Journal of Intelligent Manufacturing (2023) 34:2449–2462
https://doi.org/10.1007/s10845-022-01944-x

A novel data-driven method based on sample reliability assessment and improved CNN for machinery fault diagnosis with non-ideal data

Xin Zhang[1] · Haifeng Wang[2] · Bo Wu[1] · Quan Zhou[1] · Youmin Hu[1]

Received: 19 October 2021 / Accepted: 24 March 2022 / Published online: 20 April 2022

Abstract

Recently, numerous new data-driven methods have been proposed. But most of them focused on the innovation of models and algorithms, and rarely discussed and optimized from the perspective of data and samples. However, the reliability of sample quality directly determines the effectiveness of machine learning models. In this paper, a novel data-driven method based on sample reliability assessment (SRA) and improved convolutional neural network (ICNN) for mechanical fault diagnosis was designed. First, multinomial logistic regression (MLR) was conducted to construct the assessment model and a statistical approach named influence function was used to compute the sample weights efficiently. Then, ICNN with the improved loss function was proposed based on the strategies of sample weights, class weights and early-stopping. Compared with traditional deep learning models, ICNN can better eliminate the negative impact of the problems during the model training including sample quality imbalance, class imbalance, and overfitting phenomenon. Therefore, the fault diagnosis performance can be improved. Finally, the trained ICNN can automatically extract the fault characteristics and achieve the fault diagnosis with the input of compressed time–frequency images. Experiments on a benchmarking dataset and a gear dataset from a practical experimental platform verified the superiority of the proposed fault diagnosis method.

Keywords Data-driven · Sample quality · Fault diagnosis · Sample weight · Convolutional neural network

Introduction

Rolling bearings and gears are key components in industrial production. Since they are often in a harsh working environment, damage and faults are easy to occur and threaten the safety of equipment and personnel (Lei et al., 2020; Zhao et al., 2019). Benefit from the development of sensor technology and data analysis theory, data-driven fault diagnosis methods for mechanical equipment can effectively recognize the health status of the equipment, which attracted a lot of attention in recent years (Chen et al., 2019; Liu et al., 2018; Mao et al., 2021).

Traditional data-driven models such as support vector machine (Cui et al., 2021; Yin & Hou, 2016) and principal component analysis (Cao et al., 2021) keep a simple architecture and low requirements of sample amount. But their shallow structures also result in the poor extraction effect of sophisticated fault characteristics. As an alternative, deep learning methods have attracted many researchers' attention due to the superior ability of feature extraction and classification (Hoang & Kang, 2019; Hu et al., 2022; Zhou et al., 2021). Several deep learning methods such as deep autoencoder (Zhang et al., 2020a, 2020b), deep belief network (Zhang et al., 2020a, 2020b), and convolutional neural network (CNN) (Hu et al., 2021; Jing et al., 2017) have been widely used for mechanical fault diagnosis. Among them, CNN is a classical deep learning method, which has been studied and accomplished many applications in the field of fault diagnosis. Benefitting from its multi-layer convolution and pooling operation, CNN can better analyze the obscure fault information. Based on the difference in network structure and form, CNN can be divided into two types, i.e. one-dimensional CNN (1DCNN) (Kiranyaz et al., 2021) and two-dimensional CNN (2DCNN) (Wen et al., 2018). The main difference between 1 and 2DCNN is that the input of the former is one-dimensional series, while the latter is a

✉ Youmin Hu
youmhwh@163.com

[1] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[2] The Company of Little Swan Electric Appliance Co., Ltd., Wuxi 214000, China

two-dimensional image. Also, the scale of the convolution kernel of 1DCNN is one-dimensional, while 2DCNN adopts a two-dimensional convolution kernel. Previous research shows that 2DCNN can extract more useful information than 1DCNN because 2DCNN can be designed to analyze the constructed time–frequency or grey images that contain more intuitive fault characteristics (Zhou et al., 2020).

Along with the development of deep learning methods, some challenges also emerged. On the one hand, data is the core and driving force of data-driven methods, but the obtained data cannot be ideal in many cases. On the other hand, the effectiveness of deep learning models depends heavily on how well the models were trained. Thus, insufficient or unreliable data has a negative impact on the accuracy and generalization ability of deep learning models. Generally, there are three common situations about the type of non-ideal data in fault diagnosis: (1) The imbalance of the training set or the problem of small samples can easily lead to the model's under-learning of the faulty samples. To solve these types of problems, transfer learning (Zhao et al., 2020), data augmentation (Li et al., 2018) and the small sample classifier (Kumar et al., 2021) with a specially designed structure have been presented. (2) In the actual fault diagnosis scenario, external interferences can easily cause the non-ideal data distribution, such as low signal-to-noise ratio (SNR), partial data points missing. And the quality of the obtained signals is uneven. Some signal processing methods such as wavelet transform (Wang et al., 2018) and variational modal decomposition (Li et al., 2019a, 2019b) can be used for denoising to improve the availability of the noisy samples to a certain extent. (3) Human data collation errors or subjective misjudgments may lead to incorrect training labels, which can largely reduce the effectiveness of model training. Aiming at this problem, a solution to implement fault diagnosis under noisy labels is proposed in (Zhang et al., 2021a, 2021b). Moreover, some research of data cleaning can process known types of outliers effectively, but a uniform standard has not been established to quantitatively describe the reliability of samples (Wang et al., 2020; Xu et al., 2020).

Although the latter two types of non-ideal data mentioned above can be handled separately in different ways, the solution generally limits to a known type of non-ideal data. If a general paradigm about evaluating the non-ideal data can be established, non-ideal data of different types can be measured by the same metric, i.e., the sample reliability value. Predictably, the processing procedure of the non-ideal data will be simplified based on a uniform standard. Although (Pang Wei Koh, 2017) did not directly study the above issues, it still provides a direction to inspire other research. It converts the reliability of the training samples into the influence on the recognition effects of the test samples. To simplify the solution, influence function (IF) (Cook & Weisberg, 1980; Debruyne et al., 2008) was also adopted as an estimation tool to avoid using the "leave one out" retraining strategy (after removing a sample from the training set, retrain and verify the model each time), which greatly reduces the computational complexity. As a result, the IF values can be used as an approximate measure of training sample reliability. However, how to effectively use the sample reliability values to optimize the model training process still needs further exploration, especially in the background of imbalanced data.

In this paper, a new data-driven method based on sample reliability assessment (SRA) and improved CNN (ICNN) is proposed to solve the fault diagnosis problem with non-ideal data. The proposed method is designed to adapt the three situations of non-ideal data mentioned above. To the best of our knowledge, it is the first attempt to optimize the model performance for mechanical fault diagnosis from the perspective of sample reliability. The main contributions of this article are as follows: (1) The traditional training process considers rarely the sample quality imbalance, so the model is more likely to learn the wrong information of non-ideal samples. The proposed sample reliability assessment model based on multinomial logistic regression (MLR) (Cannarile et al., 2019) can achieve a general sample evaluation process and optimize the learning process of the fault diagnosis model by the obtained sample weights. (2) The sample weight can be computed fast and approximately by the influence function, which is an innovation in the field of fault diagnosis. Thus, the computational overhead of the SRA can be reduced. (3) In the case of sample quality imbalance and class imbalance, the improved loss function combining sample weights and class weights can improve the fault identification performance of ICNN. At the same time, the early-stopping is used to avoid the model overfitting.

The remainder of this paper is organized as follows. Section "The proposed new data-driven fault diagnosis method" introduces the main process and the implementation details of the proposed approach. Experimental results are given and discussed in sections "Experimental validation" and "Discussion". Finally, the conclusions are drawn in Section "Conclusions".

