Mathematical Biosciences and Engineering

MBE, 16(5): 3311–3330.
DOI: 10.3934/mbe.2019165
Received: 23 January 2019
Accepted: 10 April 2019
Published: 17 April 2019

http://www.aimspress.com/journal/MBE

Research article

A negative correlation ensemble transfer learning method for fault diagnosis based on convolutional neural network

Long Wen[1], Liang Gao[1], Yan Dong[2] and Zheng Zhu[3,*]

[1] The State Key Laboratory of Digital Manufacturing Equipment & Technology, School of Mechanical Science & Engineering, Huazhong University of Science & Technology, Wuhan, 430074, China
[2] School of Electronic Information & Communications, Huazhong University of Science & Technology, Wuhan, 430074, China
[3] Department of Pathology, Longgang Central Hospital of Shenzhen City, Guangdong, 518116, China

* **Correspondence:** Email: zhfo@yahoo.com; Tel: +860755848069332231; Fax: +860755848069332233.

Abstract: With the development of the smart manufacturing, data-driven fault diagnosis has receiving more and more attentions from both academic and engineering fields. As one of the most important data-driven fault diagnosis method, deep learning (DL) has achieved remarkable applications. However, the DL based fault diagnosis methods still have the following two drawbacks: 1) One of the most major branch of deep learning is to construct the deeper structures, however the deep learning models in fault diagnosis is very shadow. 2) As stated by the no-free-lunch theorem, no single model can perform best on every dataset, and the individual deep learning model still suffers from the generalization ability. In this research, a new negative correlation ensemble transfer learning method (NCTE) is proposed. Firstly, the transfer learning based ResNet-50 is proposed to construct a deep learning structure that has 50 layers. Secondly, several fully-connected layers and softmax classifiers are trained cooperatively using negative correlation learning (NCL). Thirdly, the hyper-parameters of the proposed NCTE are determined by cross validation. The proposed NCTE is conducted on the KAT Bearing Dataset, and the prediction accuracy of NCTE is as high as 98.73%. This results show that NCTE has achieved a good results compared with other machine learning and deep learning method.

Keywords: negative correlation learning; transfer learning; ensemble learning; convolutional neural network, fault diagnosis

1. Introduction

Prognostic Health Management (PHM) system has become a vital part in modern industry. The goals of PHM are to reduce the risks to avoid the dangerous situations and improve the safety and reliability of the smart equipment and the systems [1]. Over the past decades, various attempts have been made to design effective methods to achieve the superior diagnosis performance. With the development of the smart manufacturing, the machines and equipment are more automatic and complicate, the intelligent fault diagnosis of these smart machines and equipment became necessary [2]. The data from the machine are boosting, and it can be collected much faster and more widely than ever before, the data-driven fault diagnosis has attracted more and more attentions from both academic and engineering fields [3].

Traditional learning-based approaches need to extract features of signals from time, frequency, and time-frequency domains [4]. The feature extraction is an essential step and the upper-bound performances of the leaning methods rely on the feature extraction process [5]. However, the traditional handcrafted feature extraction techniques need considerable domain knowledge, and the feature extraction process is very time-consuming and labor-intensive [6]. In recent years, deep learning (DL) has achieved huge success in image recognition and speech recognition [7]. It can learn the feature-representation from raw data automatically, and the key aspect is that this process is not depended on human engineers, which can eliminate the experts' effect as more as possible. DL has been widely applied in the machine health-monitoring field [3].

Even though the applications of deep learning have achieved remarkable results in fault diagnosis, there are still some problems for the further improvements. Firstly, the deep learning models implemented by many researchers only have less than five hidden layers [8], which limits their final prediction accuracies. However, the well-trained deep learning can reach hundreds of layers on ImageNet. How to bridge the gap between the deep models in fault diagnosis and those in ImageNet could promote the performance of deep models in fault diagnosis. Secondly, the individual deep learning models for fault diagnosis still suffers from the generalization ability [9]. As stated by the no-free-lunch theorem [10–12], no single model can perform best on every dataset. To improve the generalization ability of deep learning method is essential.

To overcome these two drawbacks, a new ensemble version of deep learning method is proposed. Firstly, the transfer learning is applied to bridge the network gap between fault diagnosis and ImageNet. TL can learn a learning system from a dataset (source domain) and then applies this system to solve a new problem (target domain) more quickly and effectively. It should be noted that the new target domain can be irrelative with the source domain [13]. So the ResNet-50 which is pre-trained on the ImageNet can also perform well in fault diagnosis. The ResNet-50 has the depth of 50 layers, which is much deeper than traditional DL model applied in fault diagnosis, and it could improve the predication accuracy on fault diagnosis field. Secondly, the ensemble learning is also investigated in this research. Ensemble learning is an effective way to improve the generalization ability. Several classifiers are trained cooperatively using negative correlation learning (NCL), and then these classifiers are combined to form a powerful fault classifier. In this research, the transfer

learning technique and the NCL technique are combined, and a new negative correlation transfer ensemble model (NCTE) is proposed for fault diagnosis.

The rest of this paper is organized as follows. Section 2 discusses literature review. Section 3 presents the methodologies of negative correlation learning. Section 4 presents the proposed NCTE. Section 5 presents the case studies. The conclusion and future researches are presented in Section 6.

2. Literature review

2.1. Data-driven fault diagnosis

With the development of smart manufacturing, the data-driven fault diagnosis has received more and more attentions. It is very suitable for the complicated industrial systems, since the data-driven fault diagnosis applied the learning-based approaches to learn from the historic data without the models about the system [14–16]. The learning-based approaches can be classified into statistical analysis, machining learning methods and their joint paradigm. Principal component analysis (PCA), partial least squares (PLS), and independent Component Correlation (ICA) have received considerable attentions on the industrial process monitoring [17]. The machine learning methods also achieved good applications in fault diagnosis, such as support vector machine (SVM) [18,19], artificial neural network (ANN) [20], Bayesian network [21].

Since deep learning (DL) methods can obtain the feature-representations of raw data in an automatically way, it has shown a great potential in machine health monitoring field [3,22]. Wang et al. [23] investigated an adaptive deep CNN model, and the main parameters were determined by particle swarm optimization. Shao et al. [2] studied deep belief network based fault diagnosis on rolling bearing. Wang et al. [24] studied a new type of bilateral long short-term memory model (LSTM) for the cycle time prediction of re-entrant manufacturing system. Pan et al. [25] proposed a LiftingNet for mechanical data analysis and the results showed that LiftingNet has a good performance on different rotating speeds. Li [26] studied IDSCNN with D-S evidence for bearing fault diagnosis. This method is also an ensemble CNN, and it has a good adaptability on different load conditions. Lu et al. [27] applied Convolutional Neural Network (CNN) to fault diagnosis, and the comparison experiments showed that the accuracy of greater than 90% was achieved with fewer computational resource. Zhang et al. [28] studied the intelligent fault diagnosis under varying working conditions using domain adaptive CNN method.

