Neurocomputing 541 (2023) 126257

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An explainable intelligence fault diagnosis framework for rotating machinery

Daoguang Yang [a], Hamid Reza Karimi [a,*], Len Gelman [b]

[a] Department of Mechanical Engineering, Politecnico di Milano, via La Masa 1, Milan 20156, Italy
[b] School of Computing and Engineering, University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK

ARTICLE INFO

Article history:
Received 24 September 2022
Revised 12 January 2023
Accepted 22 April 2023
Available online 3 May 2023
Communicated by Zidong Wang

Keywords:
Rotating machinery
Intelligent fault diagnosis
Explainable artificial intelligence
Interpretability
Convolutional neural networks
Classification activation mappings

ABSTRACT

Convolutional neural networks (CNNs) are considered black boxes due to their robust nonlinear fitting capability. In the context of fault diagnosis for rotating machinery, it may happen that a standard CNN makes a final decision based on a mixture of significant and insignificant features, therefore, it is required to establish a trustworthy intelligence fault diagnosis model with the controllable feature learning capability to identify fault types. In this paper, an explainable intelligence fault diagnosis framework is proposed to recognize the fault signals, using data obtained through short-time Fourier transformation, which is easily modified from a standard CNN. The post hoc explanation method is used to visualize the features the model learned from a signal. The experimental results show that the proposed explainable intelligence fault diagnosis framework provides 100% testing accuracy and visualizations, the Average Drop and the Average Increase from a classification activation mappings method demonstrate the interpretability of the proposed framework.

1. Introduction

Rotating machinery is a key component in the mechanical systems [1–3]. Many researchers focus on the field of condition monitoring and fault diagnosis based on traditional signal processing techniques in the last two decades [4,5]. With the rapid development of artificial intelligence, many algorithms are applied to deal with rotating machinery fault diagnosis problems due to their strong capability in high-dimensional data processing and easily used by researchers in a variety of disciplines, such as convolutional neural networks(CNNs) [6,7], auto-encoders(AEs) [2,8], recurrent neural networks (RNNs) [9,10], deep Q networks(DQNs) [11], etc.

Especially, the CNNs, which is one of the most popular deep learning algorithms, is utilized to identify one-dimensional signals and two-dimensional signals in the rotating machinery fault diagnosis field with high accuracies [12,13]. Jiang et al. [14] proposed a multi-scale CNNs fault diagnosis framework for wind turbine gearboxes. Liu et al. [15] developed a multi-kernel multi-scale CNNs to identify the one-dimensional signal under nonstationary conditions. Xie et al. [16] developed a CNNs framework with multisensor fusion to identify the three-channel red-green–blue images. Shao et al. [17] proposed a modified CNNs framework to recognize the thermal images of the rotor-bearing system under varying operating conditions. The above methods have shown the superior performance of the CNNs.

In addition, Wang et al. [11] proposed a human-like intelligence fault diagnosis framework based on Deep Q networks (DQNs) framework with a deep CNNs, which owns a better generalization and stability than other existing methods. Ding et al. [18] developed a promising end-to-end intelligent fault diagnosis framework based on deep reinforcement learning and auto-encoder, which could mine the relationship between raw vibration signal and the fault types effectively. Zhang et al. [19] proposed a promising generative adversarial networks (GANs)-based intelligent fault diagnosis framework to deal with imbalance dataset problem. Zhou et al. [20] developed a global optimization GAN-based framework for fault diagnosis with an unbalanced dataset, which illustrates its superiority in the classification performance over existing deep learning algorithms.

However, CNNs and other deep learning algorithms are usually seen as the black box, which is unclear what features a model is using for fault diagnosis decisions [21]. Hence, it is necessary to visualize the CNNs model in order to be sure that it has learned the most important features to make final decisions [22]. In addition, the post hoc classification activation mapping (CAM) methods are proposed to visualize the attention of CNNs model, including CAM [23], Gradient-based CAM(Grad-CAM [24–26] and Grad-

* Corresponding author.

https://doi.org/10.1016/j.neucom.2023.126257
0925-2312/

CAM++(the improvement of Grad-CAM) [27]), and Gradient-free CAM (Score-CAM [28], Ablation-CAM [29]). But, there are several disadvantages of the CAM and the gradient-based CAM. For instance, the former has to change the structure of the learning model when implementing the CAM, but the latter would be easily saturated which would fail to display the saliency maps and it has a coarse localization precision [26] due to the size difference between the saliency maps and input data as shown in Fig. 1. On the other hand, post hoc visualization methods are usually used to evaluate the features that the deep learning models learned, which is not effective sometimes. It is necessary to develop a model that has interpretability and high-level recognition performance in the field of rotating machinery fault diagnosis, just like the interpretable CNNs in image classification [30]. Based on the characteristics of the time–frequency domain signals under stationary conditions, the frequencies of main fault features do not change over time. If a model can make a decision based on the significant features, it could increase the confidence of this model in fault diagnosis to some extent.

In order to deal with the above-mentioned issues, an explainable intelligence fault diagnosis framework is proposed to identify a fault signal via significant features. In addition, the Smoothed Score-CAM is used to visualize the attention of the explainable intelligence fault diagnosis framework [31], which would have a better visualization performance than the CAM and Gradient-based CAM.

The main contributions of this paper are highlighted as follows:

(1) This work proposes a novel explainable convolutional neural network, based on a located loss, which is first introduced in worldwide terms for intelligent fault diagnosis.
(2) This work proposes firstly how to increase the interpretability of CNN models by learning the significant features for decision-making.
(3) This work makes a novel comprehensive comparison of the performance of the proposed framework with the existing traditional convolutional neural networks based on two extensive experimental datasets (i. e. a gearbox dataset and a bearing dataset).

The remainder of this paper is organized as follows: Section 2 presents the basic theory of the signal processing technique, the CNNs and the post hoc visualization methods. Section 3 defines the details of the proposed interpretable intelligence fault diagnosis framework. The detailed experiment results are demonstrated in Section 4. Section 5 outlines the main conclusion of this work. The nomenclatures used in this paper are summarized in Table 1.

Table 1
Descriptions of notations.

Nomenclature		H	Feature map upsample into the same size of the input data
W	Convolutional kernel	M	New input with noisy feature map
b	bias	λ	Constant
$BN(\cdot)$	Batch normalization operation	β	Constant
a	The output of convolutional layer	$U(\cdot)$	Upsampling operation
s	The output of Batch normalization		
h	The output of activation function		
$MaxPooling$	Max pooling operation		
y	The output of pooling layer		
Y^c	The final score of the target class c		
c	The label of class		
A_k	The kth feature map in the last convolutional layer	**Abbreviations**	
w	The weight of kth feature map for the class c	CNNs	Convolutional Neural Networks
Z	Constant	AEs	Auto-Encoders
$L^c_{Grad-CAM}$	Saliency map for the class c produced by Grad-CAM	RNNs	Recurrent Neural Networks
α	Gradient weight computed by the gradient	DQNs	Deep Q Networks
D	Training set	GANs	Generative Adversarial Networks
$Mask$	Filter masks based on the input data	CAM	Classification Activation Mapping
$LocatedLoss$	The located loss	Grad-CAM	Gradient Based-CAM
$EntropyLoss$	The entropy loss	Grad-CAM++	The improvement of the Grad-CAM
$Loss$	The total loss	SS-CAM	Smoothed score-CAM
$\circ$	Hadamard product	STFT	short-time Fourier transform
$\odot$	Convolutional product	DFT	discrete Fourier transform
X_b	The baseline input	CIC	channel-wise increase of confidence
$f(\cdot)$	The output of the model's softmax layer	CWRU	Case Western Reserve University
$C(\cdot)$	The channel-wise increase of confidence		
$s(\cdot)$	Normalization the feature map		

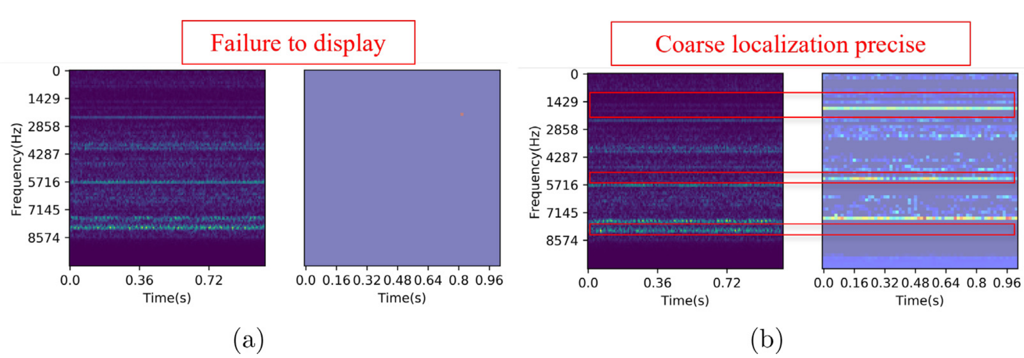

Fig. 1. The existing problems of the gradient-based CAM based on a gearbox dataset with a standard CNNs: (a) Failure to display (Grad-CAM); (b) Coarser location (Grad-CAM++).