The proposed new data-driven fault diagnosis method

In this study, a new data-driven fault diagnosis method based on SRA and improved CNN is developed. The general procedure of the proposed method is displayed in Fig. 1, which can be divided into four parts. (1) Sample acquisition: Samples are constructed at a certain sampling length from the raw signals and divided into a training set, a validation set and a test set; (2) Sample reliability assessment: In order to achieve the reliability assessment of the samples, the MLR algorithm is used to construct the assessment model with

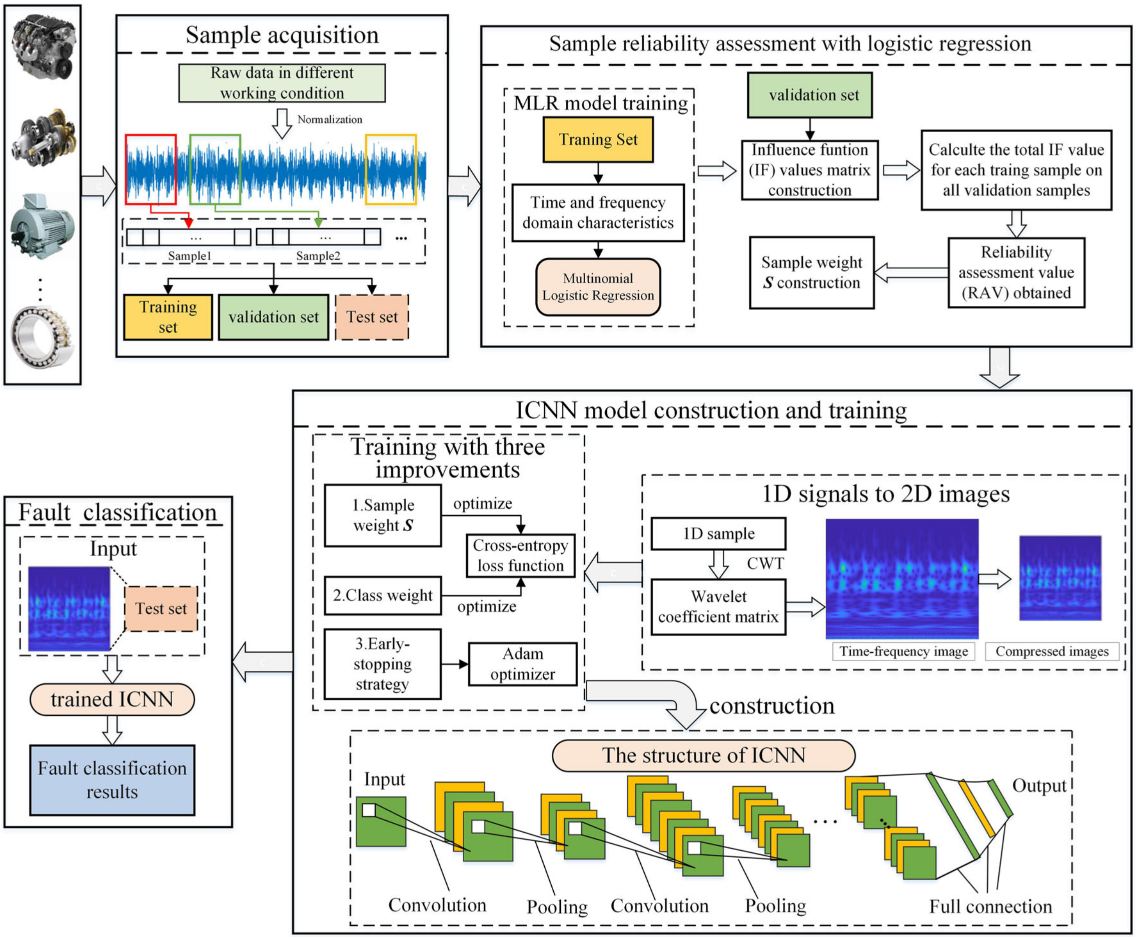
Sample acquisition
Raw data in different working condition
Normalization
Sample1
Sample2
Training set
validation set
Test set
Sample reliability assessment with logistic regression
MLR model training
Traning Set
Time and frequency domain characteristics
Multinomial Logistic Regression
validation set
Influence funtion (IF) values matrix construction
Calculte the total IF value for each traing sample on all validation samples
Reliability assessment value (RAV) obtained
Sample weight S construction
ICNN model construction and training
Training with three improvements
1.Sample weight S
optimize
Cross-entropy loss function
2.Class weight
optimize
3.Early-stopping strategy
Adam optimizer
1D signals to 2D images
1D sample
CWT
Wavelet coefficient matrix
Time-frequency image
Compressed images
construction
The structure of ICNN
Input
Convolution
Pooling
Convolution
Pooling
Output
Full connection
Fault classification
Input
Test set
trained ICNN
Fault classification results

Fig. 1 Procedure of the proposed fault diagnosis method

analyzing the time–frequency input of the original samples. It can effectively analyze the impact of sample quality on validation accuracy with a simple structure. And the statistical method of influence function is introduced to further accelerate the process of sample assessment. As a result, the sample weight (SW) vector can be established based on the obtained RAVs; (3) Construction and training of ICNN: To better utilize the feature mining performance of the CNN, continuous wavelet transform (CWT) (Gou et al., 2020) is employed to transform the one-dimensional samples into time–frequency images that can also be further compressed. In the design of the loss function of ICNN, sample weight can help the model focus more on high-quality samples and ignore low-quality samples. Moreover, to solve the problem of model accuracy degradation under class imbalance, the strategy of class weight (CW) is also introduced into the loss function. And the overfitting phenomenon can also be solved to some extent with the help of early-stopping; (4) Fault classification: Compared with traditional CNN, the trained ICNN model can better mine fault characteristics under the background of non-ideal data and improve the identification reliability for mechanical fault diagnosis.

To facilitate subsequent introduction and discussion, some notations are defined firstly. Let x_i^{tr} $(i = 1, 2, \ldots, m)$ denote training samples, x_j^{va} $(j = 1, 2, \ldots, n)$ indicate validation samples. And $h_\theta(x)$ expresses the assessment model with parameters θ, $L(h_\theta(x))$ is the model loss. $LD(x_i^{tr}, x_j^{va})$ represents the loss difference of the assessment model on x_j^{va} after and before the deletion of x_i^{tr}, as same as Eq. (1).

$$LD(x_i^{tr}, x_j^{va}) = L(h_{\theta_{wo\_x_i^{tr}}}(x_j^{va})) - L(h_{\theta_{w\_x_i^{tr}}}(x_j^{va})), \qquad (1)$$

where $h_{\theta_{w\_x_i^{tr}}}(\cdot)$ indicates that the model uses x_i^{tr} in the training process, $h_{\theta_{w\_x_i^{tr}}}(\cdot)$ means that x_i^{tr} is not used.

Sample reliability assessment model construction

The role of the SRA model is to find those “unfavorable” training samples. By computing the loss difference $LD(x_i^{tr},$

x_j^{va}) of the assessment model, the positive or negative influence of x_i^{tr} on the recognition of x_j^{va} can be expressed to some extent. If the $LD(x_i^{tr}, x_j^{va})$ value is a positive value, it means that x_i^{tr} has a positive effect on the recognition of x_j^{va}. To evaluate the training samples more comprehensively and rigorously, the loss difference for the training sample x_i^{tr} on the validation set can be used as a sort of assessment norm, which is shown as Eq. (2)

$$LD(x_i^{tr}, x_{all}^{va}) = \sum_{j=1}^{n} LD(x_i^{tr}, x_j^{va}), \tag{2}$$

where x_{all}^{va} means all validation samples. Thus, $LD(x_i^{tr}, x_{all}^{va})$ expresses the influence of x_i^{tr} on the recognition of the validation set.

For the improvement of calculation efficiency, the structure of the assessment model should be simple and effective. Multinomial logistic regression method is a simple and effective machine learning method, which is suitable for constructing the SRA model. It is actually constructed from several binary logistic regression (BLR) models. The output of the BLR is a probability value, which is shown in Eq. (3),

$$h_\theta(x) = \frac{1}{1 + e^{-z}}, \tag{3}$$

$$z = \theta^T x, \tag{4}$$

where x is the input sample of the model, θ^T is the model parameters. L-BFGS method was adopted to train the model, and the maximum likelihood estimation was used to construct the cross-entropy loss function, as shown in Eq. (5),

$$L(h_\theta(x)) = \sum_{k=1}^{p} (y_k \log h(x_k) - (1 - y_k) \log(1 - h(x_k))), \tag{5}$$

where p is the number of input samples, y_k is the predicted category for the k th input sample x_k.

To avoid underfitting, the key features should be further extracted from the raw high dimensional signals. Generally, the time domain and frequency domain features of raw signals can represent their main information effectively and directly. And 10 time-domain and 13 frequency-domain statistical characteristics are widely used for feature extraction in the field of fault diagnosis(Qu et al., 2016; Xiang et al., 2020). In this part, these 23 characteristics are also used as the model input. The detailed calculation equations of these time–frequency statistical features are listed in Table 1.

Approximate estimation by influence function

To reduce the time consumption of model retraining, a statistic tool named influence function is adopted to represent the influence of training samples on model validation. In this process, the assessment model does not to be retrained thus the computational complexity is greatly reduced.

