EKSPLOATACJA I NIEZAWODNOŚĆ MAINTENANCE AND RELIABILITY

Eksploatacja i Niezawodnosc – Maintenance and Reliability

Volume 24 (2022), Issue 4

journal homepage: http://www.ein.org.pl

Article citation info:
Shu Z, Zhang S, Li Y, Chen M. An anomaly detection method based on random convolutional kernel and isolation forest for equipment state monitoring. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2022; 24 (4): 758–770, http://doi.org/10.17531/ein.2022.4.16

An anomaly detection method based on random convolutional kernel and isolation forest for equipment state monitoring

Indexed by: Web of Science Group

Xinhao Shu [a], Shigang Zhang [a,*], Yue Li [a], Mengqiao Chen [a]

[a] National University of Defense Technology, Laboratory of Science and Technology on Integrated Logistics Support, College of Intelligence Science and Technology, Deya str., Changsha, 410073, Hunan, China

Highlights

- Random convolution kernel applied in anomaly detection for automatic feature extraction.
- The establishment of the initialization strategy of the 2-D random kernel.
- The anomaly-sensitivity evaluation is based on time series decomposing method.
- Method effectiveness are verified on varying dataset and with result analysis.

Abstract

Anomaly detection plays an essential role in health monitoring and reliability assurance of complex system. However, previous researches suffer from distraction by outliers in training and extensively relying on empiric-based feature engineering, leading to many limitations in the practical application of detection methods. In this paper, we propose an unsupervised anomaly detection method that combines random convolution kernels with isolation forest to tackle the above problems in equipment state monitoring. The random convolution kernels are applied to generate cross-dimensional and multi-scale features for multi-dimensional time series, with combining the time series decomposing method to select abnormally sensitive features for automatic feature extraction. Then, anomaly detection is performed on the obtained features using isolation forests with low requirements for purity of training sample. The verification and comparison on different types of datasets show the performance of the proposed method surpass the traditional methods in accuracy and applicability.

Keywords

anomaly detection, random convolutional kernel, isolation forest, multi-dimensional time series, equipment state monitoring.

1. Introduction

Anomaly detection is a fundamental technology to ensure the safety and reliability of systems. By analyzing the massive multi-dimensional data generated during the operation for system status monitoring and assessment can significantly improve the efficiency of system maintenance and reliability. However, there are still challenges in the current anomaly detection, such as the difficulty in distinguishing normal and abnormal samples in historical data, which makes it very complicated to manually split enough normal data for training deep learning models. The extraction of anomaly-sensitive features from these collected multi-dimensional data also strongly depends on empirics, resulting in a time-consuming process and increasing uncertainty. Therefore, it is necessary and valuable to propose a detection method that is capable of detecting under the anomaly-mixed data condition with automatic feature extraction.

Deep neural network (DNN) models are usually able to achieve higher detection precision compared to traditional machine learning (ML) models. Among previous approaches of unsupervised learning DNN models like Autoencoder [8] and its variants like VAE[11], DAE[21] and SAE[1] have made fruitful progress in anomaly detection field. For example, Zong et al. combined the Autoencoder with gaussian mixture model to jointly consider reconstruction error and the distribution of intermediate hidden layer's variables for anomaly detection [25]. Li et al. combined the intermediate layer of a Variational Autoencoder with the reconstructed error for anomaly detection [13]. DNNs large-scale parameters and iterative optimization enable the models to obtain accurate representations by learning from deep relationships and patterns automatically in training data. However, the existence of abnormal samples in the training set will cause the model to deviate from the potential distribution or low-dimensional representation of the normal samples since the data are not labeled, thus the model failing to precisely characterize the normal state and leading to inaccurate detection. Cheng et al. proposed a solution by using a combination of detection loss and reconstruction loss to optimize the learning process, reducing the model's ability of reconstructing the outliers [4]. Some traditional ML models also have similar problems, such as the OCSVM suffering from the anomalies in training which lead to bias of the support vectors [20]. Guo et al. used LOF clustering methods to preprocess the training data of OCSVM to reduce the influence of anomalous samples [7]. Most of the traditional ML based anomaly detection models are on the assumption that outliers

(*) Corresponding author.

E-mail addresses: X. Shu (ORCID: 0000-0003-3326-0573): sxinhao@163.com, S. Zhang: shigang391@foxmail.com, Y. Li: liyue@nudt.edu.cn, M. Chen: chenmengqiao21@nudt.edu.cn

are the minority and distributed away from the center of dataset, such like the widely used method isolation forest proposed by Fei et al. [14, 15, 24]. In addition, the researches conduct detection from the perspective of distance [12], angle [9] and density [22] also achieve favorable results. However, when the abnormal data differ insignificantly from the normal data or under the complex data conditions, these methods can hardly achieve accurate detection by only using the superficial features on its specific perspective rather than analyzing the underlying patterns of the data like DNN methods. Therefore, this type of methods tends to require feature engineering to obtain suitable features for detection based on different business characteristics, which heavily relying on practitioner experience-based analysis. For example, Calheiros et al. encodes time as an additional feature in isolation forests anomaly detection [2]. Puggini et al. proposed a dimension reduction method based on forward selection group analysis and used the processed features for isolation forest detection, which has superior interpretability compared to the traditional PCA [16].

To summarize, DNN with deep feature extraction capabilities are sensitive to the purity of the training dataset, while some traditional ML models are not constrained by outliers in training but have difficulty in mining deep features and require manual feature engineering [3]. Thus, we attempt to combine the advantages of both, using the DNN approach for feature extraction and the ML for anomaly detection. The rocket series methods are successful examples based on this concept, which have performed well on time series classification tasks in recent years [5, 6, 19]. The method extracts features through massive 1-D random convolution kernels and uses these features as a high-dimensional description of the original data, and then using a linear classifier to classify the descriptions. Overall, the framework can be regarded as a single-layer convolutional neural network without feedback. Benefited from the random setting of the convolutional kernel parameters, the method avoids the back-propagation of the kernels optimization and achieve state of the art accuracy at a significantly reduction in time expense. Previous to the Rocket, feature extraction methods based on random convolution concept have been widely researched [1]. Jimenez et al. proposed a sequence similarity measure by convolving the target series with a random sequence [10]. In the study of Saxe et al., random convolution kernels are used in the feature extraction of images and the obtained features are used as objects for SVM classification [17], which indicates the possibility of applying random convolution methods on multidimensional objects. In the anomaly detection task of multidimensional time series, the fault pattern is complex including the vary of the value and the relation of the features, therefore the extracted fusion feature is more sensitive in representation of the equipment status [3]. By the above studies we can conclude that the random convolution kernel has great potential of extracting features from different dimensions and scales, and the random setting of the convolution kernel parameters provides an automatic approach for feature extraction.