2. Theoretical foundation

This section mainly addresses a popular signal processing technique which is used to obtain time–frequency spectrum with distinctive features, the short-time Fourier transform (STFT). It also introduces the basic theory of Convolutional Neural Networks and several types of explanation artificial intelligence algorithms, for instance, gradient-based classification activation mapping and gradient-free classification activation mapping.

2.1. The short-time Fourier transform view

In the context of rotating machinery fault diagnosis, the discrete Fourier transform (DFT) is widely used to deal with numerous fault diagnosis problems.

Although DFT has a good efficiency in obtaining the frequency spectra for stationary signals, there would be a loss of information transfered from time domain to frequency domain. Hence, the STFT is introduced to extract the time–frequency information from the raw vibration signals. The fundamental issue of the STFT is to apply the DFT to short segments of the signal. The longer signal segments, the better frequency resolution and lower time resolution. There would be a tradeoff between the time domain information and frequency domain information.

2.2. The basic theory of the convolutional neural networks(CNNs)

CNN is one of the most popular deep learning algorithms, which has been applied in a wide range of applications [32]. A standard CNN consists of four basic layers, including convolutional layers, batch normalization layers, pooling layers and fully connected layers. Convolutional layers are usually used to extract significant features from the raw data; normalization layers are used to speed up the convergence rate and to avoid overfitting; the pooling layers are usually utilized to decrease the computational complexity and to avoid overfitting, including max pooling, average pooling, and etc [33]; fully connected layer is also used to extract features for classification problems.

In one standard CNNs model, there are several convolutional blocks, generally consists of at least two convolutional blocks, each of them contains one convolutional layer, one normalization layer, one activation function and one pooling layer. In this paper, the output of one convolutional block is expressed as follows:

$$a = W \odot x + b \tag{1}$$

$$s = BN(a) \tag{2}$$

$$h = ReLU(s) = max(0, s) \tag{3}$$

$$y = MaxPooling(h) \tag{4}$$

where $\odot$ refers to the convolutional operation, W and b refer to the convolutional kernel and the bias in this convolutional layer, respectively. $BN(a)$ means the batch normalization operation for the input a [34]. ReLU is one of the most popular activation functions in the CNNs that could also speed up the convergence rate. *MaxPooling* is the Max pooling operation [35]. a, s, h and y are the outputs of the convolutional layer, batch normalization layer, activation function and pooling layer, respectively.

2.3. Explainable artificial intelligence

One of the main limitations of the CNNs in the rotating machinery fault diagnosis is the interpretability, which is usually known as black box [36]. It is unclear whether the CNNs model learned the key information in the input dataset or not to make a final decision. Hence, some gradient-based post hoc visualization methods are developed to explain what the CNNs learned, such as the gradient-based classification activation mapping (Grad-CAM) [24] and the improvement, Grad-CAM++ [27].

The core idea of Grad-CAM is to generate the saliency map by calculating the gradient of each feature map in the certain convolutional layer corresponding to the classification result, which could avoid to change the structure of the CNNs model [37]. The process of calculating the Grad-CAM is given as follows:

$$Y^c = \sum_k w_k^c \cdot \sum_i \sum_j A_{ij}^k \tag{5}$$

$$w_k^c = Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}, \ \forall \{i, j | i, j \in A^k\} \tag{6}$$

$$L_{Grad-CAM}^c = ReLU\left(\sum_k w_k^c \cdot A_{ij}^k\right) \tag{7}$$

where Y^c is the final score of the target class c, w_k^c is the weight of kth feature map for the class c. i, j are the spatial location in the class-specific saliency map. A^k refers to the kth feature map in the last convolutional layer. Z is a constant corresponding to the data points in the feature map. $L_{Grad-CAM}^c$ is the saliency map corresponding to the target class c.

However, it is difficult to generate an effective saliency map based on the Grad-CAM if there are many essential features in the input data. Hence, its improvement (Grad-CAM++) is developed to improve the interpretability of the CNNs model for complex input data. The main improvement of the Grad-CAM++ is modification of the Eq. (6), which is given as follows:

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot ReLU\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \tag{8}$$

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2}}{2 * \frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2} + \sum_a \sum_b A_{ab}^k \left\{\frac{\partial^3 Y^c}{\left(\partial A_{ij}^k\right)^3}\right\}} \tag{9}$$

$$L_{Grad-CAM++}^c = ReLU\left(\sum_k w_k^c \cdot A_{ij}^k\right) \tag{10}$$

where α_{ij}^{kc} is a gradient weight computed by the gradient for the target class c and the feature maps.

In addition, the methods based on the gradient have several other drawbacks. On the one hand, gradient-based CAM for a CNNs may focus on the unrelated parts due to the gradient saturation in the flat zero-gradient region of the ReLU. On the other hand, the weights obtained from the gradient-based CAM do not provide right confidence scores for the feature maps, which would generate a coarse localization saliency map [28].

3. Proposed framework for time–frequency spectrum based visualization

In this section, an explainable intelligent fault diagnosis framework is proposed that contains an standard CNNs classifier with additional located loss to learn significant features of a signal instead of learning insignificant features of time–frequency spectrum and the post hoc explainable artificial intelligence algorithm that is used to visualize the classification criteria in the time–frequency domain. In Section 3.1, a novel interpretable intelligence fault diagnosis framework is proposed which could be easily modified from the standard CNNs. In addition, a gradient-free classification activation mappping method is introduced in Section 3.2 to verify whether the model is reliable.

3.1. The structure of the interpretable intelligence fault diagnosis framework

An overview of the proposed interpretable intelligence fault diagnosis framework is demonstrated in Fig. 2. The main structure of the proposed method is based on the standard CNNs. In order to make a trustworthy intelligent fault diagnosis framework, the located loss is first introduced in the training process, which is used to penalize if the model learned some insignificant fault features in the training process.

$$LocatedLoss = \sum_{i \in D} \|Mask_i \circ U(f(x_i))\|_2 \tag{11}$$

$$Mask_i(j,k) = \begin{cases} 1, & x_i(j,k) < mean(x_i) + \lambda * std(x_i). \\ 0, & x_i(j,k) \geqslant mean(x_i) + \lambda * std(x_i). \end{cases} \tag{12}$$

$$Loss = EntropyLoss + \beta * LocatedLoss \tag{13}$$

where D stands for the samples in the training set. $Mask_i$ is the filter mask based on the i_{th} input data x_i, the data points equal to 1 where the value of the data points located in the x_i is lower than its average plus λ multiply by standard deviation of x_i. j,k are the coordinates of the data point. β is a constant which is a constraint for the located loss. In the rotating machinery fault diagnosis problem, λ and β are set to 0.1 and 0.0003, respectively, which is based on the testing performance of the framework. The larger λ makes more values equal to 1 in the mask, which will make the model learn more focused features that could lead to overfitting. The larger β makes the model pay more attention on the features in the 0-valued region of the mask.

In order to show the performance of the proposed framework, a small standard CNNs is used in the framework. As shown in Fig. 3, there are four convolutional layers, four normalization layers, four pooling layers and one fully connected layer. The main parameters of the CNNs model are also given in Fig. 3.

3.2. Smoothed Score-CAM(SS-CAM)

Because of the limitation of the gradient-based post hoc visualization in the rotating machinery fault diagnosis, it is necessary to introduce an enhanced visual explanation algorithm, called smoothed score-CAM. The pipeline of the smoothed score-CAM [31] is demonstrated in Fig. 4. It is used the confidence score of each feature map in the last convolutional layer, which is similar to the CAM method [23]. The key idea of Smoothed score-CAM is that it uses the average score based on the output of the CNNs model which the inputs are the feature maps in the last convolutional layer, which is called the channel-wise increase of confidence (CIC), $C(\cdot)$. The CIC is obtained by:

$$C\left(A_l^k\right) = f\left(X \circ H_l^k\right) - f(X_b) \tag{14}$$

$$H_l^k = s\left(U\left(A_l^k\right)\right) \tag{15}$$

$$s\left(A_l^k\right) = \frac{A_l^k - min\left(A_l^k\right)}{max\left(A_l^k\right) - min\left(A_l^k\right)} \tag{16}$$

where X is the input data, $f(X)$ is the output of the model's softmax layer. A_l^k is the kth feature map in the lth convolutional layer(here is the last convolutional layer). $s(\cdot)$ is the normalization operation which data values in the mapping are within $[0,1]$. $U(\cdot)$ is the upsampling operation that is used to generate a new feature map with the same size of the input data. X_b is the baseline input.