According to the formula derivation in (Pang Wei Koh, 2017), the influence value $IF(x_i^{tr}, x_j^{va})$ can be used to approximate the loss difference $LD(x_i^{tr}, x_j^{va})$. And the definition of the influence function can be inferred as:

$$IF(x_i^{tr}, x_j^{va}) = -s_j \nabla_\theta L(h_\theta(x_i^{tr})), \tag{6}$$

$$s_j = \nabla_\theta L(h_\theta(x_j^{va}))^T H_\theta^{-1}, \tag{7}$$

$$H_\theta^{-1} = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta^2 L(h_\theta(x_i^{tr})), \tag{8}$$

where H_θ^{-1} is the inverse of the Hessian matrix, s_j is difficult to calculate directly. Thus, Hessian-vector products are used to approximate to it. More details can be found in (Pang Wei Koh, 2017). Then, the influence function values matrix (IFM) for all training samples can easily be computed and established as below,

$$IFM = \begin{vmatrix} IF(x_1^{tr}, x_1^{va}), & \cdots, & IF(x_1^{tr}, x_n^{va}) \\ IF(x_2^{tr}, x_1^{va}), & \cdots, & IF(x_2^{tr}, x_n^{va}) \\ \vdots\,\vdots\,\vdots & & \\ IF(x_m^{tr}, x_1^{va}), & \cdots, & IF(x_m^{tr}, x_n^{va}) \end{vmatrix}, \tag{9}$$

and the total influence function (TIF) value for each training sample on the verification set can be obtained by summing the values from the corresponding row. Understandably, the $TIF(x_i^{tr})$ value is an approximate solution to $LD(x_i^{tr}, x_{all}^{va})$ in Eq. (2), which can reduce much computation burden. To facilitate the subsequent discussion, a min–max normalization operation is adopted for the $TIF(x_i^{tr})$ vector, ($i = 1, 2, \ldots, m$), thus the $RAV(x_i^{tr})$ can be obtained. As a result, the RAV with a range [0, 1] for each training sample can be obtained and adopted as the final reliability assessment norm. And the above can be summarized as:

$$RAV(x_i^{tr}) = \frac{TIF(x_i^{tr}) - \min(TIF)}{\max(TIF) - \min(TIF)}, \tag{10}$$

$$TIF(x_i^{tr}) = \sum_{j=1}^{n} IF(x_i^{tr}, x_j^{va}) \approx LD(x_i^{tr}, x_{all}^{va}), \tag{11}$$

where min(.) and max(.) are respectively the minimum and maximum values of the *TIF* vector.

Table 1 Twenty-three time–frequency feature parameters

No	Feature expression	No	Feature expression	No	Feature expression	No	Feature expression
1	$I_1 = \frac{\sum\limits_{n=1}^{N} x(n)}{N}$	7	$I_7 = \frac{I_4}{I_3}$	13	$I_{13} = \frac{\sum\limits_{k=1}^{K} (p(k)-I_{11})^3}{K(\sqrt{I_{12}})^3}$	19	$I_{19} = \frac{\sum\limits_{k=1}^{K} f_k^2 p(k)}{\sqrt{\sum\limits_{k=1}^{K} p(k) \sum\limits_{k=1}^{K} f_k^4 p(k)}}$
2	$I_2 = \sqrt{\frac{\sum\limits_{n=1}^{N} (x(n)-I_1)^2}{N-1}}$	8	$I_8 = \frac{I_4}{\left(\frac{1}{N} \sum\limits_{n=1}^{N} \sqrt{\|x(n)\|}\right)^2}$	14	$I_{14} = \frac{\sum\limits_{k=1}^{K} (p(k)-I_{11})^4}{K(I_{12})^2}$	20	$I_{20} = \frac{I_{16}}{I_{15}}$
3	$I_3 = \sqrt{\frac{1}{N} \sum\limits_{n=1}^{N} (x(n))^2}$	9	$I_9 = \frac{I_3}{\frac{1}{N} \sum\limits_{n=1}^{N} \|x(n)\|}$	15	$I_{15} = \frac{\sum\limits_{k=1}^{K} f_k p(k)}{\sum\limits_{k=1}^{K} p(k)}$	21	$I_{21} = \frac{\sum\limits_{k=1}^{K} (f_k-I_{15})^3 p(k)}{K I_{16}^3}$
4	$I_4 = \max\|x(x)\|$	10	$I_{10} = \frac{I_4}{\frac{1}{N} \sum\limits_{n=1}^{N} \|x(n)\|}$	16	$I_{16} = \sqrt{\frac{\sum\limits_{k=1}^{K} (f_k-I_{15})^2 p(k)}{K}}$	22	$I_{22} = \frac{\sum\limits_{k=1}^{K} (f_k-I_{15})^4 p(k)}{K I_{16}^4}$
5	$I_5 = \frac{\sum\limits_{n=1}^{N} (x(n)-I_1)^3}{(N-1)I_2^3}$	11	$I_{11} = \frac{\sum\limits_{k=1}^{K} p(k)}{K}$	17	$I_{17} = \sqrt{\frac{\sum\limits_{k=1}^{K} f_k^2 p(k)}{\sum\limits_{k=1}^{K} p(k)}}$	23	$I_{23} = \frac{\sum\limits_{k=1}^{K} \sqrt{\|f_k-I_{15}\|} p(k)}{\sqrt{I_{16}} K}$
6	$I_6 = \frac{\sum\limits_{n=1}^{N} (x(n)-I_1)^4}{(N-1)I_2^4}$	12	$I_{12} = \frac{\sum\limits_{k=1}^{K} (p(k)-I_{11})^2}{K-1}$	18	$I_{18} = \sqrt{\frac{\sum\limits_{k=1}^{K} f_k^4 p(k)}{\sum\limits_{k=1}^{K} f_k^2 p(k)}}$		

$x(n)$ is a signal series for $n = 1, 2,\ldots, N$, where N is the number of sampling data points. $p(k)$ is the frequency spectrum for $k = 1,2,\ldots,K$, where K is the length of spectrum, and f_k means the frequency value of the k th spectrum line

Sample weight vector construction

In the above procedure, the *TIF* values obtained from our assessment can reflect the negative or positive influence of the sample on model training. A direct idea is to remove all training samples with *TIF* < 0, but it is unrealistic in the background of the small sample fault diagnosis. In this paper, a sample weight vector is constructed based on the RAVs to reflect different quality degrees of the training samples. Therefore, the ideal training samples are given bigger weights than the non-ideal samples, so that more useful information can be learned during the model training process. Meanwhile, the influence of non-ideal samples on the training process is weakened or even removed.

Since a few non-ideal samples have extremely negative effects, the best way is to remove them from the training set. In practical operations, it can be replaced by setting the sample weight to 0 for the non-ideal samples with quite low RAV. Although the RAVs of training samples may be not subject to a normal distribution, we can still use Pauta Criterion as a reference standard to determine whether the training samples should be removed. After all, our purpose is to remove only a few of the most extreme non-ideal training samples (which have a large negative impact on model performance), while removing all non-ideal samples is impossible. In practice, the removing strategy can be determined by the model performance on the validation set, but it is time-consuming and laborious. Therefore, the Pauta Criterion is adopted as the reference standard in this paper, although it may not be the optimal choice. As a result, the weights of samples whose RAVs are lower than $\mu - 3\sigma$ should be set to 0, where μ and σ are the mean and standard deviation for all training samples' RAVs, respectively. In this way, the sample weight vector S based on RAVs can be obtained and used for the training process of the improved CNN.

Image construction with CWT

Next, the one-dimensional signals need to be converted to images before the DL process. Figure 1 gives the details of the converted way. First, the technology of CWT was utilized to decompose the 1D sample into a 2D wavelet coefficient matrix. The details of the formula of CWT can be found in (Guo et al., 2018). Because the complex Morlet wavelet can match the actual fault responses well (Gu et al., 2017). In the proposed method, the complex Morlet wavelet is determined as the wavelet basis function. And the scale parameter is set to equal to the length of each sample (1024). Then, the time–frequency images can be generated with the size of 1024 × 1024. Last, the operation of bilinear interpolation

(Kim et al., 2019) is implemented to compress the time–frequency images, thus the compressed images with the size of 100 × 100 can be obtained, as the input of ICNN.

Fault identification with improved CNN model

As a classical deep learning method, CNN has a flexible construction way to adapt to different sample conditions. Moreover, 2DCNN has superior identification performance on the images, which can more fully learn the fault information contained in time–frequency images. Therefore, 2DCNN is used as the basic model for fault diagnosis in this paper. It is constructed by an input layer, two convolutional and pooling layers, a fully connected layer and a classifier layer. More details are described in experiments. The convolutional layer employs a series of filters to extract the fault information. Every filter is convolved with its input, and nonlinear mapping can be implemented by calculating with an activation function *f,* as described as follows:

$$(I * K)_i = f(\sum_{j=0}^{d_K-1} K_j I_{i+j} + b), \tag{12}$$

where I is the input of the convolution layer, K denotes several trainable filters with length d_K, * means the convolutional computation, b expresses bias, and f is a nonlinear activation function.

Max-pooling can compress the feature maps and improve the robustness of the model. And after the last pooling layer, a fully connected layer is established to further dig and reflect the previous flattened features. At last, a soft-max regression model is often placed at the end of 2DCNN as a classifier.