In this regard, we propose an unsupervised anomaly detection method that combines random convolution kernels with isolation forest. To the best of our knowledge, few researches have utilized random kernels in anomaly detection task and combine these two methods. Specifically, we establish an initialization strategy for the parameters of the random convolution kernels, and generate feature series through sliding these kernels with dot production in the multidimensional sequence. To filter the invalid feature series during the generation, we propose a selection method using time series decomposing algorithm. This method evaluates anomaly sensitivity of feature series by analyzing the similarity of its split points, and the several most sensitive feature series are used as the object for isolation forest. The effectiveness of the proposed method is evaluated on a turbine engine simulation dataset and bearing vibration datasets with comparing to other anomaly detection methods.

2. Method

The framework of the proposed method is shown in Fig. 1. In the model initialization stage, the random convolution kernels are initialized to generate massive feature series which describes the original data from different perspectives. Then the feature series selection method sorts the generated feature series by anomaly sensitivity and uses several most sensitive features as the detecting objects of the isolation forest. The anomaly threshold is set by the sample's anomaly score given by the isolation forest. In the test phase, the feature series of the input data are generated by the kernels corresponding to the anomaly-sensitive features in the initialization, and anomaly detection is performed by the same isolation forest had been trained.

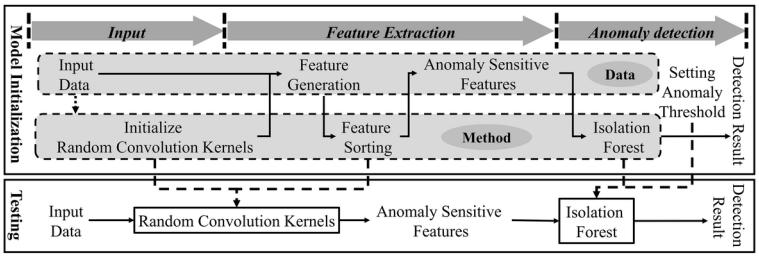


Fig. 1. Framework of the proposed method

2.1. Feature series generation by random convolution kernels

Random convolution kernel is similar to the convolution layer in the neural network, with the difference that the parameters of the former are randomly generated and the training process is not required. From the perspective of signal analysis, the operation of 1-D convolution kernel is similar to the wavelet analysis process, except that wavelet analysis uses a wavelet function, whereas the function of the random convolution method is a random sampling of specific distributions. Fourier transform is also commonly used in signal analysis, but it can only indicate the frequency characteristics of anomalous signals, while wavelet analysis can indicate the location of anomalies. By using wavelet functions with different parameters, wavelet analysis is able to extract signals of different frequencies [5]. The 1-D random convolution method is extended further on this basis, by randomly setting parameters like kernel weights, kernel length and dilation to extract many types of features at different frequencies and scales.

Intuitively, the occurrence of anomalies will inevitably lead to differences in the time series at different frequencies and scales, which have been the object of previous researches by using traditional signal analysis methods. In multidimensional time series, anomalies also lead to variations in the relationship between different features. As an example, Fig. 2 depict a simple situation.

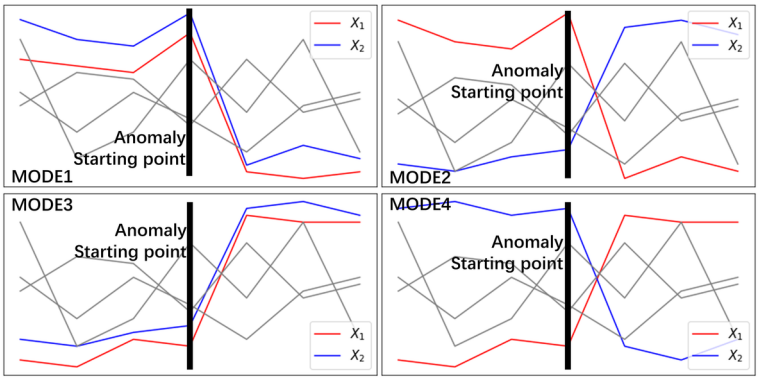


Fig. 2. Anomalous patterns

As shown in Fig. 2, the occurrence of an abnormality causes a change in the values of features x_1, x_2, with four different combination patterns. If summing the two signals in modes 1 and 3, and differencing the two signals in modes 2 and 4, the result of this kind linear combination can enlarge the amplitude of value change between normal and abnormal state. Based on the above description, the proposed feature extracting method of 2-D convolutional kernel is depicted in Fig. 3, where a large number of kernels with different parameters are successive sliding over the original data with dot production to generate a sequence of features. The dot production over different dimensions can be regarded as a linear combination of the features extracted by multiple 1-D convolution kernels, and the result not only expresses the extracted features variation of individual dimension, but also takes into account the relative variation of features in different dimensions. The random setting of the parameters of the convolution kernel implies the randomization of the weights of features participating in the linear combinations, and by generating massive kernels will enable the method to obtain plentiful features from diverse perspective.

Massive convolutional kernels will be randomly initialized

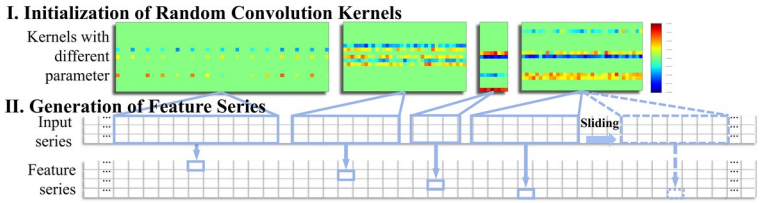


Fig. 3. Random convolution kernel based feature extraction of multidimension time series

whether the detection object is 1-D or multidimensional time series. In kernels initialization, we referred to the Rocket [5] approach and proposed a 2-D random convolutional kernel initialization strategy for multidimensional time series anomaly detection tasks. Specifically, it can be divided in to shape initialization and weights initialization.