Hence, the significance of the A_l^k could be computed by the Eq. (14), which is similar to the idea of the gradient-based post hoc visualization methods [24,27]. In addition, in order to avoid the influence of the noise information of input data on the feature map, the Gaussian noise is added into the feature map A_l^k to generate N noisy input samples, the calculation process is demonstrated in the Fig. 4 (Phase 2). The equations are given by:

$$L^c_{SS-CAM} = ReLU\left(\sum_k \eta^c_k A^k_l\right) \tag{17}$$

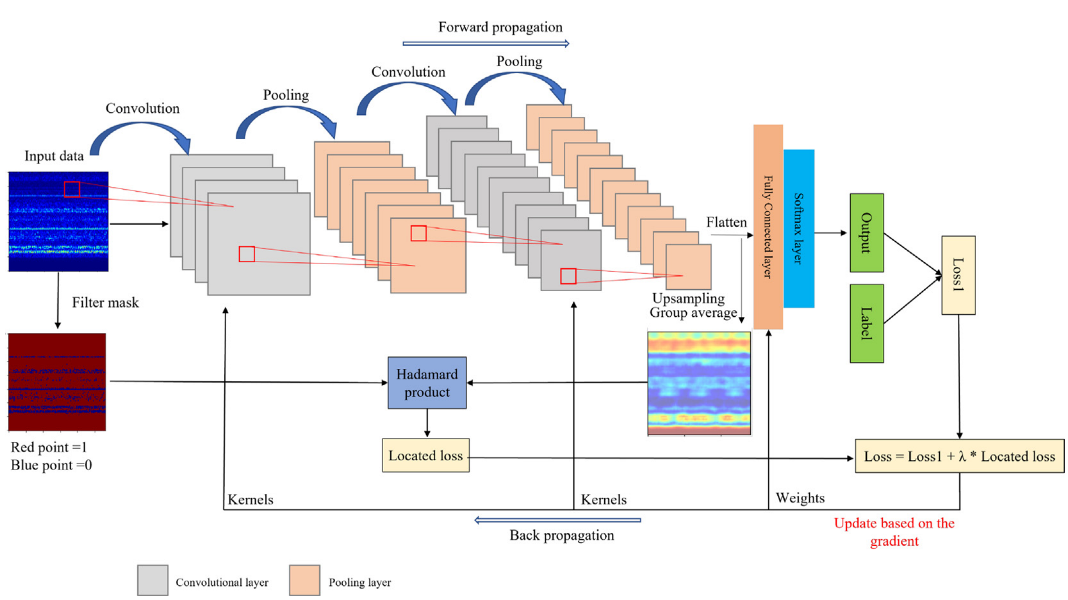

Fig. 2. The training process of the proposed intelligence fault diagnosis method.

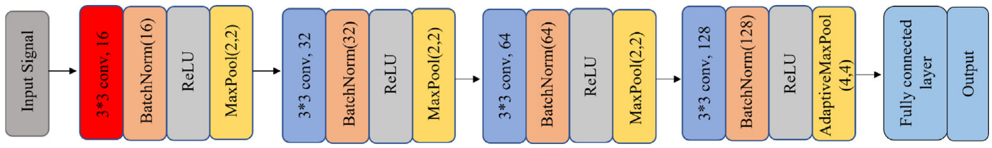

Fig. 3. The structure of the standard CNNs.

$$\eta_c^k = \frac{\sum_1^N (C(M))}{N} \tag{18}$$

$$M = \sum_1^N \left(X * \left(A_l^k + N(0, \sigma^2) \right) \right) \tag{19}$$

where c is the class of interest. α_c^k is the average score based on the N noisy input M combined the input data X with the additive noise feature map $A_l^k + N(0, \sigma^2)$.

4. Experimental validation and analysis

In this section, we use two different rotating machinery fault datasets to verify the interpretable capability of the proposed explainable CNNs. One is the gearbox dataset based on the same fault type with different fault severity [38,39]. The bearing dataset with multiple bearing fault types is taken from an open-source Case Western Reserve University(CWRU) bearing dataset [40]. Then, the training performance of the proposed CNNs is compared with the standard CNNs. Lastly, several evaluation metrics are used to test whether the model is trustworthy.

4.1. Datasets

4.1.1. Gearbox dataset

The gearbox signals are collected from a 91.5 *mm* back-to-back gearbox test rig that allows to control and create a very early stage of natural micro-pitting development [38]. The pictures of the test rig and the installation of the sensors are shown in Fig. 5. The gearbox test rig contains two identical gearboxes connected through a torsionally compliant shaft which the torque controlled by a servo-hydraulic torque actuator. Each gearbox contains two gears, 16 teeth pinion and 24 teeth gear. The tooth surface of the pinion is diagnosed in this experiment. The micro-pitting fault is a classical fault in the gears, which has some bad effects on the gear mesh and leads to a decrease in gear reliability. The original sampling rate of 40 kHz is used to record the acceleration signals and based on the frequency information in the signals, the downsampling method is used to change the sampling rate to 20 kHz. The vibration signals were collected by 3 mono-axial accelerometers (KCF-107) at 3 orthogonal directions and the condition of teeth surfaces on the pinion was diagnosed after every 10^6 cycles, total $5 * 10^6$ cycles. During the test, the pinion was spinning at 3000 rpm (50 Hz), under a load of $(500 \pm 5)\, N \cdot m$ [39]. Hence, there are five different fault severity signals.

The collected vibration data is divided into around one-second segments (19600 points), 80 segments of each fault severity are transformed by the STFT into 141*141 time–frequency spectrum. 320 of the 400 STFT spectra are randomly selected as the training set, and the other 80 time–frequency spectra are applied as the testing set.

4.1.2. Bearing dataset

This bearing dataset could be seen as the baseline dataset in the rotating machinery fault diagnosis field [40,41]. The bearing vibration signals are collected by an accelerometer located on the drive end of a motor using 12 kHz sampling frequency under different bearing loads (0-3hp). The health conditions of the bearing are divided into four different health states, including normal state, inner ring fault, ball fault and outer ring fault. Here, vibration signals used to evaluate the performance of the proposed framework

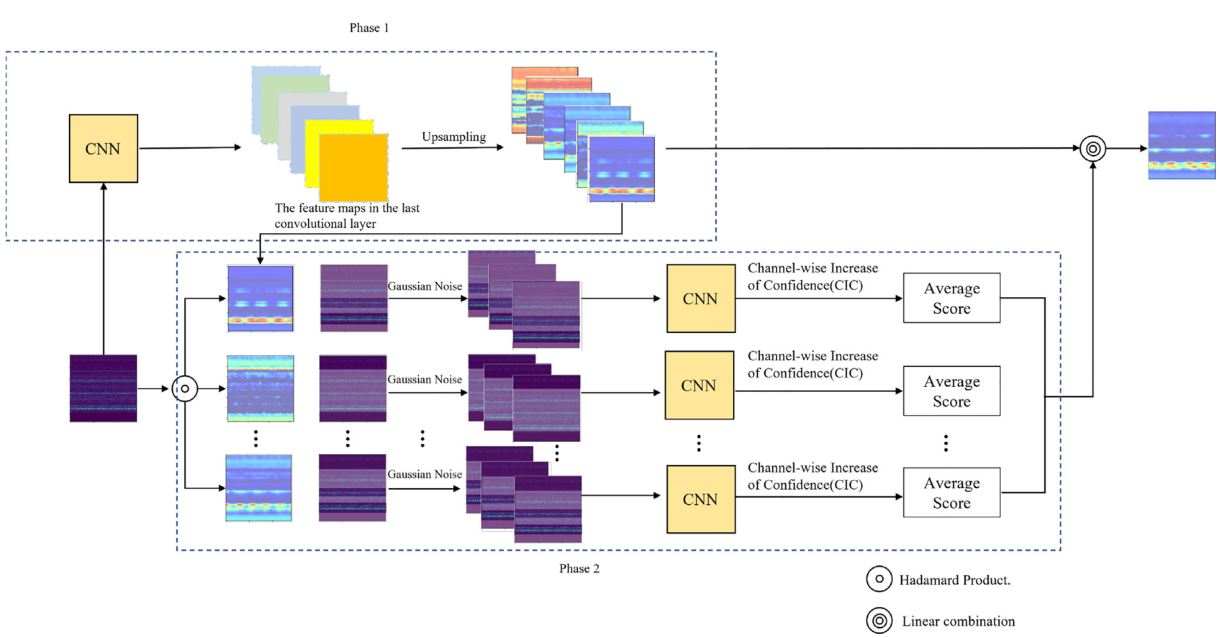

Fig. 4. The flowchart of the smoothed score-CAM.

