Article citation info:
LI Y, WANG K. Modified convolutional neural network with global average pooling for intelligent fault diagnosis of industrial gearbox. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2020; 22 (1): 63–72, http://dx.doi.org/10.17531/ein.2020.1.8.

Yaxin LI
Kesheng WANG

MODIFIED CONVOLUTIONAL NEURAL NETWORK WITH GLOBAL AVERAGE POOLING FOR INTELLIGENT FAULT DIAGNOSIS OF INDUSTRIAL GEARBOX

DIAGNOSTYKA BŁĘDÓW PRZEKŁADNI PRZEMYSŁOWYCH Z WYKORZYSTANIEM ZMODYFIKOWANEJ SPLOTOWEJ SIECI NEURONOWEJ Z GLOBALNYM UŚREDNIENIEM WARTOŚCI DLA POSZCZEGÓLNYCH KANAŁÓW

Gearboxes are key transmission components and widely used in various industrial applications. Due to the possible operational conditions, such as varying rotational speeds, long period of heavy loads, etc., gearboxes may easily be prone to failure. Condition Monitoring (CM) has been proved to be an effective methodology to improve the safety and reliability of gearboxes. Deep learning approaches, nowadays, further enable the CM with more powerful capability to exploit faulty information from massive data and make intelligently diagnostic decisions. However, for most of conventional deep learning models, such as Convolutional Neural Network (CNN), a large amount of labelled training data is a prerequisite, while to obtain the labelled data is usually a laborious and time-consuming job and sometimes even unattainable. In this paper, to handle the case of only a limited labelled data is available, a modified convolutional neural network (MCNN) is proposed by integrating global average pooling (GAP) to reduce the number of trainable parameters and simplify the architecture of deep learning model. The proposed MCNN improves the traditional CNN's ability in fault diagnosis with limited labelled data. Two experimental gearbox datasets are utilized to demonstrate the effectiveness of the proposed MCNN method. Compared with traditional deep learning approaches, namely LSTM, CNN and its variant methods, the experimental results show that the proposed MCNN with higher discrimination and generalization ability in fault classification and diagnostics under the scenario of limited labelled training samples.

***Keywords**: modified convolutional neural network, global average pooling, intelligent fault diagnosis, industrial Gearbox.*

Przekładnie stanowią kluczowe elementy układów napędowych i jako takie znajdują szerokie zastosowane w przemyśle. Ze względu na warunki eksploatacji, takie jak różne prędkości obrotowe czy długie okresy pracy pod dużym obciążeniem itp., przekładnie mogą łatwo ulegać uszkodzeniom. Udowodniono, że monitorowanie stanu skutecznie poprawia bezpieczeństwo i niezawodność przekładni. Podejścia oparte na uczeniu głębokim umożliwiają ponadto monitorowanie stanu z większym wykorzystaniem informacji o błędach pochodzących z dużych zbiorów danych i podejmowanie inteligentnych decyzji diagnostycznych. Jednak w przypadku większości konwencjonalnych modeli uczenia głębokiego, takich jak splotowe sieci neuronowe (convolutional neural networks, CNN), wymagana jest duża ilość etykietowanych danych uczących, których pozyskanie jest zwykle zadaniem praco- i czasochłonnym, a czasem wręcz niemożliwym do wykonania. W niniejszej pracy, przedstawiono zmodyfikowaną splotową sieć neuronową (modified convolutional neural network, MCNN), która rozwiązuje problem dostępności danych etykietowanych poprzez zastosowanie globalnego uśrednienia względem kanałów (global average pooling), co pozwala na zmniejszenie liczby możliwych do wyuczenia parametrów i uproszczenie architektury modelu głębokiego uczenia. W porównaniu do tradycyjnych sieci CNN, proponowana sieć MCNN zwiększa możliwości diagnozowania błędów przy ograniczonych danych etykietowanych. Skuteczność proponowanej metody wykazano na przykładzie dwóch zbiorów danych doświadczalnych dotyczących błędów przekładni. Wyniki eksperymentalne pokazują, że, w porównaniu z tradycyjnymi metodami uczenia głębokiego, takimi jak LSTM, CNN oraz warianty tej ostatniej, proponowane podejście MCNN daje większe możliwości rozróżniania i uogólniania podczas klasyfikacji i diagnostyki błędów w przypadku ograniczonej dostępności etykietowanych danych uczących.

***Słowa kluczowe**: zmodyfikowana splotowa sieć neuronowa, globalne uśrednienie względem kanałów, inteligentna diagnostyka błędów, przekładnia przemysłowa.*

1. Introduction

High transmission ratio, strong load-bearing and high efficiency makes modern gearboxes are always considered to be critical components in various industrial applications, such as wind turbine generator system, helicopter main speed reducer, aerospace engineering and etc [27, 38]. In real practice, however, gearboxes will inevitably be subjected with dynamic heavy-duty loads under complex operating conditions, making the breakdown or even accidents of the engineering system [3, 4, 8, 32, 39]. Therefore, it is of great significance to develop the condition monitoring and fault diagnostic techniques for the gearboxes.

Most of the modern gearbox fault diagnostic methods utilize vibration analysis to extract the fault features, and then make decision according to sophisticated signal processing techniques or expert knowledge of diagnosticians [1, 2, 6, 9, 18, 33]. For instance, Feng et al. [10] successfully introduced the Vold-Kalman filter into time-frequency analysis to extract fault features of the planetary gearbox under unstable operation conditions. Tang et al. [28] firstly presented a novel fault detection method to identify the categories of gearbox failures on the strength of hierarchical instantaneous energy density dispersion entropy (HIEDDE) and dynamic time warping (DTW). However, for these approaches based on vibration analysis, large amounts of signal processing efforts and abundant expert diagnostic experience are generally required to extract and analyze fault characteristics from the measured vibrations.

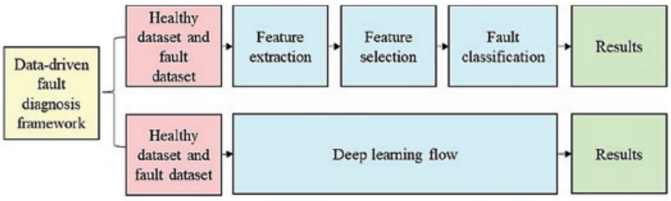

Data-driven fault diagnosis framework
Healthy dataset and fault dataset
Feature extraction
Feature selection
Fault classification
Results
Healthy dataset and fault dataset
Deep learning flow
Results

Fig. 1. Data-driven fault diagnosis framework.

Thanks to the recent advanced progress made by artificial intelligence and machine learning techniques, intelligent fault diagnosis receives an increasing research attention in the field of condition monitoring and fault diagnosis [12, 13, 25, 39, 43]. Methods, such as artificial neural network (ANN), back-propagation neural network (BPNN) and support vector machine (SVM) have become research focus. For instance, Tyagi et al. [29] successfully constructed a hybrid artificial neural network (ANN) classifier for gearbox diagnosis. The hybrid classifier consists of data preprocessing with discrete wavelet transform (DWT), genetic algorithm (GA) and back-propagation neural network (BPNN). Zhang et al. [42] developed a multivariable ensemble-based incremental support vector machine (MEISVM) and applied it into the compound failure detecting of roller bearings. Despite their successes, the outstanding performance of the intelligent diagnostic methods alike heavily count on the accuracy of the manually extracted and selected features. This typical route of intelligent method is shown in the upper half in Fig. 1. For this route, advanced signal processing techniques are usually required for data pre-processing. Moreover, for the shallow learning model (such as BPNN, SVM) employed for gearbox fault diagnosis, the diagnostic ability of the model is relied heavily on the quality of the extracted fault features. Unfortunately, for the cases of large amount of measured industrial data with unguaranteed data quality, its diagnostic capability will naturally exhibit insufficient with the increase of the data amount.

