Kaveh A[17] used reliability-based design optimization (RBDO) to deal with these uncertainties and proposed to apply the sequential optimization with reliability assessment-dual meta-heuristic (SORA-DM) framework applied to RBDO of frame structures.

When high speed and heavy loads are applied to gears, the relative sliding between the tooth surfaces creates a lot of frictional heat, which raises the temperature of the gear teeth and alters the thermal expansion characteristics of the gear material. This leads to changes in the theoretical involute of the tooth profile and heat transfer errors of the gear, which have a negative impact on the accuracy, smoothness, and noise level of the transmission. In order to solve such problems, an adaptive surrrogate model-based reliability optimisation method based on an improved SORA optimisation algorithm with an adaptive PC-Kriging model is proposed. Firstly, a new adaptive structural reliability analysis method (referred to as APC-Kriging) is constructed by combining the PC-Kriging model with an adaptive k-means clustering method. Secondly, the proposed adaptive PC-Kriging model is used to solve the reliability part of the SORA optimisation algorithm and then to optimise the design by SORA. In order to achieve a multiobjective optimised design for the reliability of high speed heavy load gears with lighter volume, better transmission smoothness and anti-galling capability.

2 High-speed heavy-duty gear model considering the effect of temperature rise

To improve the transmission accuracy, smoothness and load carrying capacity of high-speed heavy-duty gears, a finite element analysis of gears considering temperature rise is carried out, i.e. a thermal analysis of gears based on thermal-structural coupling. For the thermal analysis, Solid70 three-dimensional solid units are used, and the material properties such as heat transfer coefficient, heat flow density, modulus of elasticity, specific heat capacity and related boundary conditions are set. The model is divided by sweeping the mesh, and only the mesh refinement of its mesh surface, while the rest of the mesh density can be relatively sparse, divided into the master gear and driven gear of the finite element model as shown in Figure 1, the relevant loading coefficients are shown in Table 1.

Table 1 Loading factors

Modulus of elasticity	Poisson's ratio	Thermal conductivity	Heat flow density	Linear expansion coefficient
206	0.3	29.7	542.2	10.36

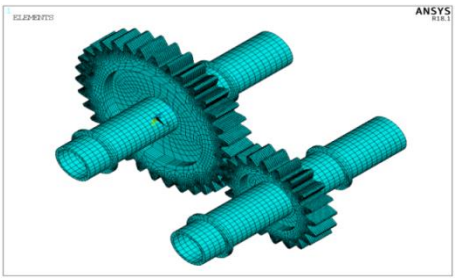
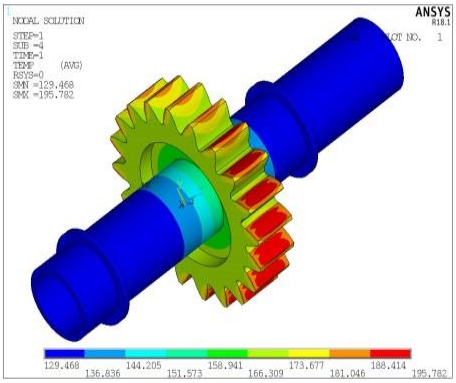
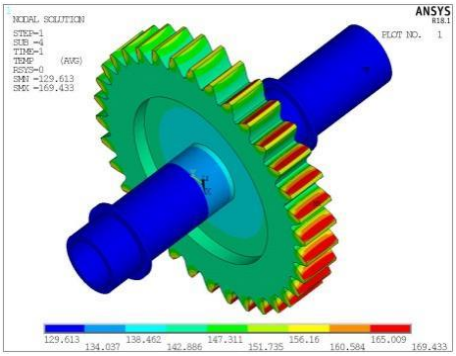


Figure 1 The three-dimensional solid model of the gear pair after meshing

According to the literature [18], the heat flow density and convective heat transfer coefficients are calculated for each face of the gear and are applied simultaneously to the meshing face of the gear, and the convective heat transfer coefficients are applied to the non-meshing, top and root faces of the gear as well as the end face of the gear with relevant settings, and finally the steady-state temperature field of the gear is solved. The steady-state temperature field clouds of the driven and driven gears are shown in Figure 2(a) and (b).



Cloud of active wheel temperature field distribution



Cloud of driven wheel temperature field distribution

Figure 2 Steady Temperature Distribution of gear

As can be seen from Figure 2, the maximum temperature on the meshing tooth surface of the active gear is (195.702 ℃) and the maximum temperature on the meshing tooth surface of the driven gear is (169.433 ℃), and the high temperature areas on the meshing tooth surface of both the active and driven gears are found near the root and the top of the teeth, which is due to the fact that the contact compressive stress between the teeth and the relative sliding speed between the teeth during the transmission process of the gears is greater in these two areas. Because in the process of gear transmission, the product of the contact compressive stress between the tooth surfaces and the relative sliding speed between the teeth is large at these two places, and then more friction heat input is generated, which is consistent with the fact that gear gluing usually occurs near the root of the driving wheel or near the top of the driven wheel. It can be proved that the analysis method of gear steady-state temperature field is correct and effective.