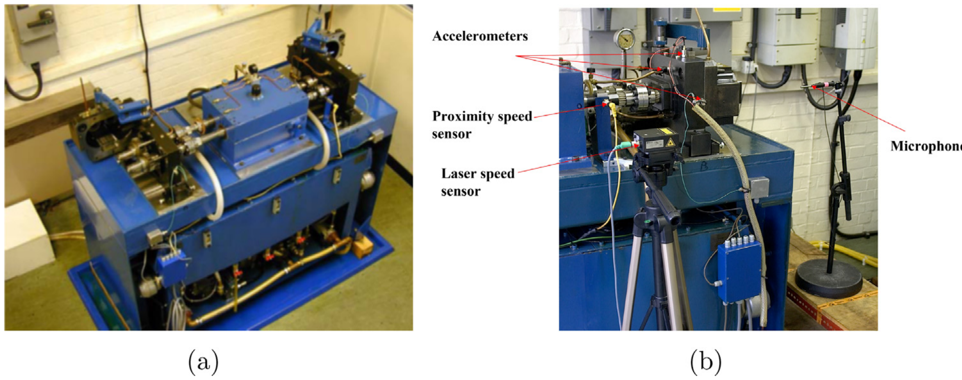

Fig. 5. The gearbox test rig and installation of sensors [38,39]: (a) gearbox test rig; (b) installation of sensors.

were collected from the bearings with 0.007 inches fault diameters of the ball fault, inner ring fault and outer ring fault under an operation load of 0hp. Due to the data limitation (vibration of each health state was collected during 20 s), the augmentation technique (shifting window) is applied to increase the training samples and testing samples, each sample contains 12100 points. The STFT is used to transform the one-dimensional time-domain signal into a two-dimensional time–frequency spectrum (111*111). Each type of signal contains 100 samples, of which 80% are randomly selected for training and 20% for testing.

4.2. The performance of the proposed CNNs

The deep learning open-source framework, Pytorch (Version 1.13.0), is used to build and train the proposed CNN model and the standard CNNs in Python (Version 3.8.16) on Windows 11. Due to the characteristic of the proposed located loss, there are not any trainable parameters added in the proposed intelligent fault diagnosis framework. Hence, the computation complexity of the proposed model is the same as the standard CNNs. The diagnosis accuracies of both models are 100% (the average results of five times running) for the gearbox dataset and bearing dataset, which means that the proposed model not only increases the interpretability but also does not compromise the accuracy of the model. Due to the existence of a filter mask, the model would more focus on the major fault information in the input data instead of the minor fault information. Hence, the proposed model has a faster convergence rate than the standard CNNs, convergence rates are shown in Fig. 6 and Fig. 7 to verify the effectiveness of the proposed framework. As shown in Fig. 6, the proposed framework could reach convergence in 10 iterations instead of 20 iterations for the standard CNNs convergence rate for the gearbox dataset. According to the Fig. 7, the proposed model could converge into 100 % testing accuracy within 2 iterations instead of 5 iterations for the standard CNNs with the bearing dataset.

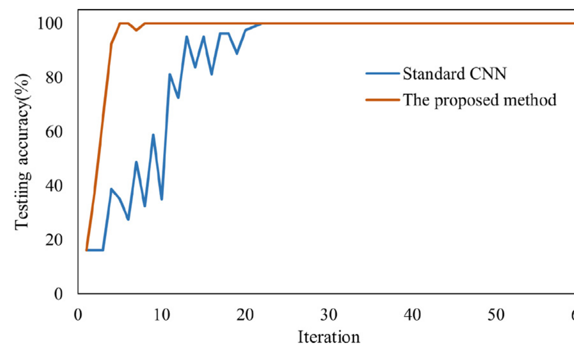

Fig. 6. The testing accuracy changes vs. iterations for the gearbox dataset.

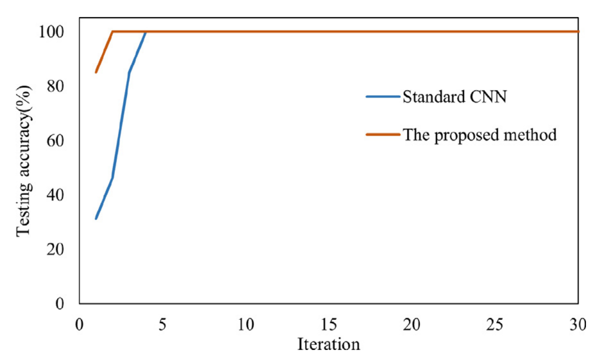

Fig. 7. The testing accuracy changes vs. iterations for the bearing dataset..

4.3. Quality evaluation via signal recognition

In this section, the results of classification activation mappings are demonstrated here. But it is not sufficient to evaluate the interpretability quality of the model only by the classification activation mappings generated by the SS-CAM. Hence, the Average Drop and Average Increase metrics are used to evaluate the quality of the model [27].

Average Drop: saliency maps indicate the crucial information for a particular type of fault in a signal. The model's confidence would be mostly lowered if parts of the signal were omitted. This drop is expected to be low. After an occlusion in a signal, this metric shows the average drop in the model's confidence for a particular fault type [27].

Average Increase: sometimes, the deep CNNs looks for the entire pattern in the most discriminative part highlighted by the saliency maps. The confidence scores of the model increases for that particular class in this situation. This metric measures the number of times that the model's confidence increased after excluding unimportant signals in its entirety [28].

Point-wise multiplication is used to mask the original input data with saliency maps in order to observe changes in the predicted score on the target class. The equations of the Average Drop and the Average Increase are given as follows:

$$Average\ Drop = \sum_{i=1}^{N} \frac{Y_i^c - O_i^c}{Y_i^c} \times 100 \quad (20)$$

Table 2
Average Drop (the lower the better) and Average Increase (the higher the better) across the gearbox dataset.

Metrics	SS-CAM	
	The proposed method	Standard CNNs
Average Drop(%)	33.35	38.77
Average Increase(%)	12.5	10

Table 3
Average Drop (the lower the better) and Average Increase (the higher the better) across the bearing dataset

Metrics	SS-CAM	
	The proposed method	Standard CNNs
Average Drop(%)	0.47	28.99
Average Increase(%)	5	1.25

$$Average\ Increase = \sum_{i=1}^{N} \frac{Func(Y_i^c < O_i^c)}{Y_i^c} \times 100 \quad (21)$$

where Y_i^c and O_i^c mean the final prediction scores on class c using the original input data i, and using the classification activation mapping point-wise multiplication on the original input data i, respectively. $Func(\cdot)$ indicates a boolean function when the condition in the brackets is true, the function returns 1; otherwise, it returns 0.

Table 2 shows results of the Average Drop and Average Increase of the proposed method and the standard CNNs for the gearbox dataset (Average result of ten tests). The results demonstrate that the proposed method is more focused on the information points than the standard CNNs across the gearbox dataset. Table 3 shows that the performance of the proposed method is much better than the standard CNNs with the bearing dataset (Average result of ten tests). The Average Drop of the proposed method is lower than 1 % which means that the decisions made by the proposed method are mainly based on the data points shown by the saliency map.

4.4. Localization evaluations

The localization ability of the attention map is important because the saliency map can be applied to localization tasks in the frequency domain of the rotating machinery signals, to verify decisions made by the intelligence fault diagnosis framework. Here, in order to pinpoint the information in the signal that the model actually learned, the point-wise mask has been applied which is based on the value of the points in the classification activation mappings, which the value is changed into 1 if its value is higher than its mean plus 1 multiply by the standard deviation of the heatmap.

With the gearbox dataset, as shown in Fig. 8, the proposed method could pay more attention on the fault information instead of the irrelevant information, especially for data with fault severity 4. By comparing the results with the standard CNNs shown in Fig. 9, the proposed model is a more trustworthy model than the standard CNNs. For example, firstly, the proposed model focuses on the fault information in the signal more intensively than the standard CNNs, such as the middle picture of Fig. 8 (a) and Fig. 9 (a), but the learned features are almost close, like the right picture of Fig. 8 (a) and Fig. 9 (a). Secondly, the decisions made by the proposed model are based on the fault information in the signal, not