However, due to the fact that the volume of labeled samples in fault diagnosis is relatively small compared with ten million annotated images in ImageNet, the DL models for fault diagnosis are shallow compared with benchmark deep learning models in ImageNet. However, it is hard to train a deep model without the large amount of well-organized training dataset like ImageNet, so to train a very deep model on fault diagnosis field is almost impossible. To deal with this challenge, by applying transfer learning and taking the deep CNN model trained on ImageNet as the feature extractor, the deep learning model that trained on ImageNet can also perform well on small data in another domain.

2.2. Transfer learning

Transfer learning (TL) is a new paradigm in machine learning field. TL can learn a learning

system from a dataset (source domain) and then applies this system to solve a new problem (target domain) more quickly and effectively. It should be noted that the new target domain can be irrelative with the source domain [13].

TL has been studied by many researchers. Donahue et al. [29] investigated the generic tasks, which may be suffered by insufficient labeled data for training a deep DL model, and they released DeCAF as generic image features across many visual recognition tasks. Based on DeCAF, Ren et al. [30] studied a feature transferring learning method using pre-trained DeCAF for Automated Surface Inspection, as shown in Figure 1. They tested the proposed methods on NEU surface defect database, weld defect database, wood defect database and micro-structure defect dataset, and the results showed that the proposed algorithm outperforms several best benchmarks in literature.

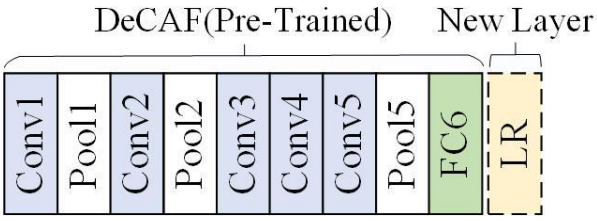

Figure 1. Structure of DeCAF Based automatically surface inspection method.

Many other famous CNN models that trained on ImageNet are also investigated for transfer learning, such as CifarNet, AlexNet, GoogleNet, ResNet and so on. Wehrmann et al. [31] studied a novel approach for adult content detection in videos and applied both pre-trained GoogleNet and ResNet architectures as the feature extractor. The results shown that the proposed method outperformed the current state-of-the-art methods for adult content detection. Shin et al. [32] applied CifarNet, AlexNet and GoogLeNet for the computer-aided detection in medical imaging tasks. They also investigated when and why transfer learning from pre-trained ImageNet (via fine-tuning) can be useful, and the results have achieved the state-of-the-art performance. Rezende et al. [33] investigated the transfer learning on ResNet-50 on the classification of malicious software, and the results showed that this approach can effectively classify the malware families with the accuracy of 98.62%.

Applying the pre-trained CNN models that trained on ImageNet to fault diagnosis has investigated by many researchers. Janssens et al. [34] selected the pre-trained VGG-16 as the feature extractor and fine-tuning all the weights of the network. The proposed transfer learning method has been applied to use the infrared thermal video to automatically determine the condition of the machine. Shao et al. [8] proposed a VGG-16 based deep transfer learning fault diagnosis and the structure of their method has been shown in Figure 2. The proposed method is applied on induction motors, gearboxes, and bearings dataset and the results showed that it has achieved a significant improvement by using the transfer learning technique. The application of transfer learning on fault diagnosis has great potential to improve the prediction accuracies.

The advantage of TL on fault diagnosis can be summarized as two aspects. Firstly, the labeled data in fault diagnosis is also small, and it is hard to train deep models in fault diagnosis, which could limit the prediction of deep learning in fault diagnosis. With transfer learning, the deep models can extract better features on fault diagnosis and then improve the accuracy on fault diagnosis.

Secondly, the deeper models has much more parameters than shallow models. The training of a deep model requires considerable computational and time resources as well as a large amount of labelled data. However, by using transfer learning, only the fine-tuning process is necessary, which could reduce the requirements on hardware and training process.

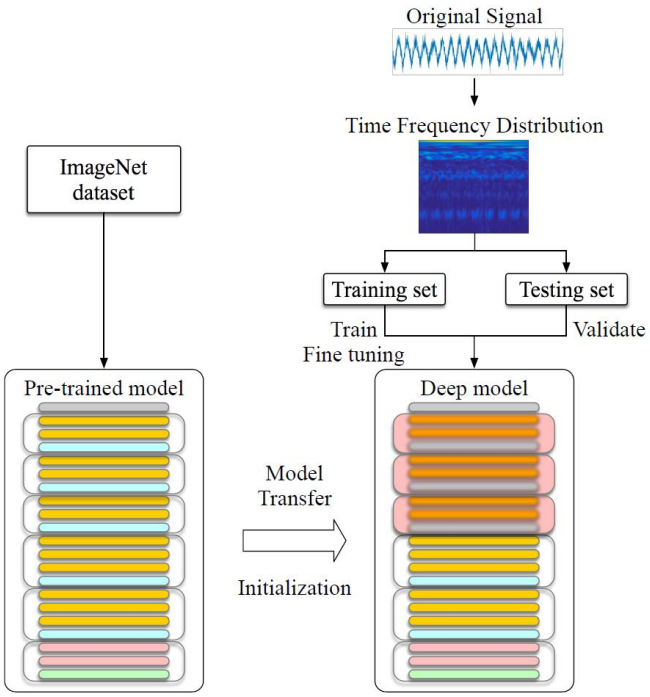

Figure 2. The deep transfer learning using VGG-16 on fault diagnosis [8].

Even the great improvement has been achieved by the transfer learning on fault diagnosis field, the application of transfer learning on fault diagnosis is only at the beginning. The further investigation and improvement on the transfer learning is necessary. In this research, a new ensemble transfer learning by using negative correlation ensemble is proposed.

2.3. Ensemble method in fault diagnosis

Ensemble method is a learning pattern in which a group of base learners is trained for the same task, and they worked together as the committee to give the final results. As stated by the no-free-lunch theorem [10,35,36], no single model can perform best on every dataset. The ensemble learning becomes an effective way to improve the performance. The ensemble learning was proposed by Hansen and Salamons [37], and their results provided the solid support that the generalization ability of a neural network can be significantly improved through combining a number of neural networks.

Ensemble learning has been studied by many researchers, and these ensemble algorithms can be classified into three categories [38]. In the first category, each base learner is trained with a subset of training samples, and then these base learners are combined at advance. The typical ensemble algorithm is Bagging and its variants. In the second category, the weights are introduced on the training samples and the training samples that are misclassified by the former base learner would be

paid more attention in the next training stage. The algorithms in the second categories include adaboosting and its variants. In the third category, the interaction and cooperation among the base learners are necessary to generate a more diverse group of base learner. One of the typical algorithm in the third category is the negative correlation learning (NCL). NCL emphasizes the cooperation and specialization among different base learners during the base learner design. It provides an opportunity for different base learner to interact with each other to solve one single problem. The accuracy and the diversity of the group of base learner, and the results of NCL has shown a good potential [39].