The shape of the kernel is controlled by three parameters: its width is determined by the dimension of the input data, and its length is jointly determined by the original length and the dilation.

- The original length of the convolution kernel is chosen randomly with equal probability within a candidate set, and the candidate value depends on the type of signal and its frequency of the detecting object. For signals with high frequency such as acceleration sensors, a larger candidate value is usually given such like $\{10,20,40,80\}$, while for slowly varying signals with low frequency such as temperature and pressure, a smaller candidate value is usually selected, such like $\{5,9,13,17\}$.
- The dilation is sampled from an exponentially distributed sequence to acquire convolutional kernels with varying sparsity, and to avoid the convolutional kernels from being too sparse, we add a factor $\alpha \in [0,1]$ to control them. Let the length of the input data be l_{in}, the original length of the convolution kernel be and l_r, the dilation d and the kernel length after dilation be l_k.

$$d = \left\lfloor 2^x \right\rfloor, x \sim U\left(0, \log_2\left(\alpha \frac{l_{in}-1}{l_r-1}\right)\right) \tag{1}$$

$$l_k = l_r + (l_r - 1) \times d \tag{2}$$

The number of the dimension of original series participated in the linear combination of the kernel is controlled by the parameter m, which is an integer randomly sampled in the range $[1,n]$, where n is the number of the original dimension of input data. Here only the number of selected dimensions is set, while the specific dimensions are randomly selected, so that convolution kernels with the same parameter m will extract features from different combinations of dimensions. Fig. 4 depicts the role played by the above parameters in the convolution kernel.

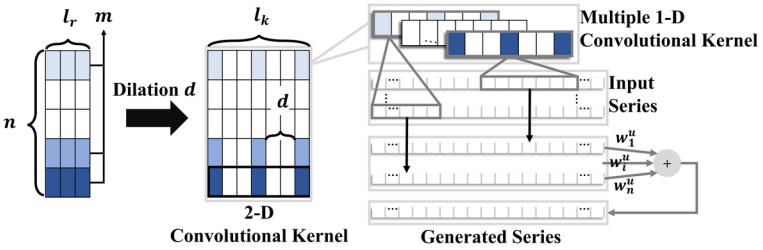


Fig. 4. Convolutional Kernel

The initialization strategy of weight largely determines the quality of the generated feature series. Operationally, the calculating result of the 2-D kernel can be regarded as the linear combination of the results of 1-D kernel, as shown in Fig. 4. Reflecting on the weight matrix, the weight vectors of the 1-D kernel are arranged vertically on the matrix of 2-D kernel, and the diversity of linear combinations is achieved by assigning different weights to each vector's result. Let the length of the original kernel be l_r and the number of dimensions of the original data be n, from which m dimensions are randomly selected to participate in the operation. Suppose the weight matrix of 2-D kernel is denoted as $W = \left[w_1, w_2, ..., w_n\right]^T \left(w \subset R^{n \times l_r}\right)$, where $w_i = \left[w_{i1}, w_{i2}, ..., w_{il_r}\right] \left(w_i \subset R^{lr}\right)$ represents 1-D kernel's weight vector. The specific weight initialization process can be divided into two steps, as follows:

(1) Assignment of weights to the 1-D kernels in the linear combination: The corresponding m weights represent by $w_u = \left[w_1^u, w_2^u, ..., w_m^u\right]$, and we analyze the influence of the weights on the generated features series from data distribution. Assuming that the sequence of m dimensions involved in the operation is $X = \left[X_1, X_2, ..., X_m\right], X_i \sim N\left(u_i, \sigma_i\right)$, the distribution of the generated series X' can be described as follows:

$$\begin{aligned} P(X') &= \textstyle\sum_{i=1}^{m} w_i^u N(u_i, \sigma_i^2) \\ &= N\left(\textstyle\sum_{i=1}^{m} w_{i=1}^u u_i, \sum_{i=1}^{m} (w_i^u)^2 \sigma_i^2\right) \end{aligned} \tag{3}$$

We limit that the weight vector w_u is sampled from the distribution $U(-1/\sqrt{m}, 1/\sqrt{m})$, so that the variance of X' is $\sigma'^2 \le (1/m)\sum_{i=1}^{m} \sigma_i^2$. This ensures that the variance of the generated series always smaller than the mean value of the variance of all the dimensions involved in the linear combination. Thus, avoiding the abnormal signal being swamped by the increasing variance. Meanwhile, checking whether the condition $\left|\sum w_u\right| < \varepsilon (\varepsilon \approx 0)$ is satisfied, otherwise re-initialize w_u. This constraint ensures that w_u is as even as possible over the positive and negative intervals and avoids assigning too much computational weight to single dimension in the linear combination.

(1) Initialization of the weight vector of the 1-D kernel. The weight vector w_i of the 1-D kernel corresponding to the dimensions

involved in the operation is sampled from $N(w_i^u,\sigma^2)$. For the remaining unselected dimensions, the weight vectors are $w_i = \mathbf{0}$. Where the variance σ^2 is set to 0.01 in this paper. For the special case that the input data is 1-D series, the 2-D weight matrix of the kernel is taken to be compressed into one dimension by $w = \sum w_i$. The initialization of kernel weights is summarized in Algorithm 1.

