Nonlinear Dyn
https://doi.org/10.1007/s11071-020-06014-6

ORIGINAL PAPER

Permutation entropy-based 2D feature extraction for bearing fault diagnosis

Mantas Landauskas · Maosen Cao · Minvydas Ragulskis

Received: 6 April 2020 / Accepted: 9 October 2020

Abstract Bearing fault diagnosis based on the classification of patterns of permutation entropy is presented in this paper. Patterns of permutation entropy are constructed by using non-uniform embedding of the vibration signal into a delay coordinate space with variable time lags. These patterns are interpreted, processed and classified by employing deep learning techniques based on convolutional neural networks. Computational experiments are used to compare the accuracy of classification with other methods and to demonstrate the efficacy of the presented early defect detection and classification method.

Keywords Permutation entropy · Convolutional neural network · Feature extraction · Fault detection

1 Introduction

Rolling and other types of bearings play an important role in different kind of machinery. Excess vibrations in bearings might induce other mechanical faults, increase the wear of devices, or even be a serious safety threat. However, a direct inspection of bearings is usually an unfeasible approach due to the complexity of the machinery, work safety problems or costs related to time constraints. Non-intrusive early fault diagnosis of bearings is usually based on intelligent computational analysis of experimental vibration data. Methods developed in this paper use vibration data in order to make informed decisions about the identification and classification of early defects in rotational ball bearings.

Machine learning (ML) algorithms have been successfully used for early fault detection in rotational bearings. Support vector machines (SVM) and artificial neural networks (ANN) are applied on features extracted from vibration data in [16] (these features are mainly basic statistical measures). The fault classification performance of this approach is evaluated by the confusion matrix [16]. Confusion matrices are a common tool for evaluating the fault classification quality. More extensive approach of fault detection in rotating machinery is discussed in [2]. Support vector classification (SVC) analysis with a number of other ML techniques are applied on features of centrifugal pump vibration data; the McNemar's test on confusion matrix is used to compare different ML methods and rank them in terms of their performance in [2]. A review of broader class of methods (Artificial intelligence (AI) in particular) for fault diagnosis can be found in [9,25]. A short

M. Landauskas · M. Ragulskis (✉)
Center for Nonlinear Systems, Kaunas University of Technology, Studentu 50-146, Kaunas, LT 51368, Lithuania
e-mail: minvydas.ragulskis@ktu.lt

M. Landauskas
e-mail: mantas.landauskas@ktu.lt

M. Cao
Department of Engineering Mechanics, Hohai University, Hohai 210098, China
e-mail: cmszhy@hhu.edu.cn

Published online: 28 October 2020

overview of the main ANN architectures used for fault diagnosis is discussed in [14].

Advanced ANN architectures are often utilized for bearing fault identification and classification [6,10,33]. Sparse representation can be used to enhance the functionality of fault diagnosis algorithms [41]. Fault detection could be considered as an optimization problem which can be solved by using dictionary-based learning [30,40]. Well-known methods as stochastic resonance have recent applications too [21,27]. Transient component extraction from vibration data is employed in [35].

It is commonly agreeable that various feature extraction techniques do serve as a common first step in fault identification algorithms. These techniques are usually classified as time, frequency, or time–frequency domain approaches [25,28]. Fault data might be below the detection threshold of the feature extraction technique especially in early stages of defect evolution. In such a scenario techniques like stochastic or vibrational resonance might be employed to amplify fault information [36]. For example, externally excited Duffing system can be used for weak feature extraction [37]. Next step is often classification of the features extracted. The most straightforward way would be to use the fault class as output of the model [8] but usually one-hot encoding the class variable is more accurate approach [20]. One-hot encoding of the output is not required if the class itself is representing the extent of the fault, in other words—when one deals with the regression task.

A wide class of fault detection algorithms are entropy based. Permutation entropy (PE) is a numeric tool for estimating how different patterns of orderings are distributed throughout a time series. PE is successfully applied in bearing fault diagnosis. The extraction of probability mass functions for PE from vibration data for anomaly detection is thoroughly presented in [19]. Multiscale PE (MPE) is also applied in fault detection ([23,42]), where the selection criterion for wavelet mother transform is used as a complexity measure. Improved version of MPE were introduced and fed to the classifier also [24]. However, direct application of PE for feature extraction may not avoid the side effects of noise in the data—prediction accuracy in such cases tend to be smaller [15]. Therefore, PE is often used as a fault detection technique only after initial feature extraction techniques. Such an approach is usually based on empirical mode decomposition [18,22,29]. Nevertheless, this strongly confirms that PE does play an important role in feature selection and/or classification problems.

Note that PE itself has inherent imperfections as a method: The amplitude information of the time series is not considered by PE. Thus other kinds of entropy measures like multiscale dispersion entropy (MDE) [43] or generalized refined composite multiscale sample entropy (GRCMSE) are often preferred [31]. For example, Ref. [34] presents fault diagnosis technique based on improved multiscale dispersion entropy (IMDE) which is used for feature extraction. Then max-relevance min-redundancy (mRMR) is used for selecting the most relevant features and the classifier is then trained. But then again, different inherent imperfections exist (like dependency on the length of the time series) in different approach. If one is trying to obtain universally optimal model for fault detection, the problem is ill-posed by nature. Some of the other entropy-based algorithms include fuzzy measure entropy [44], application of information entropy to multiscale morphological filtering [5].

PE itself is part of the novelty of the fault detection algorithm presented here. But if considering only the concept of entropy, it has served as an important feature extraction technique in other works. For example, wavelet packet-characteristic entropy combined with ANNs were presented in [39]. The entropy was computed for wavelet energy values for different nodes of decomposed signals.

The main objective of this paper is to propose a new approach towards early fault diagnosis based on PE. The proposed approach is also based on feature extraction techniques. However, the feature extraction step is directly related to properties of PE. Non-uniform embedding of the original time series into the delay coordinate space (with variable time lags) helps to construct the pattern of PE. Such an approach is in a stark contrast to data driven fault diagnosis methods based on feature extraction from raw data [32].

The paper is organized as follows: the concept of the proposed feature extraction method is presented in Sect. 2. Vibration data are described in Sect. 3. The enhancement of PE patterns is discussed in Sect. 4; image classification by convolutional neural networks (CNN) is described in Sect. 5. Results of computational experiments are presented in Sect. 6; concluding remarks are given in the last Section.

2 The concept of the feature extraction method

Let us consider a scalar time series $\{x_1, x_2, \ldots, x_N\}$. Non-uniform embedding to a 3-dimensional delay coordinate space with time delays τ_1 and τ_2 yields a trajectory matrix $y_i = \{y_i, y_{i+\tau_1}, y_{i+\tau_1+\tau_2}\}$, $i = 1, 2, \ldots, (N - \tau_1 - \tau_2)$. Each 3-dimensional row vector of the trajectory matrix is converted into a corresponding ordered permutation vector $\pi_i = \left\{\pi_1^{(i)}, \pi_2^{(i)}, \pi_3^{(i)}\right\}$, $i = 1, 2, \ldots, (N - \tau_1 - \tau_2)$ where $\pi_j^{(i)} \in \{1, 2, 3\}$, $j = 1, 2, 3$ are ranks of the elements in that vector. For example, a vector $\{0.1, -0.2, 0.3\}$ is converted to the permutation $\{2, 1, 3\}$. Thus there exist a total of $3! = 6$ different permutations (or classes of patterns) in the embedded trajectory matrix.

Ordered permutation vectors $\pi_i = \left\{\pi_1^{(i)}, \pi_2^{(i)}, \pi_3^{(i)}\right\}$, $i = 1, 2, \ldots, (N - \tau_1 - \tau_2)$ are used to compute the PE in the next step. Relative frequencies of the ordered permutations are used for that purpose:

$$p_i^{(\tau_1,\tau_2)} = \frac{n\,(\pi_i)}{N - \tau_1 - \tau_2},\ i = 1, 2, \ldots, 3!,$$

where $n\,(\pi_i)$ denotes number of occurrences of the permutation π_i in the sequence. Finally, the normalized PE is defined as:

$$E\,(\tau_1, \tau_2) = -\frac{1}{\ln 3!} \sum_{i=1}^{3!} p_i^{(\tau_1,\tau_2)} \ln p_i^{(\tau_1,\tau_2)}\ .$$

Note that $E\,(\tau_1, \tau_2)$ is the normalized classical Shannon entropy of ordered permutation vectors (numerical values of PE fit into the interval [0, 1]).