The ensemble learning in fault diagnosis has also been investigated. Hu et al. [40] proposed a new ensemble approach for the data-driven remaining useful life estimation. Their ensemble method is the first category, and the member algorithms are weighted to form the final ensemble algorithm. The accuracy-based weighting, diversity-based weighting and optimization-based weighting are applied and the results showed that the ensemble approach with any weighting scheme gives more accurate RUL predictions compared to any sole member algorithm. Wang et al. [9] studied the selective ensemble neural networks (PSOSEN) for the fault diagnosis of bearings and pumps. In their method, the adaptive particle swarm optimization (APSO) is developed for not only determining the optimal weights but also selecting superior base learners. The results demonstrated that PSOSEN has achieved desirable accuracies and robustness under the environmental noise and working condition fluctuations. Wu et al. [41] proposed the Easy-SMT ensemble algorithm based on synthesizing SMOTE-based data augmentation policy. The method is tested on the PHM 2015 challenge datasets and the results showed that it could achieve good performance on multi-class imbalance learning task.

However, even though the ensemble learning has achieved remarkable results in the fault diagnosis field, as far as I know, the NCL technique has not been applied on fault diagnosis. In this research, the NCL is combined with transfer learning to construct the high accuracy classifier for fault diagnosis.

3. Negative correlation learning

NCL introduces a correlation penalty term to the error function of each individual network in the ensemble so that all the networks can be trained interactively on the same training dataset. Given the training dataset $\{x_n, y_n\}_{n=1}^{N}$, NCL combines M neural networks $f_i(x)$ to constitute the ensemble.

$$f_{ens}(x_n) = \frac{1}{M}\sum_{i=1}^{M} f_i(x_n) \tag{1}$$

To train network f_i, the cost function e_i for network i is defined by Eq 2. Where λ is a weighting parameter on the penalty term p_i as shown in Eq 3.

$$e_i = \sum_{n=1}^{N}\left(f_i(x_n) - y_n \right)^2 + \lambda p_i \tag{2}$$

$$p_i = \sum_{n=1}^{N}\left\{\left(f_i(x_n) - f_{ens}(x_n)\right)\sum_{j\neq i}\left(f_j(x_n) - f_{ens}(x_n)\right)\right\} = -\sum_{n=1}^{N}\left(f_i(x_n) - f_{ens}(x_n)\right)^2 \tag{3}$$

From Eq 2, it can be seen that NCL uses a penalty term in the error function to produce base learners whose errors tend to be negatively correlated. So the NCL model can cooperate the training of base learner and the whole ensemble model simultaneously. λ control the degree of the negatively correlated. If set $\lambda = 0$, then the error Eq 2 will become Eq 4, and each individual models would be trained separately. When set $\lambda = 1$, then error Eq 2 can be trained as a whole ensemble model.

$$e_i = \sum_{n=1}^{N}\left(f_i(x_n) - y_n\right)^2 \tag{4}$$

$$e_i = \sum_{n=1}^{N}\left(f_i(x_n) - y_n\right)^2 - \sum_{n=1}^{N}\left(f_i(x_n) - f_{ens}(x_n)\right)^2 = \sum_{n=1}^{N}\left(f_{ens}(x_n) - y_n\right)^2 \tag{5}$$

In this research, the NCL technique is applied with transfer learning technique to obtain a new ensemble method for fault diagnosis.

4. Proposed negative correlation transfer ensemble model for fault diagnosis

In this section, a new negative correlation transfer ensemble model (NCTE) is proposed.

4.1. The flowchart of the proposed NCTE

The whole flowchart of the proposed NCTE consists of four parts, the data preprocessing part, the feature transferring part, the fine-tuning part and the hyper-parameter selection part.

(1) Data preprocessing part: Since the input of ResNet-50 is the RGB images, it is essential to convert the time-domain signals to 3D matrix in order to use the pre-trained ResNet-50 network.

(2) Feature transferring part: Establish the structure of ResNet-50, and keep the layers weights in ResNet-50 unchanged. Since the output of ResNet-50 is 2048, the feature obtained by ResNet-50 is also a 2048 vector.

(3) Training part: Adding the several separated fully-connected (FC) layers at the end of ResNet-50, and then training these FC layers using the NCL technique.

(4) Hyper-parameter selection part: It is vital to select the key parameter, λ, in the NCL technique. In this research, the cross validation is applied to test the most proper λ.

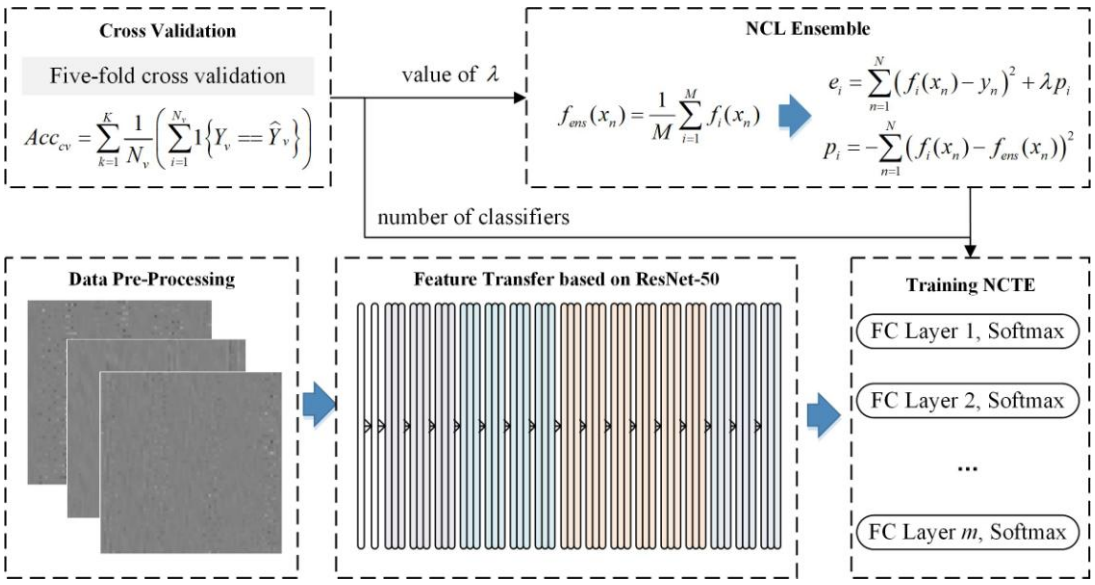

Figure 3. The Flowchart of the proposed NCTE.