Algorithm 1 Initialization of Kernel Weights

Input: Kernel length l_k, Dimension of input data n, Number of dimensions involved in the operation m, Variance σ, Error $\varepsilon(\varepsilon \approx 0)$

Output: Weight Matrix w

Initialize $w[1,2,...,n], w_u[1,2,...,m], idx[1,2,...,m]$ (array)
$idx \leftarrow sample([0,1,...,n],m)$
while True **do**
  $w_u \leftarrow sample\left(\left\lfloor U\left(-1/\sqrt{m},1/\sqrt{m}\right)\right\rfloor,m\right)$
  if $\left|\sum w_u\right| < \varepsilon$ **then**
    break
  end if
end while
for $i = 1 \rightarrow n$ **do**
  if $i \in idx$ **then**
    $w[i] \leftarrow sample(N(w_u[i],\sigma^2),l_k)$
  else
    $w[i] \leftarrow \mathbf{0}$
  end if
end for
return w

The preprocessing includes standardization to avoid the impact of different units and padding to ensure the generated series by different kernels are aligned in time. On the assumption that there are no anomalies in the earlier data, for kernel of length l_k and dilation d, repeating the first $l_k + (l_k - 1) \times d$ in front of the input data to pad the sequence. During the convolution, the kernel slides over the preprocessed data and conduct dot production. For a n dimensions, L samples multidimensional time series $X = [X_1, X_2, ..., X_n]^T (X \subset R^{n \times L})$, preprocessed as $\bar{X}$, kernel $K(W, l_k)$ produces a feature series T':

$$T' = [t'_1, t'_2, ..., t'_L] \tag{4}$$

$$t'_i = \Sigma_{j=1}^{n} w_j \bullet \bar{X}_{j,[i,i+l_k]} \tag{5}$$

If k kernels are initialized, generation with X produces a k-dimensional time series $T = [T'_1, T'_2, ..., T'_k]^T (T \subset R^{k \times L})$, each T'_i represent a feature series generated by a specific kernel.

2.2. Abnormal sensitive feature series selection

The random initialization strategy of the parameters of the random convolution kernel can extract different features of the original data, but it will inevitably generate many irrelevant features at the same time. The isolation forest itself has a certain anti- irrelevant feature capability, but since the proportion of irrelevant features in the generated features is difficult to estimate, preprocessing the generated features can further enhance the stability of the combined method. Since abnormal event cause changes in the content of series segments, we use the time series decomposing method, which split series into segments by its content, to obtain the split points caused by anomalies. By analyzing the similarity of the series split points distribution with that of anomalous split points, the anomaly-sensitivity of generated series can be evaluated and then sort the generated series with the sensitivity. In this section, we combine the relevant research on time series decomposition methods by Zhao et al. [23]. The feature series selection can be divided into three parts specifically: ①Converting the generated numerical series into symbolic series, which characterize the result of numerical analysis and describe the data in a more abstract way to facilitate the following content-based decomposition. ②Using time series decomposing method to search the split points of each symbolic series and calculating the integrated distribution of all split points. ③Evaluate the anomaly sensitivity of the generated series by the similarity between the distribution of the split points of each sequence and the distribution of the integrated split points. Then rank the generated series based on sensitivity and select several top-ranking series as the object of anomaly detection. The main framework of the method is depicted in Fig. 5, and the specific technical details are described in the following.

2.2.1. Symbolic series conversion

Symbolic series describe the numerical series in a more abstract approach. In Zhao et al.'s study [23], he used a 1-D directional gradient histogram (HOG) to describe the shape character of the sampled sequences, and clustered the description vectors into a certain number of categories. The conversion of series from numerical to symbolic is accomplished by replacing the numerical values of subsequences using the labels of each category. In this paper, the description of the original series has been completed within the generation of feature series by random convolution kernel, which not only considers multiple dimensions but also avoid the limitation of specific description methods to describe different types of data. Among the description approaches of anomalies, the simplest form is to directly express them as numerical magnitudes rather than shapes or frequencies. Therefore,

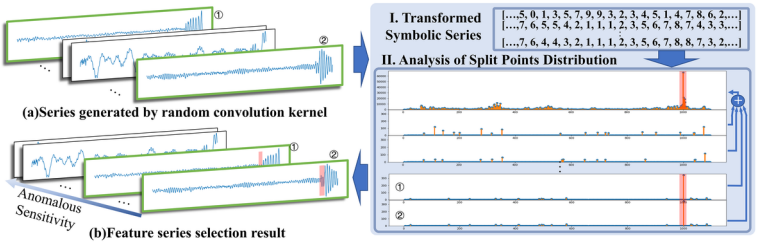


Fig. 5. Feature series selection framework

we only research the pattern that describe anomalies in numerical values by using the simplest equal-interval division method for symbolic transformation of the series data. This method evenly divides the interval between maximum and minimum of the object sequence into n segments and uses the symbols $c_i \in \{1,2,...,n\}$ to describe the value falling in, thereby realizing the transition from the time series $T = t_1, t_2, ..., t_L (t_i \in R)$ to the symbolic series $C = c_1, c_2, ..., c_L$. The conversion process of $T \rightarrow C$ is shown in Fig. 6.

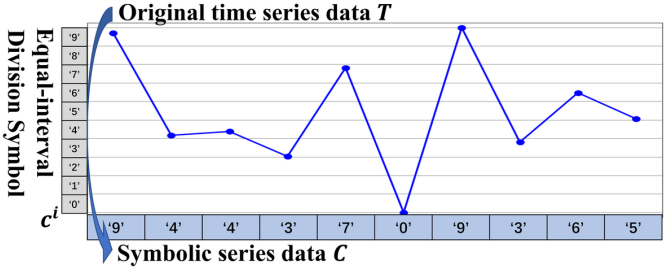


Fig. 6. Equal-interval symbolic series conversion

2.2.2. Split points search

The purpose of the split points searching is to obtain the location of possible abnormality on the time axis. As above analysis, anomalous events can cause changes in the content of the series segments, which can be reflected as an increase in the disorganization of the symbol series C. And this disorganization can be quantified by using information entropy. In the searching process, the location of anomalous events is obtained by the target that maximizing the information gain of each split point in series C, so that the character descriptions within each segment are homogeneous and between segments are heterogeneous. This part refers to the search method for split points in the study of Zhao et al. which is briefly outlined below [23]. Specifically, the symbolic sequence C is recursively decomposed until all its subsequences reach the minimum length l_s or the number of split points reaches the limit β. Each split point s cuts the parent sequence C_p into two subsequences $\{C_l, C_r\}$. Using the information gain as the criterion for the selection of the segmentation points $\Delta E(s, C_p)$:

$$\Delta E(s, C_p) = E(C_p) - \left[\frac{|C_l|}{|C_p|} E(C_l) + \frac{|C_r|}{|C_p|} E(C_r) \right] \quad (6)$$

$$E(C) = -\sum_{i=1}^{n} p_i^C \log p_i^C \quad (7)$$

where p_i^C is the frequency of symbol i in the sequence C. The split point with the maximum information gain is denoted by s^*. The searching for best split points can be transformed into the maximization problem:

$$\arg\max_{s \in C_p} \{\Delta E(s, C_p)\} \quad (8)$$

And the weight $v_{s^*}^{C_p}$ for the split point s^* is:

$$v_{s^*}^{C_p} = |C_p| \bullet \Delta E(s^*, C_p) \quad (9)$$

The split point search is summarized in Algorithm 2 [23].