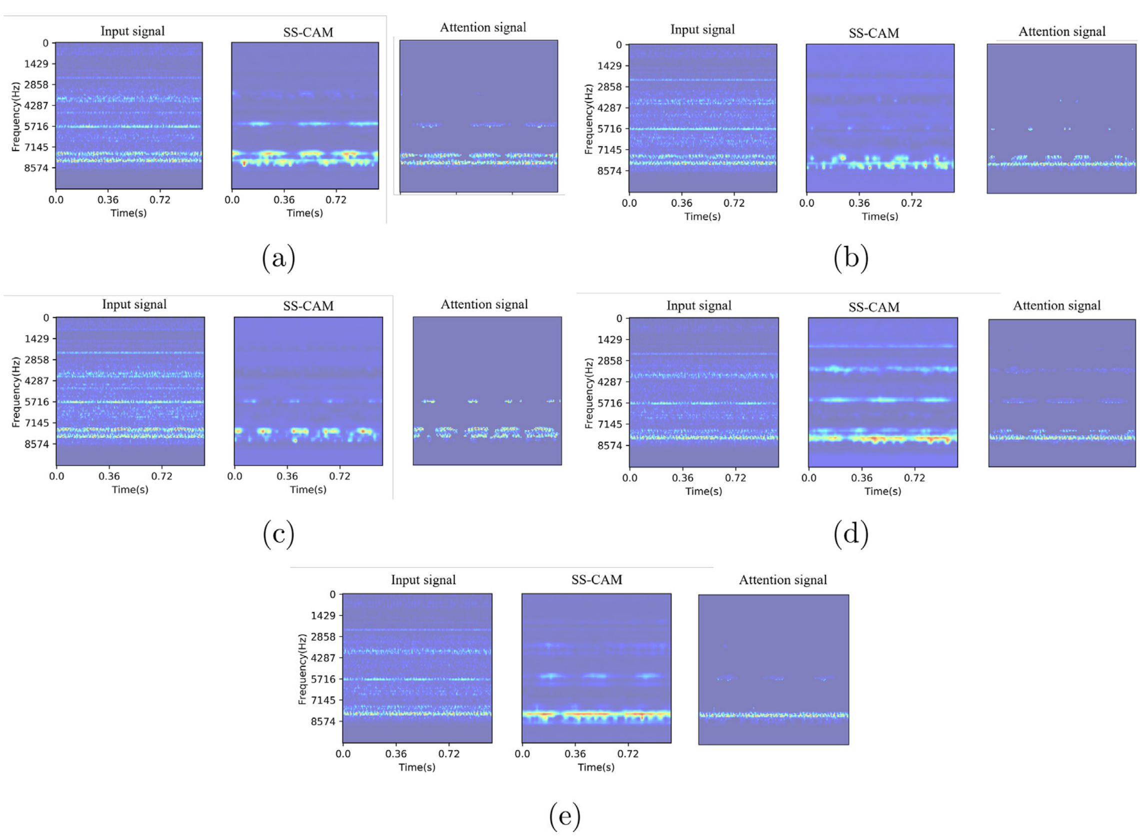

Fig. 8. Localization evaluation results of the proposed model for the gearbox fault dataset: (a) fault severity level 1; (b) fault severity level 2; (c) fault severity level 3; (d) fault severity level4; (e) fault severity level 5.

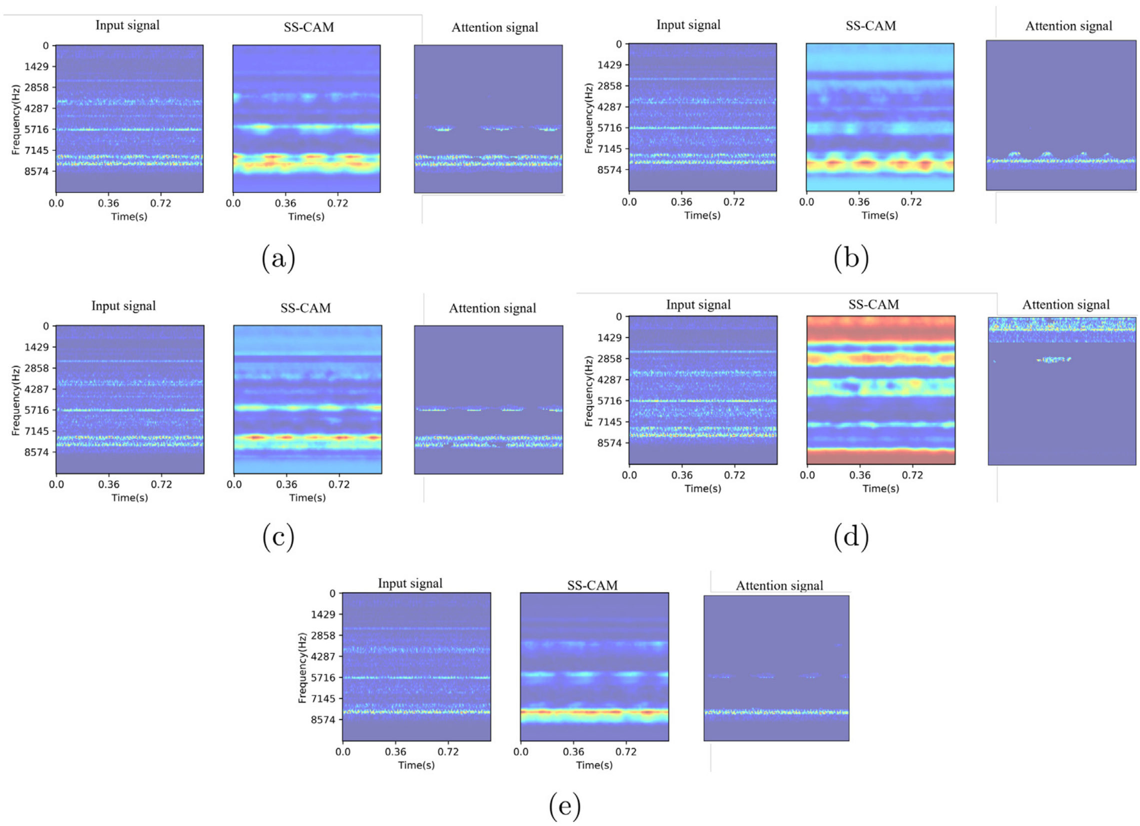

Fig. 9. Localization evaluation results of the standard CNNs for the gearbox fault dataset: (a) fault severity level 1; (b) fault severity level 2; (c) fault severity level 3; (d) fault severity level4; (e) fault severity level 5.

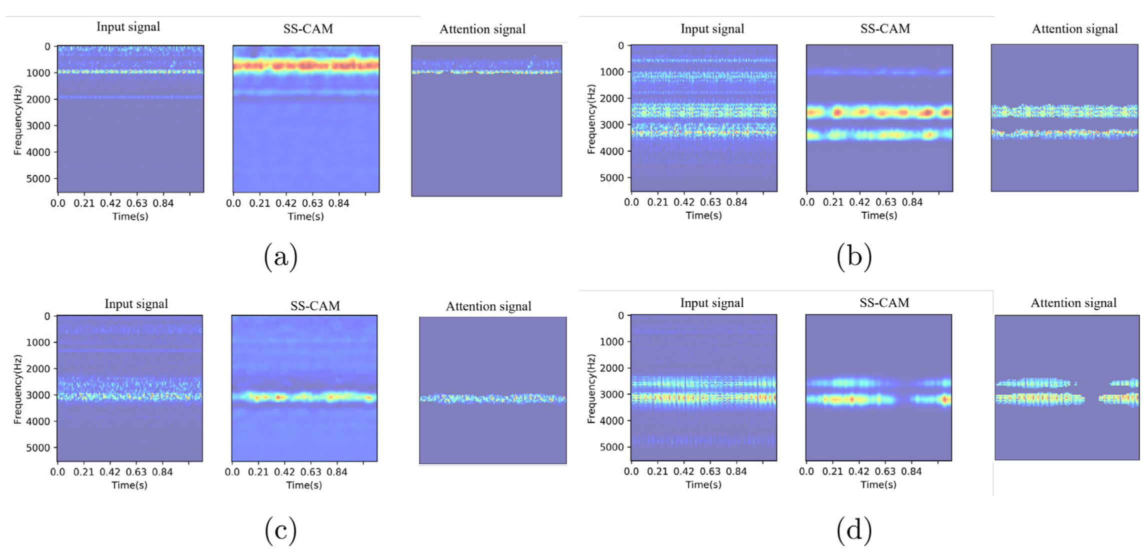

Fig. 10. Localization evaluation results of the proposed model for the bearing fault dataset: (a) normal; (b) inner ring fault; (c) ball fault; (d) outer ring fault.

the unrelated features, such as the middle picture and right picture of Fig. 8 (d) and Fig. 9 (d).

For the bearing dataset, comparing Fig. 10 with Fig. 11, it is easy to distinguish which model is more trustworthy. For example, the proposed model pays more attention to the key information among all the input data, like the middle picture of Fig. 10 (b) and Fig. 11 (b), and the standard CNNs makes its diagnosis decision based on its minor fault information, which cannot be a valid classification result, like the middle picture of Fig. 10 (a) and Fig. 11 (a).

4.5. Frequency-domain localization evaluations

In the field of rotating machinery fault diagnosis, frequency-domain signals are easily identified by fault diagnosis experts,

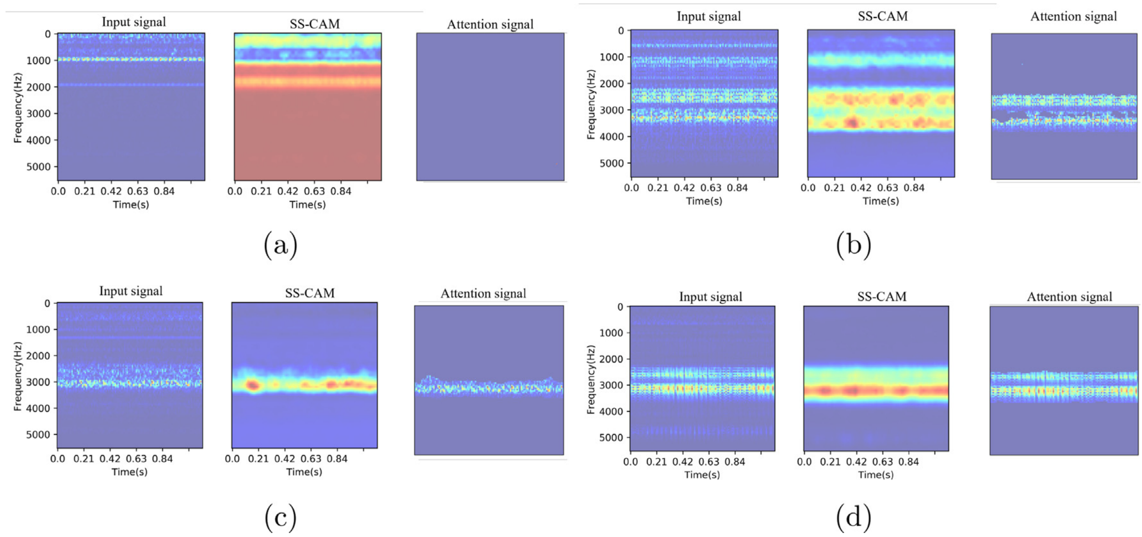

Fig. 11. Localization evaluation results of the standard CNNs for the bearing fault dataset: (a) normal; (b) inner ring fault; (c) ball fault; (d) outer ring fault.