The flowchart of the proposed NCTE is presented in Figure 3. The details of these four parts are given as following:

4.2. Data preprocessing

Data preprocessing is the essential part in the data-driven fault diagnosis. Since the input of ResNet-50 is the 3D natural image, so it is essential to transfer the time-domain signals to the 3D format. Chong [42] proposed the data preprocessing methods to convert the time-domain raw fault signals to 2D images. Wen et al [43] studied a new time domain signal to gray image method. Suppose the raw fault signals of all fault types are collected and then segmented to obtain the data samples. Let $m \times m$ denote the gray image size and $L_i(a)$, $i=1...N$, $a=1...m^2$, denote the strength value of signal samples. N the number of samples. $GP(j,k)$, $j=1...m$, $k=1...m$ is matrix of 2D gray images. The time domain signals to gray images can be formulated by Eq 6.

$$GP(j,k) = \frac{L((j-1)\times m + k) - Min(L)}{Max(L) - Min(L)} \times 255 \tag{6}$$

However, RGB image is 3D matrix format. Let $RP(j,k,p)$, $p=1,2,3$ presents this 3D matrix. The third elements of the RGB image are the strength of red (p=1), green (p=2) and blue (p=3) channels. In this research, the data preprocessing method that transfers the time domain raw signals to 3D RGB images is presented as Eqs 7–10.

$$NM_i(j,k) = \frac{L_i((j-1)\times M + k) - Min_{i,j,k}\left(L_i((j-1)\times M + k)\right)}{Max_{i,j,k}\left(L_i((j-1)\times M + k)\right) - Min_{i,j,k}\left(L_i((j-1)\times M + k)\right)} \tag{7}$$

$$RP_i(j,k,1) = NM_i(j,k) \times 255 \tag{8}$$

$$RP_i(j,k,2) = NM_i(j,k) \times 255 \tag{9}$$

$$RP_i(j,k,3) = NM_i(j,k) \times 255 \tag{10}$$

The difference between Eq 6 and Eq 7 is that Eq 6 applies the maximum and minimum values of the data sample while Eq 7 selects the maximum and minimum values of the whole samples. Then scale the normalized matrix (*NM(j,k)*) to 0-255 and copy the scaled results to *RP(j,k,p)*, as shown in Eq 8–10.

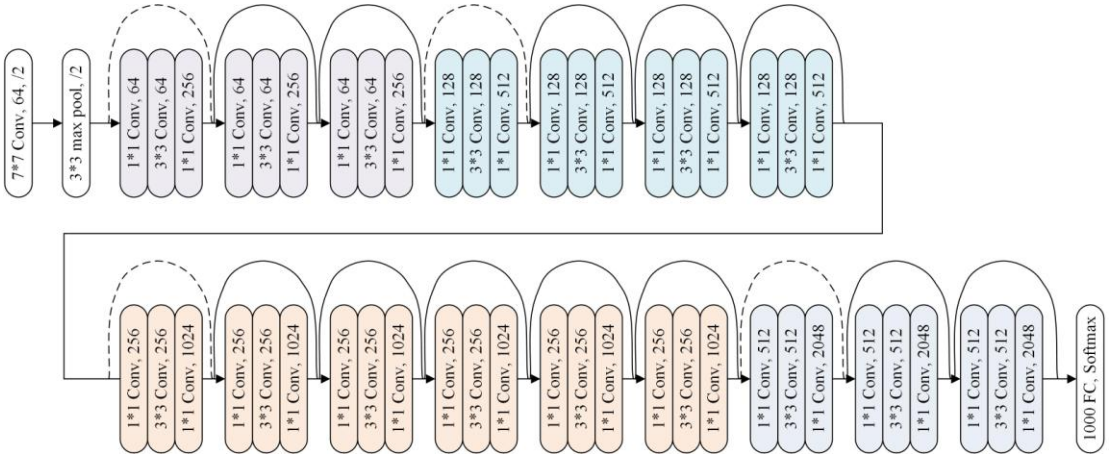

Figure 4. The Structure of ResNet-50 Network.

4.3. Feature transfer based on ResNet-50

Residual Networks (ResNet) [44] is a very famous Convolutional Neural Network developed in recent years. Since the vanishing/exploding gradient problem is also found in deep learning algorithms using gradient-based learning methods and backpropagation [45], the ResNet applied the shortcut connections to construct the deep networks to avoid this problem, and it has shown a great performance in image recognition.

ResNet-50 is a released version of ResNet, which has 50 layers. The input of ResNet-50 is 224 × 224, and the detail structure of ResNet-50 is shown in Figure 4. The output of ResNet-50 is 1000. In this research, the transfer learning is combined with ResNet-50 and the NCL technique is applied to train several newly constructed FC layers and softmax classifiers.

4.4. The training method of NCTE

Based on the ResNet-50, a new structure of NCTE is proposed. For most transfer learning method, there are only one softmax classifier. However, in this research, total *M* and softmax classifier are conducted in order to form the inherit ensemble version of transfer learning. As shown in Figure 5, along with the sofmax classifiers, one FC layer is also constructed for each softmax classifier, and the hidden neurons are 128 for all FC layers. FC layers of each softmax are separate and they have no interaction to each other.

Since there are *M* classifiers in the structure, the final output of the NCTE is the ensemble version of all *M* classifiers, and the bagging ensemble is applied, as shown in Eq 1. The training of these *M* classifiers are based on the NCL training process. For the training of each softmax classifier, there are two parts in the error function. The first part is the error function between the output of softmax classifier and the labels. The second part is the diversity term, and it tries to make *M* classifiers to be as diversity as possible. The second part worked as the penalty term in the loss function. The training method of NCTE is presented in Algorithm (1).

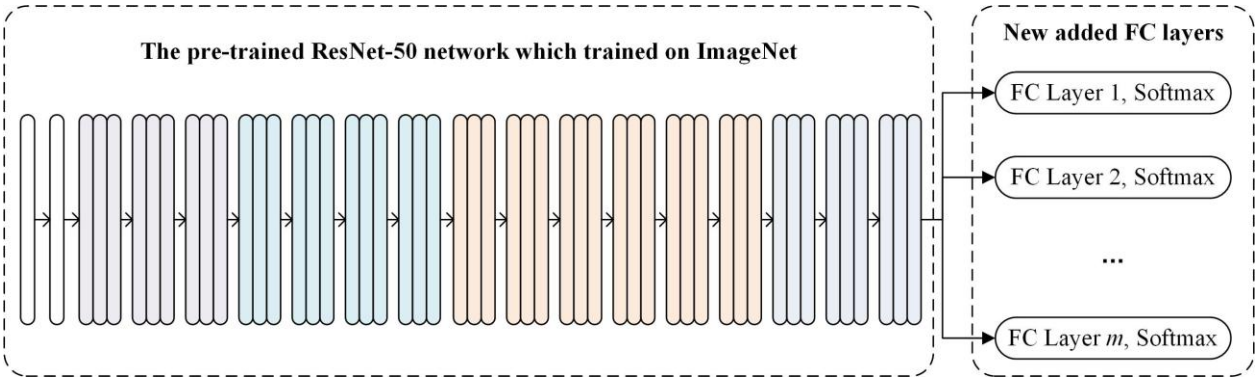

Figure 5. The structure of the proposed NCTE.

Algorithm (1), Training method for NCLE

Step 1: Let *M* be the final number of classifiers

Step 2: Take a training dataset $\{x_n, y_n\}_{n=1}^{N}$ and the hyper-parameter λ.