To tackle these issues, deep learning route shown in the lower half in Fig. 1 with the structure of deep learning model of multi-layer nonlinear modelling solutions, comes into the recent research focuses and provides a straightforward end-to-end learning process from the measured input signal to the output diagnostic results, which completely eliminate the challenges of manual feature extractions and selections [13, 44]. Deep learning methods utilize the deep architectures to constitute hierarchical feature representations to discover the distributed feature representations of data. Due to its powerful capability of perception, self-learning, modelling and characterization, recent years have witnessed the tremendous progress of deep learning techniques in various fields, mainly including image processing, speech recognition and fault diagnosis. For examples to fault diagnosis, Yu et al. [37] exploited stacked denoising auto-encoder (SDAE) and gated recurrent unit neural network (GRUNN) to enhance the capability of anti-noise and the adaptive ability to time-varying rotational speed during the fault diagnosis of the planetary gearbox. Liu et al. [22] automatically extracted features from vibration signal by combining batch normalization with deep belief network (DBN), which achieves more precise performance than DBN and other conventional approaches in wind turbine gearbox diagnosis. Furthermore, Other deep neural network such as deep residual network (DRN) [40], recurrent neural network (RNN) [21], generative adversarial networks (GAN) [26], long short-term memory networks (LSTM) [19, 36] and especially convolutional neural networks (CNN) are seriously and widely investigated in gearbox fault diagnosis. Specifically, to CNN, Chen et al. [7] applied CNN to adaptively learn fault features and classify fault patterns with extreme learning machine (ELM) for mechanical faults. Jiao et al. [17] developed a deep coupled dense convolutional network (CDCN) to diagnose the faults of planetary gearbox, which could relieve gradient vanishing in deep architecture and realize two-stage information fusion.

Even though various successful cases on the applications of CNN in fault diagnosis have been reported, the works generally employed massive labelled measured samples to train a deep network. Nevertheless, it is difficult to acquire an enough number of labelled samples in real industrial application, especially for certain faulty scenarios which seldomly occur. Recently, transfer learning may be promising in this problem, and some works about transfer learning-based fault diagnosis have been reported [15]. The unsupervised domain adaptation is the major branch of this framework [5, 14, 34]. By adapting the feature distribution between two domains, the diagnostic model can generalize well to the unseen conditions where no labelled data can be used for model training. Although these methods avoid to use labelled data in target domain, a large amount of unlabelled data (similar to the data in source domain) are generally necessary. In addition, due to the complicated and deep multi-layer structures, the parameter optimization of CNN model may lead to a huge computational burden. In this scenario, promoting the precision of diagnosis, accelerating the training speed of CNN as well as boosting the generalization ability and robustness under the small number of labelled samples become a critical issue to research. To this end, in this paper, a modified convolutional neural network (MCNN) for the fault diagnosis was proposed in which the global average pooling (GAP) is introduced into the internal structure of CNN model to replace the traditional fully connected layer where the majority of parameters for training are contained. By doing so, compared with traditional CNN, it enables the MCNN an improved capability to deal with fault classification problem with limited labelled samples. The MCNN is validated by two gearbox datasets from PHM 2009 conference data challenge and measured experimental data at University of Electronic Science and Technology of China. The experimental results demonstrate the advantages of the proposed MCNN method in accuracy of fault classification, time-saving of training model, more important, the method is more effective for fault diagnosis under condition of limited number of labelled samples.

The structure of the paper is arranged as follows. In section 2, the basic theory of CNN is briefly described. The method of MCNN is introduced in section 3. In section 4, the two experimental application of proposed method is analysed. Finally, the summary is concluded in section 5.

2. Methods

2.1. Traditional architecture of CNN

As one of the most representative deep learning algorithms, CNN is a combination of convolutional computation and deep structure, which is generally composed of three parts, i.e., input layer, the feature descriptor and the classifier. The feature descriptor consists of multiple convolution layers, activation layers, pooling layers. The input signal is mapped to the feature space of the CNN hidden layers to extract the features of the input data. The classifier is composed of one or several fully connected layers, namely a multi-layer perceptron classifier, for fusion and classification of the extracted features. The input layer of CNN is to pre-process multidimensional data, usually referring to one-dimensional data, two-dimensional data, or three-dimensional data. As the core of CNN, the convolution layer, containing multiple convolution kernels, is to perform feature learning and extraction from the input signal. Each convolution kernel corresponds to a weight matrix and a bias vector, similar to a neuron of a feed-forward neural network. The convolution kernels sweep through the input features with a pre-set stride, and obtain the activated feature maps in the receptive field.

Supposing that the input signal, and the filer $w\epsilon R^n$, the convolution process can be depicted as:

$$X_i^{(k)}=\sum_{c=1}^{C} W_i^{(c,k)} * X_{i-1}^{(c)}+B_i^{(k)} \tag{1}$$

where i represents the index of convolutional layer, k represents the index of feature map in i^{th} layer, c means the number of convolutional kernel in i^{th} layer, * means the convolution operation, $X_{i-1}^{(c)}$ is the input feature map of $(i-1)^{th}$ layer, $X_i^{(k)}$ is the output feature map, $W_i^{(c,k)}$ and $B_i^{(k)}$ donates the weights and bias of the convolution kernels respectively.

By introducing nonlinear activation function into the network model, the ability of feature representation will be further enhanced. The generally utilized activation functions include sigmoid, tanh, rectified linear units (ReLU), etc. These functions are listed in equation (2). Among them, ReLU is one of the most noteworthy functions with efficient gradient descent ability and avoiding gradient explosion and disappearance during the training process:

$$\begin{cases} sigmoid: f=\dfrac{1}{1+e^{-y}} \\ tanh: f=\dfrac{e^{y}-e^{-y}}{e^{y}+e^{-y}} \\ ReLu: f=max(0,y) \end{cases} \tag{2}$$

After the convolutional operations, the output feature maps are delivered to the pooling layer for down-sampling. The widely used max-pooling is to divide the feature maps into a series of blocks without overlapping and extract the maximum value in each block as the eigenvalue of the window while discarding other points. The max-pooling process can be defined as equation (3):

$$X_{i+1}^{(k)}=max \underset{(i-1)l+1\le t\le il}{\overset{max}{}} X_i^{(k)}(t) \tag{3}$$

where $X_i^{(k)}(t)$ represents the feature map after convolution operation of the kth neuron at the i^{th} layer, l denotes the width of a local area for max-pooling, $X_{i+1}^{(k)}(i)$ is the output feature map after max-pooling.

The fully connected layers are located at the last part of CNN with the purpose of nonlinearly combining the extracted features and mapping them into output labels. The softmax function is used at the final output layer to calculate the probability distribution for each label, whose mathematical expression can be described in equation (4):

$$\mathrm{H}_{,}\left(\mathrm{x}^{(\mathrm{i})}\right)=\begin{bmatrix} p\left(\mathrm{y}^{(\mathrm{i})}=1|\mathrm{x}^{(\mathrm{i})};,\right) \\ p\left(\mathrm{y}^{(\mathrm{i})}=2|\mathrm{x}^{(\mathrm{i})};,\right) \\ \cdots \\ p\left(\mathrm{y}^{(\mathrm{i})}=\mathrm{m}|\mathrm{x}^{(\mathrm{i})};,\right) \end{bmatrix}=\frac{1}{\sum_{\mathrm{j}=1}^{\mathrm{k}}\exp\left(,_{\mathrm{j}}^{\mathrm{T}}\mathrm{x}^{(\mathrm{i})}\right)}\begin{bmatrix} \exp(,_{1}^{\mathrm{T}}\mathrm{x}^{(\mathrm{i})}) \\ \exp(,_{2}^{\mathrm{T}}\mathrm{x}^{(\mathrm{i})}) \\ \cdots \\ \exp(,_{\mathrm{k}}^{\mathrm{T}}\mathrm{x}^{(\mathrm{i})}) \end{bmatrix} \tag{4}$$

where the interregional of $y^{(i)}$ is $\{1,2,\ldots m\}$, m is the number of classifications, and θ is the assemblage of the arguments of the model.