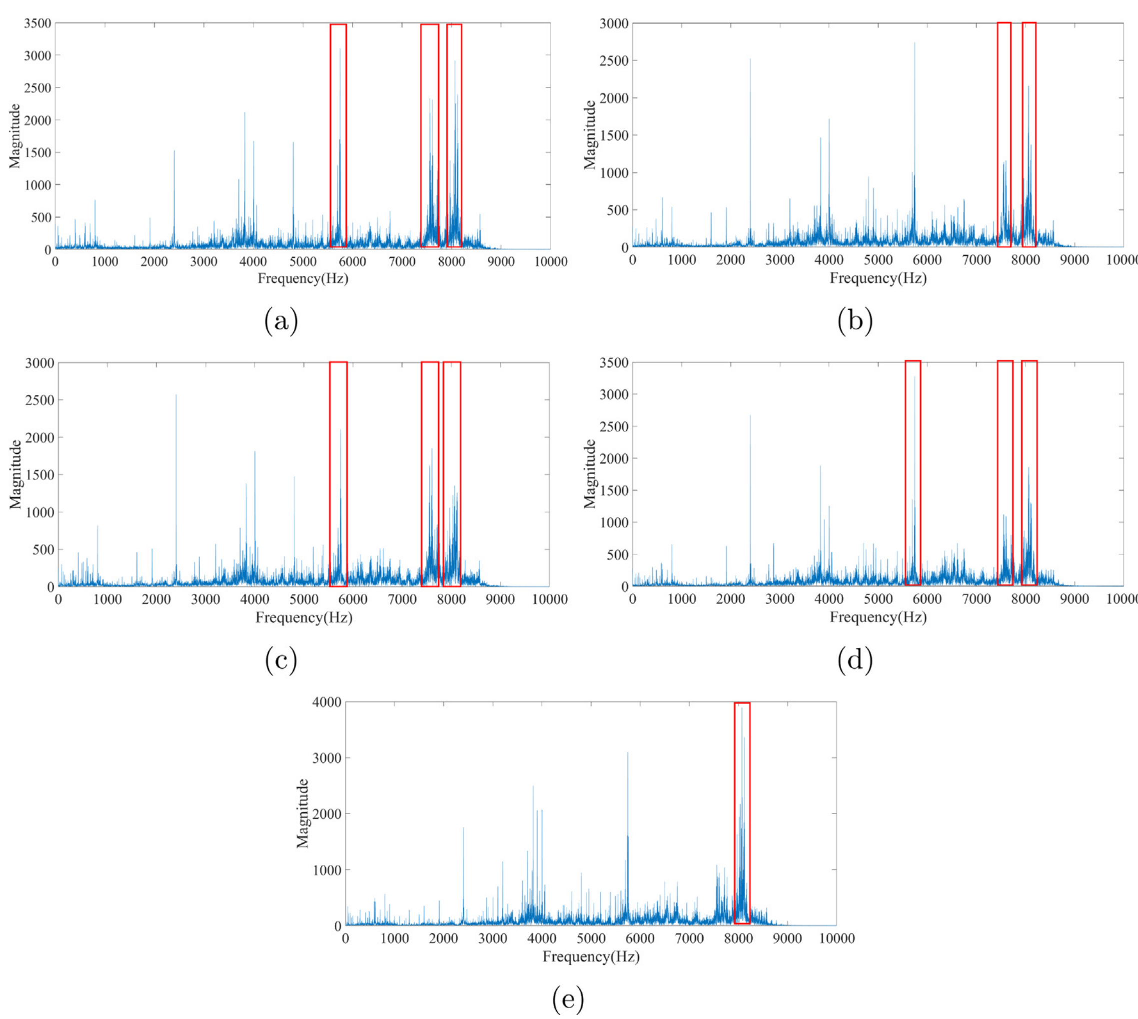

Fig. 12. Frequency-domain localization of the proposed model for the gearbox fault dataset: (a) fault severity level 1; (b) fault severity level 2; (c) fault severity level 3; (d) fault severity level 4; (e) fault severity level 5.

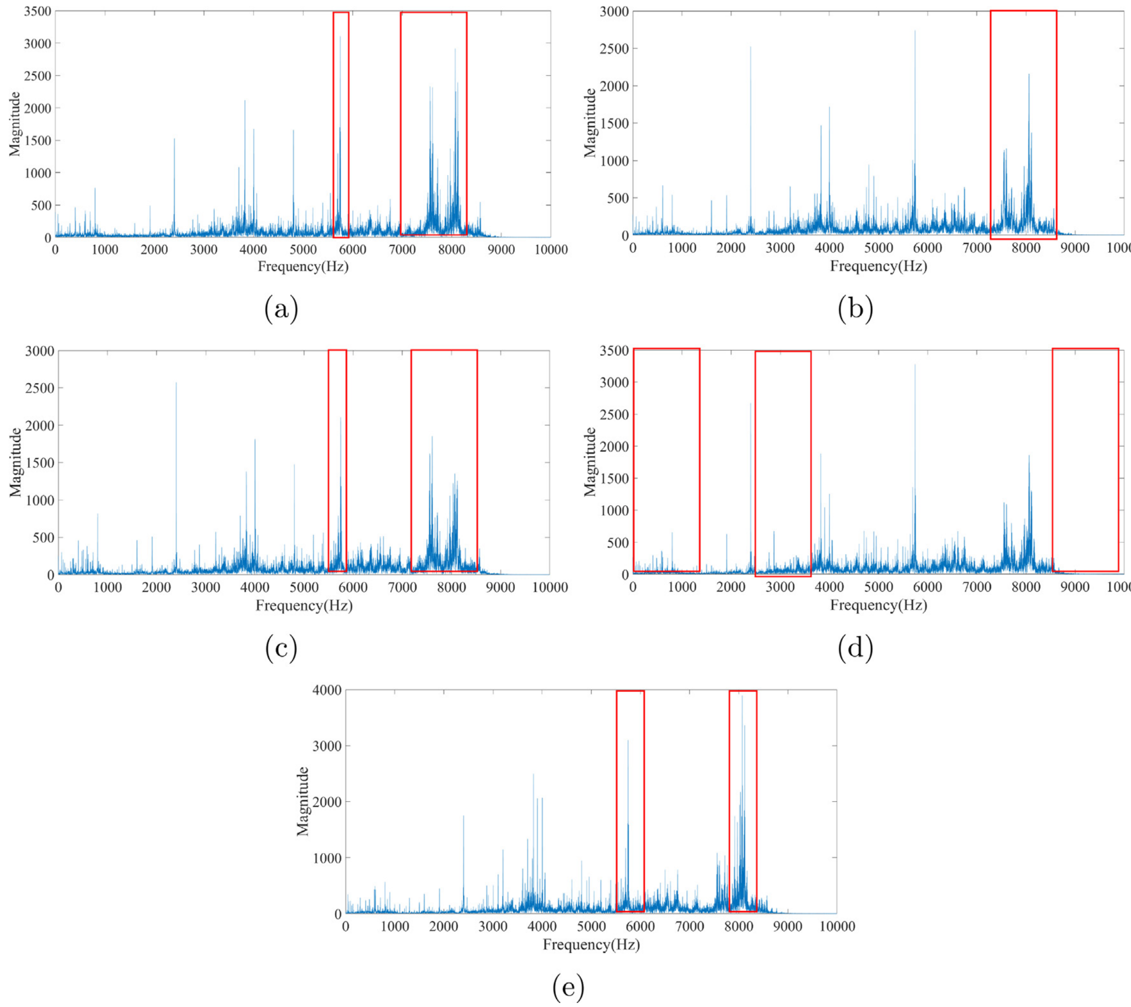

Fig. 13. Frequency-domain localization of the standard CNNs for the gearbox fault dataset: (a) fault severity level 1; (b) fault severity level 2; (c) fault severity level 3; (d) fault severity level 4; (e) fault severity level 5.