3 Mathematical model of APCK-SORA

An adaptive structural reliability analysis method based on a combination of PC-Kriging and adaptive k-means (referred to as APC-Kriging) is proposed. Firstly, PC-Kriging is an improved Kriging algorithm whose regression basis function uses a sparse polynomial optimal truncation set to approximate the global behaviour of the numerical model, and Kriging is used to handle local variations in the model output, which improves the computational efficiency while ensuring accuracy. Secondly, while common reliability methods collect sample points one by one, the adaptive k-means clustering analysis in this paper divides the space into several regions and selects an optimal sample point from each region, thus enabling multiple regions to simultaneously achieve the aim of improving the accuracy of the PC-Kriging model, thus again improving the computational efficiency of the model. Finally, the proposed adaptive PC-Kriging model is combined with SORA to construct a multi-objective reliability optimisation method based on the adaptive surrogate model.

3.1Adaptive PC-Kriging reliability model

The adaptive PC-Kriging method uses a k-means cluster analysis approach to ensure that a number of sample points that contribute significantly to the probability of failure are added at each iteration. The main steps of the proposed adaptive PC-Kriging method for selecting sample points are as follows:

Step 1 t=0. The initial experimental design sample points are generated by random sampling of the Latin hypercube and the corresponding functional response values are computed exactly, i.e. $\hat{G}_0(\boldsymbol{x})$, $\hat{P}_{f,0}$. Let the initial number of sample points be M_0, then we have $\Omega_0=\{(\boldsymbol{x}_{0,i},y_{0,i}),i=1,2,...,M_0\}$, $\boldsymbol{X}_0=\{\boldsymbol{x}_{0,1},\boldsymbol{x}_{0,2},...,\boldsymbol{x}_{0,M_0}\}$.

Step 2 t=t+1, generating K points by Markov chain Monte Carlo simulation (MCMC) on $\hat{G}_{t-1}(\boldsymbol{x})=0$. Given $\hat{G}_{t-1}(\boldsymbol{x})$, a random vector of M dimensions obeying f(x) (f(x) is the joint probability density function of x) and satisfying equation (1) is generated using the MCMC method, then the randomly selected points are considered to be on $\hat{G}_{t-1}(\boldsymbol{x})=0$. The random extraction process stops when the number of extracted points reaches K. The random vector generated is $\tilde{\boldsymbol{X}}_{t-1}=\{\tilde{\boldsymbol{x}}_{t-1,1},\tilde{\boldsymbol{x}}_{t-1,2},...,\tilde{\boldsymbol{x}}_{t-1,K}\}$, and in this paper, let K be 2000 and [ε] be 0.01.

$$|\hat{G}_{t-1}(\boldsymbol{x})|\le[\varepsilon] \tag{1}$$

Step 3 The k-means cluster analysis method is used to divide $\tilde{\boldsymbol{X}}_{t-1}$ into k categories and map the centroids of these k categories onto $\hat{G}_{t-1}(\boldsymbol{x})=0$. Let $\{s_{t-1,1},s_{t-1,2},...,s_{t-1,k}\}$ denote the k clustering centres. When $\hat{G}_{t-1}(\boldsymbol{x})=0$ is a non-linear surface, it is not guaranteed that the centroids of the k categories are all on $\hat{G}_{t-1}(\boldsymbol{x})=0$, and it is necessary to map the centroids that are not on $\hat{G}_{t-1}(\boldsymbol{x})=0$ to $\hat{G}_{t-1}(\boldsymbol{x})=0$. The mapping is done by finding the points that satisfy equation (2) and obtaining $S_{t-1}=\{\hat{\boldsymbol{s}}_{t-1,0},\hat{\boldsymbol{s}}_{t-1,1},...,\hat{\boldsymbol{s}}_{t-1,k}\}$, where $\hat{\boldsymbol{s}}_{t-1,0}$ is the design point for $\hat{G}_{t-1}(\boldsymbol{x})$.

$$\left.\begin{array}{l} min\ \|\boldsymbol{x}-\boldsymbol{s}_{t-1,i}\| \\ s.t.\hat{G}(\boldsymbol{x})=0 \end{array}\right\} \tag{2}$$

where i=1,2,...,k.

Step 4 Adjust the positions of the points in the set S_{t-1}. Define the distance D_0 as shown in equation (3) and assume that if the distance between any two sample points in the set S_{t-1} is less than D_0 it is considered unacceptable and the location of individual points in the set S_{t-1} needs to be adjusted.

$$D_0=e\cdot\left(\frac{2}{M(M-1)}\sum_{i<j}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|\right) \tag{3}$$

where e is a given constant.