The PE pattern (a two-dimensional digital image of $E\,(\tau_1, \tau_2)$ is produced by varying time lags τ_1 and τ_2. The PE pattern (a two-dimensional digital image of $E\,(\tau_1, \tau_2)$) is produced by varying time lags tau_1 and tau_2 (used for the embedding of a single scalar time series). In other words, the algorithm presented in Fig. 1 transforms a scalar time series into a unique PE pattern. The upper bound of time delays in this paper is set to 100: $\tau_1, \tau_2 = 1, 2, \ldots, 100$. Such an approach

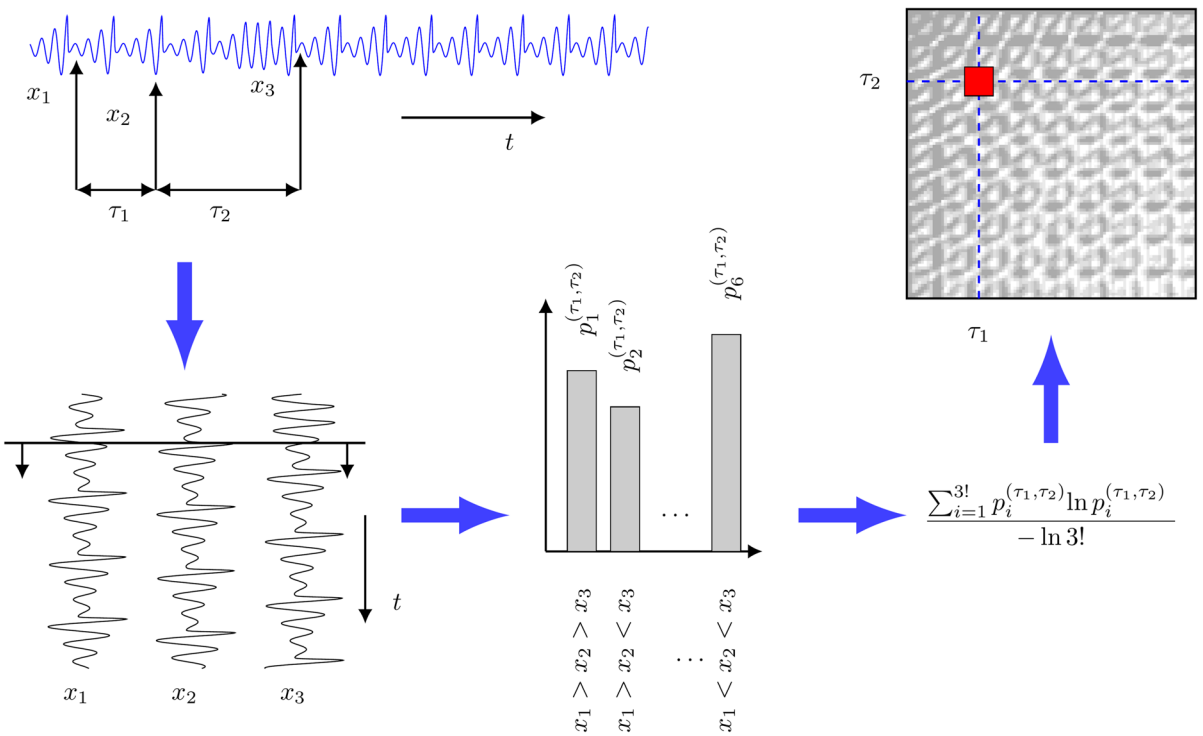

Fig. 1 A schematic diagram illustrating the computation of the PE pattern for a given time series. Non-uniform embedding transforms a scalar time series into a three-dimensional delay coordinate space (with time delays τ_1 and τ_2). Probabilities of different patterns of orderings yield a scalar value of PE at fixed τ_1 and τ_2. Varying time delays produce a PE surface which is visualized as a PE pattern

can be considered as a feature extraction technique. Different PE patterns are produced by different time series. We do raise a hypothesis that PE patterns produced by rolling bearings vibration data could serve as an effective fault diagnosis feature. Slight changes in the vibration data caused by developing defects result in spatial modifications of the PE pattern. PE pattern analysis transforms intelligent early fault identification and classification problem into an image recognition problem. Clearly, the changes in the PE patterns must be detected automatically. The objective of this paper is to use deep learning technology (convolutional neural networks) for PE pattern analysis.

3 Vibration data

Data used in this paper is obtained from the Case Western Reserve University Bearing Data Center database [1]. The database has become a standard test set for intelligent defect detection algorithms in rolling bearings systems. The data come from the experimental test rig comprised of a 2-hp motor, a torque sensor/encoder, a power meter, accelerometers, and electronic control unit. The faults are created by electrical discharge machining. There are four different health states of rolling bearings—the inner race fault, the outer race fault, the ball fault, and no fault. Vibration data is classified into fault data when the defect is on drive end (sampled at 12k), fault data when the defect is on drive end (sampled at 48k), fault data when the defect is on fan end, and normal baseline data. We do not use data sampled at 48k and do use all other subsets as they are all sampled at the same rate of 12k. Hierarchical view of the data set is depicted in Fig. 2.

Each type of vibration signal is then further grouped into four different loads and four different damage degrees. The four kinds of loads provided by the motor are 0, 1, 2, or 3 HP. The four kinds of damages are 0.1178, 0.3556, 0.5334, and 0.7112 mm, respectively. The bearing type is SKF 6205. The fault location is used as the class of the defect in computational experiments in this paper.

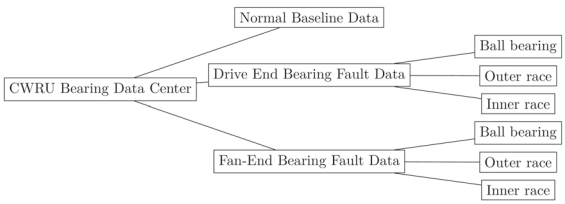

Fig. 2 Hierarchical view of the fault data classification in Case Western Reserve University Bearing Data Center database

4 The enhancement of PE patterns

Some typical PE patterns produced by vibration data representing the inner race fault, the outer race fault, the ball fault and no fault are shown in the first row of Fig. 3 (the length of the observation window is limited to 6000 data points). Clearly, the extracted features in PE images can be enhanced by unifying different color ranges and improving the contrast of the patterns.

The pattern enhancement technique is set an integral part of the preprocessing step of the PE pattern analysis strategy. First of all, the lower quartile Q_1, the median Q_2 and the upper quartile Q_3 are computed for each individual PE pattern. Then, the color intensity of each pixel $E(\tau_1, \tau_2)$ in the PE pattern is transformed into one of 4 discreet values according to the following rule:

- $E^{(enh)}(\tau_1, \tau_2) = 0$ if $E(\tau_1, \tau_2) < Q_1$;
- $E^{(enh)}(\tau_1, \tau_2) = 1$ if $Q_1 \leq E(\tau_1, \tau_2) < Q_2$;
- $E^{(enh)}(\tau_1, \tau_2) = 2$ if $Q_2 \leq E(\tau_1, \tau_2) < Q_3$;
- $E^{(enh)}(\tau_1, \tau_2) = 3$ if $E(\tau_1, \tau_2) > Q_3$.

Histograms of PE patterns (gray dashed lines depict the lower quartile, the median and the upper quartile accordingly) are shown in the middle line of Fig. 3. Enhanced PE patterns are depicted in the third row of Fig. 3. Clearly, enhanced PE patterns do reveal more pronounced features of the signal—what is helpful for the subsequent image classification task.

5 Image classification by CNN

As mentioned previously, the proposed approach to intelligent fault diagnosis transforms the problem of vibratory signal analysis into a problem of the classification of PE patterns. The CNN input is comprised of a set of two-dimensional digital images, while the output is a set of one-hot encoded vectors corresponding to each class of vibration signal. One-hot encoding is preferred to direct id of a class because the classes are not ranked. We prevent the CNN from incorporating such ranking into the learning process.