Before the fault diagnosis for unknown samples, the 2DCNN model needs to be fully trained. Aiming at the background of small sample and non-ideal data, three improvements and optimizations are implemented for the training procedure in the proposed method. First, the sample weight vector S based on RAVs is established to optimize the training process of the 2DCNN model. Specifically, the model is designed to learn more information from high-weight samples compared with the low-weight samples. Second, the strategy of class weight is introduced to the training process of 2DCNN to adapt to the imbalance between normal and fault samples. The above two optimizations are achieved in the design of the loss function, and the improved cross-entropy loss function can be described as:

$$Loss(x_i) = -S(x_i)C(x_i)\left[y_i \log \widehat{y}_i + (1 - y_i)\log(1 - \log \widehat{y}_i)\right], \tag{13}$$

where y_i, $\widehat{y}_i$ are desired and actual output of the input x_i, respectively. The vectors of S and C express the sample weight and class weight of x_i, respectively. And both S and C are normalized to [0,1]. Besides, the values of class weight C are determined with the ratio of the number of samples in different machine states.

Third, the early-stopping strategy is used to avoid the overfitting problems that are common in the diagnosis of small sample. For instance, the training process would be stopped when the validation loss does not decrease over the previous 500 epochs. And the best model parameters that can keep the smallest validation loss will be restored from previous training records. Based on the above three improvements, the improved CNN can better identify the mechanical fault type under the non-ideal data.

Experimental validation

Experimental setup

To validate the effectiveness of the proposed method, a public bearing fault dataset from Case Western Reserve University (CWRU) (Smith & Randall, 2015) was used. The dataset was acquired from a rolling bearing rig, which was composed of a 1491.4 W three-phase motor, a loading motor, and a torque sensor. The signals were acquired in the following working conditions: The rotation speeds were 1730, 1750, 1772, and 1797 rpm, and the corresponding motor loads were 3, 2, 1, 0 Hp, respectively. The sampling frequency of the acceleration sensor was 12 kHz. With the above conditions, the signals under four different bearing states were acquired, including normal (N), inner race fault (IRF), outer race fault (ORF), and rolling element fault (REF). For each fault type, it has four fault sizes (7, 14, 21, and 28 mils). More details can be found in (Smith & Randall, 2015).

In this experiment, ten bearing states with different fault types and fault diameters were constructed and labeled, which are N, IRF7 (inner race fault with fault size 7 mils), IRF14, IRF21, ORF7, ORF14, ORF21, REF7, REF14, REF21. For each bearing state, the signals from four working conditions were all adopted to conduct the subsequent experiment. In this paper, all experimental codes were written by Python 3.6 with TensorFlow 2.0. A computer with an Intel®CoreTM i7-8550U processor and 16 GB of RAM was used.

Non-ideal sample sets construction

Since the bearing data of CWRU is recognized as a relatively desirable dataset in the industry, the non-ideal sample sets need to be further constructed to validate the proposed method. In this experiment, three different situations of non-ideal data were considered, i.e. unbalanced dataset (meanwhile keeping a small sample set), signal noise and noisy labels. And the general performance of the proposed method was validated with the eleven constructed sample sets (S0–S10), where S0 is the original sample set and S1-S10 are non-ideal. Specifically, the gaussian white noise with SNR = 0 dB was added into some samples of S1-S5, and some noisy labels were injected into S6-S10. Moreover, a ratio of the number of the normal samples and other fault samples was set to 10:1 in the training set, thus the situation of unbalanced data distribution can be simulated. Meanwhile, the validating set and test set were balanced for ten bearing states. And sequential 1024 data points were used as one sample.

The details of the constructed eleven sample sets are shown in Table 2. For the training set, validation set and test set in each sample set of S1-S5, the ratio of the number of total samples to the number of non-ideal samples is an equal ratio with the range of 0 to 25%. Only the training sets have non-ideal samples for S6-S10. In addition, S0-S10 were independently sampled from the sensing signal and constructed to ensure randomness.

Samples reliability assessment procedure

Before the operation of samples reliability assessment, some preparatory work including three phases needs to be finished. First, 10 time-domain and 13 frequency-domain features were calculated for each group in sample set S0-S10. Then, the MLR model was established and the input dimension, output dimension, batch size, learning rate were set to 23, 10, 64, 1×10^{-3}, respectively. Last, the MLR model was trained and validated by S0-S10 to test whether the MLR model is suitable as SRA model.

By calculating the IF values between the training samples and the validation samples, the IFM can be constructed, as shown as Fig. 2 (For ease of presentation, Fig. 2 only draws the points with its absolute value of IF > 0.05). Predictably, several points with high IF values may have much greater influence on the validation samples than the others. The IFM contains much useful information that can help the process of SRA. By summation computation, the TIF values and RAVs can be obtained, which can more efficiently express the sample reliabilities.

Taking the training set of S1 for instance, Fig. 3 gives the distribution of RAVs. The color of the circles in the figure represents the type of the samples, which was unknown during the experiment. And the size of the circles indicates the unreliable degree of the sample, i.e. the smaller the RAV, the larger the size of the circle. From the perspective of the overall distribution, the RAVs of non-ideal samples are lower than the original samples. From the view of local distribution, the RAVs of several non-ideal samples are extremely low while all RAVs of the original samples are higher than 0.5. Moreover, it can be seen that the RAV of the original samples ranged from 0.5 to 1.0, which shows that the SRA processing also has a certain assessment effect on the original sample. After all, the original dataset is also not absolutely ideal. As a result, the proposed indicator of RAV can be used as a valid assessment metric to reflect the influence of samples on model training.

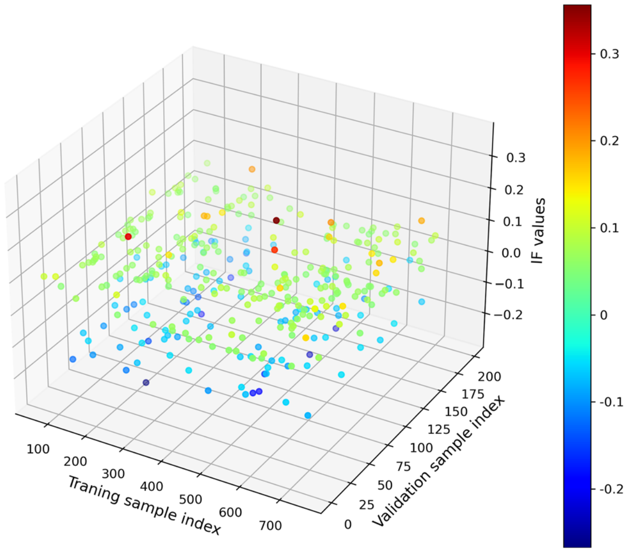

Fig. 2 The IFM information on S1

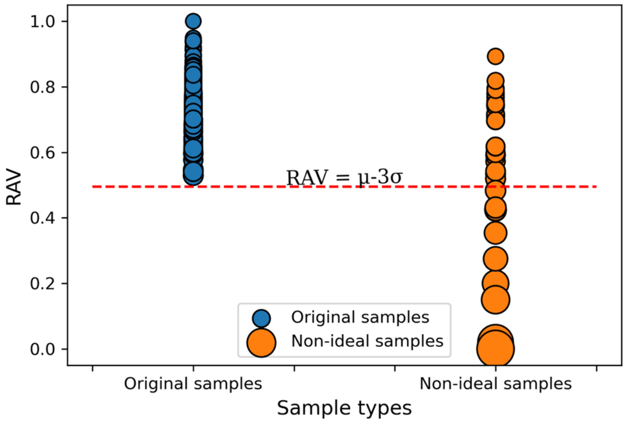

Fig. 3 The detailed assessment results on sample set S1

As discussed previously, the sample weights should be revised to 0 when the RAV value is lower than $\mu - 3\sigma$, and the boundary is also marked with a red dashed line in Fig. 3. Subsequently, the sample weights were established with the revised RAV vector.