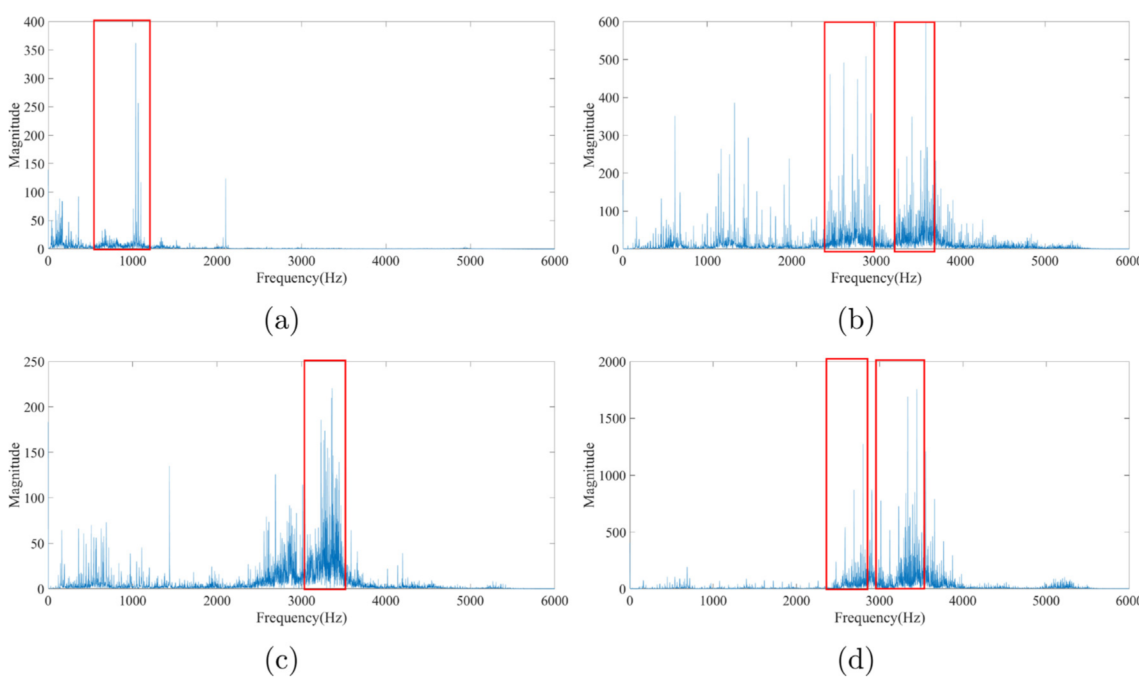

Fig. 14. Frequency-domain localization of the proposed framework for the bearing fault dataset: (a) normal; (b) inner ring fault; (c) ball fault; (d) outer ring fault.

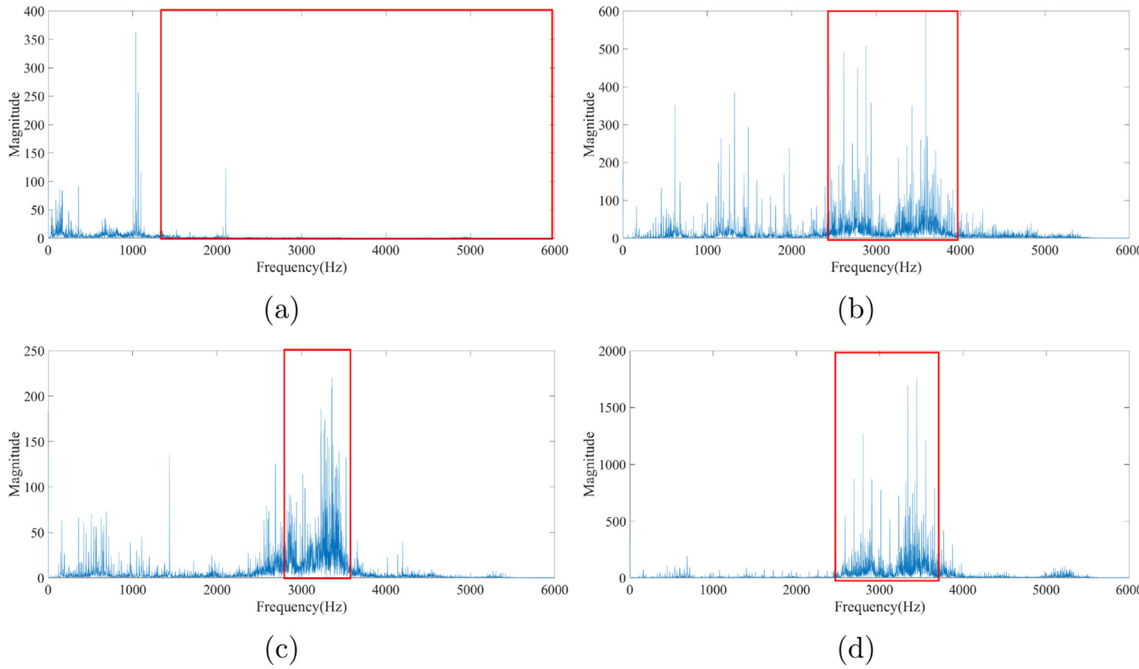

Fig. 15. Frequency-domain localization evaluation results of the standard CNN with bearing fault dataset: (a) normal; (b) inner ring fault; (c) ball fault; (d) outer ring fault.

and the Fourier transform is a popular signal processing technique. Hence, there is a need to use frequency-domain spectra to evaluate the proposed framework which actually learns the valid features from different types of input. Based on the saliency maps of the time–frequency domain signal by the proposed framework and the standard CNNs, it is easy to identify which frequency is the focus frequency for identification, and to explain what the black-box learned. The width of red boxes in these figures are depended on the focused frequency of the above saliency maps.

In the gearbox fault dataset, the frequency-domain signals of different fault severities of the gearbox are also close to each other. As shown in Fig. 12 and 13, the frequency component of around 6000 *Hz* and 7000 to 8500 *Hz* are the main components related to the fault severity. The proposed framework focuses on the main valid information instead of the unrelated information. For instance, if the fault severity level 4 is considered, it is obvious that the standard CNNs model focuses on the frequency parts with less fault information, while the proposed CNNs model could focus on the frequency parts with major fault information.

In the CWRU bearing dataset, there are four different bearing healthy condition signals, the main frequencies of fault components are different. Hence, it could be seen as an effective comparative experiment to evaluate the performance of the proposed framework. As shown in Figs. 14 and 15, the proposed framework mainly focuses on the powerful frequency domain signal instead of the irrelevant part, the standard CNNs focuses on the weak frequency domain signal which could be easily disturbed by the noise and would lead to the fault diagnosis model to make wrong decisions. In addition, the proposed framework would be more likely to focus on the differences between similar signals, such as inner race fault frequency domain signals and outer race fault frequency domain signals, shown in Fig. 10 and Fig. 11, respectively.

5. Conclusion

This paper proposed a new interpretable intelligence fault diagnosis framework based on a novel located loss that could push filters in the last convolutional layer to learn major features without any annotations for supervision, which is easy to be modified from the standard CNNs. The experimental results have shown better convergence speed and interpretability of the proposed model than the standard CNNs based on a gearbox dataset and a bearing dataset. For future work, it is necessary to develop an intelligent fault diagnosis framework, which could directly explain which part of the input signal is interpreting the fault component to make the final decision.

CRediT authorship contribution statement

Daoguang Yang: Conceptualization, Methodology, Software, Data curation, Writing – original draft. **Hamid Reza Karimi:** Methodology, Visualization, Formal analysis, Resources, Supervision, Writing – review & editing. **Len Gelman:** Methodology, Investigation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is supported in part by the scholarship from the China Scholarship Council (CSC), China under Grant CSC N201906050158, in part by the Italian Ministry of Education, University and Research, Italy for the support provided through the Project "Department of Excellence LIS4.0 – Lightweight and Smart Structures for Industry 4.0" and in part by the Horizon Marie Skłodowska-Curie Actions program (101073037).

References

[1] Y. Jin, L. Hou, Y. Chen, A time series transformer based method for the rotating machinery fault diagnosis, Neurocomputing 494 (2022) 379–395.