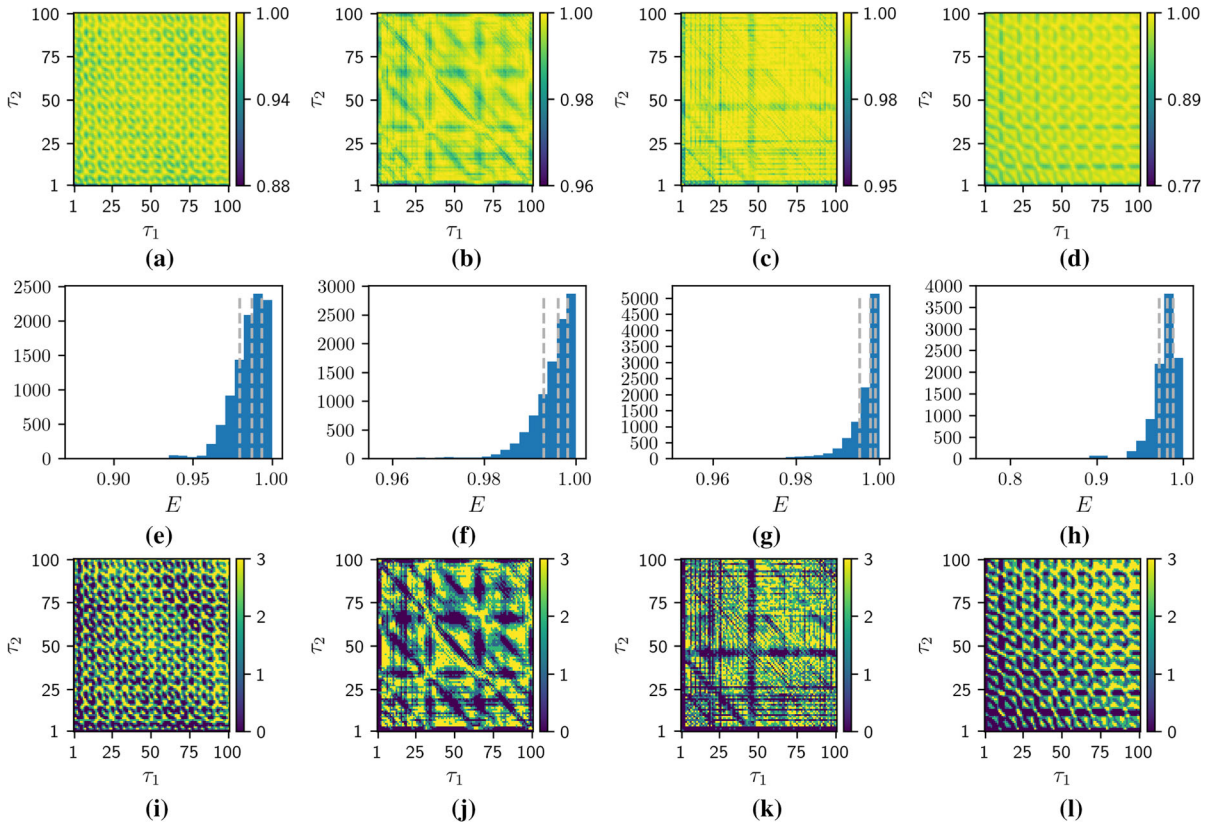

Fig. 3 A schematic diagram illustrating the enhancement of PE patterns. Typical PE patterns are presented in the first row. Histograms of PE patterns (gray dashed lines depict the lower quartile, the median and the upper quartile accordingly) are shown in the middle row. Enhanced PE patterns are illustrated in the third row. Parts **a**, **e** and **i** represent a fault on the inner race; **b**, **f** and **j**—a fault on a ball; **c**, **g** and **k**—a fault on the outer race; **d**, **h** and **l**—normal baseline data

A typical structure of a CNN dedicated for image recognition is employed in this paper (Table 1). Repetitive training of different CNN variants is used for the estimation of average prediction errors. For deeper understanding of different layers in CNN, one can refer to [26].

The architecture of the CNN is optimized by varying the number of filter groups (2–3) and the size of filters in convolutional layers. Initial experiments show that 4×4 convolutional filters do produce a better classification output. As for other parameters, *adadelta* optimizer was selected for the training, binary cross-entropy acted as the loss function and categorical accuracy served as accuracy metrics.

Feature extraction is the first phase of the network. A PE pattern is put through a set of convolutional filters (Conv 1) first; small-scale features do activate different filters. Consequently, the batch normalization (BN) layer follows. Normalization of data makes the learning process faster. Rectified linear unit (ReLU) preserves positive values and acts as an activation of features. Max pooling downsamples the images and thus reduces the total number of parameters of the network. A set of all mentioned layers is called a filter. Having more than one such filter enables CNN to learn multi-scale features. The last filter may not have MP layer as there are no more convolutional layers which would require an introduction of a number of parameters again.

Next, the classification stage starts. The output is flattened and fed to the fully connected layer. Softmax layer follows right after it in order to improve the learning rate of the classification task. Finally, the id of the class exits at the output.

Table 1 The architecture of the CNN used for the classification of PE patterns

Group name	Layer name	Layer type	Description
Input	Input	Image Input	100 × 100 × 1 images with 'zerocenter' normalization
Filter 1	Conv 1	Convolution	8 4 × 4 convolutions with stride [11] and padding [0000]
	BN 1	Batch normalization	Batch normalization
	ReLU 1	ReLU	ReLU
	MP2D 1	Max Pooling	2x2 max pooling with stride [22] and padding [0000]
Filter 2	Conv 2	Convolution	16 4 × 4 convolutions with stride [11] and padding [0000]
	BN 2	Batch Normalization	Batch normalization
	ReLU 2	ReLU	ReLU
	MP2D 2	Max Pooling	2 × 2 max pooling with stride [22] and padding [0000]
Filter 3	Conv 3	Convolution	32 4 × 4 convolutions with stride [11] and padding [0000]
	BN 3	Batch Normalization	Batch normalization
	ReLU 3	ReLU	ReLU
Classification	FC	Fully Connected	Fully connected layer of 4 neurons
	Softmax	Softmax	Softmax
	Class	Classification output	One hot encoded vector

Preprocessing of the digital images (conversion from vibrational data to PE images) were carried out by batch scripts written in C. All other computations were coded in Python. CNNs were trained with Keras and TensorFlow libraries using GPU hardware acceleration (NVIDIA GeForce GTX 1070).

6 Experiments with the vibrational data

Each class of bearing fault data contains vibration data time series of different lengths. As mentioned previously, we use time series segments containing 6000 data points for producing a PE pattern. In other words, all vibration data time series are segmented into non-overlapping segments. Segmentation is performed in order to increase the amount of data for more efficient learning process of CNN. Four different classes (based on the defect type) are assigned and marked by C1 through C4. The number of different PE patterns in each class are:

1. The defect on the inner race (C1)—1520 PE patterns.
2. The defect on the ball (C2)—1520 PE patterns.
3. The defect on the outer race (C3)—2940 PE patterns.
4. The normal baseline data (C4)—560 PE patterns.

The training data set is filled by 392 PE patterns from each class in order to balance class sizes in the training set (making a total of 1568 PE patterns). This corresponds to 70% of the all available PE patterns from the smallest class C4. A separate set of randomly chosen 84 PE patterns from each class are used for the validation. Training is repeated for 10 times in order to compare the classification quality as well as the training performance. Training and validation sets are randomly selected again at the beginning of each epoch in order to reduce the classification error. Such kind of learning is similar to the so called bagging approach and it is a known solution for imbalanced datasets too [11].

Note that CNN filters small-scale features in the original enhanced PE pattern in the first stage. Later these features are joined to higher scale patterns by the filters in further layers (Fig. 4).