ICNN model establishment and training

In the previous section, the SRA procedure has been accomplished, thus the sample weight vector was obtained. Then, the technology of CWT was implemented to transfer the 1D signals into 2D images with three channels, as the input of the

Table 2 The descriptions of eleven sample sets used in this case

Sample set	Training num		Validating num	Test num	Ratio of non-ideal sample (%)
	Normal	Fault			
S0	400	40 × 9	200	2000	0
S1-S5	400	40 × 9	200	2000	5,10,15,20,25
S6-S10	400	40 × 9	200	2000	5,10,15,20,25

The signals from four working conditions were all used averagely, e.g., the number of normal samples for S0 is 400, which means the number of samples from each working condition is 100

Table 3 The detailed structure of ICNN

Layer	Parameter name	Activation function	Output size
input	–	–	100 × 100 × 3
conv1	Convolutional kernel	Relu	(100,100,32)
pool1	Max pooling kernel	–	(20, 20, 32)
conv2	Convolutional kernel	Relu	(20, 20, 64)
pool2	Max pooling kernel	–	(4, 4, 64)
FC1	–	–	1024 × 1
FC2	Weight Matrix	Relu	256 × 1
DR	–	–	256 × 1
output	Weight Matrix	Softmax	10 × 1

Table 4 The parameter settings of the training process

Parameter description	Value
Loss function	Categorical_crossentropy loss
Batch size	64
Dropout rate	0.2
Max iterations	2000
Early stopping patience	500
Early stopping monitor	Val_loss
Sample weight	The obtained vector *S* from the SRA
Class weight	The ratio between normal and fault type is 1:10

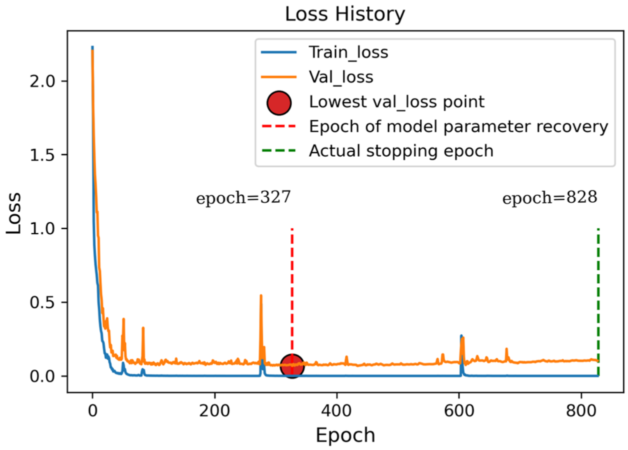

Fig. 4 The training and validating loss curves of ICNN

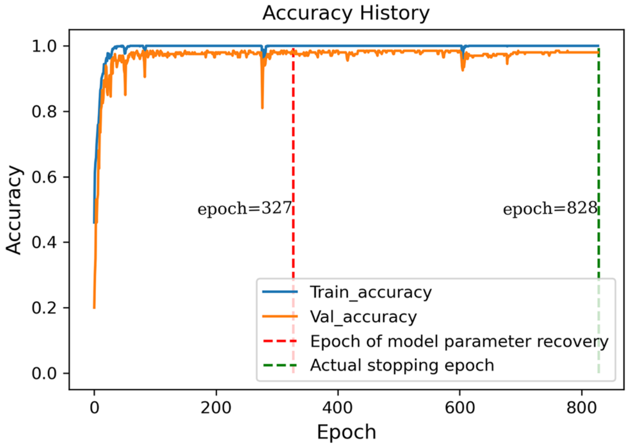

Fig. 5 The training and validating accuracy curves of ICNN

ICNN model. Before the fault identification, the ICNN model should be established firstly based on the hyperparameters in Table 3.

During the training process, some classical settings such as batch processing and dropout were used. All parameters about training settings are illustrated in Table 4. Three improved strategies were also carried out in the ICNN model. First, the obtained sample weight vector *S* from the SRA was used to adjust the loss function, which makes the ideal training samples can provide more impact on model fitting. Second, the class weight vector *C* was also injected into the loss function to optimize the learning effect of the fault samples with a small amount. Last, the strategy of early-stopping was utilized to avoid overfitting.

Figures 4 and 5 show an example training and validating process on sample set S1. Because the patience of early-stopping was set to 500, the actual stopping epoch was 501

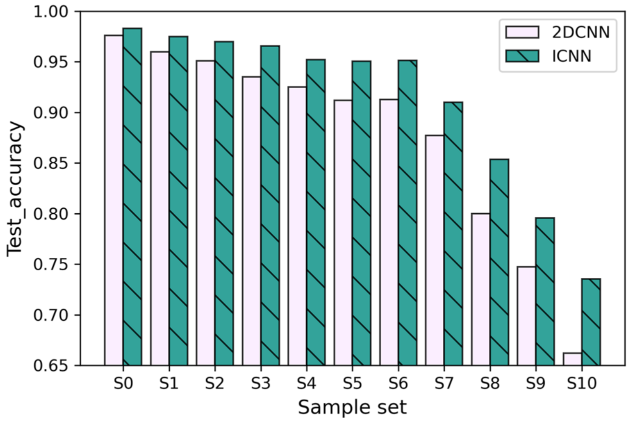

Fig. 6 Test accuracies of the ICNN models on S0-S10

generations later than the epoch of model parameters recovery. After the training process is stopped, the early-stopping strategy can restore the best model parameters based on the lowest validating loss overall process. It can be seen that the validating accuracy from the epoch of model parameters recovery is significantly higher than the validating accuracy from the actual stopping epoch. Therefore, with the help of early stopping, the ICNN model can guarantee the best generalization performance and avoid the influence of fluctuation in the training process.

Fault identification by ICNN

After the training process, the ICNN model can be used to identify the mechanical states. To further demonstrate the effectiveness of the proposed method, the test sets of S0–S10 were all used separately in this section. To illustrate the superiority of ICNN, a 2DCNN (without the three improved strategies) was used as a comparison method. On all eleven sample sets, the proposed method can achieve higher recognition accuracy, as shown in Fig. 6. Especially when the training set contains more non-ideal samples, such as S5 and S10, the improvement of test accuracy is more substantial for the proposed ICNN. Moreover, the ICNN model can also improve the accuracy facing the test set of S0 (original sample without artificially added non-ideal samples), although the rise was not significant enough. It indicates that the proposed ICNN can also improve the fault diagnosis performance in practical scenarios.

Discussion

Effectiveness of the SRA procedure

Two aspects were discussed separately for the effectiveness of the SRA procedure. One aspect is about the feasibility of the SRA model, and the other is the efficiency of the IF tool. In this paper, the SRA model should have good convergence and stability for the training process. Otherwise, the perturbations brought by the model training will bring great uncertainty to the determination of the RAV values, especially when facing non-ideal data. Thus, most deep learning methods are not suitable to be a choice of SRA models. Fortunately, classical machine learning methods generally have good convergence and stability, such as KNN, SVM. Table 5 gives the accuracies for KNN, SVM and the used MLR on the test set of S1. It can be inferred that the MLR method is a pretty good choice for the SRA model. Because it has better fault identification performance for samples compared with the other two methods. Meanwhile, the convergence of MLR is also acceptable.

Table 5 The comparison of different choices for the SRA model

Methods	Parameters	Accuracy (%)
KNN	$k = 11$	56.75
SVM	Rbf kernel (3 times)	41.30
MLR	learning rate $= 0.001$, batch size $= 64$	89.50

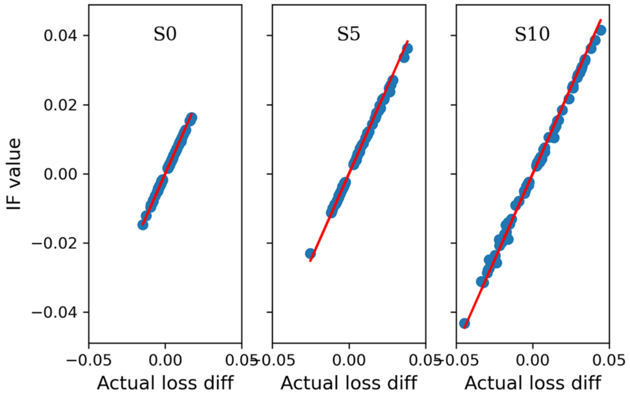

Fig. 7 The IF values match the actual loss differences

For the proposed method, the improvement effect depends on the reliability of the IF tool to some extent. Therefore, it is necessary to discuss the efficiency of the IF tool during the SRA procedure. Compared with using the actual loss differences of the “leave one out” retraining strategy, the IF tool can save a lot of computation time and maintain a high approximation accuracy. By conducting five tests, the mean consuming time of IF tool for constructing the IFM matrix was 175 s, but using the actual loss differences needed 2198 s. To illustrate the effectiveness of the IF tool, the influence of some training samples on a validation sample was calculated in two ways (IF tool and “leave one out” retraining), as shown in Fig. 7. And the used samples of three subfigures were from

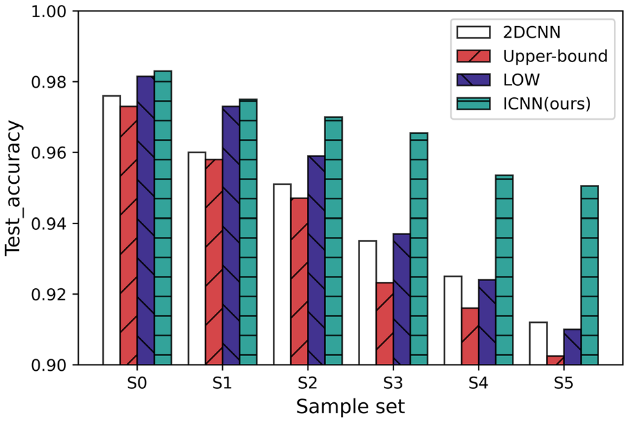

Fig. 8 The comparison with other sample weighting methods

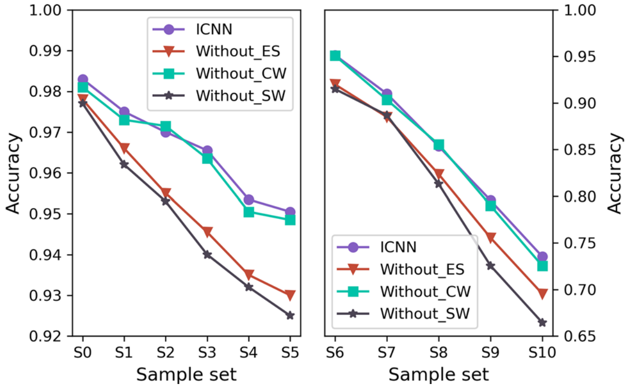

Fig. 9 The comparison for the impact of three improvements of ICNN on test accuracy

S0, S5 and S10. From Fig. 7, it shows that the IF values are very close to the actual loss differences so that the points in the figure are distributed along the line $y = x$. Therefore, the IF tool can save more than 90% time compared with “leave one out” retraining meanwhile keeping superior accuracy.