Step 3: For the training dataset, repeat the following (a) to (d) steps until the maximal epochs is reached:

(a) Calculate the ensemble output of *M* softmax classifiers.

$$f_{ens}(x_n) = \frac{1}{M}\sum_{i=1}^{M} f_i(x_n)$$

(b) For each softmax classifiers, from *i*=1 to *M*, for each weight w_{ij} in FC layer and softmax classifiers *i*, perform the update of the *i*-th FC layer and softmax classifiers:

$$e_i = \sum_{n=1}^{N}\left(f_i(x_n) - y_n\right)^2 - \lambda\sum_{n=1}^{N}\left(f_i(x_n) - f_{ens}(x_n)\right)^2$$

$$\frac{\partial e_i}{\partial w_{ij}} = 2\sum_{n=1}^{N}\left(f_i(x_n) - y_n\right)\frac{\partial f_i}{\partial w_{ij}} - 2\lambda\sum_{n=1}^{N}\left(f_i(x_n) - f_{ens}(x_n)\right)\left(1-\frac{1}{M}\right)\frac{\partial f_i}{\partial w_{ij}}$$

(c) Calculate the new output of the *i*-th softmax classifiers.

(d) Repeat (a)-(c) until all *M* FC layer and softmax classifiers are updated.

Step 4: Combine all softmax classifiers to formulate the final ensemble classifiers.

4.5. Hyper-parameter selection using cross validation

As shown in Eq 2, hyper-parameter λ control the degree of the negative correlate rate of the NCTE, so to select a proper hyper-parameter λ is vital for NCTE. In this research, the λ is selected according to its model performance. In many data-driven fault diagnosis methods, the performance is evaluated by the testing dataset, and the model that has the best performance on the testing dataset are selected. However, this model selection method has the following shortcomings: (1) It requires the testing dataset in addition to the training data. However, the testing dataset should be untouched during the training method and model selection period. (2) The selected standalone algorithm may not be robust, since no statistical analysis of the results are conducted. To overcome the above shortcoming, the cross validation technique is applied in these researches to obtain a reliable performance evaluation method for the model selection.

Cross validation (CV) is a popular technique to obtain a reliable model [46]. The CV technique divides the training dataset into two parts, and they are the training part and the validation part. The typical CV techniques includes Leave-one-out CV, Generalized CV, K-Fold CV and so on [47]. *K*-fold CV is the most popular technique of CV techniques. It divides the whole data into *K* subsamples with approximately equal cardinality *N/K* samples. Each subsample successively plays the role of validation part, while the rest *K*-1 subsamples are used for train part. However, the selection of *K* has no theoretical analysis [48], and the popular value of *K* are set to be 3, 5 and 10. In this research, the five-fold cross validation is applied.

Suppose Y_v and $\hat{Y}_v$ denote the actual and prediction labels on the validate part, and N_v is the sample number of validate dataset. The accuracy of CV (Acc_{cv}) is the mean of five-fold accuracy, and it can be shown by Eq 11.

$$Acc_{cv} = \sum_{k=1}^{K} \frac{1}{N_v} \left(\sum_{i=1}^{N_v} 1\left\{ Y_v == \hat{Y}_v \right\} \right) \tag{11}$$

The Acc_{cv} is applied to the selection of the proper λ. After finishing this selection, the obtained fault diagnosis classifier would be tested on a separated testing dataset, and the accuracy of testing dataset is the final results (*Acc*) of NCTE for comparison.

5. Case studies: KAT bearing dataset

5.1. Data description

The KAT bearing damage dataset provided by KAT datacenter in Paderborn University [45]. The hardware of this experiment is shown in [45], and there are 15 datasets and they can be categorized as three healthy classifications as shown in Table 1. The K0-series (K001–K005) are the healthy condition, the KA-series (KA04, KA15, KA16, KA22, KA30) are the outer bearing ring with damage and the KI-series (KI04, KI14, KI16, KI18, KI21) are the inner bearing ring with damage. The experiments are conducted with four different operating parameters, and the operating

parameters are shown in Table 2. Each experiment is conducted 20 repeated and the vibrations signals are collected for analysis, and the sampling rate is 64 kHz. It should be noted that the damage of this dataset is real damages caused by accelerated lifetime test.

Table 1. Categorization of datasets.

Healthy (Class 1)	Outer ring damage (Class 2)	Inner ring damage (Class 3)
K001	KA04	KI04
K002	KA15	KI14
K003	KA16	KI16
K004	KA22	KI18
K005	KA30	KI21

Table 2. Four operation parameters.

No.	Rotational speed	Load torque	Radial force
0	1500	0.7	1000
1	900	0.7	1000
2	1500	0.1	1000
3	1500	0.7	400

5.2. Hyper-parameter selection using CV technique

During the experiments, the algorithm is written in python 3.5 using Tensorflow. The hidden neurons in the FC layers are set to be 128, the L2 regulations rate is 1e-5, m is set to be 64. The learning rate scheduler is the momentum optimizer and the initial learning rate is 0.005 and the momentum value is 0.9. The batch size is 200, and the total epoch is 40. In this research, the five-fold cross validation is applied for selection the proper λ. The tested λ are from 0 to 1 with the increment of 0.1.

Table 3. The results of cross validation (Acc_{cv}) on the hyper-parameter λ.

λ	0	0.1	0.2	0.3	0.4	0.5
max	98.67%	98.62%	**98.71%**	98.68%	98.68%	98.66%
mean	98.52%	98.56%	98.49%	98.52%	**98.62%**	98.55%
min	98.14%	98.46%	98.13%	98.21%	**98.59%**	98.44%
std	0.0022	0.0006	0.0024	0.0018	**0.0004**	0.0009
λ	0.6	0.7	0.8	0.9	1.0	
max	98.68%	98.63%	98.64%	98.64%	98.67%	
mean	98.60%	98.52%	98.47%	98.29%	98.56%	
min	98.50%	98.27%	97.98%	97.76%	98.48%	
std	0.0008	0.0016	0.0028	0.0039	0.0007	

During the cross validation process, the number of the softmax classifiers are set to be 2, and the effect of the λ on the cross validation process is presented in Table 3 and Figure 6. From Table

3, it can be seen that the mean (*mean*), the minimum (*min*) and the stand deviation (*std*) of the Acc_{cv} is the best on all the values of λ. Since the results of λ=0.4 have the best *mean* and *std*, the selection of λ is 0.4 in this round. Figure 6 presents the mean value of Acc_{cv} along with the increase of λ. It can be seen that the whole curve like an inverse 'U' type, and the peak of this curve is also at λ=0.4.