Table 2 shows the minimum, average, and maximum of the number of patterns for confusion matrices for fault classification when the unified set of PE patterns is constructed from vibrational signals recorded both at the drive end and at the fan end of the test rig at different loads. Confusion matrix is constructed as follows. We start at zero-valued matrix. If PE pattern corresponding to class C1 is classified as being of class C1, the element at the first row and the first column is increased by one. If at the other epoch a pattern of class C2 is classified as of being of class 3, the element at the

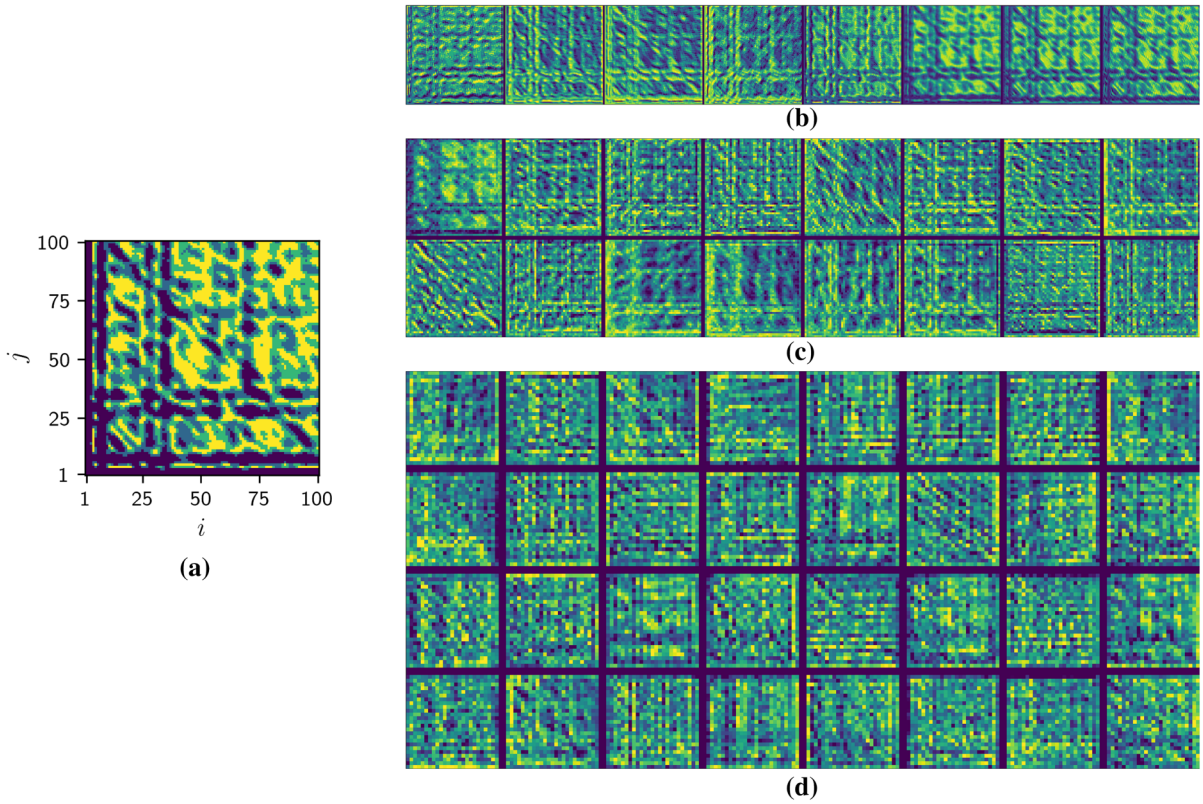

Fig. 4 A schematic diagram illustrating the activation at different locations inside the CNN. A typical PE pattern (from class C3) is shown in part **a**. The output from the first layer of convolution filters in depicted in part **b**. The output from the second layer of convolution filters is shown in **c**. The result of the last layers of convolution filters is demonstrated in **c**

second row and the third column is increased by one. In the perfect scenario, when all the patterns are classified correctly, we observe the number of samples in the test set on the main diagonal while all the other elements are zero. Since we repeat the training process for 100 time (each time randomly selecting training and testing samples), we show the summary of confusion matrices. Thus each cell has three numbers: the minimum, average and maximum of the values of that particular cell observed during all 100 epochs.

In the experiments below we talk about prediction (or rather classification) accuracy which is the ratio between successfully identified classes of images and total number of images. Thus we obtain it by summing the main diagonal of confusion matrix and divide the result by total number of images used for training or testing.

In order to test the functionality of the proposed method against noise, all computational experiments are repeated with the same dataset but with a different amount of the Gaussian noise added to the vibrational signals. The signal to noise ratio (SNR) is used to quantify the amount of the additive noise. SNR is described by the ratio of the averaged values of the squares of the signal (S) and the noise (N): $SNR = \frac{E(S^2)}{E(N^2)}$.

Computational experiments show that the average number of the successful defect detections is between 130.93 and 141.11 (out of 168 PE patterns) when the CNN comprises 3 convolutional layers and no noise is added to the vibrational signal (Table 2). The sensitivity measure ranges from 0.779 to 0.840 (Table 2).

A small amount of the additive noise ($SNR = 10$) changes the range of the sensitivity measure (0.730 to 0.809). Such a change is not unexpected because the noise can change the distribution of permutations of the vibrational signal. A larger amount of the additive noise ($SNR = 1$) distorts this interval even further (0.730–0.807). The actual value of the acceptable level

Table 2 Summary of confusion matrices for the classification of PE patterns

No noise	C1	C2	C3	C4	Sensitivity	Specificity	Accuracy
(a) Results produced by 2 convolutional layers							
C1	0 **70.43** 168	0 32.92 168	0 27.40 168	0 37.25 168	0.419	0.831	**0.468**
C2	0 28.91 168	0 **81.79** 168	0 22.02 168	0 35.28 168	0.487	0.815	
C3	0 31.18 168	0 31.63 168	0 **68.38** 168	0 36.81 168	0.407	0.862	
C4	0 25.20 168	0 28.56 168	0 20.20 168	0 **94.04** 168	0.560	0.783	
$SNR = 10$							
C1	0 **62.05** 168	0 31.73 168	0 29.88 168	0 44.34 168	0.369	0.871	**0.459**
C2	0 21.85 168	0 **81.95** 168	0 25.42 168	0 38.78 168	0.488	0.816	
C3	0 24.55 168	0 31.20 168	0 **70.04** 168	0 41.85 168	0.419	0.840	
C4	0 18.70 168	0 30.02 168	0 25.20 168	0 **94.08** 168	0.560	0.752	
$SNR = 1$							
C1	0 **68.60** 168	0 39.80 168	0 27.57 168	0 32.03 168	0.408	0.863	**0.464**
C2	0 23.52 168	0 **87.34** 168	0 25.22 168	0 31.92 168	0.520	0.768	
C3	0 23.68 168	0 40.36 168	0 **71.90** 168	0 32.06 168	0.428	0.845	
C4	0 21.84 168	0 36.96 168	0 25.20 168	0 **84** 168	0.500	0.810	
(b) Results produced by 3 convolutional layers							
C1	0 **141.11** 168	0 10.08 168	0 5.05 168	0 11.76 168	0.840	0.910	**0.812**
C2	0 15.12 168	0 **136.08** 168	0 5.04 168	0 11.76 168	0.810	0.940	
C3	0 15.19 168	0 10.12 168	0 **130.93** 168	0 11.76 168	0.779	0.970	
C4	0 15.12 168	0 10.08 168	0 5.04 168	0 **137.76** 168	0.820	0.930	
$SNR = 10$							
C1	0 **122.64** 168	0 12.00 168	0 18.81 168	0 14.55 168	0.730	0.950	**0.780**
C2	0 8.40 168	0 **130.91** 168	0 15.25 168	0 13.44 168	0.780	0.936	
C3	0 8.46 168	0 10.15 168	0 **135.95** 168	0 13.44 168	0.809	0.902	
C4	0 8.40 168	0 10.08 168	0 15.12 168	0 **134.40** 168	0.800	0.918	
$SNR = 1$							
C1	0 **127.67** 168	0 20.16 168	0 13.45 168	0 6.72 168	0.760	0.920	**0.764**
C2	0 13.44 168	0 **135.51** 168	0 12.33 168	0 6.72 168	0.807	0.880	
C3	0 13.50 168	0 20.24 168	0 **127.54** 168	0 6.72 168	0.759	0.926	
C4	0 13.44 168	0 20.16 168	0 11.76 168	0 **122.64** 168	0.730	0.960	

The unified set of PE patterns is constructed from vibrational signals recorded both at the drive end and at the fan end of the test rig at different loads. The three numbers in each cell do correspond to the minimum, average and maximum of the number of patterns. Average values for the sensitivity, the specificity and the accuracy are depicted in the last columns. The test sample contains 168 randomly selected PE patterns. Training is repeated for 100 times
Mean values are shown in bold

of the sensitivity measure is subjective. Nevertheless, the results of the computational experiments show that the additive noise reduces the sensitivity just by second digit after the decimal point at $SNR = 1$.