After implementing the SRA process, the sample weights of the entire training set can be calculated and used to optimize the subsequent model training. The above process has formal similarities with some deep learning studies using important sampling (Angelos & Fleuret, 2018; Johnson & Guestrin, 2018; Santiago et al., 2021), because they both achieve training optimization by setting the training sample weights. Differently, the SRA procedure determines the sample weights of training samples by the model performance on the validating set, while the sample weights of important sampling are determined based on samples’ own contribution to the gradient descent during the model fitting process. To investigate the superiority of the SRA process, two importance sampling-based sample weighting methods, namely upper-bound (Angelos & Fleuret, 2018) and LOW (Santiago et al., 2021), are used for comparison. Moreover, a baseline method (2DCNN) without sample weighting is also used. The average test accuracies were obtained by conducting five trials on S0-S5 independently, as shown in Fig. 8.

It can be observed from Fig. 8 that ICNN has the highest fault identification accuracies on all six sample sets. At the same time, as the unreliability of the sample set gradually increases, the proposed ICNN has a more pronounced improvement effect. This indicates that the proposed method can handle the non-ideal samples more effectively compared with others. For the method of LOW, its accuracies are significantly improved compared with 2DCNN on S0, S1 and S2. It means that LOW can help the model fitting and improve the test accuracies when the sample sets are relatively ideal. But the fault recognition effect of LOW is unsatisfied on S4 and S5, which contain more non-ideal samples. This may be because when the number of non-ideal samples increases, some non-ideal samples of one batch can also achieve rapid gradient descent. In this case, LOW would learn the wrong information from those non-ideal samples more easily, which can reduce the recognition performance of the model. In contrast, the proposed SRA determines the sample weights by the model loss on the validation set, which is equivalent to introducing additional supervisory information to set the sample weights that can achieve better generalization performance. Moreover, the method of upper-bound has poor performance in terms of recognition accuracy, because it sacrifices some accuracy while achieving fast model training.

Implementation effect of three improvements of ICNN

The respective effect for the three improved strategies, i.e. sample weight (SW), class weight (CW) and early-stopping (ES) were further explored. Each strategy is separately removed from the ICNN model to determine its influence on model performance. And the results are shown in Fig. 9. It can be concluded that the strategy of SW has the biggest contribution to the performance of ICCN because the accuracy is the lowest without it. It indicates that the SW strategy can optimize the training process of the 2DCNN network, although the sample weights were calculated from the raw one-dimensional samples. This is because the sample quality was not altered significantly before and after the input image was constructed. ES strategy also has a significant impact, and it is even equivalent to the SW strategy in S6 and S7. And the CW strategy can also improve the recognition performance of the model on most sample sets, although this improvement is not significant. The reason is that the RAVs included in the SW strategy already consider the influence of the samples on the model fitting to some extent and give higher weights to the faulty samples with scarce quantity. In summary, the three improvement strategies in the proposed ICNN method are all effective and necessary.

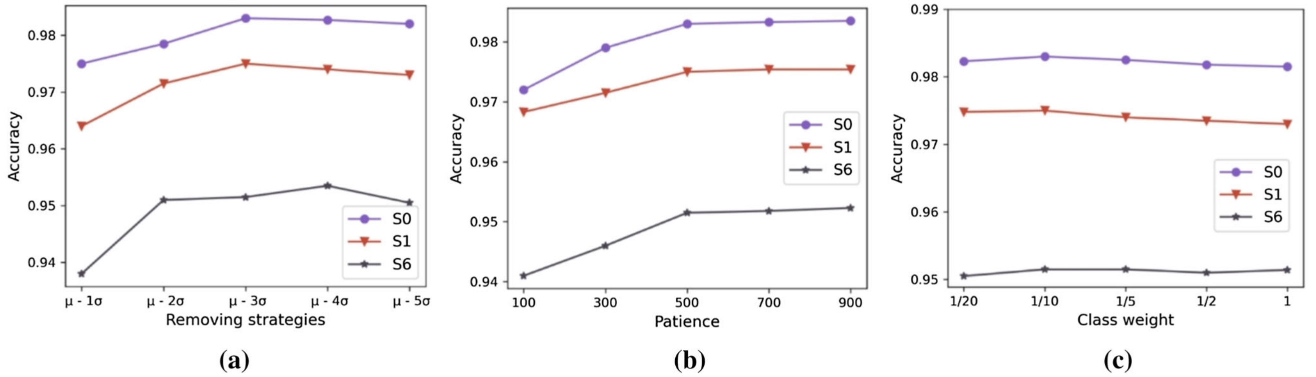

Fig. 10 The robustness validation of three kinds of model hyper-parameters on test accuracy. **a** The comparison of the removing strategies. **b** The comparison of the patience settings of early-stopping. **c** The comparison of the class-weighted settings

Robustness validation of model parameters

In the proposed method, some hyperparameters need to be selected artificially, and improper selection may influence the fault diagnosis effect of the model to a certain extent. Generally, the hyper-parameter settings can be determined by the model performance on the validating set. In this part, the robustness of the three most critical hyper-parameters that are removing strategy, patience and class weight are discussed. Because these three hyper-parameters correspond to the three improvements of the proposed model, i.e. sample weight, early-stopping and class weight, thus the robustness of the proposed three improvements can also be validated. Specifically, the three hyper-parameters were adjusted numerically several times and the average test accuracies were obtained by conducting ten trials on S0, S1, S6. The details are shown in Fig. 10.

From Fig. 10a, five different removing strategies were selected to validate the model performance. For example, the removing strategy of $\mu - 5\sigma$ means that the training samples with RAV $< \mu - 5\sigma$ would be removed. It can be seen that there is no r excessive change on the test accuracy when the removing strategy uses $\mu - 3\sigma$, $\mu - 4\sigma$ or $\mu - 5\sigma$. It indicates that the model performance can be improved after only removing a few extreme non-ideal samples. But the test accuracy decreases significantly when more training samples were removed (using $\mu - \sigma$ or $\mu - 2\sigma$). This is because when the removing strategy is set too strict, more useful samples can also be deleted by mistake. And the model fitting would become more difficult with a small sample set. Therefore, in addition to determining the removing strategy by the model performance on the verification set, it is also feasible to remove only a few extreme non-ideal samples. From Fig. 10b, it can be observed that the smaller values of patience would result in lower test accuracies, which is due to the model stopping training too early. Thus, the patience values are more inclined to be set larger if the training duration is not considered. In Fig. 10c, the class weight setting with 1:10 can achieve the highest test accuracy, which is also consistent with the unbalanced ratio of the training set.

Comparison with other state-of-the-art methods

Further, the proposed method is verified and compared with other state-of-the-art methods on the CWRU dataset, as shown in Table 6. Five trials were applied and the mean accuracy was calculated for each method. On the sample sets containing less non-ideal data such as S0 and S1, ResNet-APReLU and CWT + 2DCNN can achieve high accuracies comparable to our method. But with the increasing influence of abnormal data, the accuracy reduction speed of our method is significantly slower than other methods. Even on S10, the accuracy rate of our method can still be maintained at 73.50%. Because some comparison methods such as CNNEPDNN and ResNet-50 are not suitable for small sample classification, thus they cannot obtain good accuracy performance. And in the total time-consuming comparison on one sample set, the performance of the proposed method is also acceptable.