2.2. A discussion on the shortcoming of the CNN

LeNet, as the pioneering and widely used architecture of CNN, basically established by convolutional layers, pooling layers and fully-connected layers. Most of researches for fault diagnosis utilized the simplified and improved LeNet-5 (containing 5 convolutional layers). For CNN with LetNet architecture, though it has been widely acknowledged, a large amount of training data for a proper modelling is a prerequisite and this may be largely attributed to the architecture of the LeNet with numerous network parameters. Specifically, in this architecture, the last set of feature maps are flattened into one-dimensional feature vector, and each feature is connected to each neuro in the first fully-connected layer. In this manner, the extracted features will be mapped into label space. However, it should be noted that, even at the end of feature maps, considerable amount of network parameters still exists and need to be trained. As a result, it makes the proper training of the model with small of amount of data becomes a tough issue. To give an intuitive presentation and quantitative analysis, three famous CNN architectures in computer vision, i.e., LeNet-5, AlexNet and VGG-16, as examples, the distributions of parameters between front convolutional block and later fully-connected layers are shown in Table 1. It is clear that the vast majority of parameters are distributed in the fully-connected layers. Consequently, to remove or modify the complex part of the model by reducing the number of trainable parameters within the CNN structure, at the same time as large as possible to remain the feature extraction representation results, can be a promising solution to improve the capability of the model, especially enabling the model to deal with small amount of labelled data samples which will be beneficial to the real engineering practice.

Table 1. The distributions of parameters in CNN with three typical architectures

Architecture	**Parameters distribution (%)**	
	Convolutional block	Fully-connected layer
LeNet-5	2.8	97.2
AlexNet	3.8	96.2
VGG-16	10.6	89.4

2.3. Global average pooling (GAP)

The novel global average pooling (GAP) [20] is, therefore, introduced in this section with which fully connected layers of CNN is superseded. For each feature map at the end of pooling layers, we take the average value of each feature vector directly maps to a category label or an output node. This process was called global average pooling. The original fully-connected layers are replaced. By doing so, a tremendously reduction of the number of parameters needed to be trained is realized and the computational burden of training the model is decreased. More important, though it simplified the CNN model, it still completely remains the key convolutional layers and therefore the ability of feature representation still remained. Furthermore, it gives an extra capability to the model to deal with the training problems with small amount of data samples. The detailed illustrations of the traditional fully-connected layers and global average pooling layer are shown in Fig. 2.

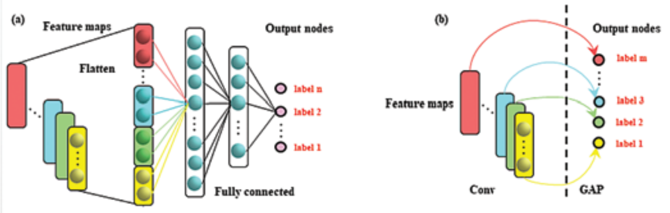


Fig. 2. An illustration of fully connected layer and global average pooling layer. (a) fully-connected layer; (b) global average pooling layer

2.4. Modified CNN

The structure of MCNN consists of input layers, convolutional layers, max-pooling layers, dropout layers, global average pooling layers. The deep network has a total of 19 layers. We utilize 5 convolutional layers to learn features from raw data (follow the architecture of LeNet-5). The ReLU function is employed as the activation function. A max-pooling layer is performed after per convolution operation. Since the number of last sets of feature maps is equal to the category labels or output nodes where the global average operation will be performed. Thus, an additional convolutional layer, called task specific layer, is added to revise the number of the out-feature maps. The number of convolutional kernels in the task specific layer equals to the category labels. The detailed structure and parameter setting of MCNN constructed in this paper are illustrated in Fig. 3. Referring to the literature [41], larger values, such as 128 and 64, are selected as the kernel sizes for the front two convolutional layers to capture essential features and reduce high frequency noise for 1D vibration signal. With the increase of network depth, the number of kernels also increase from 16 to 256, which helps to learn hierarchical feature representation. It worth noting that, as a regularization approach, dropout layers are integrated after each specific layer to prevent overfitting.

In the following comparative analysis, the same construction and parameters in the feature descriptor are chosen in the traditional CNN in order to conduct a fair comparison. The following flatten layer is employed for transforming the high dimensional feature maps to one-dimension feature vector. Afterwards, 3 fully connected layers are established to integrate local information with class discrimination in convolutional blocks and map the learned distributed feature representation to the sample label space. The detailed CNN and MCNN structures can be seen in Fig. 3.

2.5. The intelligent fault diagnosis framework with MCNN

In this paper, an intelligent fault diagnosis framework for gearbox with MCNN is proposed and listed in figure 4. There are 3 steps in total: (1) data processing, (2) train the MCNN model, and (3) fault diagnosis of gearbox.

(a) Data processing

Measured vibration signals are collected from accelerators and directly input into the fault diagnostic framework without any manual feature extraction and selection. In this way, an end-to-end fault diagnostic framework is realized and the loss of data information caused by human interventions or advanced signal processing techniques are minimized. According to the generally used sample size of deep learning researches [30, 41], the raw signal is partitioned into a series of fixed-length segments by shifting the window with a constant stride, and then the training and testing data sets are selected from the whole measured vibrations without repetition.

(b) MCNN Model learning

Model training consists of two stages: forward calculation and loss backward propagation. In the forward calculation stage, training samples are fed into the MCNN model and predicted outputs. Then, the loss between the predicted outputs and the real outputs are backward propagated to optimize the network parameters layer by layer. The method of the optimizer utilized is stochastic gradient descent (SGD) technique. The optimization process can be represented as:

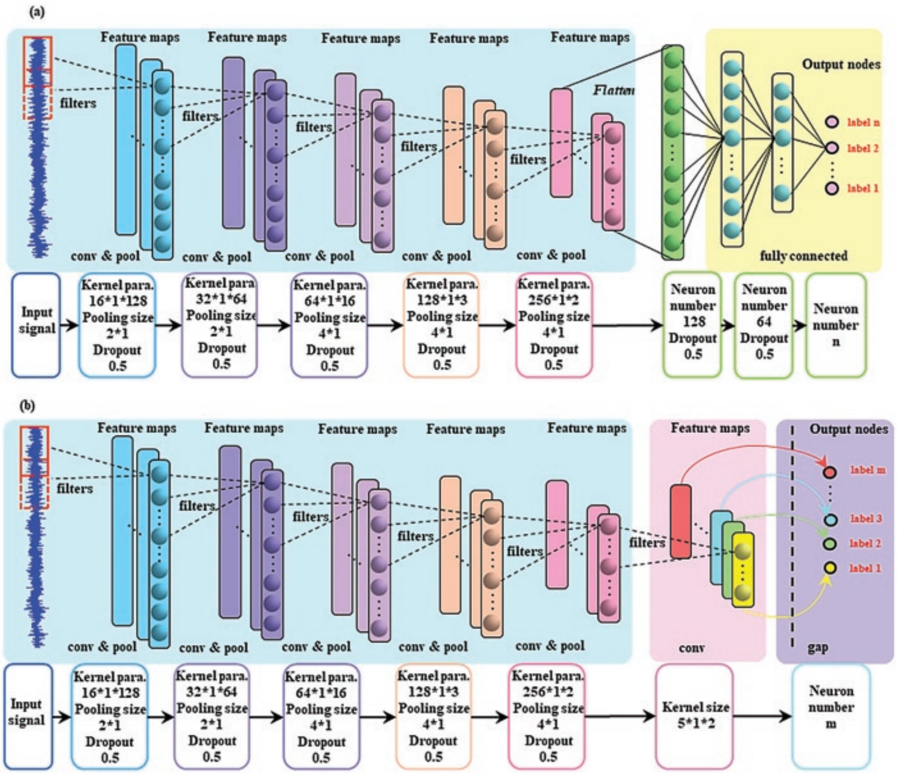

(a)
Feature maps
Feature maps
Feature maps
Feature maps
Feature maps
Flatten
Output nodes
filters
filters
filters
filters
filters
label n
label 2
label 1
conv & pool
conv & pool
conv & pool
conv & pool
conv & pool
fully connected
Input signal
Kernel para. 16*1*128 Pooling size 2*1 Dropout 0.5
Kernel para. 32*1*64 Pooling size 2*1 Dropout 0.5
Kernel para. 64*1*16 Pooling size 4*1 Dropout 0.5
Kernel para. 128*1*3 Pooling size 4*1 Dropout 0.5
Kernel para. 256*1*2 Pooling size 4*1 Dropout 0.5
Neuron number 128 Dropout 0.5
Neuron number 64 Dropout 0.5
Neuron number n
(b)
Feature maps
Feature maps
Feature maps
Feature maps
Feature maps
Feature maps
Output nodes
filters
filters
filters
filters
filters
filters
label m
label 3
label 2
label 1
conv & pool
conv & pool
conv & pool
conv & pool
conv & pool
conv
gap
Input signal
Kernel para. 16*1*128 Pooling size 2*1 Dropout 0.5
Kernel para. 32*1*64 Pooling size 2*1 Dropout 0.5
Kernel para. 64*1*16 Pooling size 4*1 Dropout 0.5
Kernel para. 128*1*3 Pooling size 4*1 Dropout 0.5
Kernel para. 256*1*2 Pooling size 4*1 Dropout 0.5
Kernel size 5*1*2
Neuron number m

Fig. 3. An illustration of CNN and MCNN: (a) CNN; (b) MCNN

$$\theta = \theta - \cdot *\nabla_{,} J\left(\theta; x^{(i)}; y^{(i)}\right) \tag{5}$$

where θ is the collection of network parameters, $x^{(i)}$ and $y^{(i)}$ represent the input sample and corresponding label, $J(\cdot)$ is the loss function, η is the learning rate and ∇_{θ} denotes the gradients.

(c) Fault diagnosis of gearbox

After completing the training of the MCNN model, the diagnostic model is deployed for fault classification. And the testing samples are fed into the model for validation. The diagnostic accuracy of model is defined to evaluate the performance of the network.

$$Accuracy = \frac{\left|x : x \in D \wedge \hat{y}(x) = y(x)\right|}{\left|x : x \in D\right|} \tag{6}$$

where D is the set of test data, x is the input sample, $y(x)$ is the truth label of x, $\hat{y}(x)$ is the label predicted by the diagnosis model.

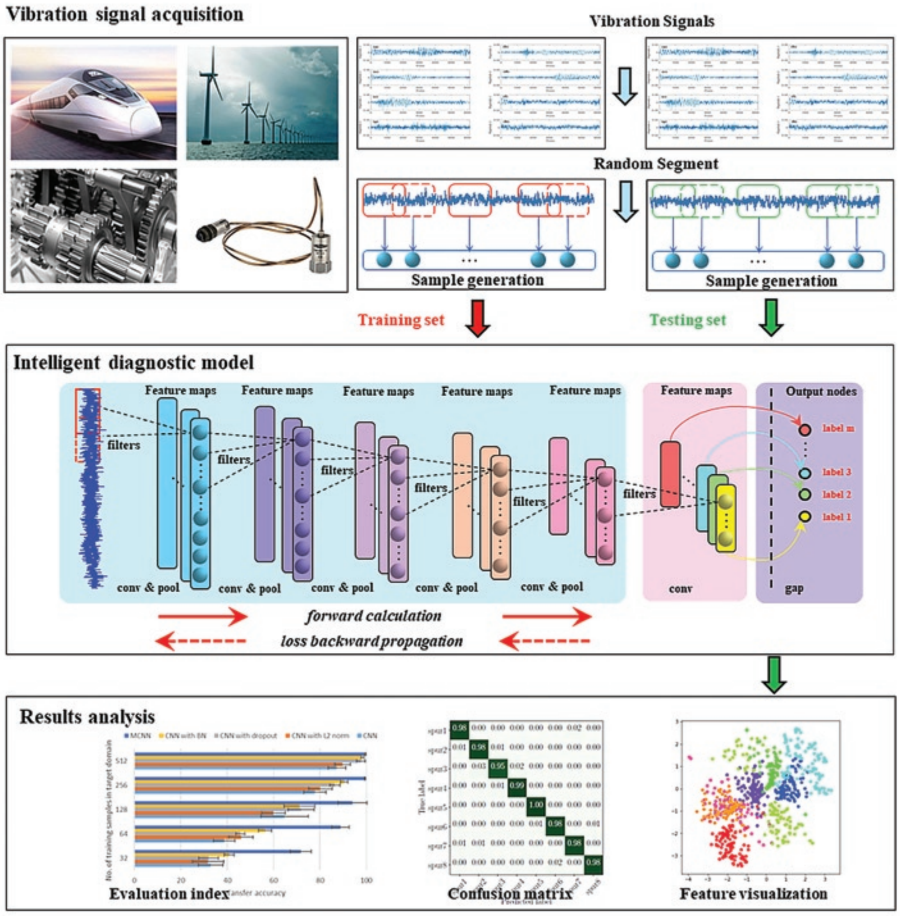


Fig. 4. The intelligent fault diagnosis framework with MCNN

3. Experimental studies

3.1. Description of dataset from 2009 challenge dataset of the Prognostics and Health Management (PHM) society

The 2009 challenge dataset from Prognostics and Health Management (PHM) society is first used to validate the proposed MCNN [23]. The datasets are measured from the gearbox shown in Fig.5. Fig. 5 (a) illustrates the constitution of the fixed shaft gearbox and the position of accelerometers and tachometer. The gearbox consists of 3 shafts, 4 spur gears and 6 bearings, as is shown in Fig. 5 (b). Vibration signals of the spur gear with 8 health conditions, 5 shaft speed conditions including 30, 35, 40, 45 and 50 Hz, under the same amount of high load are collected. The Table 2 lists the specific 8 failure modes of the gearbox. These experiments can fundamentally cover the frequently occurred faults in gearbox. These faults are artificially introduced to machines so as to simulate diverse health conditions. The sampling frequency is 66.67 kHz. There are 533312 points under each fault mode and operation condition, hence 6144 data points with a 4096 stride are collected for one sample to guarantee that there is abundant fault information for each sample. There are 5208 samples in total. The time waveforms for each health condition are shown in

Table 2. Descriptions of detailed fault patterns

Label	Gear				Bearing						Shaft	
	32T	48T	80T	96T	IS:IS	ID:IS	OS:IS	IS:OS	ID:OS	OS:OS	Input	Output
1	G	G	G	G	G	G	G	G	G	G	G	G
2	C	G	E	G	G	G	G	G	G	G	G	G
3	G	G	E	G	G	G	G	G	G	G	G	G
4	G	G	E	Br	B	G	G	G	G	G	G	G
5	C	G	E	Br	In	B	O	G	G	G	G	G
6	G	G	G	Br	In	B	O	G	G	G	Im	G
7	G	G	G	G	In	G	G	G	G	G	G	Ks
8	G	G	G	G	G	B	O	G	G	G	Im	G

IS = input shaft; :IS = input side; ID = idler shaft; OS = output shaft; :OS= output side. G: good; C: chipped; E: eccentric; Br: broken; B: ball; In: innerrace; O: outer race; Im: imbalance; Ks: keyway sheared.

Fig. 6. The training and testing data set are randomly selected from the whole dataset without repetition. The number of samples in the training is selected as 128, 256, 512, 1024, and 4096 successively, while the number of testing sets is 1000.