Note that this effect is less pronounced for the specificity measure. The main reason for such an effect is based on the fact that the one-against-all approach used in the multiclass classification problem does evaluate

Table 3 Summary of confusion matrices for the classification of PE patterns

	C1	C2	C3	C4	Sensitivity	Specificity	Accuracy
(a) Results produced by 2 convolutional layers							
C1	0 **27.05** 48	0 7.48 48	0 5.79 48	0 7.68 48	0.564	0.823	**0.586**
C2	0 10.19 48	0 **26.66** 48	0 4.6 48	0 6.55 48	0.555	0.864	
C3	0 9.28 48	0 8.7 48	0 **22.73** 48	0 7.29 48	0.474	0.909	
C4	0 5.96 48	0 3.36 48	0 2.68 48	0 **36.** 48	0.750	0.851	
(b)Results produced by 3 convolutional layers							
C1	0 **46.55** 48	0 0.48 48	0 0.49 48	0 0.48 48	0.970	0.990	**0.972**
C2	0 0.48 48	0 **46.52** 48	0 0.52 48	0 0.48 48	0.969	0.993	
C3	0 0.52 48	0 0.48 9	0 **46.52** 48	0 0.48 48	0.969	0.990	
C4	0 0.48 48	0 0 0	0 0.48 48	0 **47.04** 48	0.980	0.990	

The data used corresponds to the case when there is 1 horsepower (hp) load on the system. The three numbers in each cell correspond to the minimum, average and maximum of the number of patterns. Average values for the sensitivity, the specificity and the accuracy are depicted in the last columns. The test sample contains 48 randomly selected PE patterns. Training is repeated for 100 times
Mean values are shown in bold

Table 4 Summary of confusion matrices for the classification of PE patterns

	C1	C2	C3	C4	Sensitivity	Specificity	Accuracy
(a) Results produced by 2 convolutional layers							
C1	0 **13.36** 24	0 4.82 24	0 3.94 24	0 1.88 24	0.557	0.873	**0.648**
C2	0 3.8 24	0 **14.47** 24	0 4.49 24	0 1.24 24	0.603	0.844	
C3	0 4.12 24	0 6.14 24	0 **12.26** 24	0 1.48 24	0.511	0.876	
C4	0 1.2 24	0 0.24 24	0 0.48 24	0 **22.08** 24	0.920	0.936	
(b) Results produced by 3 convolutional layers							
C1	0 **23.28** 24	0 0.01 1	0 0.56 24	0 0.15 15	0.970	0.995	**0.974**
C2	0 0.17 3	0 **23.15** 24	0 0.67 24	0 0.01 1	0.965	0.994	
C3	0 0.22 19	0 0.42 24	0 **23.32** 24	0 0.04 4	0.972	0.980	
C4	0 0 0	0 0 0	0 0.24 24	0 **23.76** 24	0.990	0.997	

The data used corresponds to the case when there is 1 hp load on the system and only fan end sensors are considered. The three numbers in each cell correspond to the minimum, average and maximum of the number of patterns. Average values for the sensitivity, the specificity and the accuracy are depicted in the last columns. The test sample contains 24 randomly selected PE patterns. Traini ng is repeated for 100 times
Mean values are shown in bold

the assignment of the class of interest, but does not consider if the classification is correct with respect of all other remaining classes.

If the experiment is repeated only for the data corresponding to the system with 1 hp load it can be observed that on average more defects are classified correctly (note diagonal of the confusion matrix in Table 3). And again: 3 convolutional layers are needed to achieve that.

Similar accuracy is achieved when only the fan end bearings are considered (Table 4).

Figure 5 shows detailed dynamics of the training performance of the CNN for training and testing samples when the unified set of PE patterns is constructed from vibrational signals recorded both at the drive end and at the fan end of the test rig at different loads. The fault classifications accuracy here is worse compared to two other cases which will be discussed below. It is natural to expect that if one considers bearings positioned on different places on the rotating shaft, they might experience different vibrational dynamics. This

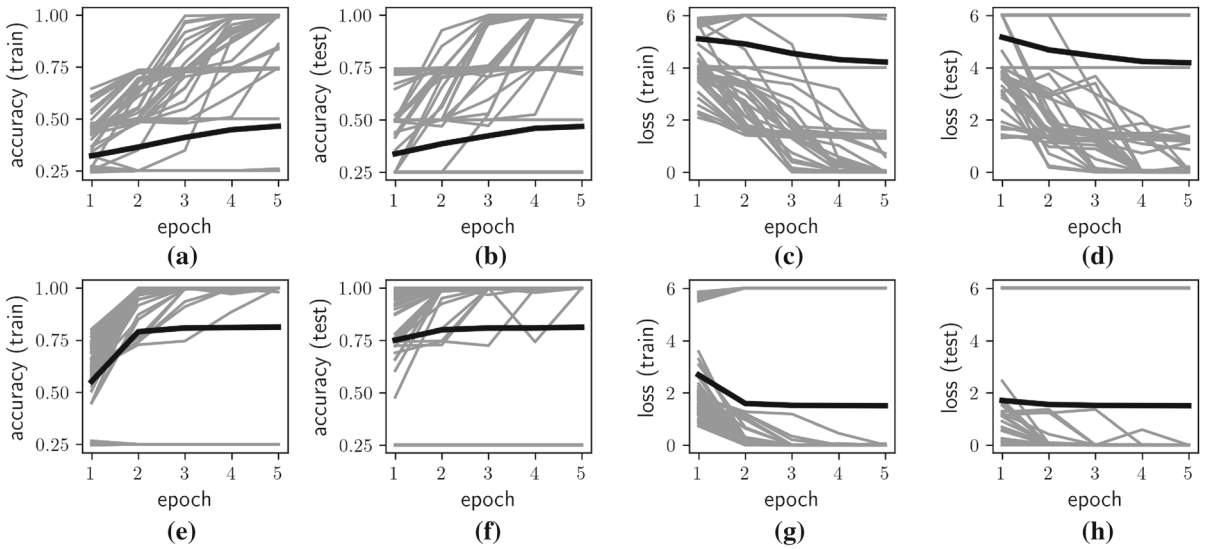

Fig. 5 Detailed dynamics of the training performance of the CNN for training and testing samples. The unified set of PE patterns is constructed from vibrational signals recorded both at the drive end and at the fan end of the test rig at different loads. Parts **a**–**d** correspond to CNN with two convolutional layers; **e**–**h**—to CNN with 3 convolutional layers. Each case is repeated 100 times and the blach thick line shows average training performance. Simultaneous flattening of these two parameters usually signal an overfitting scenario

heuristic argument complies with lower fault prediction accuracy.

By looking deeper into the training process (Fig. 5) one can observe that essentially after every training epoch accuracy is slowly improving if only 2 convolutional layers are used. More training cycles could be performed but due to low accuracy the end result might still be poor. Completely different training dynamics is observed when 3 convolutional layers are used. At the second epoch the accuracy increases significantly but then improves just by a small amount if at all. Same effect is present at the training and at the testing samples which are primarily influenced by rather distinctive patterns throughout different classes of the signals.

The network reaches optimal performance both in the set of the training and the testing samples at the third epoch (Fig. 5). This suggests that optimal training time is three epochs for a CNN comprising three convolutional layers. Also, this allows to avoid the overfitting.

Numerical experiments showed that when the data used corresponds to the case when there is 1 hp load on the system, the fault prediction accuracy is significantly higher (Fig. 6). Heuristically thinking, the system investigates less different classes of PE images which enables CNN to train more effectively provided the same system architecture is used.