There are two possible scenarios for the points in the set S_{t-1}: 1. the distance between some points within S_{t-1} may be too small. $\hat{\boldsymbol{s}}_{t-1,0}$ It is more likely that the distance to some point in $\hat{\boldsymbol{s}}_{t-1,1},...,\hat{\boldsymbol{s}}_{t-1,k}$ is small; 2. The distance between some point in S_{t-1} and some point in X_{t-1} is too small. If case 1 occurs, e.g. the distance between $\hat{\boldsymbol{s}}_{t-1,1}$ and $\hat{\boldsymbol{s}}_{t-1,2}$ is less than D_0, the position of the point with the smaller probability density function is to be changed and the position of the point with the larger probability density function is to be left unchanged; if case 2 occurs, the corresponding point in S_{t-1} is to be changed. The

sample points are adjusted by assuming that the position of $\hat{\boldsymbol{s}}_{t-1,1}$ is to be changed first. The points in $\tilde{\boldsymbol{X}}_{t-1}$ are arranged in ascending order of distance from $\hat{\boldsymbol{s}}_{t-1,1}$ and $\hat{\boldsymbol{s}}_{t-1,1}$ is changed to the new sequence of points in turn until the distance between $\hat{\boldsymbol{s}}_{t-1,1}$ and all points in S_{t-1} and X_{t-1} is greater than D_0 .

Step 5 Calculate the value of the function corresponding to each sample point in the set S.

$$\Omega_{t-1}^{0}=\{(\hat{\boldsymbol{s}}_{t-1,i},y_{t-1,i}),i=0,1,...,k\}\ ,\Omega_t=\Omega_{t-1}\cup\Omega_{t-1}^{0} \quad (4)$$

Step 6 Calculate $\hat{G}(\boldsymbol{x})$, $P_{f,t}$ based on Ω_t and combining Eqs. (1) and (2). If the convergence condition of equation (4) is satisfied, the iterative process stops and $\tilde{P}_{f,t}$ is the estimated value of P_f ; otherwise, return to step 2 until $\tilde{P}_{f,t}$ satisfies the convergence condition.

$$\frac{N_{un}}{N_{fail}}\le\alpha \quad (5)$$

where N_{un} is the total number of samples with sign prediction errors, N_{fail} represents the total number of failed samples, and α is the permissible error of $\hat{P}_f$, where

$$N_{un}=2\left(\frac{N_{U<P}}{2}+N_{P\le U\le Q}\cdot\Phi(-U)\right) \quad (6)$$

$$N_{fail}=\sum_{i=1}^{N_{MC}}I_{\hat{G}}(\boldsymbol{x}_i) \quad (7)$$

In this case, sample points with a high probability of symbol prediction error are indicated by $N_{U<P}$, and such points can be considered as certain failure points. Sample points with a high probability of symbol prediction error are indicated by $N_{P\le U\le Q}$, and the total number of failure prediction errors for such sample points can be indicated by $N_{P\le U\le Q}\cdot\Phi(-U)$. In this study, P=1, Q=2 and α=0.03.

3.2 Development of the APCK-SORA mathematical model

Firstly, PC-Kriging is used to approximate the global behaviour of the numerical model, and Kriging is used to deal with local variations in the model output. Secondly, adaptive k-means cluster analysis is used to divide the space into several regions and select an optimal sample point from each region, so that multiple regions can simultaneously achieve the objective of improving the accuracy and computational efficiency of the PC-Kriging model. Finally, in combination with the SORA optimisation strategy (i.e. separating the reliability assessment from the optimisation design), the model converges and obtains an optimal solution with a small number of cycles, making the solution of complex optimisation design problems simple and efficient.

The mathematical model for APCK-SORA based reliability optimization is:

$$\begin{aligned}
&\mathrm{DV}=(\boldsymbol{d},\boldsymbol{X}^M)\\
&\min\quad f(\boldsymbol{d},\boldsymbol{X}^M,\boldsymbol{Y}^M)=\left\{f_1(\boldsymbol{d},\boldsymbol{X}^M,\boldsymbol{Y}^M),f_2(\boldsymbol{d},\boldsymbol{X}^M,\boldsymbol{Y}^M),\cdots,f_m(\boldsymbol{d},\boldsymbol{X}^M,\boldsymbol{Y}^M)\right\}\\
&s.t.\quad G_i(\boldsymbol{d},\boldsymbol{X}^M-\boldsymbol{s}_i,\boldsymbol{Y}_{MPPi})\ge 0,i=1,2,\cdots,n\\
&\qquad g_j(\boldsymbol{d},\boldsymbol{X}^M,\boldsymbol{Y}^M)\le 0,\ j=1,2,\cdots,p\\
&\qquad h_k(\boldsymbol{d},\boldsymbol{X}^M,\boldsymbol{Y}^M)=0,k=1,2,\cdots,q\\
&\qquad \boldsymbol{d}^L\le\boldsymbol{d}\le\boldsymbol{d}^U\\
&\qquad \boldsymbol{X}^{M,L}\le\boldsymbol{X}^M\le\boldsymbol{X}^{M,U}
\end{aligned} \quad (8)$$

where $\boldsymbol{s}_i=\boldsymbol{X}^{M(k-1)}-\boldsymbol{X}_{MPPi}^{(k-1)}$

The reliability analysis section is then solved using the APC-Kriging proposed in section 3.1. It is determined whether the reliability requirements are met and if not, a deterministic optimisation model is constructed for the next cycle.