Verification on another practical case

To further validate the effectiveness of the proposed method, another practical case was conducted based on a gearbox experimental platform (Hu et al., 2019; Zhang et al., 2021a, 2021b). It comprised a one-stage reduction gearbox, a servo motor, a magnetic power brake, a torque sensor, and a brake controller, as shown in Fig. 11. The detailed parameters of driving and driven gears are listed in Table 7. To simulate different fault sizes, wire-electrode cutting technology was used to construct four kinds of gear crack conditions (non-crack, 1/4 crack, 1/2 crack, and 3/4 crack). The length of the gear crack can be computed by $L_c{=}i \times (R_c{-}r_h)/4$, i

Table 6 The comparison of different methods

Method	Test accuracy on different sample set (%)											Time (s)
	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
ResNet-APReLU (Zhao et al., 2021)	98.65	98.20	96.75	96.15	95.45	93.50	92.55	86.10	83.75	67.90	52.35	3708.60
SuperGraph (Yang et al., 2022)	81.50	79.10	78.50	76.50	73.0	76.10	76.50	80.0	76.0	75.0	72.10	95.80
CNNEPDNN (Li et al., 2019a, 2019b)	61.75	53.40	50.35	45.75	43.15	38.45	48.95	39.30	38.75	34.75	29.65	2245.80
M2DCNN (Gong et al., 2019)	94.75	95.25	90.75	89.15	88.25	81.45	91.55	87.45	75.95	79.10	68.25	120.50
ResNet-50 (Wen et al., 2019)	82.10	80.45	73.25	71.55	66.45	61.10	67.50	45.65	43.05	38.95	29.90	224.60
CWT + 2DCNN (Zhou et al., 2020)	98.35	98.05	94.15	95.25	94.65	91.90	92.85	81.25	79.65	67.05	61.45	1222.90
Ours	98.30	97.50	97.00	96.55	95.35	95.05	95.15	91.00	85.35	79.55	73.50	1327.50

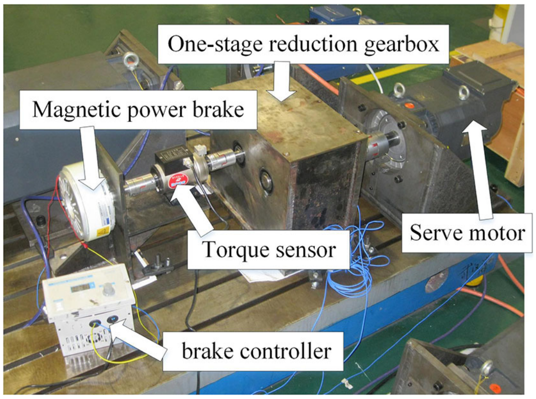

Fig. 11 Composition of the monitoring platform

Table 8 Parameters of experimental conditions

Types	Value
Crack length	0, 5, 10, and 15 mm
Input shaft speed	300, 600, 900, 1200, and 1500 r/min
Load	0 Nm, 4 Nm
Number of training samples	800
Number of validating samples	80
Number of test samples	800
Points of each sample	1024

Table 7 Parameters of experimental gears

Gear type	Teeth number	Gear module	Teeth width
Driving gear	50	2 mm	20 mm
Driven gear	80	2 mm	20 mm

= 0, 1, …, 3, where R_c and r_h are the radius of the root circle of the main driving wheel and the center hole, with values of 27.5 mm and 47.5 mm, respectively. The sampling frequency is 5 kHz and each sample contains 1024 sampling points. Five input shaft speed conditions of the driving gear and two kinds of loads were used in this experiment. Thus, the data was sampled under 10 operating working conditions. For each working condition, 168 samples can be collected and the sample amount under different crack sizes was equal. More details are listed in Table 8.

Without any artificial noise injected, the obtained samples are purely obtained from the actual testing environment. Thus, the robustness and generalization of the proposed method can be further validated. Because the length of the sample is set to 1024 as same as in the bearing experiment, the model can refer to the previous settings. Seven methods were adopted separately, including DBN, DAE, 1DCNN, ResNet-50 (Wen et al., 2019), LeNet-5, M2DCNN (Gong et al., 2019), the proposed ICNN. Among these methods, the inputs of the first four methods are 1D signals, and the inputs of the last three methods are 2D images generated by CWT. To improve the reliability of the fault diagnosis results, ten trials were applied for the comparison, as shown in Fig. 12. It can be seen that the test accuracy of the proposed ICNN is significantly higher than other methods. And the reason why the first four methods in Fig. 12 have quite low accuracies is that the time-domain input signals are difficult to be fully extracted and learned in the small sample case. On the contrary, the time–frequency images contain more direct fault characteristics, so the accuracy of the latter three methods is significantly better.

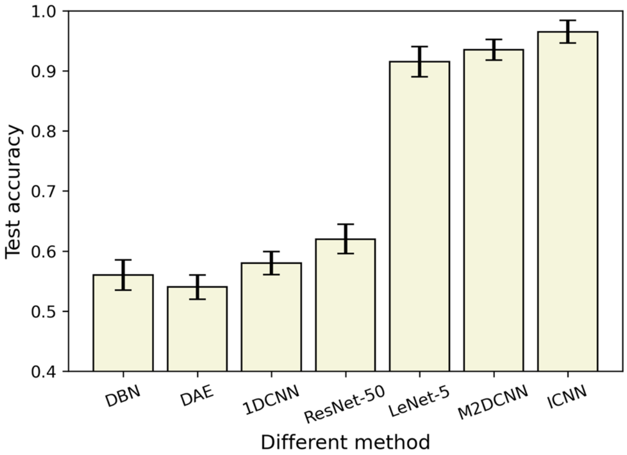

Fig.12 Comparison of different methods

Conclusions

In this paper, a novel data-driven method based on the sample reliability assessment process and improved ICNN is proposed, which can improve the fault diagnosis performance of models under the non-ideal data. First, the original training samples were evaluated based on the MLR assessment model. Meanwhile, the influence function was used to simplify the computational burden and the RAVs of all training samples can be obtained. Then, three strategies about sample weight based on RAVs, class weight, and early-stopping were utilized in the improvements of ICNN. Finally, the trained ICNN can automatically extract the characteristics and achieve the fault diagnosis with the input of compressed time–frequency images. Experiments showed that the proposed method can effectively optimize the model training process and thus improve the performance of fault identification. By comparing with other advanced methods on two datasets, the superiority and robustness of the proposed method have been discussed and verified.

However, this paper only considered the way of evaluating and weighting non-ideal samples to improve the performance of fault diagnosis models. Combining data denoising or data recovery to improve the reliability of model training is also a future research trend.

Acknowledgements This study was financially supported by the National Key R&D Program of China (No. 2017YFD0400405).

References

Angelos, K., & Fleuret, F. (2018, Feb). Not all samples are created equal: Deep learning with importance sampling. In International conference on machine learning. <Go to ISI>://WOS:000585302200012

Cannarile, F., Compare, M., Baraldi, P., Diodati, G., Quaranta, V., & Zio, E. (2019). Elastic net multinomial logistic regression for fault diagnostics of on-board aeronautical systems. *Aerospace Science and Technology*. https://doi.org/10.1016/j.ast.2019.105392

Cao, S., Hu, Z., Luo, X., & Wang, H. (2021). Research on fault diagnosis technology of centrifugal pump blade crack based on PCA and GMM. *Measurement*. https://doi.org/10.1016/j.measurement.2020.108558

Chen, H., Jiang, B., Chen, W., & Yi, H. (2019). Data-driven detection and diagnosis of incipient faults in electrical drives of high-speed trains. *IEEE Transactions on Industrial Electronics, 66*(6), 4716–4725. https://doi.org/10.1109/tie.2018.2863191

Cook, R. D., & Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics, 22*(4), 495–508. https://doi.org/10.1080/00401706.1980.10486199

Cui, M., Wang, Y., Lin, X., & Zhong, M. (2021). Fault diagnosis of rolling bearings based on an improved stack autoencoder and support vector machine. *IEEE Sensors Journal, 21*(4), 4927–4937. https://doi.org/10.1109/jsen.2020.3030910

Debruyne, M., Hubert, M., & Suykens, J. A. (2008). Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research, 9*, 2377–2400.