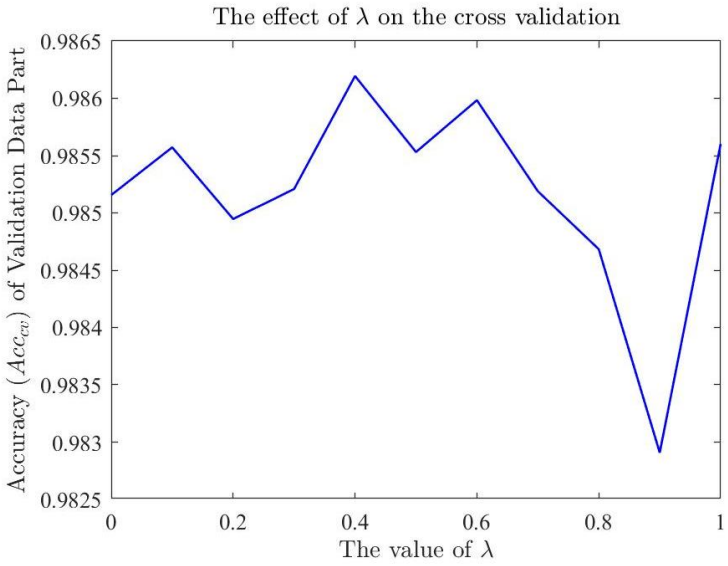

Figure 6. The effect of λ on the cross validation process.

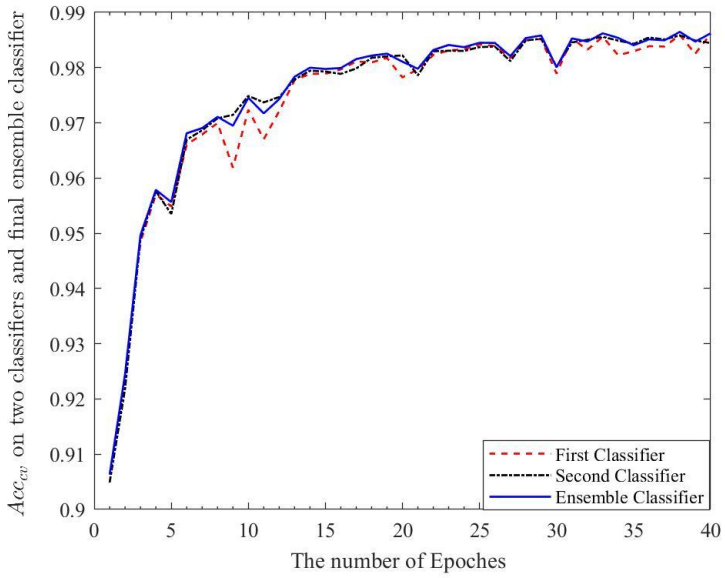

Figure 7. The convergence of two classifiers and the final ensemble classifier (%).

The convergence of two classifiers and the final ensemble classifier (NCTE) are plotted in Figure 7. From the results, it can be seen that both two classifiers have similar convergence speed, and the final ensemble classifier outperforms the two classifiers at most time. These results validate that the ensemble of these two classifiers can promote the performance than the individual single classifiers.

5.3. Sensibility analysis of classifier number

The number of the classifiers is also an important hyper-parameter for NCTE. In this subsection, the effect of the number of classifiers on the final results is analyzed. The number of classifiers in the experiments are set to be 2, 3, 5, 7, 9, 11, 13, and 15. The NCTE with the large number of classifiers are discussed as well, and the number of classifiers are 20, 30 and 50. The baseline method is using the NCTE with only one number of the classifiers.

The results in this experiment are presented in Table 4 and Figure 8. The best λ of cross validation, Acc_{cv}, *Acc* and the training time are presented in Table 4. For each results, only the best λ value and Acc_{cv} is presented. From the results, it can be seen that the best number of the classifiers is 13. And the Acc_{cv} is 98.73% while the performance of this version of NCTE in the testing dataset *Acc* is also the best among these methods, and it is as high as 98.72%.

Table 4. The results of cross validation (Acc_{cv}) on the number of the classifiers.

Number of classifiers	**1 (Baseline)**	**2**	**3**	**5**	**7**	**9**
λ value	-	0.4	0.8	0.5	0.4	1.0
Acc_{cv}	98.41%	98.62%	98.65%	98.64%	98.68%	98.70%
Acc	98.38%	98.62%	98.64%	98.63%	98.67%	98.66%
Time	261.31	429.27	608.82	930.67	1320.73	1670.69
Number of classifiers	**11**	**13**	**15**	**20**	**30**	**50**
λ value	0.8	**0.4**	0.1	0.2	0.2	0
Acc_{cv}	98.69%	**98.73%**	98.71%	98.69%	98.69%	98.69%
Acc	98.67%	**98.72%**	98.69%	98.67%	98.67%	98.68%
Time	1932.04	2389.01	2626.02	3447.68	4706.05	8082.40

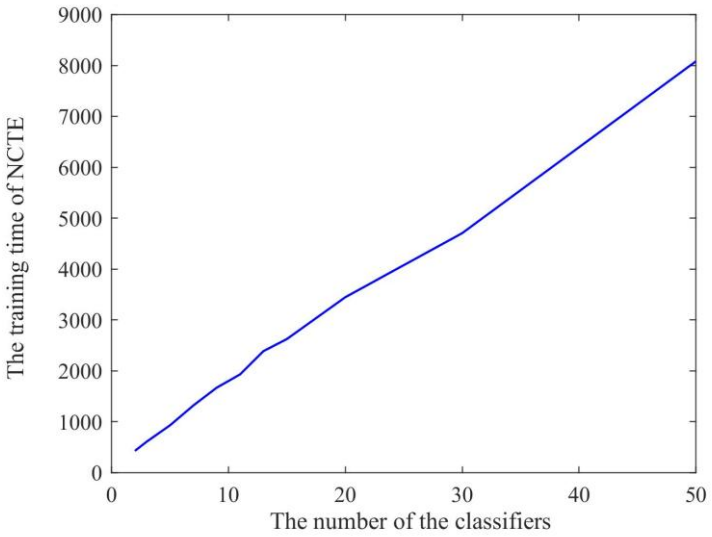

Figure 8. The training times of NCTE with different number of the classifiers (second).

On the other side, the training time increases sharply along with the number of the classifiers, as

shown in Figure 8. From the Figure 8, it can be seen that the number of the classifiers should be keep in a proper size. A large number of classifiers don't help to increase the final accuracy while it would increase the computation resource largely. However, taking the baseline into consideration, the *Acc* of baseline is only 98.41%, all NCTE variants are better than this result.

5.4. The analysis on TL and NCL

In this subsection, the NCTL is compared with traditional bagging method and the ResNet-50. The bagging is select as the k-fold bagging [1,50]. The ResNet-50 are random initialized and there are used to show the effect of TL. The comparison results are shown in TABLE 5. It should be noted that the bagging method is also based on TL, and it replace the ensemble method from NCL to Bagging. The ResNet-50 uses the same data-preprocessing process with NCTL, but it trained from the raw data without TL.