3.3 Solving steps for APCK-SORA

In each cycle of SORA, deterministic optimisation is first carried out, followed by reliability analysis. the basic flow of the APCK-SORA method (Figure 4) and the main steps are shown below:

(1) The initial experimental design sample points were first generated by random sampling of the Latin hypercube, and the corresponding functional response values were calculated exactly, i.e. $\hat{G}_0(\boldsymbol{x})$, $\hat{P}_{f,0}$. Let the initial number of sample points be M_0 .

(2) Solve for deterministic optimization. Set the initial value of the optimisation design variable to $d^{(0)},\boldsymbol{X}^{M(0)}$ and the superscript 0 to indicate that no reliability analysis has been performed. Starting from the 1st cycle, get $\boldsymbol{X}_{\mathrm{MPP}i}^{(k)}$ and $\boldsymbol{s}_i$ in the deterministic optimization to build a completely new optimization calculation model.

(3) Perform a reliability analysis at the optimal design points $\boldsymbol{d}^{(k)}$, $\boldsymbol{X}^{M(k)}$ obtained in the deterministic optimization and find the corresponding $\boldsymbol{X}_{\mathrm{MPP}i}^{(k)}$.

(4) Test for feasibility and convergence. If all reliability constraints and deterministic constraints are satisfied and the system objective function values converge, $\left\|f^{(k)}-f^{(k-1)}\right\|\le\varepsilon$, ε to a 0.001, then the reliability optimization process stops. Instead, calculate $\boldsymbol{S}_i$ based on the current MPP, adjust the position of the design variable $\boldsymbol{X}^M$ to ensure that the constraint boundaries are within the feasible domain, and go to step (3).

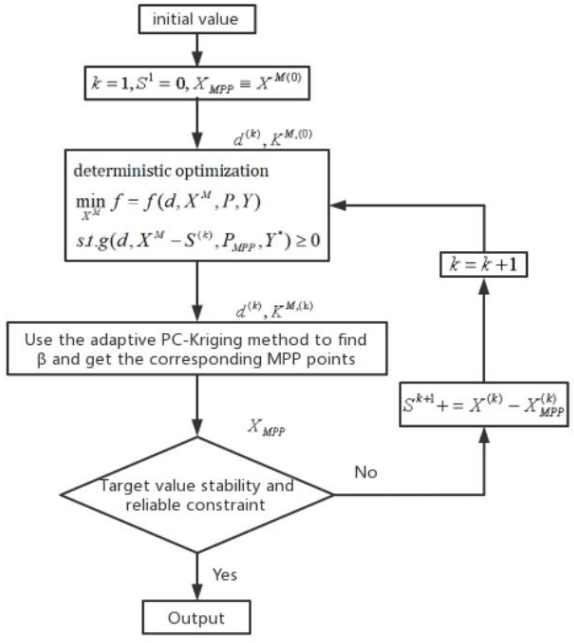


Figure 3 The flow diagram of APCK-SORA

4 Multi-objective reliability optimisation of high-speed heavy-duty gears design

4.1 Objective function

(1) Minimum sum of masses of gear pairs

Considering the involute straight cylindrical gear as a cylinder, the diameter as its index circle diameter and the height as the tooth width, the total mass of the drive system is:

$$M = M_1 + M_2 = \frac{\pi}{4}(d_1^2 b_2 + d_2^2 b_2)\rho \tag{9}$$

(2) Maximum overlap

In order to ensure the continuity of the gearing, the degree of overlap ε must be greater than or at least equal to 1. The greater the value of ε, the better the continuity of the gearing and the smoother the transmission.

$$\varepsilon = \frac{1}{2\pi}\left[z_1(\tan\alpha_1 - \tan\alpha) + z_2(\tan\alpha_2 - \tan\alpha)\right] \tag{10}$$

Where $\alpha_1 = \arccos\frac{d_{b1}}{d_{a1}}$, $\alpha_2 = \arccos\frac{d_{b2}}{d_{a2}}$, d_{a1}, d_{a2} are the top circle diameters and d_{b1}, d_{b2} are the base circle diameters.

(3) Maximum resistance to gluing

The instantaneous contact criterion (Blok flash temperature method) considers that the hot glue damage is caused by the high temperature generated by friction at the contact point, which ruptures the lubricant film and causes a sharp increase in the coefficient of friction, resulting in a higher temperature and the formation of a sticky weld between the metals, which tears open the weld joint due to the relative movement and thus forms the glue damage. In engineering, the Blok flash temperature method of hot bonding strength conditions are:

$$t_{C\max} = t_M + T_{t\max} \le t_s \tag{11}$$

The factor of safety for gear gluing strength is given by Eq:

$$S_B = \frac{t_s - t_{oil}}{t_{C\max} - t_{oil}} \tag{12}$$

Where: $t_{C\max}$ is the maximum contact temperature on the engagement surface, ℃; t_M is the body temperature, ℃; $T_{t\max}$ is the maximum flash temperature on the contact surface, ℃; t_s is the critical gelling temperature, ℃; t_{oil} is the lubricant temperature at thermal steady state, ℃. Since the limiting temperature of the lubricant is 220 ℃, the critical bonding temperature is set to 220 ℃ in this paper. Where $t_{C\max}$ and t_{oil} are obtained through finite element simulation, the greater the safety factor of gear gluing strength S_B , the greater the resistance to gluing, which means that the maximum contact temperature of the meshing surface $t_{C\max}$ should be lower.