3.2. Description of data from the Drivetrain Diagnostics Simulator (DDS) test rig at UESTC

The second dataset is from the Drivetrain Diagnostics Simulator (DDS) test rig at University of Electronic Science and Technology of China (UESTC). The layout of the test rig is shown in Fig. 7. The accelerometer is mounted on the one-stage planetary gearbox for the collection of vibration signals. The structure of the one-stage planetary gearbox is shown in Fig. 8 (a), which is constituted by a sun gear, 4 planet gear, planet carrier and ring gear. Four different kinds of faults in the sun gear of the one-stage of the planetary gearbox is shown in Fig. 8, including tooth wear, tooth broken, tooth missing and root crack. For each sun gear health condition, 6.39 seconds of data is collected under 2 different loads (0A, 1.3A) and 3 different rotational speeds (30Hz, 40Hz and 50Hz), with the sampling frequency of 30.72 kHz. 196,608 points are collected for each health condition under each operation condition, and 2048 data points with a 1000 stride are cut for each data sample. Fig. 9 shows the typical original vibration waveforms for each health condition. Therefore, 49 samples are generated, and two sets of samples with different sizes are randomly selected, one of which is set as the training set while another is set as the testing set. We set the number of samples in the training set to 128, 256, 512, 1024, and 4096 in turn, and the sample size in the testing set is 800.

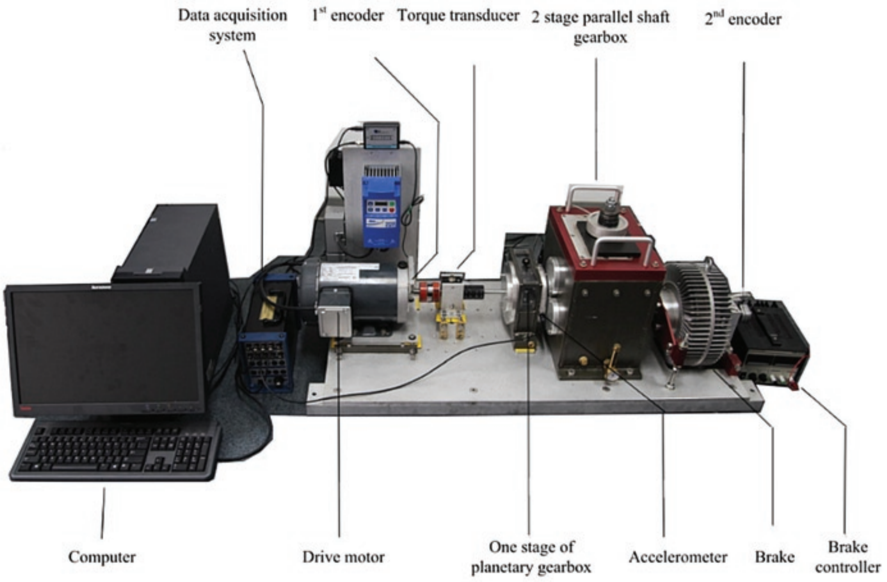


Fig. 7. DDS test rig

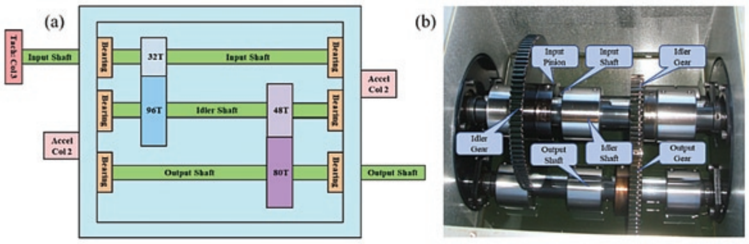


Fig. 5. The gearbox in the 2009 Challenge Data of PHM society: (a) Schematic diagram; (b) Overview of the gearbox

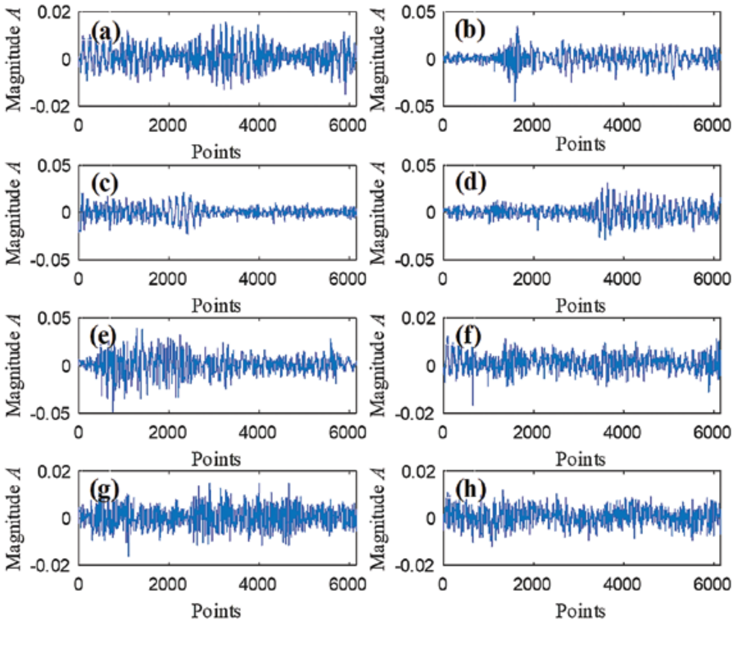


Fig. 6. Collected vibration signals of 8 machine conditions: (a) spur1 (b) spur2 (c) spur3 (d) spur3 (e) spur5 (f) spur6 (g) spur7 (h) spur8.

3.3. Comparative methods

The proposed MCNN will be compared with other intelligent fault diagnostic methods, including (1) support vector machine (SVM) [35], (2) random forest (RF) [11], (3) long short-term memory (LSTM) [19], (4) CNN [41], (5) CNN with l2-norm [24], (6) CNN with batch normalization (BN) [31]. among them, RF and SVM are two of the most commonly used models in machine learning. RF, containing multiple decision trees, is a classifier that uses multiple trees to train and predict samples. SVM is a nonlinear kernel classifier that categorizes the data based on supervised learning. LSTM is a time-cycle neural network. L2-norm and BN are two different regularization algorithms, which can effectively improve the performance of CNN and prevent overfitting.

The related architecture parameter settings of the other methods are listed as follows. (1) RF: the number of trees and random feature subset are separately set as 500 and $\sqrt{m}$. (2) SVM: radial basis function (RBF) is introduced as the kernel function of SVM, besides, the arguments of RBF and the penalty factor are intelligently optimized by genetic algorithm (GA). The maximum generation and the number of populations in GA is set to 50 and 20 respectively. And the searching range of parameters in SVM is set to [0, 100]. (3) LSTM: By referring to [19], the dimension of hidden layer is 128 and two RNN layers are stacked. (5) CNN: the architecture and parameter settings have been listed in 2.3. (6) CNN with l2-norm: l2-norm regularization is introduced to the parameters of CNN with a weight of 1e-2. It should be noted that the popular statistical features in time domain and frequency domain [11], such as

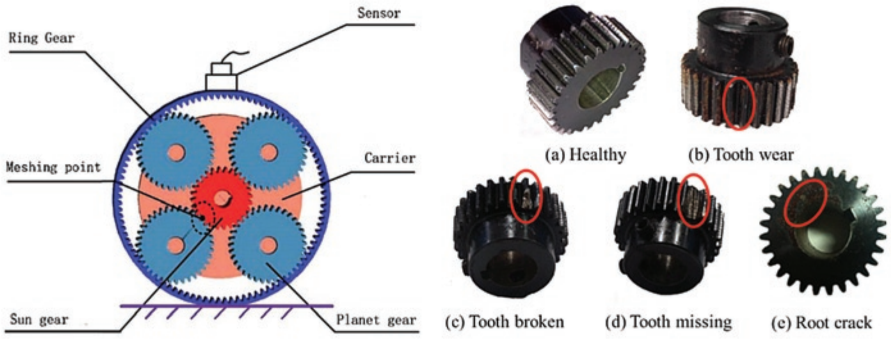


Fig. 8. The gearbox in DDS. (a) Schematic diagram; (b) 5 health conditions of sun gear