We subset the training and testing sets further by considering only the case when there is 1 hp load on the system and only fan end sensors are considered (Fig. 7). We do not observe a clear visible improvement in average training accuracy although note that it was quite high at the first place.

The performance of the proposed method is also described by the sensitivity and the specificity. The sensitivity is the ratio between the number of correct assignments to a fixed class and the total number of items in that class. The specificity is the ratio between the number of correct assignments to the other class and the total number of specimens belonging to that class. In case of the multiple class scenario (defect locations and/or their sizes) one-against-all approach for two classes is considered—the fixed class (for which the sensitivity and the specificity are computed) and the joint set of the remaining classes.

As mentioned previously, computational experiments are performed using the standard Case Western University dataset [1]. That allows straightforward and objective comparisons with other methods. However, it is important to test the functionality of the proposed approach not only on a machine examined in a laboratory environment. If the machine would be placed into an industrial environment, the measured vibration sig-

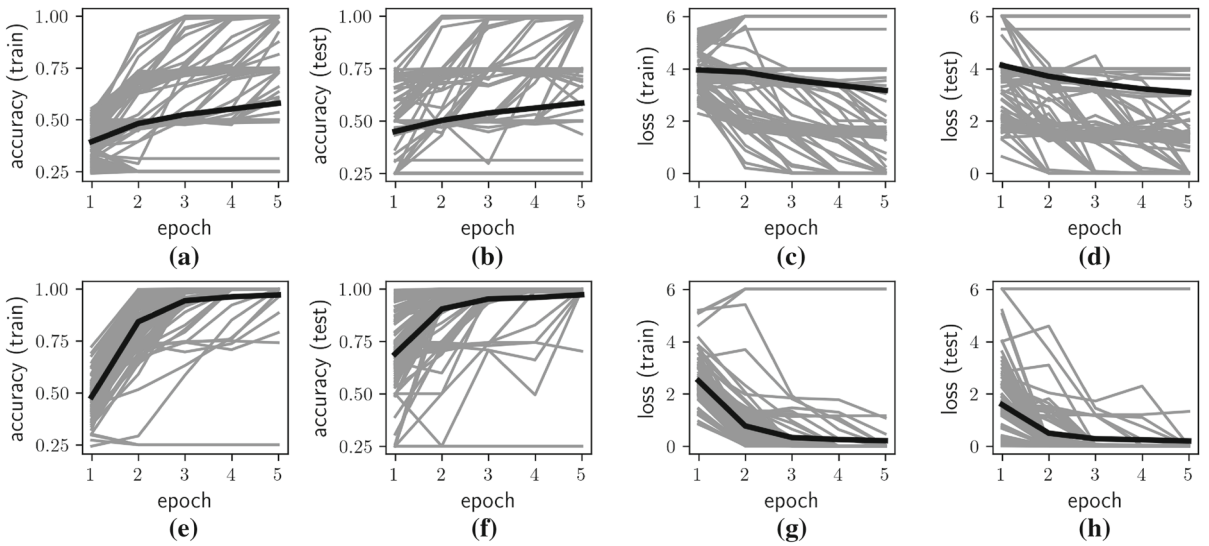

Fig. 6 Detailed dynamics of the training performance of the CNN for training and testing samples. The data used corresponds to the case when there is 1 hp load on the system. Parts **a**–**d** correspond to CNN with 2 convolutional layers; **e**–**h**—to CNN with 3 convolutional layers. Each case is repeated 100 times and the black thick line shows average training performance. Simultaneous flattening of these two parameters usually signal an overfitting scenario

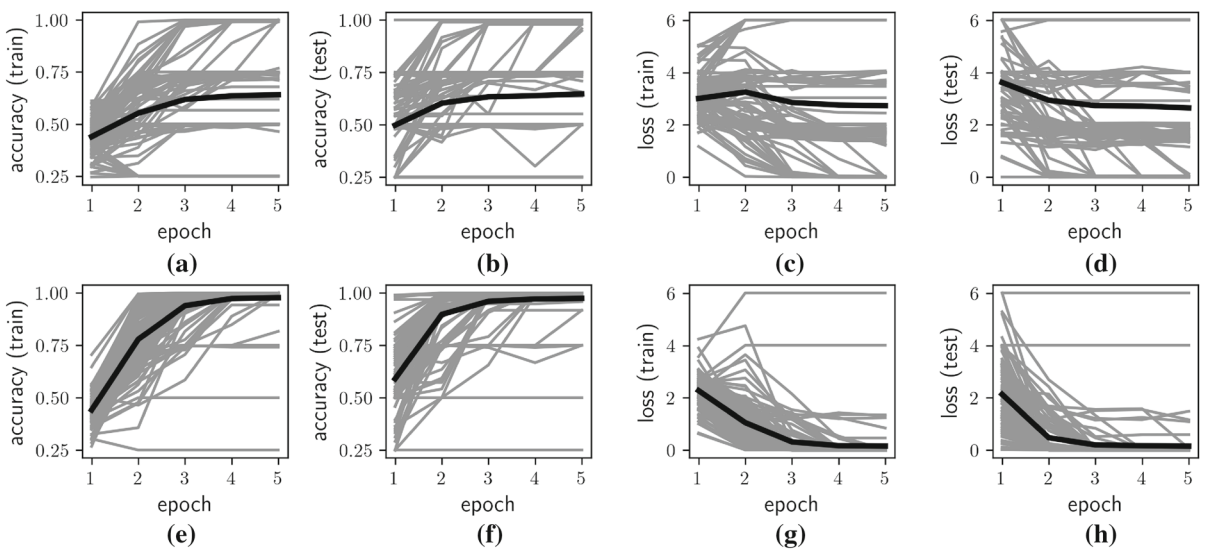

Fig. 7 Detailed dynamics of the training performance of the CNN for training and testing samples. The data used corresponds to the case when there is 1 hp load on the system and only fan end sensors are considered. Parts **a**–**d** correspond to CNN with 2 convolutional layers; **e**–**h**—to CNN with 3 convolutional layers. Each case is repeated 100 times and the blach thick line shows average training performance. Simultaneous flattening of these two parameters usually signal an overfitting scenario

nals would be contaminated by the unavoidable additive external noise. Therefore, testing the robustness of the proposed technique to the external noise is an important part of the method validation.

Additional computational experiments are performed when classes are directly linked to the size of the defect:

1. The normal baseline data (D1)—560 PE patterns.
2. The defect of 0.007 in. (D2)—2400 PE patterns.
3. The defect of 0.014 in. (D3)—1500 PE patterns.
4. The defect of 0.021 in. (D4)—1920 PE patterns.
5. The defect of 0.028 in. (D5)—160 PE patterns.

Table 5 Summary of confusion matrices for the classification of PE patterns

	D1	D2	D3	D4	D5	Sensitivity	Specificity	Accuracy
(a) Results produced by 2 convolutional layers								
D1	0 **19.18** 48	0 3.89 48	0 5.52 48	0 7.31 48	0 12.1 48	0.400	0.886	**0.414**
D2	0 6. 48	0 **15.38** 48	0. 6.54 48	0 7.99 48	0 12.09 48	0.320	0.911	
D3	0 5.6 48	0 5.39 48	0 **15.08** 48	0 8.78 48	0 13.15 48	0.314	0.882	
D4	0 5.75 48	0 4.39 48	0 5.85 48	0 **20.54** 48	0 11.47 48	0.428	0.843	
D5	0 4.57 48	0 3.5 48	0 4.72 48	0 6.13 48	0 **29.08** 48	0.606	0.746	
(b) Results produced by 3 convolutional layers								
	D1	D2	D3	D4	D5	Sensitivity	Specificity	Accuracy
D1	0 **40.2** 48	0 1.54 48	0. 1.74 48	0 2.04 48	0 2.48 48	0.838	0.954	**0.838**
D2	0 2.95 48	0 **37.76** 48	0. 2.12 48	0 2.43 48	0 2.74 48	0.787	0.979	
D3	0 2.48 48	0 1.15 48	0 **38.85** 48	0 2.96 48	0 2.56 48	0.809	0.964	
D4	0 1.96 48	0 0.9 48	0 1.54 48	0 **41.67** 48	0 1.93 48	0.868	0.951	
D5	0 1.44 48	0 0.48 48	0 1.44 48	0 1.92 48	0 **42.72** 48	0.890	0.949	

The unified set of PE patterns is constructed from vibrational signals recorded both at the drive end and at the fan end of the test rig at different loads. Classes D1–D5 are directly related to the size of the defect. The three numbers in each cell correspond to the minimum, average and maximum of the number of patterns. Average values for the sensitivity, the specificity and the accuracy are depicted in the last columns. The test sample contains 48 randomly selected PE patterns. Training is repeated for 100 times
Mean values are shown in bold

Such definition of classes helps to answer the question on what is the minimum fault extent that can be diagnosed by the proposed technique. Now the training set contains 112 digital images and the testing set comprises 48 images. The results of computational experiments are presented in Table 5.