4.2 Design variables

The number of teeth, modulus and tooth width of the active and driven gears are selected as the optimisation design variables, while the other basic design parameters of the gearing system remain unchanged:

$$\begin{aligned} x &= [x_1, x_2, x_3, x_4, x_5]^{\mathrm{T}} \\ &= [m, z_1, b_1, z_2, b_2]^{\mathrm{T}} \end{aligned} \tag{13}$$

4.3 Constraints

(1) Heat deflection constraint

In order to avoid the thermal deformation of high speed heavy duty gears leading to "jamming" of the gearing, it is necessary to ensure that the thermal deformation of the gear is less than the minimum side clearance of the gear (the amount by which the width of the tooth groove on the pitch line is greater than the tooth thickness).

$$j_{\min} = \frac{2}{3}(0.06 + 0.0005a + 0.03m) \tag{14}$$

$$G_1(X) = j_{\min} - \Delta\delta_{\max} \ge 0 \tag{15}$$

where $J_{\min}$ is the minimum side clearance of the gear, a is the centre distance and $\Delta\delta_{\max}$ is the maximum tooth deflection from the ANSYS simulation.

(2) Thermal resonance frequency constraint

According to the resonance principle, the gear rotor will resonate when the excitation frequency is close to or equal to the intrinsic frequency. According to reliability disturbance theory, the state function for random structural failure analysis is:

$$G_{2,i}(X) = |p - \omega_i|, (i = 1, 2...n) \tag{16}$$

Geared rotor systems constructed to avoid resonance

$$G_{2,i}(X) = |p - \omega_i| - \gamma > 0 \tag{17}$$

where p is the excitation frequency; ω_i is the ith intrinsic frequency, γ is 10% of the intrinsic frequency of each order of the gear rotor.

(3) Gear ratio constraint

The transmission ratio is the ratio of the angular velocities of the two involute gears. The optimised design (lightweighting) will result in a change in the number of teeth of the gear pair, which needs to be constrained in order not to affect the transmission ratio of the gear pair.

$$i_{12} = \frac{z_2}{z_1} \tag{18}$$

$$G_3(X) = 0.05 - |i_{12} - 1.7| / 1.7 > 0 \tag{19}$$

Where, i_{12} is the ratio between the active and driven gears, z_1 and z_2 are the number of teeth of the active and driven gears respectively.

(4) Centre distance constraint

The property that the centre distance between the gear pairs changes while the two ratios remain unchanged is known as the divisibility of involute gears. However, considering the overall dimensions of the gearbox, the centre distance of the gear pairs is therefore constrained.

$$Va = |a - a_0| \tag{20}$$

$$G_4(X) = Va = 0 \tag{21}$$

Where Va is the change in the initial and optimised centre distance between the gear pairs, a_0 is the initial centre distance between the gear pairs and a is the optimised centre distance between the gear pairs.

(5) Variation factor constraint

In the optimised design of gears the number of teeth and modules may be improved, resulting in the theoretical centre distance being unequal to the actual centre distance, when it is often necessary to adjust the coefficient of variation to meet the centre distance constraint. Where the total coefficient of variation is:

$$x_{\Sigma}=\frac{(\text{inv}\alpha'-\text{inv}\alpha)\cdot(z_1+z_2)}{2tg\alpha} \tag{22}$$

$$\alpha'=\arccos\left(\frac{a}{a'}\cos\alpha\right) \tag{23}$$

Where, α is tooth angle, α' is engagement angle, a isideal centre distance, a' is actual centre distance.

For the assignment of the coefficient of variation, due to the high speed of high-speed heavy-duty gears, the coefficient of variation of involute straight cylindrical gears is assigned using the method of gluing failure[18] Constraint on it

$$G_5(X)=x_1,\ \text{where}\ \begin{cases}x_1=\dfrac{x_{\Sigma}}{i+1}\text{g}\dfrac{i-1}{i+1+0.4z_2}\\ x_2=x_{\Sigma}-x_1\end{cases} \tag{24}$$

Where x_{Σ} is the total displacement factor, x_1 , x_2 are the displacement factors of the master and driven gears respectively, and i is the transmission ratio.