Gong, W., Chen, H., Zhang, Z., Zhang, M., Wang, R., Guan, C., & Wang, Q. (2019). A novel deep learning method for intelligent fault diagnosis of rotating machinery based on improved CNN-SVM and multichannel data fusion. *Sensors (basel)*. https://doi.org/10.3390/s19071693

Gou, L., Li, H., Zheng, H., Li, H., & Pei, X. (2020). Aeroengine control system sensor fault diagnosis based on CWT and CNN. *Mathematical Problems in Engineering, 2020*, 1–12. https://doi.org/10.1155/2020/5357146

Gu, X., Yang, S., Liu, Y., Deng, F., & Ren, B. (2017). Compound faults detection of the rolling element bearing based on the optimal complex Morlet wavelet filter. *Proceedings of the Institution of Mechanical Engineers, Part c: Journal of Mechanical Engineering Science, 232*(10), 1786–1801. https://doi.org/10.1177/0954406217710673

Guo, S., Yang, T., Gao, W., & Zhang, C. (2018). A novel fault diagnosis method for rotating machinery based on a convolutional neural network. *Sensors (basel)*. https://doi.org/10.3390/s18051429

Hoang, D.-T., & Kang, H.-J. (2019). A survey on deep learning based bearing fault diagnosis. *Neurocomputing, 335*, 327–335. https://doi.org/10.1016/j.neucom.2018.06.078

Hu, Z., Lv, C., Hang, P., Huang, C., & Xing, Y. (2022). Data-driven estimation of driver attention using calibration-free eye gaze and scene features. *IEEE Transactions on Industrial Electronics, 69*(2), 1800–1808. https://doi.org/10.1109/tie.2021.3057033

Hu, Z.-X., Wang, Y., Ge, M.-F., & Liu, J. (2019). Data-driven fault diagnosis method based on compressed sensing and improved multi-scale network. *IEEE Transactions on Industrial Electronics*. https://doi.org/10.1109/tie.2019.2912763

Hu, Z. X., Xing, Y., Lv, C., Hang, P., & Liu, J. (2021). Deep convolutional neural network-based Bernoulli heatmap for head pose estimation. *Neurocomputing, 436*, 198–209. https://doi.org/10.1016/j.neucom.2021.01.048

Jing, L., Zhao, M., Li, P., & Xu, X. (2017). A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox. *Measurement, 111*, 1–10. https://doi.org/10.1016/j.measurement.2017.07.017

Johnson, T. B., & Guestrin, C. (2018). Training deep models faster with robust, approximate importance sampling. *Advances in Neural Information Processing Systems 31 (Nips 2018), 31*, 7265–7275. <Go to ISI>://WOS:000461852001079

Kim, K.-H., Shim, P.-S., & Shin, S. (2019). An alternative bilinear interpolation method between spherical grids. *Atmosphere*. https://doi.org/10.3390/atmos10030123

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*. https://doi.org/10.1016/j.ymssp.2020.107398

Kumar, A., Vashishtha, G., Gandhi, C. P., Zhou, Y., Glowacz, A., & Xiang, J. (2021). Novel convolutional neural network (NCNN) for the diagnosis of bearing defects in rotary machinery. *IEEE Transactions on Instrumentation and Measurement, 70*, 1–10. https://doi.org/10.1109/tim.2021.3055802

Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing, 138*, 106587. https://doi.org/10.1016/j.ymssp.2019.106587

Li, F., Li, R., Tian, L., Chen, L., & Liu, J. (2019a). Data-driven time-frequency analysis method based on variational mode decomposition and its application to gear fault diagnosis in variable working conditions. *Mechanical Systems and Signal Processing, 116*, 462–479. https://doi.org/10.1016/j.ymssp.2018.06.055

Li, H., Huang, J., & Ji, S. (2019b). Bearing fault diagnosis with a feature fusion method based on an ensemble convolutional neural network and deep neural network. *Sensors (basel)*. https://doi.org/10.3390/s19092034

Li, X., Zhang, W., Ding, Q., & Sun, J.-Q. (2018). Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation. *Journal of Intelligent Manufacturing, 31*(2), 433–452. https://doi.org/10.1007/s10845-018-1456-1

Liu, R., Yang, B., Zio, E., & Chen, X. (2018). Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mechanical Systems and Signal Processing, 108*, 33–47. https://doi.org/10.1016/j.ymssp.2018.02.016

Mao, W., Feng, W., Liu, Y., Zhang, D., & Liang, X. (2021). A new deep auto-encoder method with fusing discriminant information for bearing fault diagnosis. *Mechanical Systems and Signal Processing, 150*, 107233. https://doi.org/10.1016/j.ymssp.2020.107233

Pang Wei Koh, P. L. (2017). *Understanding Black-box Predictions via Influence Functions*. In International conference on machine learning, Sydney, Australia.

Qu, J., Zhang, Z., & Gong, T. (2016). A novel intelligent method for mechanical fault diagnosis based on dual-tree complex wavelet packet transform and multiple classifier fusion. *Neurocomputing, 171*, 837–853. https://doi.org/10.1016/j.neucom.2015.07.020

Santiago, C., Barata, C., Sasdelli, M., Carneiro, G., & Nascimento, J. C. (2021). LOW: Training deep neural networks by learning optimal sample weights. *Pattern Recognition*. https://doi.org/10.1016/j.patcog.2020.107585

Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mechanical Systems and Signal Processing, 64–65*, 100–131. https://doi.org/10.1016/j.ymssp.2015.04.021

Wang, D., Zhao, Y., Yi, C., Tsui, K.-L., & Lin, J. (2018). Sparsity guided empirical wavelet transform for fault diagnosis of rolling element bearings. *Mechanical Systems and Signal Processing, 101*, 292–308. https://doi.org/10.1016/j.ymssp.2017.08.038

Wang, Y., Pan, Z., & Pan, Y. (2020). A training data set cleaning method by classification ability ranking for the k-nearest neighbor classifier. *IEEE Transactions on Neural Networks and Learning Systems, 31*(5), 1544–1556. https://doi.org/10.1109/TNNLS.2019.2920864

Wen, L., Li, X., & Gao, L. (2019). A transfer convolutional neural network for fault diagnosis based on ResNet-50. *Neural Computing and Applications, 32*(10), 6111–6124. https://doi.org/10.1007/s00521-019-04097-w

Wen, L., Li, X., Gao, L., & Zhang, Y. (2018). A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Transactions on Industrial Electronics, 65*(7), 5990–5998. https://doi.org/10.1109/tie.2017.2774777

Xiang, S., Qin, Y., Zhu, C., Wang, Y., & Chen, H. (2020). Long short-term memory neural network with weight amplification and its application into gear remaining useful life prediction. *Engineering Applications of Artificial Intelligence*. https://doi.org/10.1016/j.engappai.2020.103587

Xu, X., Lei, Y., & Li, Z. (2020). An incorrect data detection method for big data cleaning of machinery condition monitoring. *IEEE Transactions on Industrial Electronics, 67*(3), 2326–2336. https://doi.org/10.1109/tie.2019.2903774

Yang, C., Zhou, K., & Liu, J. (2022). SuperGraph: Spatial-temporal graph-based feature extraction for rotating machinery diagnosis. *IEEE Transactions on Industrial Electronics, 69*(4), 4167–4176. https://doi.org/10.1109/tie.2021.3075871

Yin, Z., & Hou, J. (2016). Recent advances on SVM based fault diagnosis and process monitoring in complicated industrial processes. *Neurocomputing, 174*, 643–650. https://doi.org/10.1016/j.neucom.2015.09.081

Zhang, K., Tang, B., Deng, L., Tan, Q., & Yu, H. (2021a). A fault diagnosis method for wind turbines gearbox based on adaptive loss weighted meta-ResNet under noisy labels. *Mechanical Systems and Signal Processing*. https://doi.org/10.1016/j.ymssp.2021.107963

Zhang, X., Guo, S., Li, Y., & Jiang, L. (2020a). Semi-supervised fault identification based on Laplacian eigenmap and deep belief networks. *Jixie Gongcheng Xuebao/journal of Mechanical Engineering, 56*(1), 69–81. https://doi.org/10.3901/JME.2020.01.069

Zhang, X., Huang, T., Wu, B., Hu, Y., Huang, S., Zhou, Q., & Zhang, X. (2021b). Multi-model ensemble deep learning method for intelligent fault diagnosis with high-dimensional samples. *Frontiers of Mechanical Engineering, 16*(2), 340–352. https://doi.org/10.1007/s11465-021-0629-3

Zhang, Y., Li, X., Gao, L., Chen, W., & Li, P. (2020b). Intelligent fault diagnosis of rotating machinery using a new ensemble deep auto-encoder method. *Measurement*. https://doi.org/10.1016/j.measurement.2019.107232

Zhao, K., Jiang, H., Wu, Z., & Lu, T. (2020). A novel transfer learning fault diagnosis method based on Manifold Embedded Distribution Alignment with a little labeled data. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-020-01657-z

Zhao, M., Zhong, S., Fu, X., Tang, B., Dong, S., & Pecht, M. (2021). Deep residual networks with adaptively parametric rectifier linear units for fault diagnosis. *IEEE Transactions on Industrial Electronics, 68*(3), 2587–2597. https://doi.org/10.1109/tie.2020.2972458

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing, 115*, 213–237. https://doi.org/10.1016/j.ymssp.2018.05.050

Zhou, K., Yang, C., Liu, J., & Xu, Q. (2021). Dynamic graph-based feature learning with few edges considering noisy samples for rotating machinery fault diagnosis. *IEEE Transactions on Industrial Electronics*. https://doi.org/10.1109/tie.2021.3121748

Zhou, Q., Li, Y., Tian, Y., & Jiang, L. (2020). A novel method based on nonlinear auto-regression neural network and convolutional neural network for imbalanced fault diagnosis of rotating machinery. *Measurement, 161*, 107880. https://doi.org/10.1016/j.measurement.2020.107880

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.