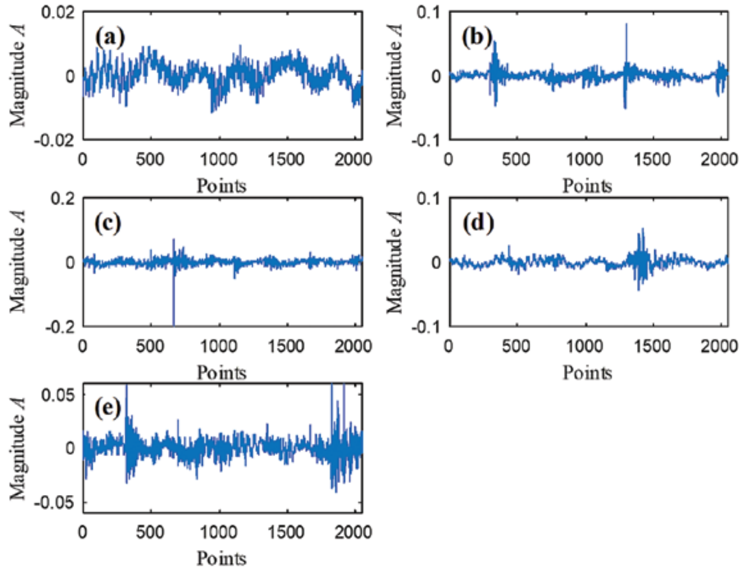


Fig. 9. Collected vibration signals of 5 machine conditions: (a) healthy(b) tooth broken(c) tooth crack (d) tooth missing (e) tooth wear.

root mean square, kurtosis, skewness, etc, are used as the input for shallow methods, i.e., SVM and RF, while the raw signal are sent to deep methods, i.e., LSTM, CNN and its variant methods, MCNN, due to the adaptive feature learning ability.

4. Results and discussions

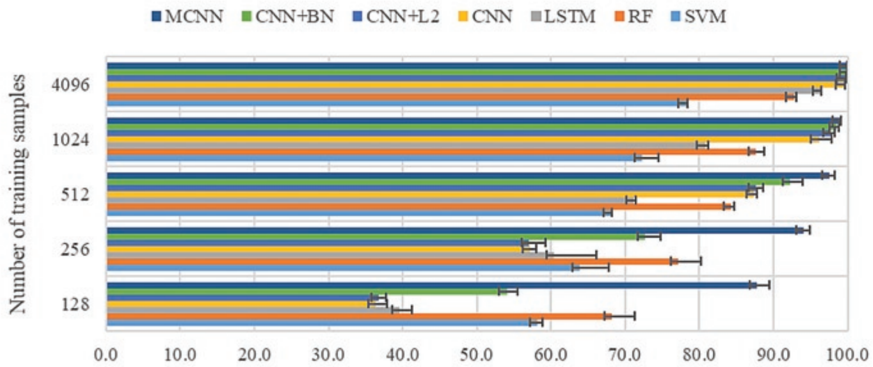
The results of two case studies versus different number of training samples for diverse methods are given in Figs. 10 and 11. Each result is an average of 10 random repeats, the average value and variance of classification accuracy of testing samples are shown in figures.

Deep and shallow learning structure comparison: As we can see in the Fig. 10, when sufficient training samples are provided, diagnostic models based on deep learning perform superior to the shallow machine learning methods, and all the intelligent diagnostic methods have achieved good classification results and it indicates that deep network structure has stronger feature learning ability than shallow network architecture.

Training sample size comparison: When the size of the samples decreases, the performance of traditional deep learning approaches presents a dramatic decline. CNN shows a worst sharp decrease in classification accuracy and it reveals that CNN is prone to over-fitting and has a weak generalization ability with small training sample size. BN and L2-norm are two frequently-used algorithms to prevent CNN from over-fitting. As is shown in the chart of Fig. 10, BN with CNN indeed has some improvements, compared to basic CNN, under small sample conditions.

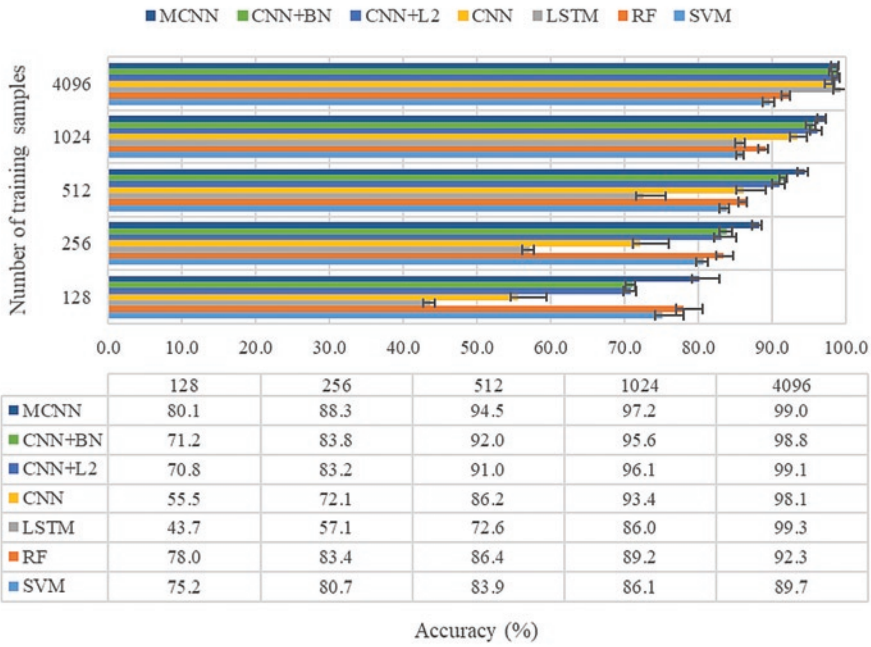
Comparisons with the proposed MCNN: As is shown in Fig. 10 and Fig. 11, MCNN exhibits a significant improvement in terms of fault classification accuracy compared with other models. Compared with CNN, CNN with L2-norm, and CNN with BN, MCNN increases classification accuracy with 51.3%, 50.9%, 33.7% respectively when using the number of training samples with 128. This remarkable improvement indicates that, with introduction of GAP into the network structure, the proposed MCNN exhibits superior advantages in feature learning and generalization with small number of training samples. It enables the proposed MCNN method can be a promising tool to deal with the real-world challenge for fault diagnosis with only limited labelled data available.