From the results, it can be seen that the accuracy of Bagging is 98.62%, which is inferior to NCTL slightly. The results of ResNet-50 is 72.31%. The results show that the NCTL has better performance than the random initialized ResNet-50. These results show that transfer learning using the pre-trained ResNet-50 could provide better results than to train a new random initialized ResNet-50.

Table 5. The analysis of NCTE on TL and NCL (%).

Methods	Mean Accuracy
NCTE	98.73
Bagging	98.62
ResNet-50	72.31

5.5. The results and comparison

In order to validate the performance of the proposed NCTE, the version of NCTE with 13 classifiers are compared with other published methods. The comparison of NCTE with traditional machine learning methods [49] are presented in Table 5, and the comparison of NCTE with deep learning methods are presented in Table 6.

Table 6. The comparison of NCTE with traditional machine learning methods (%).

Methods	Mean Accuracy
NCTE	98.73
Ensemble	98.3
CART	98.3
RF	98.3
BT	83.3
SVM-PSO	75.8
KNN	62.5
ELM	60.8
NN	44.2

In Table 6, the comparison methods are classification and regression trees (CART), random forests (RF), Boosted Trees (BT), neural networks (NN), support vector machines with parameters optimally tuned using particle swarm optimization (SVM-PSO), extreme learning machine (ELM), k-nearest neighbors (KNN) and their ensemble algorithms using majority voting (Ensemble). The details of these methods can be found in [49], and here their results are directly taken from [49]. From the results, it can be seen that NCTE has achieved a good result, and it outperforms all these traditional machine learning methods.

Table 7 presents the comparison of NCTE with other deep learning methods. These deep learning methods are deep inception net with atrous convolution (ACDIN), Convolution Neural Networks with Training Interference (TICNN), Deep Convolutional Neural Networks with Wide First-layer Kernels (WDCNN), AlexNet, ResNet and convolutional neural network based on a capsule network with an inception block (ICN). Their results can be found in [51] and [52]. The results show that the prediction accuracy of ACDIN, TICNN, WDCNN, AlexNet, ResNet and ICN are 94.5%, 54.09%, 54.55%, 79.92%, 77.52% and 82.05% respectively. These results validate the performance of NCTE.

Table 7. The comparison of NCTE with deep learning methods (%).

Methods	Mean Accuracy
NCTE	98.73
ACDIN 51	94.5
TICNN 51	54.09
WDCNN 51	54.55
AlexNet 52	79.92
ResNet 52	77.52
ICN 52	82.05

6. Conclusion

This research presents a new negative correlation ensemble transfer learning for fault diagnosis based on convolutional neural network (NCTE). The main contribution of this paper are as following: 1) On the structure aspect, the transfer learning is applied for fault diagnosis to build a deeper structure than traditional DL method for fault diagnosis; 2) On the training method aspect, the transfer learning is trained using negative correlation learning (NCL), and several softmax classifiers are added and trained cooperatively based on the transfer learning.3) The hyper-parameter of NCTE are determined by cross validation, and it could help to obtain a more reliable fault classifier. The proposed NCTE is conducted on the KAT Bearing Dataset, and the results show that NCTE has achieved good results compared with other machine learning and deep learning methods. However, the time consumption of NCTE increases sharply with the increase of the number of softmax classifiers. So it is better to keep the number of the classifiers in a proper size.

The limitations of the proposed method may include as followings: Firstly, the time consumption of NCTE increases sharply with the increase of the number of softmax classifiers. Secondly, the imbalance of the fault data and normal data in fault diagnosis is ignored in this research. Based on these limitations, the future researches can be done in the following ways. Firstly, an improve version of NCTE can be investigated to reduce the time consumption. Secondly, the

imbalance data handle techniques can be combined with NCTE.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grants 51805192, National Natural Science Foundation for Distinguished Young Scholars of China under Grant No.51825502, China Postdoctoral Science Foundation under Grant 2017M622414, Guangdong Science and Technology Planning Program under Grant 2015A020214003 and Supported by Program for HUST Academic Frontier Youth Team under Grant 2017QYTD04.

Conflict of interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

1. L. Wen, Y. Dong and L. Gao, A new ensemble residual convolutional neural network for remaining useful life estimation, *Math. Biosci. Eng.*, **16** (2019), 862–880.
2. H. D. Shao, H. K. Jiang, X. Zhang, et al., Rolling bearing fault diagnosis using an optimization deep belief network, *Meas. Sci. Tech.*, **26** (2015), 115002.
3. R. Zhao, R. Yan, Z. Chen, et al., Deep learning and its applications to machine health monitoring: A survey. *arXiv preprint arXiv:1612.07640*, 2016.
4. M. Cerrada, R. V. Sánchez, C. Li, et al., A review on data-driven fault severity assessment in rolling bearings, *Mech. Syst. Signal Proc.*, **99** (2018), 169–196.
5. Y. Bengio, A. Courville and P. Vincent, Representation learning: a review and new perspectives, *IEEE T. Pattern Anal. Mach. Intell.*, **35** (2013), 1798–1828.
6. L. Wen, L. Gao and X. Y. Li, A new deep transfer learning based on sparse auto-encoder for fault diagnosis, *IEEE T. Syst. Man Cybern. Syst.*, **49** (2019), 136–144.
7. J. L. Wang, J. Zhang and X. X. Wang, A data driven cycle time prediction with feature selection in a semiconductor wafer fabrication system, *IEEE T. Semicond. Manuf.*, **31** (2018), 173–182.
8. S. Shao, S. McAleer, R. Yan, et al., Highly-accurate machine fault diagnosis using deep transfer learning, *IEEE T. Ind. Inform.*, **15** (2019), 2446–2455.
9. Z. Y. Wang, C. Lu and B. Zhou, Fault diagnosis for rotary machinery with selective ensemble neural networks, *Mech. Syst. Signal Proc.*, **113** (2018), 112–130.
10. D. H. Wolpert and W. G.Macready, No free lunch theorems for optimization, *IEEE T. Evol. Comput.*, **1** (1997), 67–82.
11. H. Y. Sang, Q. K. Pan, J. Q. Li, et al., Effective invasive weed optimization algorithms for distributed assembly permutation flowshop problem with total flowtime criterion, *Swarm Evol. Comput.*, **44** (2019), 64–73.
12. X. Y. Li, C. Lu, L. Gao, et al., An Effective Multi-Objective Algorithm for Energy Efficient Scheduling in a Real-Life Welding Shop, *IEEE T. Ind. Inform.*, **14** (2018), 5400–5409.
13. J. Yosinski, J. Clune, Y. Bengio, et al., How transferable are features in deep neural networks? In Advances in neural information processing systems, (2014), 3320–3328.