The average sensitivity measure of the successful defect detection (when classes correspond to the depth of the defect) fits into the interval (0.787, 0.890) when 3 convolutional layers are used (Table 5). The best average sensitivity was obtained for the class D5 which corresponds to the largest defect. The results of computational experiments show that the successful detection of smaller defects (classes D2 and D3) occurs less often. The average values in the confusion matrix shows that these defects are more often assigned to any of the remaining classes including D1 and D5.

7 Discussion

As pointed out by other researchers [4], fault diagnosis might be classified into four stages. At first the fact about the existence of the fault is determined. Then the location is found and the degree of the degradation is estimated. The last step is the prediction of remaining useful life or other characteristics of future dynamics. Our paper covers the qualitative fault detection, i.e., the method determines whether the fault exist and shows where it is located. So the results have twofold purpose.

Computational results demonstrate that our approach produces good results even if in the most general distribution of classes. That allows to use standard approach for the intelligent defect detection, identification and localization. Initially one could use a small number of general classes to detect the existence of the defect (the location and the type of the defect is not known). Later, an adaptation of specifically suited classes can allow to pinpoint the defect and its severity.

Comparative study of various methods for bearing fault detection is discussed in [3]. Although the discussion is only about traditional techniques such as temporal analysis or applications as Fourier transform. Machine learning methods are not covered there.

Applications of CNN to predict or classify rolling bearing faults have been extensively studied in scientific literature. Among the most straightforward approaches is construction of 2D vibration images directly from vibrational signal [12]. More sophisticated method which proves to be highly efficient is to classify spectral images of a segmented vibrational signal using CNN [13,17,38].

Table 6 Comparison of different methods for detecting faults in rotating machinery

Method	Classes considered	Comments
Temporal analysis; Fourier transform [3]	Vibrational data is analyzed for abnormal behavior	Able to detect abnormal changes, inability to detect fault location
Scalogram + CNN with batch normalization [13]	Raw data, no feature extraction, batch normalization, 10 classes	Accuracy 0.99, more layers in convolutional network, unclear performance for noisy signals
Spectral images of a segmented vibrational signal + CNN [17]	Data from a custom test bench, 5 classes (defect location)	Accuracy: minimum 0.969 depending on operating speed, 3 convolutional layers
Resonance in the Duffing system after inputting vibrational signal [37]	Vibrational data is analyzed for abnormal behavior	Excellent amplification effect for fault characteristic
Feature extraction using multiscale dispersion entropy (MDE) + extreme learning machine (ELM) [34]	Same data, 10 classes (defect location + size)	Accuracy: 0.906, suitable for noisy signals
Fully connected winner-take-all (FC-WTA) autoencoder based [20]	Same data, 10 classes (defect location + size)	Very robust to noise

The data set used in this research was not balanced. We dealt with it by sampling same number of observations from different classes of fault data. As an alternative one can use recently introduced upsampling technique based on generative adversarial neural networks combined with CNN [7]. Near-real fault data samples can be generated by using that technique.

In this paper we define the classes C1-C4 of defects with respect of defect location and its presence as well as the calsses D1-D5 of defects with respect of defect size and its presence. The data set is extensively used in research but there is no unified way of defining classes for defect detection. One of the closest approach is used in [34] where the classes (10 in total) are almost identical except defects at the same locations but with different diameters are not unified into a single class as in our case. Same 10 classes of fault data and same test data are used in [13,20,38]. It can be noted that MDE and RCMDE supersedes multiscale PE in that approach. But we use rather different approach to construct 2D digital images as features employing PE.

Depending of the nature of the vibrational data of different defects and the quality of feature extraction technique used the unification into more general classes might be useful or not for the classifier. Our rationale of unifying a group of vibrational signals of defects in same location but of different extents to a single class is as follows. Heuristically if the defect is more pronounced the feature extraction technique will be able to better represent its class and/or presence. Features of different classes not necessary are close to each other. In other words, we leave out the task of evaluating the extent of the defect. If is very logical to do that in other stages of fault diagnosis as mentioned before.

Papers which use data from custom test setup of rotating machinery also define combined defect (at inner and outer race) as separate class [17]. The comparison of different methods for detecting faults in rotating machinery (using the same set of data) is presented in Table 6.

As for the distinction between data-driven and model-based fault detection techniques [9], our approach could be considered data-driven as it deals with 2D features extracted from vibrational data. Usage of CNN classifies that further into deep learning-based techniques. Our approach is robust to the additive noise and works well when the classes span over different motor loads, different sensors, different sizes of the defect. Moreover, our approach can be used both as a coarse-grained method to detect the existence of a fault and as a fine-grained method to identify the exact location and the severity of the defect.

8 Conclusions

A rolling bearing fault identification and classification method based on PE patterns is presented in this paper. This method uses the feature extraction functionality from the raw data. The recognition and classification

of PE patterns is employed by using deep learning techniques based on CNN.

The architecture of the CNN is optimized in order to solve the defined classification problem. Computational experiments demonstrate that 3 convolutional layers in the CNN do result into a high classification accuracy according to the type of the defect. All numerical computations are carried out with the standard vibration signal data set.

The results of computational experiments do show that 4 epochs are completely sufficient for good training performance of the CNN. If the vibration data would be more complex (would contain more fault sources, would be heavily contaminated by external noise, etc.), three dimensional non-uniform time series embedding could be insufficient to construct representative PE patterns. The construction of PE patterns in higher dimensions of the delay coordinate space (involving a larger number of different time lags) remains a definite objective of future research.

Subsetting the data by leaving only the vibrations from one end of the system or considering only 1 load (for example 1 hp) helps the CNN to classify the PE patterns better. However, it is demonstrated that the proposed technique can work well with a larger number of classes—even if those classes do correspond to different physical features of the bearing(s) and the machine. The ability of the proposed technique to perform successful classification of different faults under the presence of the additive noise makes it an attractive tool for solving complex rolling bearing fault identification and diagnosis problems.