It also should be noted that classical machine learning algorithms, such as RF and



	128	256	512	1024	4096
■ MCNN	87.7	94.0	97.6	99.0	100.0
■ CNN+BN	54.0	72.7	92.3	98.6	99.9
■ CNN+L2	36.8	57.1	87.7	97.7	99.6
■ CNN	36.4	57.2	87.4	96.1	99.4
■ LSTM	39.6	60.4	71.2	80.7	96.3
■ RF	68.3	77.1	84.2	87.7	92.7
■ SVM	58.2	63.9	68.1	72.3	78.2

Fig. 10. Diagnosis results of PHM09 gearbox using different sample sizes with 7 methods



	128	256	512	1024	4096
■MCNN	80.1	88.3	94.5	97.2	99.0
■CNN+BN	71.2	83.8	92.0	95.6	98.8
■CNN+L2	70.8	83.2	91.0	96.1	99.1
■CNN	55.5	72.1	86.2	93.4	98.1
■LSTM	43.7	57.1	72.6	86.0	99.3
■RF	78.0	83.4	86.4	89.2	92.3
■SVM	75.2	80.7	83.9	86.1	89.7

Fig. 11. Diagnosis results of DDS planetary gearbox using different sample sizes with 7 methods

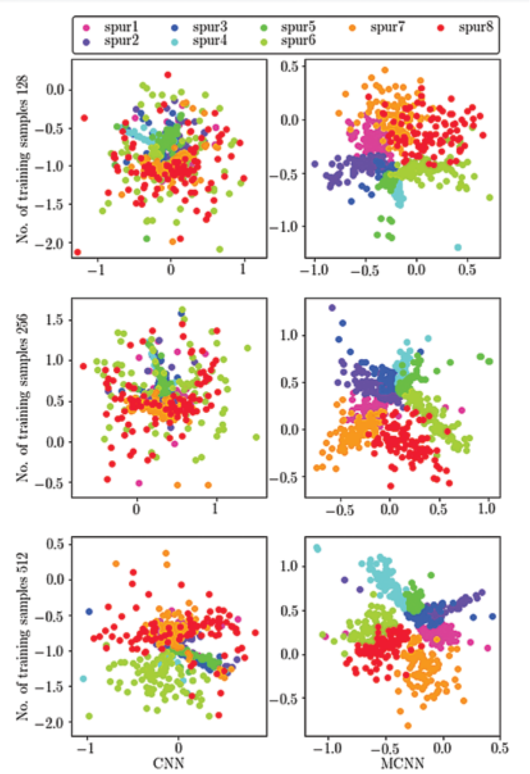


Fig. 12. Features visualization in PHM2009 dataset with CNN and MCNN

SVM, are shallow structure methods and are capable of handling small sample size. From the Fig. 10, RF and SVM are also performed better than most of the deep learning methods, such as LSTM, CNN and its variant methods, when the number of training samples are 128 and 256 respectively. Nevertheless, the proposed MCNN, though with the limited training samples of 128 and 256, it still achieves a higher diagnostic accuracy than RF and SVM. The traditional shallow methods of RF and SVM rely more on the quality of the artificial feature extraction. The classical statistical features in time-domain and frequency-domain are utilized in this paper [11]. These features have a certain sensitivity and resolving power for different faults, whereas, they possess distinct fault description capabilities for different application objectives. However, MCNN is a framework based on deep learning with remarkable adaptive feature learning capabilities. Experimental results illustrate that MCNN still achieves superior performance even under small sample conditions. Similar conclusions can also be demonstrated in Fig. 11.

In order to visually verify the effectiveness of the MCNN algorithm, t-distributed stochastic neighbour embedding (t-SNE) is applied to reduce the dimension of learned features for visualization. The features in the 5th convolutional layers are used for analysis, since this layer is the end of feature descriptor in traditional architecture and the learned features should be highly abstract and separable. Taking the data from 2009 PHM as an example, the results of feature visualization for CNN and MCNN versus small numbers of training samples (128, 256, 512) are shown in Fig. 12. It is clear that the features learned by MCNN are well clustered compared with the counterpart of CNN cases. For different number of samples, the features of CNN are mixed and overlapping, such as the spur 7(orange) and spur8 (red) and the traditional CNN fails the classification of these two kinds of faults. Again, the visualized results tell that the proposed MCNN features remarkable feature representation ability with limited number of training samples.

In addition, the training time for each training epoch and the memory usage during the training progress of the 2 datasets with different methods are recorded and presented in Table 3. As exhibited, for each training epoch, MCNN utilizes less computational time as well as low memory footprint. Compared to CNN, the computational times per training epoch of MCNN have been reduced by 0.168 seconds and 0.193 seconds. The memory footprint of MCNN have been reduced by 31.5MB and 28MB respectively in the two datasets.

5. Conclusion

In this work, a modified convolutional neural network that replacing the fully-connected layers with the global average pooling scheme, is proposed to reduce the number of trainable model parameters. The improved architecture possesses higher precision, less computational burden, superior generalization ability with limited training samples. Moreover, a MCNN-based intelligent fault diagnosis framework is presented. In order to assess the performance of the proposed method, the case studies on two industrial gearboxes are conducted from three aspects including the classification accuracy, the features visualization and the computational efficiency. These results demonstrate the

Table 3. Time cost and memory footprint by different approaches

Approaches	PHM2009 dataset		DDS dataset	
	Time(sec/epoch)	Memory (MB)	Time(sec/epoch)	Memory (MB)
CNN	0.915	383.1	0.873	232
CNN+l2	0.961	388.6	0.905	236
CNN+BN	0.914	410.3	0.909	243
MCNN	0.747	351.6	0.68	204

superiority of MCNN, compared to shallow machine learning methods i.e. SVM and RF and other popular deep learning approaches, such as CNN and its variant methods. Specifically, MCNN achieves the 51.3% and 24.6% improvements in the aspect of classification accuracy for two dataset with limited training data, i.e., 128 training samples. The impressive performances, achieved by the MCNN, show a broad prospect for intelligent fault diagnosis in the industrial gearbox.

Acknowledgement

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by the National Key Research and Development Program of China (Project No. 2017YFC0108401) and National Natural Science Foundation of China (5