Algorithm 2 Searching for split points

Input: Input series $T = t_1, t_2, ..., t_L$, Limitation number of split points β, Minimum length of the subsequence l_s

Output: Split points set S, Weights of the split points V

Symbolic series conversion $T \rightarrow C$
Initialization set $Q = \{C\}, S = \varnothing, V = \varnothing$
while $Q \neq \varnothing$ **do**
  $C_p \leftarrow Q.dequeue()$
  if $|C_p| < l_s$ **then**
    continue
  else
    Obtain s^* of C_p according to Eq.(8), $S.add(s^*)$
    Obtain $w_{s^*}^{C_p}$ of s^* according to Eq.(9), $V.add(v_{s^*}^{C_p})$
    if $|S| > \beta$ **then**
      break
    end if
    $Q \leftarrow Q.enqueue(\{C_l, C_r\})$
  end if
end while
return S, V

Fig. 7 depicts the split points obtained by the above searching algorithm, such as the distribution of split points (c), (e), (g) for feature series (b), (d), (f).

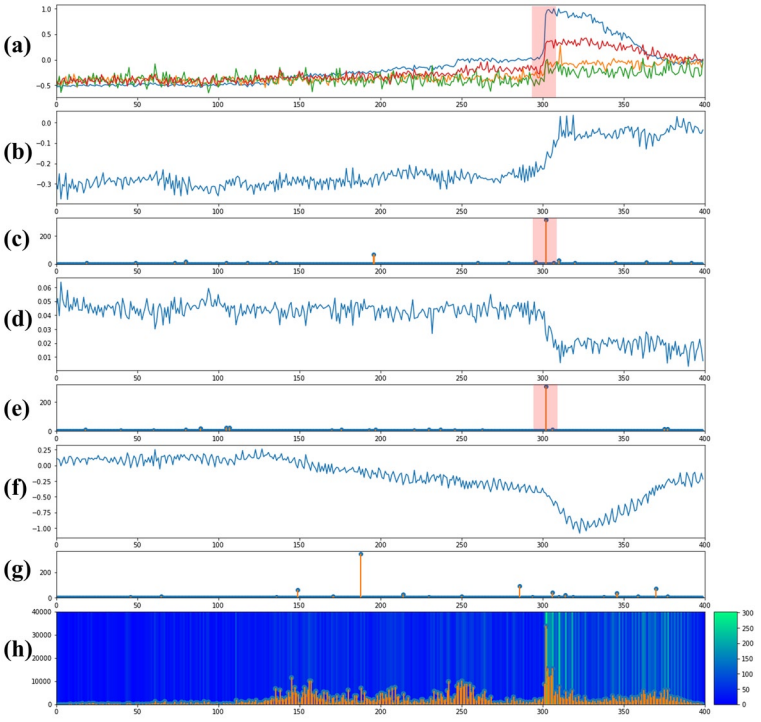


Fig. 7. Distribution of the split points location and its weight and density of part of the generated feature series. (a) is the input data. (b), (d), (f) are the feature series generated by random convolution kernel. (c), (e), (g) are the corresponding split points' location and weights. (h) is the density (depict by background color) and weight (depict by vertical axis) and location (depict by horizontal axis) of integrated split points of all generated series

2.2.3. Feature series ranking by anomalous sensitivity

As the generated feature series explain the anomalies in different perspective, the searched split points of each series are dominated by anomalies with varying strength. But the distribution of anomalous split points of different features series has similarity in location on the time axis since the anomaly is prevalent in all features. This similarity will be expressed as the density of split points caused by anomalies is higher than that caused by other reasons in the cumulative distribution among split points of all feature series, which is depict in Fig. 7(h) by background color. As it shown that the high-density location of Fig. 7(h) is consistent with the position of the anomaly start in (a). On the other hand, when the part of the split points of a series are dominated by anomalies, its weights is also significantly higher than that of the points caused by other reasons within the series. And this can be found from the generated feature series in Fig. 7(b)(d), which can significantly distinguish normal data from abnormal data. In the distribution of its split points (c) and (e), the weight of the anomalous split points are significantly higher than that of the other points, which coincides with the peak of the cumulative weight and the high-density position in the integrated distribution (h). Comparing to the feature series (f), there is also a relative high weight of its point distribution in (g) might be dominated by other reasons, thus the distribution is not consistent with the integrated distribution, and the difference between normal and abnormal in (f) is relatively obscure. Together with

the above two factors, the features with strong expression ability for anomalies have two characteristics: ① Its split points are distributed at the higher density locations in the integrated distribution. ② Its split points with high weight are also near the peak of the cumulative weight in the integrated distribution. Based on the above analysis, the task of evaluating the anomaly sensitivity of the feature series transformed into measuring the similarity of the split point locations and weights to the integrated distribution.