14. L. Wen, X. Y. Li and L. Gao, A New Two-level Hierarchical Diagnosis Network based on Convolutional Neural Network, *IEEE T. Instrum. Meas.*, (2019).
15. H. Y. Sang, Q. K. Pan, P. Y. Duan, et al., An effective discrete invasive weed optimization algorithm for lot-streaming flowshop scheduling problems. *J. Intell. Manuf.*, **29** (2018), 1337–1349.
16. X. Y. Li, L. Gao, Q. Pan, et al., An effective hybrid genetic algorithm and variable neighborhood search for integrated process planning and scheduling in a packaging machine workshop. *IEEE Trans. Syst. Man Cybern. Syst.*, (2018).
17. K. Tidriri, N. Chatti, S. Verron, et al., Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges, *Annu. Rev. Control*, **42** (2016), 63–81.
18. Z. Y. Yin and J. Hou, Recent advances on SVM based fault diagnosis and process monitoring in complicated industrial processes, *Neurocomputing*, **174** (2016), 643–650.
19. J. Zheng, L. Gao, H. B. Qiu, et al., Difference mapping method using least square support vector regression for variable-fidelity approximation modelling, *Eng. Optimiz.*, **47** (2015), 719–736.
20. T. Han, D. Jiang, Q. Zhao, et al., Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery, *Trans. Inst. Meas. Control*, (2017), 1–13.
21. B. Cai, L. Huang and M. Xie, Bayesian networks in fault diagnosis, *IEEE T. Ind. Inform.*, **13** (2017), 2227–2240.
22. L. Wen, L. Gao and X. Y. Li, A new snapshot ensemble convolutional neural network for fault diagnosis, *IEEE Access*, **7** (2019), 32037–32047.
23. F. Wang, H. K. Jiang, H. D. Shao, et al., An adaptive deep convolutional neural network for rolling bearing fault diagnosis, *Meas. Sci. Technol.*, **28** (2017), 9.
24. J. L. Wang, J. Zhang and X. X. Wang, Bilateral LSTM: A two-dimensional long short-term memory model with multiply memory units for short-term cycle time forecasting in re-entrant manufacturing systems. *IEEE T. Ind. Inform.*, **14** (2018), 748–758.
25. J. Pan, Y. Zi, J. Chen, et al., LiftingNet: A novel deep learning network with layerwise feature learning from noisy mechanical data for fault classification, *IEEE T. Ind. Inform.*, **65** (2018), 4973–4982.
26. S. B. Li, G. K. Liu, X. H. Tang, et al., An ensemble deep convolutional neural network model with improved ds evidence fusion for bearing fault diagnosis, *Sensors*, **17** (2017), 1729.
27. C. Lu, Z. Y. Wang and B. Zhou, Intelligent fault diagnosis of rolling bearing using hierarchical convolutional network based health state classification, *Adv. Eng. Inform.*, **32** (2017), 139–151.
28. B. Zhang, W. Li, X. Li, et al., Intelligent fault diagnosis under varying working conditions based on domain adaptive convolutional neural networks, *IEEE Access*, **6** (2018), 66367–66384.
29. J. Donahue, Y. Q. Jia, O, Vinyals, et al., Decaf: A deep convolutional activation feature for generic visual recognition, International conference on machine learning, (2014), 647–655.
30. R. Ren, T. Hung and K. C. Tan, A generic deep-learning-based approach for automated surface inspection, *IEEE T. Cybern.*, **48** (2018), 929–940.
31. J. Wehrmann, G. S. Simoes and R. C. Barros, et al., Adult content detection in videos with convolutional and recurrent neural networks, *Neurocomputing*, **272** (2017), 432–438.

32. H. C. Shin, H. R. Roth, M. C. Gao, et al., Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, *IEEE T. Med. Imaging*, **35** (2016), 1285–1298.
33. E. Rezende, G. Ruppert, T. Carvalho, et al., Malicious software classification using transfer learning of ResNet-50 deep neural network, 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), (2017), 1011–1014.
34. O. Janssens, R. Walle, M. Loccufier, et al., Deep learning for infrared thermal image based machine health monitoring. *IEEE-ASME Trans. Mechatron.*, **23** (2018), 151–159.
35. Y. Zhou, W. C. Yi, L. Gao, et al., Adaptive differential evolution with sorting crossover rate for continuous optimization problems. *IEEE T. Cybern.*, **47** (2017), 2742–2753.
36. H. Y. Sang, P. Y. Duan and J. Q. Li, An effective invasive weed optimization algorithm for scheduling semiconductor final testing problem. *Swarm Evol. Comput.*, **38** (2018), 42–53.
37. L. K. Hansen and P. Salamon, Neural network ensembles, *IEEE T. Pattern Anal. Mach. Intell.*, **12** (1999), 993–1001.
38. H. H. Chen and X. Yao, Regularized negative correlation learning for neural network ensembles, *IEEE T. Neural Netw.*, **20** (2009), 1962–1979.
39. J. C. Fernández, M. Cruz-Ramírez and C. Hervás-Martínez, Sensitivity versus accuracy in ensemble models of artificial neural networks from multi-objective evolutionary algorithms. *Neural Comput. Appl.*, **30** (2018), 289–305.
40. C. Hu, B. D. Youn, P. Wang, et al., Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life, *Reliab. Eng. Syst. Saf.*, **1** (2012), 120–35.
41. Z. Wu, W. Lin and Y. Ji, An integrated ensemble learning model for imbalanced fault diagnostics and prognostics. *IEEE Access*, **6** (2018), 8394–8402.
42. U. P. Chong, Signal model-based fault detection and diagnosis for induction motors using features of vibration signal in two-dimension domain. *Strojniski Vestn. J. Mech. Eng.*, **57** (2011), 655–666.
43. L. Wen, X. Y. Li, L. Gao, et al., A new convolutional neural network based data-driven fault diagnosis method, *IEEE Trans. Ind. Electron.*, **65** (2018), 5990–5998.
44. K. M. He, X. Y. Zhang, S. Q. Ren, et al., Deep residual learning for image recognition, IEEE Conference on Computer Vision and Pattern Recognition, (2016), 770–778.
45. K. M. He, X. Y. Zhang, S. Q. Ren, et al., Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. IEEE International Conference on Computer Vision, (2015), 1026–1034.
46. M. Xiao, L. Wen, X. Li, et al., Modeling of the feed-motor transient current in end milling by using varying-coefficient model. *Math. Probl. Eng.*, (2015).
47. S. Arlot and A. Celisse, A survey of cross-validation procedures for model selection, Statistics surveys, **4** (2010), 40–79.
48. I. H. Witten, E. Frank, M. A. Hall, et al., Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, (2016).
49. C. Lessmeier, J. K. Kimotho, D. Zimmer, et al., Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. European Conference of the Prognostics and Health Management Society, 05-08, (2016).

50. T. Han, D. Jiang, Q. Zhao, et al., Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. *Trans. Inst. Meas. Control*, (2017), 1–13.
51. Y. H. Chen, G. L. Peng, C. H. Xie, et al., ACDIN: Bridging the gap between artificial and real bearing damages for bearing fault diagnosis, *Neurocomputing*, **294** (2018), 61–71.
52. Z. Y. Zhu, G. L. Peng, Y. H. Chen, et al., A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis, *Neurocomputing*, **323** (2019), 62–75.

AIMS Press