[2] C. Wang, C. Xin, Z. Xu, M. Qin, M. He, Mix-vaes: A novel multisensor information fusion model for intelligent fault diagnosis, Neurocomputing 492 (2022) 234–244.
[3] Y. Miao, B. Zhang, J. Lin, M. Zhao, H. Liu, Z. Liu, H. Li, A review on the application of blind deconvolution in machinery fault diagnosis, Mech. Syst. Signal Process. 163 (2022).
[4] Y. Jiang, S. Yin, O. Kaynak, Performance supervised plant-wide process monitoring in industry 4.0: A roadmap, IEEE Open J. Ind. Electron. Soc. 2 (2020) 21–35.
[5] Y. Jiang, S. Yin, O. Kaynak, Optimized design of parity relation-based residual generator for fault detection: Data-driven approaches, IEEE Trans. Industr. Inf. 17 (2) (2020) 1449–1458.
[6] M.S. Kim, J.P. Yun, P. Park, Deep learning-based explainable fault diagnosis model with an individually grouped 1d convolution for 3-axis vibration signals, IEEE Trans. Industr. Inf. 18 (12) (2022) 8807–8817.
[7] Y. Chen, D. Zhang, H.R. Karimi, C. Deng, W. Yin, A new deep learning framework based on blood pressure range constraint for continuous cuffless bp estimation, Neural Networks 152 (2022) 181–190.
[8] D. Yang, H.R. Karimi, K. Sun, Residual wide-kernel deep convolutional auto-encoder for intelligent rotating machinery fault diagnosis with limited samples, Neural Networks 141 (2021) 133–144.
[9] H. Miao, B. Li, C. Sun, J. Liu, Joint learning of degradation assessment and rul prediction for aeroengines via dual-task deep lstm networks, IEEE Trans. Industr. Inf. 15 (9) (2019) 5023–5032.
[10] Y. Lei, H.R. Karimi, X. Chen, A novel self-supervised deep lstm network for industrial temperature prediction in aluminum processes application, Neurocomputing 502 (2022) 177–185.
[11] H. Wang, J. Xu, C. Sun, R. Yan, X. Chen, Intelligent fault diagnosis for planetary gearbox using time-frequency representation and deep reinforcement learning, IEEE/ASME Trans. Mechatron. 27 (2) (2021) 985–998.
[12] J. Jiao, M. Zhao, J. Lin, K. Liang, A comprehensive review on convolutional neural network in machine fault diagnosis, Neurocomputing 417 (2020) 36–63.
[13] S. Liu, H. Jiang, Z. Wu, X. Li, Data synthesis using deep feature enhanced generative adversarial networks for rolling bearing imbalanced fault diagnosis, Mech. Syst. Signal Process. 163 (2022).
[14] G. Jiang, H. He, J. Yan, P. Xie, Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox, IEEE Trans. Industr. Electron. 66 (4) (2018) 3196–3207.
[15] R. Liu, F. Wang, B. Yang, S.J. Qin, Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions, IEEE Trans. Industr. Inf. 16 (6) (2019) 3797–3806.
[16] T. Xie, X. Huang, S.-K. Choi, Intelligent mechanical fault diagnosis using multi-sensor fusion and convolution neural network, IEEE Trans. Industr. Inf. 18 (5) (2021) 3213–3223.
[17] H. Shao, M. Xia, G. Han, Y. Zhang, J. Wan, Intelligent fault diagnosis of rotor-bearing system under varying working conditions with modified transfer convolutional neural network and thermal images, IEEE Trans. Industr. Inf. 17 (5) (2020) 3488–3496.
[18] Y. Ding, L. Ma, J. Ma, M. Suo, L. Tao, Y. Cheng, C. Lu, Intelligent fault diagnosis for rotating machinery using deep q-network based health state classification: A deep reinforcement learning approach, Adv. Eng. Inform. 42 (2019).
[19] W. Zhang, X. Li, X.-D. Jia, H. Ma, Z. Luo, X. Li, Machinery fault diagnosis with imbalanced data using deep generative adversarial networks, Measurement 152 (2020).
[20] F. Zhou, S. Yang, H. Fujita, D. Chen, C. Wen, Deep learning fault diagnosis method based on global optimization gan for unbalanced data, Knowl.-Based Syst. 187 (2020).
[21] P. Pandey, A. Rai, M. Mitra, Explainable 1-d convolutional neural network for damage detection using lamb wave, Mech. Syst. Signal Process. 164 (2022).
[22] J. Grezmak, J. Zhang, P. Wang, K.A. Loparo, R.X. Gao, Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis, IEEE Sens. J. 20 (6) (2019) 3172–3181.
[23] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.
[24] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
[25] H. Sun, X. Cao, C. Wang, S. Gao, An interpretable anti-noise network for rolling bearing fault diagnosis based on fswt, Measurement 190 (2022).
[26] H.-Y. Chen, C.-H. Lee, Vibration signals analysis by explainable artificial intelligence (xai) approach: Application on bearing faults diagnosis, IEEE Access 8 (2020) 134246–134256.
[27] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE, 2018, pp. 839–847.
[28] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-cam: Score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 24–25.
[29] H.G. Ramaswamy, et al., Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 983–991.
[30] Q. Zhang, Y.N. Wu, S.-C. Zhu, Interpretable convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8827–8836.
[31] H. Wang, R. Naidu, J. Michael, S.S. Kundu, Ss-cam: Smoothed score-cam for sharper visual feature localization, arXiv preprint arXiv:2006.14255 (2020).
[32] S. Shao, R. Yan, Y. Lu, P. Wang, R.X. Gao, Dcnn-based multi-signal induction motor fault diagnosis, IEEE Trans. Instrum. Meas. 69 (6) (2019) 2658–2669.
[33] J. Yu, X. Zhou, One-dimensional residual convolutional autoencoder based feature learning for gearbox fault diagnosis, IEEE Trans. Industr. Inf. 16 (10) (2020) 6347–6358.
[34] N. Bjorck, C.P. Gomes, B. Selman, K.Q. Weinberger, Understanding batch normalization, Advances in neural information processing systems 31 (2018).
[35] N. Murray, F. Perronnin, Generalized max pooling, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 2473–2480.
[36] L.C. Brito, G.A. Susto, J.N. Brito, M.A. Duarte, An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery, Mech. Syst. Signal Process. 163 (2022).
[37] J.R. Lee, S. Kim, I. Park, T. Eo, D. Hwang, Relevance-cam: Your model already knows where to look, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14944–14953.
[38] L. Gelman, K. Soliński, A. Ball, Novel higher-order spectral cross-correlation technologies for vibration sensor-based diagnosis of gearboxes, Sensors 20 (18) (2020) 5131.
[39] L. Gelman, K. Soliński, A. Ball, Novel instantaneous wavelet bicoherence for vibration fault detection in gear systems, Energies 14 (20) (2021) 6811.
[40] D. Neupane, J. Seok, Bearing fault detection and diagnosis using case western reserve university dataset with deep learning approaches: A review, IEEE Access 8 (2020) 93155–93178.
[41] S. Shao, S. McAleer, R. Yan, P. Baldi, Highly accurate machine fault diagnosis using deep transfer learning, IEEE Trans. Industr. Inf. 15 (4) (2018) 2446–2455.

Daoguang Yang received the B.S. degree in thermal energy and power engineering from Hunan University, China, in 2016, and his M.S. degree in power and mechanical engineering from Chongqing University, China, in 2019. He is currently a Ph.D. Candidate with the Department of Mechanical Engineering of Politecnico di Milano. His research interests include deep learning, fault diagnosis, prognostics and health management, reinforcement learning, and explainable artificial intelligence.

Hamid Reza Karimi is currently Professor of Applied Mechanics with the Department of Mechanical Engineering, Politecnico di Milano, Milan, Italy. From 2009-2016, he has been Professor of Mechatronics-Control Systems at University of Agder, Norway. His original research and development achievements span a broad spectrum within the topic of automation/control systems, and intelligence systems with applications to complex systems such as wind turbines, vehicles, robotics and mechatronics. Karimi is an ordinary Member of Academia Europa (MAE), Distinguished Fellow of the International Institute of Acoustics and Vibration (IIAV), Fellow of The International Society for Condition Monitoring (ISCM), Fellow of the Asia-Pacific Artificial Intelligence Association (AAIA), Member of Agder Academy of Science and Letters and also a member of the IFAC Technical Committee on Mechatronic Systems, the IFAC Technical Committee on Robust Control, and the IFAC Technical Committee on Automotive Control. Prof. Karimi is the recipient of the 2021 BINDT CM Innovation Award, the 2016-2021 Web of Science Highly Cited Researcher in Engineering, the 2020 IEEE Transactions on Circuits and Systems Guillemin-Cauer Best Paper Award, August-Wilhelm-Scheer Visiting Professorship Award, JSPS (Japan Society for the Promotion of Science) Research Award, and Alexander-von-Humboldt-Stiftung research Award, for instance.

Len Gelman, PhD, Dr. of Sciences (Habilitation) joined Huddersfield University as a Professor, Chair in Signal Processing/Condition Monitoring and Director of Centre for Efficiency and Performance Engineering, in 2017 from Cranfield University, where he worked as Professor and Chair in Vibro-Acoustical Monitoring since 2002. Len developed novel condition monitoring technologies for aircraft engines, gearboxes, bearings, turbines and centrifugal compressors. Len published more than 250 publications, 17 patents and is Co-Editor of 12 Springer books. He is Fellow of: British Institute of NDT, International Association of Engineers and Institution of Diagnostic Engineers, Executive Director, International Society for Condition Monitoring, Editor-in-Chief, International Journal of Engineering Sciences (SCMR), Chair, annual International Condition Monitoring Conferences, Honorary Co-Chair, annual World Congresses of Engineering, Co-Chair, International Congress COMADEM 2019 and Chair, International Scientific Committee of Third World Congress, Condition Monitoring. Len is Editorial Board member of the International Journal "Electronics", the International Journal "Sensors", the International Journal "Energies", the International Journal "Insight", the International Journal of Acoustics and Vibration, the International Journal of Prognostics and Health Management, the International Journal of Basic and Applied Sciences, the International Journal Vibration and Acoustics Research and the International Journal Signal, Image and Video Processing, Associate Editor of International Journal "Sensors".