Specifically, for a multidimensional time series $MT = \{T_1, T_2, ..., T_k\}$ with k features series, the corresponding split points set is $MS = \{S_1, S_2, ..., S_k\}$, and by computing the concatenation of MS to obtain the integrated split point set $\mathbb{S} = S_1 \cup S_2 \cup ... \cup S_k = \{\mathbb{s}_1 \cup \mathbb{s}_2 \cup ... \cup \mathbb{s}_\gamma\}$. The weight of the points in $\mathbb{S}$ is $\mathbb{V} = \{\mathbb{v}_1, \mathbb{v}_2, ..., \mathbb{v}_\gamma\}$, where $\mathbb{v}_i$ is the sum of the weights of split point $\mathbb{s}_i$ in each set of MS:

$$\mathbb{v}_i = \sum_{\mathbb{s}_i \in \mathbb{S} \cap \mathbb{s}_i \in S_j} V_{\mathbb{s}_i}^j \tag{10}$$

where $V_{\mathbb{s}_i}^j$ is denoted as the weight value corresponding to the split point $\mathbb{s}_i$ of $\mathbb{S}_j$, and $V_{\mathbb{s}_i}^j = 0$ when $\mathbb{s}_i$ does not exist in the S_j. The frequency of each split point in $\mathbb{S}$ is denoted as $\mathbb{P} = \{\mathbb{p}_1, \mathbb{p}_2, ..., \mathbb{p}_\gamma\}$, where $\mathbb{p}_i$ is the frequency of $\mathbb{s}_i$ occurring in all split points set MS:

$$\mathbb{p}_i = \left(\sum_{\mathbb{s}_i \in \mathbb{S} \cap \mathbb{s}_i \in S_j} 1\right) \Big/ \sum_{j=1}^{k} |S_j| \tag{11}$$

Kullback-Leibler (KL) divergence can measure the extent of one distribution explains the other, which can be interpreted as the similarity of one distribution to the other, but it is asymmetric. In this study, we modify the discrete KL formula to jointly consider the split point weights. Using the factor ρ_i to represent the similarity between the distribution of split point set S_i of feature series $T_i (T_i \in R^L)$ and the integrated distribution $\mathbb{S}$:

$$\rho_i = \sum_{j=1}^{\gamma} \mathbb{v}_{\mathbb{S}_j} \frac{v_{\mathbb{S}_j}^i}{\sum v_i} \times p_{\mathbb{S}_j} \log \frac{p_{\mathbb{S}_j}}{\mathbb{p}_{\mathbb{S}_j}} \tag{12}$$

where $V_i = \{v_{\mathbb{S}_1}^i, v_{\mathbb{S}_2}^i, ..., v_{\mathbb{S}_\beta}^i\}$ is the weight of the split point in S_i, and $p_{\mathbb{S}_j} = 1/L$. Intuitively, the second part of the equation is the traditional discrete KL. The explanation for the value of $p_{\mathbb{S}_j}$: For a time series with L samples, the probability that any point is selected as a splitting point is $1/L$. Calculating the KL can be understood as the process of sampling each point in the point set $\mathbb{S}$, which explain the second part of the Eq.(12).

The first part of the Eq.(12) considers the weights of the split points. The higher the ratio of the weight $v_{\mathbb{S}_j}^i$ contributes among the sum of the weights $\sum V_i$ of the series T_i, the greater the difference between the data segments before and after the point $\mathbb{S}_j$. This ratio is then multiplied with the accumulated weight $\mathbb{v}_{\mathbb{S}_j}$, indicating that only if other features also have high weights at split point $\mathbb{S}_j$, then the ratio can contribute more similarity to integrated distribution. Thus, the similarity factor ρ_i accounting more for the prevalence of different feature series that all occur numerical changes at point $\mathbb{S}_j$. The traditional KL is numerically non-negative, but in Eq.(12), the meaning of its value is modified so that the first part of the equation is always greater than 0 and the larger the value, the more similar the weights of the two distributions are, while the second part is always smaller than 0. Therefore, factor ρ is a negative value. And the smaller the value, the more similar the two distributions are, signifying that the split point distribution of the series is similar to the anomaly distribution, which also means that the series has stronger anomaly sensitivity. Algorithm 3 summarizes the above description.

Algorithm 3 Feature selection method

Input: Multivariate time series $MT = \{T_1, T_2, ..., T_k\}$

Output: Sorted multivariate time series $\overline{MT} = \{\overline{T}_1, \overline{T}_2, ..., \overline{T}_k\}$

Initialize the set $\mathbb{S} = \varnothing, \mathbb{V} = \varnothing, \mathbb{P} = \varnothing, S = \varnothing, V = \varnothing, \rho = \varnothing$
For $i = 1 \to k$ **do**
    Searching split point of T_i according to Algorithm2,
    $S_i, V_i \leftarrow$ Algroithm2(T_i)
    $S.add(S_i)$, $V.add(V_i)$, $\mathbb{S} \leftarrow \mathbb{S} \cup S_i$
end for
Obtain $\mathbb{V}$ from $\mathbb{S}, S, V$ according to Eq.(10)
Obtain $\mathbb{P}$ from $\mathbb{S}, S$ according to Eq.(11)
For $i = 1 \to k$ **do**
    Obtain ρ_i from $\mathbb{S}, \mathbb{V}, \mathbb{P}, S, V$ according to Eq.(11), $\rho.add(\rho_i)$
end for
Sort MT according to ρ, obtain $\overline{MT} = \{\overline{T}_1, \overline{T}_2, ..., \overline{T}_k\}$
return $\overline{MT}$

The above analysis demonstrates that the impact of anomalous events on all generated features is prevalent, so the feature selection method can capture this factor and rank feature series according to anomalous sensitivity, which signifies that the method is based on the assumption that the anomaly is exist. When there are no anomalies in the data, the feature series will be sorted with other factors that dominate split points, but this kind of obvious consistency is practically impossible in the condition, since these series descript the input data on different perspective. Therefore, the ranking results have limited impact to subsequent anomaly detection in this situation. The number of finally selected features series is not specified in this section, since the proportion of irrelevant features generated by the random convolution kernel is not controllable, and the purpose of the feature series selection aims to further reduce the risk of isolation forests being affected by irrelevant feature series. The several top rank series, obtained by ranking according to the similarity of the distribution of split points and integrated split points, consider the significant variation in all series comprehensively, which means these series contain the most information embedded in the rest features series. Therefore, a wide range of the specific number of the selected feature series can be chosen. In this paper, a precise number is obtained by the balance between the detection result and number of redundant feature series in the Section 3.3.2.