(6) Number of teeth constraint

Considering the minimum number of teeth required for gearing and the distribution of ratios, the number of teeth of the master and driven gears is taken as an integer, and the minimum number of teeth for a standard straight cylindrical gear without heel tangency is 17, so the range of values is as follows: $\begin{cases}17\le z_1\le 24\\ 30\le z_2\le 38\end{cases}$, z_1, z_2 are integers

$G_6(X)=z_1-17>0$.

where $G_5(X)$ is the number of teeth of the main and driven gears respectively.

(7) Modulus constraint

Modulus m is a basic parameter of gears, the larger the modulus, the larger the tooth pitch of the gear, the modulus of standard straight cylindrical gears has been standardized, the modulus selection range is as follows:

m=(1.0,1.25,1.5,1.75,2.0,2.25,2.5,2.75,3)

Referring to the mechanical design manual take the modulus greater than or equal to 2.5, i.e: $G_7(X)=m-2.5>0$.

(8) Tooth width constraint

Calculated from the tooth width of the large gear (driven wheel) $b_2=\Phi_d * d_2$, which should be rounded off and used as the tooth width of the large gear. Required tooth width of the pinion (active wheel): $b_1 = b_2$ + (0.5 to 1.0) mm.

$$\Phi_d<1.2 \tag{25}$$

$$G_8(X)=1.2-\Phi_d\ge 0 \tag{26}$$

Where Φ_d is the tooth width factor and d_1 is the pinion indexing circle diameter.

(9) Strength constraints for gears

Strength constraints for gears include gear contact strength constraints and tooth root bending strength constraints. The tooth contact strength is related to the tooth contact stress and the permissible contact stress.

The constraint function for the contact strength of the tooth surface σ_H is: $G_9(X)=\sigma_H-[\sigma_H]<0$.

The root bending strength is related to the permissible bending stress of the gear material. The constraint functions for the root bending strengths σ_{F1} and σ_{F2} of the pinion and large gears respectively are:

$$\begin{aligned}G_{10}(X)&=\sigma_{F1}-[\sigma_F]_1<0\\ G_{11}(X)&=\sigma_{F2}-[\sigma_F]_2<0\end{aligned} \tag{27}$$

4.4 Reliability optimised design of gears

The mathematical model for reliability optimisation of thermally-structurally coupled gears is:

$$\begin{aligned}&\text{Design variables}\quad \boldsymbol{X}^M=(x_1^M,x_2^M,x_3^M,x_4^M,x_5^M)\\ &\min\quad f(\boldsymbol{X}^M)=\{f_1(\boldsymbol{X}^M),f_2(\boldsymbol{X}^M),f_2(\boldsymbol{X}^M)\}\\ &st.\quad G_i(\boldsymbol{X}^M\text{-}s)\ge 0\\ &\quad f_1(\boldsymbol{X}^M)\ge 0\\ &\quad f_2(\boldsymbol{X}^M)\ge 1.4\\ &\quad f_3(\boldsymbol{X}^M)\le S_B\\ &\quad 2.5\le x_1^M<4\\ &\quad 17\le x_2^M\le 24\\ &\quad 30\le x_3^M\le 38\\ &\quad 13\le x_4^M<16\\ &\quad 13\le x_5^M<16\end{aligned} \tag{28}$$

where $\boldsymbol{s}_i=\boldsymbol{X}^{M(k-1)}-\boldsymbol{X}_{\text{MPP}}^{(k-1)}$, $\boldsymbol{X}^{M(k-1)}$ and $\boldsymbol{X}_{\text{MPP}i}^{(k-1)}$ are the deterministic optimisation solution and the most probable point (MPP) obtained from the previous loop, respectively. A reliability analysis is performed at the most probable point to calculate the MPP point in the probability constraint $(\boldsymbol{X}_{\text{MPP}}^k,\boldsymbol{Y}_{\text{MPP}}^k)$ and the shift vector for the next loop constraint $\boldsymbol{S}^{k+1}$, which in turn constructs the deterministic optimisation model for the next iteration.

The specific steps for the optimal reliability design of thermally-structurally coupled gears are as follows:

(1) The objective function is first defined.

$$f_1(\boldsymbol{X})=M(\boldsymbol{X})=M_1+M_2=\frac{\pi}{4}(d_1^2b_2+d_2^2b_2)\rho \tag{29}$$

$$f_2(\boldsymbol{X})=\varepsilon=\frac{1}{2\pi}\left[z_1(\tan\alpha_1-\tan\alpha)+z_2(\tan\alpha_2-\tan\alpha)\right] \tag{30}$$

$$f_3(\boldsymbol{X}) = S_B = \frac{t_s - t_{oil}}{t_{C\max} - t_{oil}} \tag{31}$$

The objective is to minimise the mass of $f_1(\boldsymbol{X})$, maximise the degree of overlap $f_2(\boldsymbol{X})$ and maximise the coefficient of safety against gluing $f_3(\boldsymbol{X})$ according to the design requirements and to satisfy the constraints.