Acknowledgements This research was supported by the Research, Development and Innovation Fund of Kaunas University of Technology (project acronym DDetect). This research is also partially supported by International Science and technology cooperation project (Grant No.: BZ2018022).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Case western reserve university bearing data center. https://csegroups.case.edu/bearingdatacenter/home
2. Azadeh, A., Saberi, M., Kazem, A., Ebrahimipour, V., Nourmohammadzadeh, A., Saberi, Z.: A flexible algorithm for fault diagnosis in a centrifugal pump with corrupted data and noise based on ANN and support vector machine with hyper-parameters optimization. Appl. Soft Comput. **13**(3), 1478–1485 (2013)
3. Boudiaf, A., Moussaoui, A., Dahane, A., Atoui, I.: A comparative study of various methods of bearing faults diagnosis using the case western reserve university data. J. Fail. Anal. Prev. **16**(2), 271–284 (2016)
4. Chen, X., Wang, S., Qiao, B., Chen, Q.: Basic research on machinery fault diagnostics: past, present, and future trends. Front. Mech. Eng. **13**(2), 264–291 (2018)
5. Cui, L., Wang, J., Ma, J.: Early fault detection method for rolling bearing based on multiscale morphological filtering of information-entropy threshold. J. Mech. Sci. Technol. **33**(4), 1513–1522 (2019)
6. Eren, L., Ince, T., Kiranyaz, S.: A generic intelligent bearing fault diagnosis system using compact adaptive 1d CNN classifier. J. Signal Process. Syst. **91**(2), 179–189 (2019)
7. Gao, S., Wang, X., Miao, X., Su, C., Li, Y.: Asm1d-gan: an intelligent fault diagnosis method based on assembled 1d convolutional neural network and generative adversarial networks. J. Signal Process. Syst. **91**(10), 1237–1247 (2019)
8. Gunerkar, R.S., Jalan, A.K., Belgamwar, S.U.: Fault diagnosis of rolling element bearing based on artificial neural network. J. Mech. Sci. Technol. **33**(2), 505–511 (2019)
9. Hamadache, M., Jung, J.H., Park, J., Youn, B.D.: A comprehensive review of artificial intelligence-based approaches for rolling element bearing phm: Shallow and deep learning. JMST Adv. **1**(1), 125–151 (2019)
10. He, M., He, D.: A new hybrid deep signal processing approach for bearing fault diagnosis using vibration signals. Neurocomputing (2019)
11. Hido, S., Kashima, H., Takahashi, Y.: Roughly balanced bagging for imbalanced data. Stat Anal Data Min: ASA Data Sci. J. **2**(5–6), 412–426 (2009)
12. Hoang, D.T., Kang, H.J.: Convolutional neural network based bearing fault diagnosis. In: International conference on intelligent computing, pp. 105–111. Springer (2017)
13. Hoang, D.T., Kang, H.J.: Rotary machine fault diagnosis using scalogram image and convolutional neural network with batch normalization. In: International conference on intelligent computing, pp. 283–293. Springer (2019)
14. Hoang, D.T., Kang, H.J.: A survey on deep learning based bearing fault diagnosis. Neurocomputing **335**, 327–335 (2019)
15. Jiang, F., Zhu, Z., Li, W., Wu, B., Tong, Z., Qiu, M.: Feature extraction strategy with improved permutation entropy and its application in fault diagnosis of bearings. Shock Vib. **2018** (2018)
16. Kankar, P.K., Sharma, S.C., Harsha, S.P.: Fault diagnosis of ball bearings using machine learning methods. Expert Syst. Appl. **38**(3), 1876–1886 (2011)
17. Khodja, A.Y., Guersi, N., Saadi, M.N., Boutasseta, N.: Rolling element bearing fault diagnosis for rotating machinery using vibration spectrum imaging and convolutional neural networks. Int. J. Adv. Manuf. Technol. 1–15 (2019)
18. Kuai, M., Cheng, G., Pang, Y., Li, Y.: Research of planetary gear fault diagnosis based on permutation entropy of CEEMDAN and ANFIS. Sensors **18**(3), 782 (2018)

19. Leite, G.d.N.P., Araújo, A.M., Rosas, P.A.C., Stosic, T., Stosic, B.: Entropy measures for early detection of bearing faults. Physica A Stat. Mech. Appl. **514**, 458–472 (2019)
20. Li, C., Zhang, W., Peng, G., Liu, S.: Bearing fault diagnosis using fully-connected winner-take-all autoencoder. IEEE Access **6**, 6103–6115 (2017)
21. Li, J., Zhang, J., Li, M., Zhang, Y.: A novel adaptive stochastic resonance method based on coupled bistable systems and its application in rolling bearing fault diagnosis. Mech. Syst. Signal Process. **114**, 128–145 (2019)
22. Li, Y., Chen, X., Yu, J., Yang, X.: A fusion frequency feature extraction method for underwater acoustic signal based on variational mode decomposition, duffing chaotic oscillator and a kind of permutation entropy. Electronics **8**(1), 61 (2019)
23. Li, Y., Xu, M., Wei, Y., Huang, W.: A new rolling bearing fault diagnosis method based on multiscale permutation entropy and improved support vector machine based binary tree. Measurement **77**, 80–94 (2016)
24. Li, Y., Zhang, W., Xiong, Q., Luo, D., Mei, G., Zhang, T.: A rolling bearing fault diagnosis strategy based on improved multiscale permutation entropy and least squares svm. J. Mech. Sci. Technol. **31**(6), 2711–2722 (2017)
25. Liu, R., Yang, B., Zio, E., Chen, X.: Artificial intelligence for fault diagnosis of rotating machinery: a review. Mech. Syst. Signal Process. **108**, 33–47 (2018)
26. Lu, C., Wang, Z., Zhou, B.: Intelligent fault diagnosis of rolling bearing using hierarchical convolutional network based health state classification. Adv. Eng. Inform. **32**, 139–151 (2017)
27. Lu, S., He, Q., Wang, J.: A review of stochastic resonance in rotating machine fault detection. Mech. Syst. Signal Process. **116**, 230–260 (2019)
28. Sharma, A., Jigyasu, R., Mathew, L., Chatterji, S.: Bearing fault diagnosis using frequency domain features and artificial neural networks. In: Information and Communication Technology for Intelligent Systems, pp. 539–547. Springer (2019)
29. Tian, Y., Wang, Z., Lu, C.: Self-adaptive bearing fault diagnosis based on permutation entropy and manifold-based dynamic time warping. Mech. Syst. Signal Process. **114**, 658–673 (2019)
30. Wang, C., Gan, M., Zhu, C.: A supervised sparsity-based wavelet feature for bearing fault diagnosis. J. Intell. Manuf. **1**, 1–11 (2019)
31. Wang, Z., Yao, L., Cai, Y.: Rolling bearing fault diagnosis using generalized refined composite multiscale sample entropy and optimized support vector machine. Measurement **156**, 107574 (2020)
32. Wen, L., Li, X., Gao, L., Zhang, Y.: A new convolutional neural network-based data-driven fault diagnosis method. IEEE Trans. Industr. Electron. **65**(7), 5990–5998 (2017)
33. Xu, G., Liu, M., Jiang, Z., Söffker, D., Shen, W.: Bearing fault diagnosis method based on deep convolutional neural network and random forest ensemble learning. Sensors **19**(5), 1088 (2019)
34. Yan, X., Jia, M.: Intelligent fault diagnosis of rotating machinery using improved multiscale dispersion entropy and MRMR feature selection. Knowl.-Based Syst. **163**, 450–471 (2019)
35. Yu, G.: A concentrated time-frequency analysis tool for bearing fault diagnosis. IEEE Trans. Inst. Measurem. (2019)
36. Zhang, J., Yang, J., Litak, G., Hu, E.: Feature extraction under bounded noise background and its application in low speed bearing fault diagnosis. J. Mech. Sci. Technol. **33**(7), 3193–3204 (2019)
37. Zhang, S., Yang, J., Zhang, J., Liu, H., Hu, E.: On bearing fault diagnosis by nonlinear system resonance. Nonlinear Dyn. **98**(3), 2035–2052 (2019)
38. Zhang, Y., Xing, K., Bai, R., Sun, D., Meng, Z.: An enhanced convolutional neural network for bearing fault diagnosis based on time–frequency image. Measurement p. 107667 (2020)
39. Zhao, W., Zhang, L., Meng, X.: Rolling bearing fault diagnosis using neural networks based on wavelet packet-characteristic entropy. In: International Workshop on Computer Science for Environmental Engineering and EcoInformatics, pp. 424–429. Springer (2011)
40. Zhao, Z., Qiao, B., Wang, S., Shen, Z., Chen, X.: A weighted multi-scale dictionary learning model and its applications on bearing fault diagnosis. J. Sound Vib. **446**, 429–452 (2019)
41. Zhao, Z., Wu, S., Qiao, B., Wang, S., Chen, X.: Enhanced sparse period-group lasso for bearing fault diagnosis. IEEE Trans. Ind. Electron. **66**(3), 2143–2153 (2019)
42. Zheng, J., Cheng, J., Yang, Y.: Multiscale permutation entropy based rolling bearing fault diagnosis. Shock Vib. **2014** (2014)
43. Zheng, J., Pan, H., Liu, Q., Ding, K.: Refined time-shift multiscale normalised dispersion entropy and its application to fault diagnosis of rolling bearing. Physica A: Stat. Mech. Appl. 123641 (2019)
44. Zhu, K., Chen, L., Hu, X.: A multi-scale fuzzy measure entropy and infinite feature selection based approach for rolling bearing fault diagnosis. J. Nondestr. Eval. **38**(4), 90 (2019)