2.3. Isolation forest anomaly detection

Isolation forest assumes that the anomalies locate far from the center of the dataset with sparsely distribution, and using tree method by randomly select the feature and split point in leaf node to isolate the anomalies which usually have a short split path, as in Fig. 8 [14]. Without the feedback from the training set to update the parameters, isolation forest has limited sensitivity to the outliers in the training set. The method has linear time complexity which is suitable for multidimensional data with large volumes and with its excellent performance it is popularly used in industrial application.

The random process of tree node attributes equipped the method robustness to noisy features in the high-dimensional data. However, the number of interfering features in multi-dimensional series generated by random convolution kernel is difficult to estimate, and the excessive amount nonsense features will result unstable in detection result. By the propose feature series selection method to preprocess the mas-

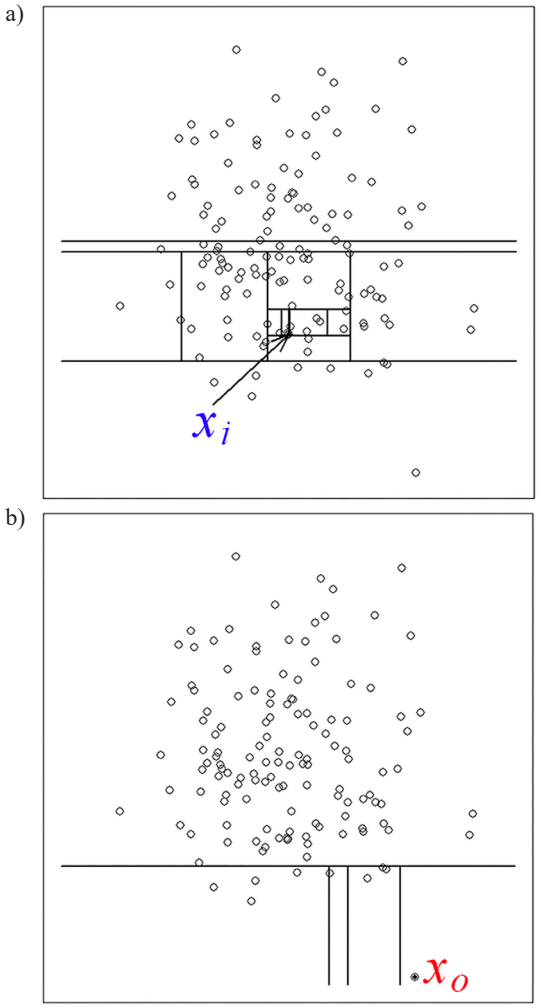


Fig. 8. Isolation forest anomalies detection: a) Isolation of normal point x_i, b) Isolation of outliner x_o

sive series can enhance the robustness of the isolation forest in the input side. Let the m be the number of feature series which are selected in the rank in Section2.2.3, the L samples selected feature series are represented by $\overline{M} = \left[\overline{T}_1, \overline{T}_2, ..., \overline{T}_m\right]$, $\overline{T}_i = \left[t_{i1}, t_{i2}, ..., t_{iL},\right]^T$, then the detection object can be represented by $X_j = \left[t_{1j}, t_{2j}, ..., t_{mj}\right]$.

3. Experimental results

In this paper, we conduct evaluation in three datasets, which include two different data types and varying dimensions to test the effectiveness of the method. And the proposed method is compared with traditional methods under the same data conditions.

3.1. Dataset

The three datasets used in this paper are the turbine engine simulation dataset and the public bearing vibration dataset. Among them, the simulation data of the aero-engine is generated by the GSP software, which is sampled at a low frequency (HZ). The bearing vibration dataset includes bearing failure data from Case Western Reserve University (CWRU) (https://engineering.case.edu/bearingdatacenter/welcome) [18] and bearing degradation data provided by NASA (http: //ti.arc.nasa.gov/project/prognostic-data-repository) which are sampled in high frequency (kHz). The details of the dataset are shown in Table 1.

The detail of the dataset is as follow:

- Gas Turbine Simulation Dataset: Gas turbine Simulation Program (GSP) is a software that can simulate various parameters of turbines engine and is often used to assist in structural design and operating condition analysis. Since it is difficult to consistently collect operating data under fault conditions in practical, this study uses efficiency degradation at the inlet fan as an abnormal event and simulates abnormal data at five different degradation rates by GSP. The generated dataset includes a total of 24 dimensions (sensors) such as temperature, pressure, air flow and thrust. The simulation model and generated data of Dataset1 is shown in

Table 1. Statistics of the datasets

	Dimension	Name	Type	Length	Anomaly ratio	Degradation rate
Gas Turbine Simulation Dataset (GSP)	24	#Dataset1	train	500	10%	4%
			test	120	16.67%	
		#Dataset2	train	500	10%	2%
			test	120	16.67%	
		#Dataset3	train	500	10%	1%
			test	120	16.67%	
		#Dataset4	train	500	10%	0.50%
			test	120	16.67%	
		#Dataset5	train	500	10%	0.25%
			test	120	16.67%	
		#Dataset6	train	500	15%	0.50%
			test	400	25%	
		#Dataset7	train	500	15%	0.25%
			test	400	25%	
Bearing Failure Dataset (CWRU)	1	#Dataset8	train	5000	20%	-
			test			
		#Dataset9	train			
			test			
		#Dataset10	train			
			test			
Bearing Degradation Dataset (NASA)	4	#Dataset11	-	984	46.14%	
			-			