Considering the 3 design requirements simultaneously, uses a linear weighted combination method to convert the 3 sub-objective functions into a single objective optimization function $F(\boldsymbol{X})$, such that

$$F(\boldsymbol{X}) = \lambda_1 f_1(\boldsymbol{X}) + \lambda_2 f_2(\boldsymbol{X}) + \lambda_3 f_3(\boldsymbol{X}) \tag{32}$$

where λ_i (i=1,2,3) denotes the weighting factors , λ_1=0.3, $\lambda_2 = -0.3$, $\lambda_3 = -0.4$ and therefore $\min F(\boldsymbol{X})$ as the final optimisation target.

(2) The initial sampling in the design space is carried out using the Latin square design of experiments method and the corresponding response values are calculated at $F(\boldsymbol{X})$. For $f_1(\boldsymbol{X})$ and $f_2(\boldsymbol{X})$, the equations are calculated and for $f_3(\boldsymbol{X})$ $t_{C\max}$ is used to obtain the maximum contact temperature of the meshing surfaces by a MATLAB call to ANSYS software for coupled thermal-structural analysis of the gear.

(3) Construct PC-Kriging approximation models for the objective and constraint functions. The deterministic optimal design is then based on the current approximate model to obtain the current optimal solution. At the initial first iteration step, set $\boldsymbol{u}_{MPTP} = \boldsymbol{0}$.

(4) Given μ_X, σ_X, a reliability analysis is performed by the adaptive PC-Kriging reliability method to obtain the $\boldsymbol{u}_{MPP}$ point. The optimised solution is transformed from the design space into the coordinate space where the random variable X is located to obtain its response value i.e. the current minimum value G_{min} If more than one constraint exists, each constraint function will obtain a minimum value $G_{i,min}$ (i=1,...,N).

(5) The best sample points are selected using the adaptive PC-Kriging method and added to the initial sample to reconstruct the PC-Kriging model.

(6) Update all sample points to construct the PC-Kriging approximation model and find all constrained MPP points. If at this point all $\hat{G}_{i,MPP} \geq 0, i = 1,..., N$, and the objective function value does not change much, then stop the iterative process that is the final result; otherwise if there is any $\hat{G}_{i,MPP} \geq 0$, otherwise, $k = k + 1$, then based on the current result return to step (4) to re-optimize the solution.

The final optimisation results obtained through two iterations of the optimisation design are shown in Table 2.

As can be seen from Table 2, the adaptive agent model's optimised design approach (APCK-SORA) is used to optimise the reliability design of high-speed heavy-duty gears. Firstly, a feasible domain of gear modulus, number of teeth and tooth width satisfying the conditions is selected based on the constraints of gear modulus, number of teeth, transmission ratio and strength. On the basis of this, optimisation is carried out for the objective functions of minimising the sum of the masses of the gear pairs, maximising the degree of overlap and maximising the resistance to galling. The optimisation results are shown in Table 2. It can be seen that the number of teeth of the master and driven wheels has been increased, but the modulus has been reduced from 3 mm to 2.75 mm, and the width of the master and driven wheels has been reduced. In order to meet the centre distance constraint, the optimised master and driven gears were machined using a positive displacement method. The result of the optimisation is a reduction in mass of 0.13kg, an increase in the degree of overlap of 0.016 and an increase in the coefficient of safety against galling of 0.19. This achieves the optimised design objectives of light weight, smoothness of transmission and maximum reliability against galling.

Table 2 Comparison of results before and after optimization

Reference items	Initial value	Results of one iteration	Optimal results
m(mm)	3	2.75	2.75
z_1	20	20	22
z_2	34	34	36
b_1 (mm)	15	14.7	14.0
b_2 (mm)	14.5	13.9	13.5
Mass (kg)	1.26	1.15	1.13
Overlap	1.618	1.628	1.634
Anti-glueing safety factor	1.24	1.38	1.43

Table 3 Comparison of different optimization results

Reference items	SAP with PMA	SORA	SLA	The proposed method
Number of iterations	308+5	296+2	265+4	53+2
m(mm)	2.75	2.75	2.75	2.75
z1	20	20	20	20
z2	34	34	34	34
b1 (mm)	14.8	14.7	14.8	14.7
b2 (mm)	14.0	13.9	13.8	13.9
Mass (kg)	1.16	1.15	1.15	1.15
Overlap	1.628	1.628	1.628	1.628
Anti-glueing safety factor	1.39	1.38	1.37	1.38

The first row of data in Table 3 is 308/5, which means that the number of optimisation iterations is 5 and the number of iterations to build the maximum contact temperature model for the meshing surface is 308. It can be seen from the comparison that the optimisation results obtained using the four algorithms are basically the same, but the proposed method only requires 53 iterations to complete the construction of the maximum contact temperature model for the meshing surface, and the parameters are optimised by two optimisation iterations, which has a higher computational efficiency.

5 Conclusion

An adaptive PC-Kriging model is proposed to improve the reliability part of the SORA optimization algorithm for multi-objective reliability optimization design of high-speed heavy-duty gears with an adaptive agent model. The objective functions of minimizing the sum of gear pair masses (lightweighting), maximizing degree of overlap (ensuring smooth transmission), and maximizing strength against galling are utilized to achieve multi-objective reliability optimization design for high-speed heavy-duty gears.