References

1. Cai B P, Huang L, Xie M. Bayesian networks in fault diagnosis. IEEE Transactions on industrial informatics 2017; 13(5): 2227 - 2240, https://doi.org/10.1109/TII.2017.2695583.
2. Cai B P, Liu Y H, Fan Q ,Zhang Y W, Liu Z K, Yu S L, Ji R J. Multi-source information fusion based fault diagnosis of ground-source heat pump using Bayesian network. Applied Energy 2014; 114: 1-9, https://doi.org/10.1016/j.apenergy.2013.09.043.
3. Cai B P, Liu H L, Xie M. A real-time fault diagnosis methodology of complex systems using object-oriented Bayesian networks. Mechanical Systems and Signal Processing 2016; 80: 31-44, https://doi.org/10.1016/j.ymssp.2016.04.019.
4. Cai B P, Shao X Y, Liu Z K, Kong X D. Remaining useful life estimation of structure systems under the influence of multiple causes: Subsea pipelines as a case study. IEEE Transactions on Industrial Electronics 2019; 99:1-1, https://doi.org/10.1109/TIE.2019.2931491.
5. Chen D, Yang S, Zhou F. Transfer learning based fault diagnosis with missing data due to multi-rate sampling. Sensors 2019; 19(8):1826, https://doi.org/10.3390/s19081826.
6. Chen Yuejian, Liang Xihui, Zuo Ming J. Sparse time series modeling of the baseline vibration from a gearbox under time-varying speed condition. Mechanical Systems and Signal Processing 2019; 134: 106342, https://doi.org/10.1016/j.ymssp.2019.106342.
7. Chen ZY, Gryllias K, Li WH. Mechanical fault diagnosis using Convolutional Neural Networks and Extreme Learning Machine. Mechanical Systems and Signal Processing 2019; 133: 106272, https://doi.org/10.1016/j.ymssp.2019.106272.
8. Duan R, Lin Y, Zeng Y. Fault diagnosis for complex systems based on reliability analysis and sensors data considering epistemic uncertainty. Eksploatacja i Niezawodnosc - Maintenance and Reliability 2018; 20(4): 558-566, https://doi.org/10.17531/ein.2018.4.7.
9. Feng K, Wang K, Ni Q. A phase angle based diagnostic scheme to planetary gear faults diagnostics under non-stationary operational conditions. Journal of Sound and Vibration 2017; 408:190-209, https://doi.org/10.1016/j.jsv.2017.07.030.
10. Feng Z P, Zhu W P, Dong Zhang. Time-Frequency demodulation analysis via Vold-Kalman filter for wind turbine planetary gearbox fault diagnosis under nonstationary speeds. Mechanical Systems and Signal Processing 2019; 128: 93-109, https://doi.org/10.1016/j.ymssp.2019.03.036.
11. Han T, Jiang D, Qi Z, Lei W, Kai Y. Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. Transactions of the Institute of Measurement and Control 2018; 40: 2681-93, https://doi.org/10.1177/0142331217708242.
12. Han, T, Jiang, D, Sun, Y, Wang, N, Yang, Y. Intelligent fault diagnosis method for rotating machinery via dictionary learning and sparse representation-based classification. Measurement 2018; 118:181-193, https://doi.org/10.1016/j.measurement.2018.01.036.
13. Han T, Liu C, Wu Lj, Sarkar S, Jiang DX. An adaptive spatiotemporal feature learning approach for fault diagnosis in complex systems. Mechanical Systems and Signal Processing 2019; 117:170-187, https://doi.org/10.1016/j.ymssp.2018.07.048.
14. Han T, Liu C, Yang WG, Jiang DX. Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application. ISA Transactions 2019, https://doi.org/10.1016/j.isatra.2019.08.012.
15. Han T, Liu C, Yang WG, Jiang D. Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions. ISA Transactions, https://doi.org/10.1016/j.isatra.2019.03.017.
16. Huang W Y, Cheng J H. An improved deep convolutional neural network with multi-scale information for bearing fault diagnosis. Neurocomputing 2019, https://doi.org/10.1016/j.neucom.2019.05.052.
17. Jiao J, Zhao M, Lin J. Deep Coupled Dense Convolutional Network with Complementary Data for Intelligent Fault Diagnosis. IEEE Transactions on Industrial Electronics 2019; 66(12): 9858 - 9867, https://doi.org/10.1109/TIE.2019.2902817.
18. Kaluer S, Fekete K, Jozsa L, Klai Z. Fault diagnosis and identification in the distribution network using the fuzzy expert system. Eksploatacja i Niezawodnosc - Maintenance and Reliability 2018; 20(4): 621-629, https://doi.org/10.17531/ein.2018.4.13.
19. Lei J, Liu C, Jiang D. Fault diagnosis of wind turbine based on Long Short-Term memory networks. Renewable Energy 2019; 133: 422-432, https://doi.org/10.1016/j.renene.2018.10.031.
20. Lin M, Chen Q, Yan S C, Network in Network, Neural and Evolutionary Computing. arXiv:1312.4400.
21. Liu H , Zhou J , Zheng Y. Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders. ISA Transactions 2018; 77: 167-178, https://doi.org/10.1016/j.isatra.2018.04.005.
22. Liu XL, Zhang XY, Wang LY. Fault Diagnosis Method of Wind Turbine Gearbox Based on Deep Belief Network and Vibration Signal. Society of Instrument and Control Engineers of Japan.
23. PHM, Phm data challenge 2009., https://www.phmsociety.org/competition/PHM/09, 2009.
24. Rezaei M, Yang H J, Meinel C. Deep Neural Network with l2-norm Unit for Brain Lesions Detection. arXiv:1708.05221.
25. Shao H D, Jiang H K, Zhao K. A novel tracking deep wavelet auto-encoder method for intelligent fault diagnosis of electric locomotive bearings. Mechanical Systems and Signal Processing 2018; 110: 193-209, https://doi.org/10.1016/j.ymssp.2018.03.011.
26. Shao SY, Wang P, Yan R Q. Generative adversarial networks for data augmentation in machine fault diagnosis. Computers in Industry 2019; 106: 85-93, https://doi.org/10.1016/j.compind.2019.01.001.
27. Sikora M, Szczyrba K, Wróbel, Michalak M. Monitoring and maintenance of a gantry based on a wireless system for measurement and analysis of the vibration level. Eksploatacja i Niezawodnosc - Maintenance and Reliability 2019; 21(2): 341-350, https://doi.org/10.17531/ein.2019.2.19.

28. Tang G J, Pang B. Gearbox Fault Diagnosis Based on Hierarchical Instantaneous Energy Density Dispersion Entropy and Dynamic Time Warping. Entropy 2019; 21(6): 593, https://doi.org/10.3390/e21060593.
29. Tyagi S, Panigrahi S K. A Hybrid Genetic Algorithm and Back-Propagation Classifier for Gearbox Fault Diagnosis. Applied Artificial Intelligence 2017; 1-20, https://doi.org/10.1080/08839514.2017.1315502.
30. Verstraete D, Ferrada A, Droguett E.L, Meruane V, Modarres M. Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings. Shock and Vibration 2017; 2017: 1-17, https://doi.org/10.1155/2017/5067651.
31. Wang J, Li S, An Z. Batch-normalized deep neural networks for achieving fast intelligent fault diagnosis of machines. Neurocomputing 2019; 329: 53-65, https://doi.org/10.1016/j.neucom.2018.10.049.
32. Wang K S, Heyns P S. Application of computed order tracking, Vold-Kalman filtering and EMD in rotating machine vibration. Mechanical Systems and Signal Processing 2011; 25(1): 416-430, https://doi.org/10.1016/j.ymssp.2010.09.003.
33. Wang K S, Heyns P S. The combined use of order tracking techniques for enhanced Fourier analysis of order components. Mechanical Systems and Signal Processing 2011; 25(3): 803-811, https://doi.org/10.1016/j.ymssp.2010.10.005.
34. Wen L, Gao L, Li X, Wen L, Gao L, Li X. A new deep transfer learning based on sparse auto-encoder for fault diagnosis. IEEE Transactions on systems, man, and cybernetics: systems 2017; 1-9.
35. Xu H and Chen G. An intelligent fault identification method of rolling bearings based on LSSVM optimized by improved PSO. Mechanical Systems and Signal Processing 2013; 35(1-2): 167-175, https://doi.org/10.1016/j.ymssp.2012.09.005.
36. Yang J, Guo Y Q, Zhao W L. Long short-term memory neural network based fault detection and isolation for electro-mechanical actuators. Neurocomputing 2019; 360: 85-96, https://doi.org/10.1016/j.neucom.2019.06.029.
37. Yu J, Xu YG, Liu K. Planetary gear fault diagnosis using stacked denoising autoencoder and gated recurrent unit neural network under noisy environment and time-varying rotational speed conditions. Measurement Science and Technology 2019; 30: 095003, https://doi.org/10.1088/1361-6501/ab1da0.
38. Zhang M, Wang K, Li Y. Motion Periods of Planet Gear Fault Meshing Behavior. Sensors 2018; 18(11), https://doi.org/10.3390/s18113802.
39. Zhang M, Wang K S, Wei D D. Amplitudes of characteristic frequencies for fault diagnosis of planetary gearbox. Journal of Sound and Vibration 2018; 432:119-132, https://doi.org/10.1016/j.jsv.2018.06.011.
40. Zhang W, Li X and Ding Q. Deep residual learning-based fault diagnosis method for rotating machinery. ISA Transactions 2018, https://doi.org/10.1016/j.isatra.2018.12.025.
41. Zhang W, Peng G, Li C, Chen Y, Zhang Z. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors 2017; 17(2): 425, https://doi.org/10.3390/s17020425.
42. Zhang X L, Wang B J, Chen X F. Intelligent fault diagnosis of roller bearings with multivariable ensemble-based incremental support vector machine. Knowledge-Based Systems 2015; 89: 56-85, https://doi.org/10.1016/j.knosys.2015.06.017.
43. Zhang Z Z, Li S M. General normalized sparse filtering: A novel unsupervised learning method for rotating machinery fault diagnosis. Mechanical Systems and Signal Processing 2019; 124: 596-612, https://doi.org/10.1016/j.ymssp.2019.02.006.
44. Zhao X L, Jia M P. A new Local-Global Deep Neural Network and its application in rotating machinery fault diagnosis. Neurocomputing, https://doi.org/10.1016/j.neucom.2019.08.010.

Yaxin LI
Kesheng WANG

Department of Mechanical and Electrical Engineering
University of Electronic Science and Technology of China
Chengdu, 610059, China

E-mail: yaxinli@std.uestc.edu.cn, keshengwang@uestc.edu.cn