Comparing the proposed method with other optimization algorithms, it can be seen that while the final optimization results are essentially the same, the overall efficiency of modeling and optimization has been greatly improved. The proposed method has significant engineering value and is particularly well-suited for addressing reliability optimization problems in practical engineering applications, which typically require substantial investments of time and resources to construct an optimal model.

High-speed heavy-duty gears are inevitably affected by a variety of random factors during operation, such as fluctuations in external loads, changes in the environment and changes in the thermophysical properties of gear materials. Future research should consider the multi-objective optimization of gear design for dynamic reliability of relevant parameters over time.

Author Contributions: For Conceptualization, methodology, analysis, and writing original draft preparation, Yu Zhenliang; writing review and full-text editing, Wang Shuo; writing—original draft preparation, Zhao Fengqin; writing—original draft preparation, Li Chenyuan.

Conflicts of interest: The authors declare no conflict of interest.All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The research work financed with the means of Yingkou Institute of Technology Introduction of doctors to start the fund project(YJRC202109).

References

[1] Omar D. M, Akshay D.S. Bhat, Peter Falk. Robust multi-objective optimization of gear microgeometry design [J]. Simulation Modelling Practice and Theory, 2022(119):102593.

[2] Li Z , Wang S , Li F, et al. Research on Multiobjective Optimization Design of Meshing Performance and Dynamic Characteristics of Herringbone Gear Pair Under 3D Modification [J].Journal of Mechanical Design, 2022(144):10340.

[3] Zhao J, Hou L, Li Z, et al. Prediction of tribological and dynamical behaviors of spur gear pair considering tooth root crack [J]. Engineering Failure Analysis, 2022(135):106145.

[4] Daniel M, Zezelj, et al. Multi-objective spur gear pair optimization focused on volume and efficiency [J]. Mechanism and Machine Theory: Dynamics of Machine Systems Gears and Power Trandmissions Robots and Manipulator Systems Computer-Aided Design Methods, 2018(125):185-195.

[5] Dixit Y, Makarand S, Kulkarni. Multi-objective optimization with solution ranking for design of spur gear pair considering multiple failure modes [J]. Tribology International, 2023(180):108284.

[6] Maruti P, Ramkumar P, Shankar K. Multi-objective optimization of the two-stage helical gearbox with tribological constraints [J].. Mechanism and Machine Theory, 2019(138)38-57.

[7] Edmund S.M. Rajesh A. Multi-objective optimization of a 2-stage spur gearbox using NSGA-II and decision -making methods [J].. Journal of the Brazilian Society of Mechanical Sciences and Engineering,2020(42):477.

[8] Emna C, Cc A, Jb B, et al. Multi-objective optimization of gear unit design to improve efficiency and transmission error [J].. Mechanism and Machine Theory, 2021,167:03247458

[9] Ekansh C, Pinar A, Corina S. Multi-objective macrogeometry optimization of gears: Comparison between sequential quadratic programing and genetic algorithm [J]. Mechanics based design of structures and machines, 2023.51(1):97-112.

[10] Zhang J, Zhang B. A Collaborative Approach for Multidisciplinary Systems Reliability Design and Optimization [J]. Advanced Materials Research, 2013(4_:911-914.

[11] Cho K K, Lee I, Zhao L. Sampling-based RBDO using the stochastic sensitivity analysis and Dynamic Kriging method [J]. Journal of Mechanical Design, 2011(2): 71-80..

[12] Du X, Chen W. Sequential Optimization and Reliability Assessment Method for Efficient Probabilistic Design [J]. Journal of Mechanical Design, 2003, 126(2): 871-880.

[13] Cheng J, LI, Q.S. Reliability analysis of structures using artificial neural network based genetic algorithms [J]. Computer Methods in Applied Mechanics & Engineering, 2008, 197(45): 3742-3750.

[14] Ilchi Ghazaan M, Saadatmand F. Decoupled reliability-based design optimization with a double-step modified adaptive chaos control approach [J]. Structural and Multidisciplinary Optimization, 2022, 65(10):1-20.

[15] Ilchi Ghazaan M , Saadatmand F . A new performance measure approach with an adaptive step length selection method hybridized with decoupled reliability-based design optimization [J].Structures 2022(44):977-987.

[16] Kaveh A, Zaerreza A. A new framework for reliability-based design optimization using metaheuristic algorithms [J].Structures 2022(38):1210 -1225.

[17] Kaveh A, Zaerreza A. Reliability-based design optimization of the frame structures using the force method and SORA-DM framework [J].Structures 2022(45):977 -987.

[18] Yu Z L , et al. A new Kriging-based DoE strategy and its application to structural reliability analysis [J].Advances in Mechanical Engineering, 2018, 10(3):1-13.

Publisher: Viser Technology Pte. Ltd.

URL: www.viserdata.com

Add.:21 Woodlands Close, #08-18,

Primz Bizhub SINGAPORE (737854)