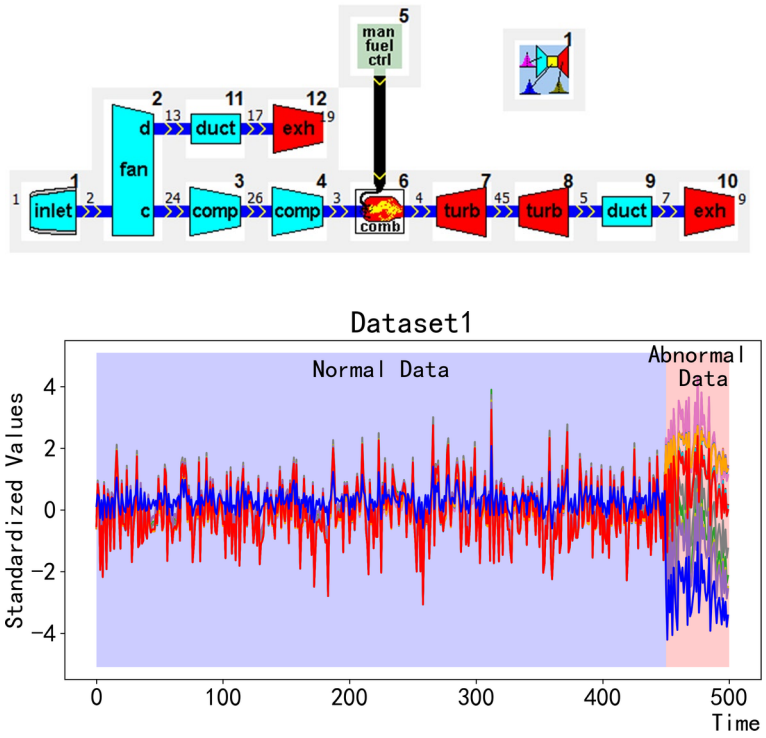


Fig. 9. GSP model and data generation: a) GSP simulation model, b) GSP simulation data

Dataset1 has the largest degradation rate among all the simulation datasets, and it can be seen from Fig. 9(b) that the difference between normal and abnormal is relatively obvious. The normal data are sampled from the engine operating under no degradation, and the abnormal data are sampled under the efficiency is degraded by 4% in the fan inlet. The simulation dataset is used to test the effectiveness of the proposed method on multi-dimensional data sampled at a low frequency.

- Bearing Failure Dataset: Consists by concatenation of normal and abnormal data from the same operating conditions in the original dataset. Abnormal events including inner race, outer race and ball damage. The dataset contains only one dimension, which is the sensor at the drive-side of the equipment. The dataset is used to validate the effectiveness on one-dimensional high-frequency signals.
- Bearing Degradation Dataset: In this dataset, four vibration sensors were used to measure the vibration of four bearings respectively and recorded the degradation process collectively, where the degradation of each bearing was inconsistent. We analyzed the abnormal starting points of the dataset and labeled all the succeeding data as abnormal. The dataset is used to test the effectiveness on multidimensional high-frequency signals.

3.2. Evaluation Metrics

This paper considered *precision* , *recall* and *F1score* for anomalous samples to evaluate the performance of the methods. The anomaly detection algorithm usually provides the anomaly score for each sample. Thus, the threshold setting determines the result of the detection. In order to focus on the performance of algorithm itself, we set the anomaly threshold directly by the proportion of anomalous samples in each training dataset. For example, the percentage of outliers in the training set of Dataset1 is 10%, then the top 10% of samples with the highest abnormal scores given by the detection method will be marked as abnormal and the smallest score among them is set as the threshold for test set. However, in practical the proportion of abnormal is not known in advance, so we also use receiver operating characteristic (ROC) curve and the area under curve (AUC) to avoid the influence of the threshold setting. In addition, we repeat each test for several times to obtain the average performance to avoid the effect of randomness.

3.3. Analysis and Test

Initialization parameters of random convolution kernels: The number of convolution kernels is set to 1000. For the gas turbine simulation dataset, which is sampled by a low frequency, the set of the original kernel length candidates is set to $[2,3,4,5,8]$, for the CWRU bearing data is set to $[8,15,30,50,80,100]$, and for the NASA bearing degradation data set to . These two sets of vibration data are sampled at a higher frequency, so a larger original length of the kernel is considered. The number of features series participate in each kernel, that is parameter in section 2.1, is chosen randomly among . The dilation controller in Section 2.1 is set to 0.2. The top 10 features in the anomaly sensitivity rank are selected for all datasets to isolation forest according to the analysis in Section 3.3.2. The number of random trees in the isolation forest is set to 100.

3.3.1. Effectiveness of feature series selection

In the first part of the experiment, we validate the proposed feature filtering method. Taking the Dataset5 as an example, the filtered features obtained by the proposed method in section 2.2 are shown in Fig. 10.

As shown in Fig. 10(b), the nonsense feature series among the massive generated series are ranked in the bottom and anomaly sensitive feature series are ranked in the top, which proves that

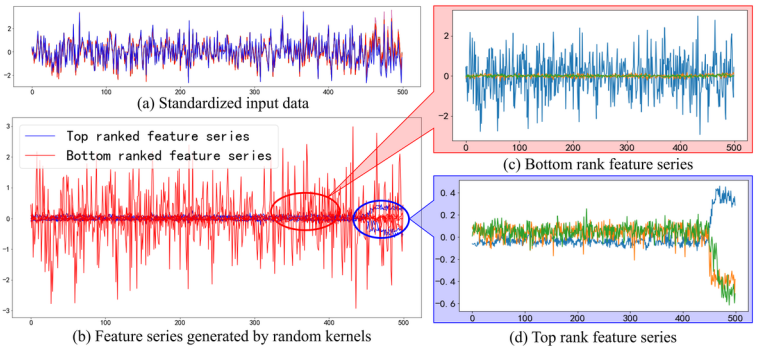


Fig. 10. Features series generated by random kernels and anomaly-sensitivity rank

the method is capable of evaluating anomaly sensitivity. Although there might be a few series ranked in a wrong position, since the object of subsequent anomaly detection is a pack of several feature series with high ranks, which impact on the final result can be ignored.

Dataset5 and Dataset7 has the lowest degradation rate in the simulation thus the outliers are similar to the normal data. We assess the performance in separating abnormal and normal data of the proposed method by using the t-SNE to visualize the low-dimensional distribution of the original data and the top 5 generated anomaly-sensitive data in these two datasets, and the results are shown in Fig. 11.

As depicted in Fig. 11(a)(c), the distributions of normal and abnormal data in the original data are overlapped and hard to separate. After processing of random convolution kernel and anomaly sensitivity selection, the normal and abnormal data are clearly separated, as shown in Fig. 11(b)(d), indicating that feature series obtained are with high sensitivity to abnormalities.

3.3.2. Parameters impact on the detection result

In the second part of the experiment, we analyze the effect of the key parameters of the proposed method on the detection result, as shown

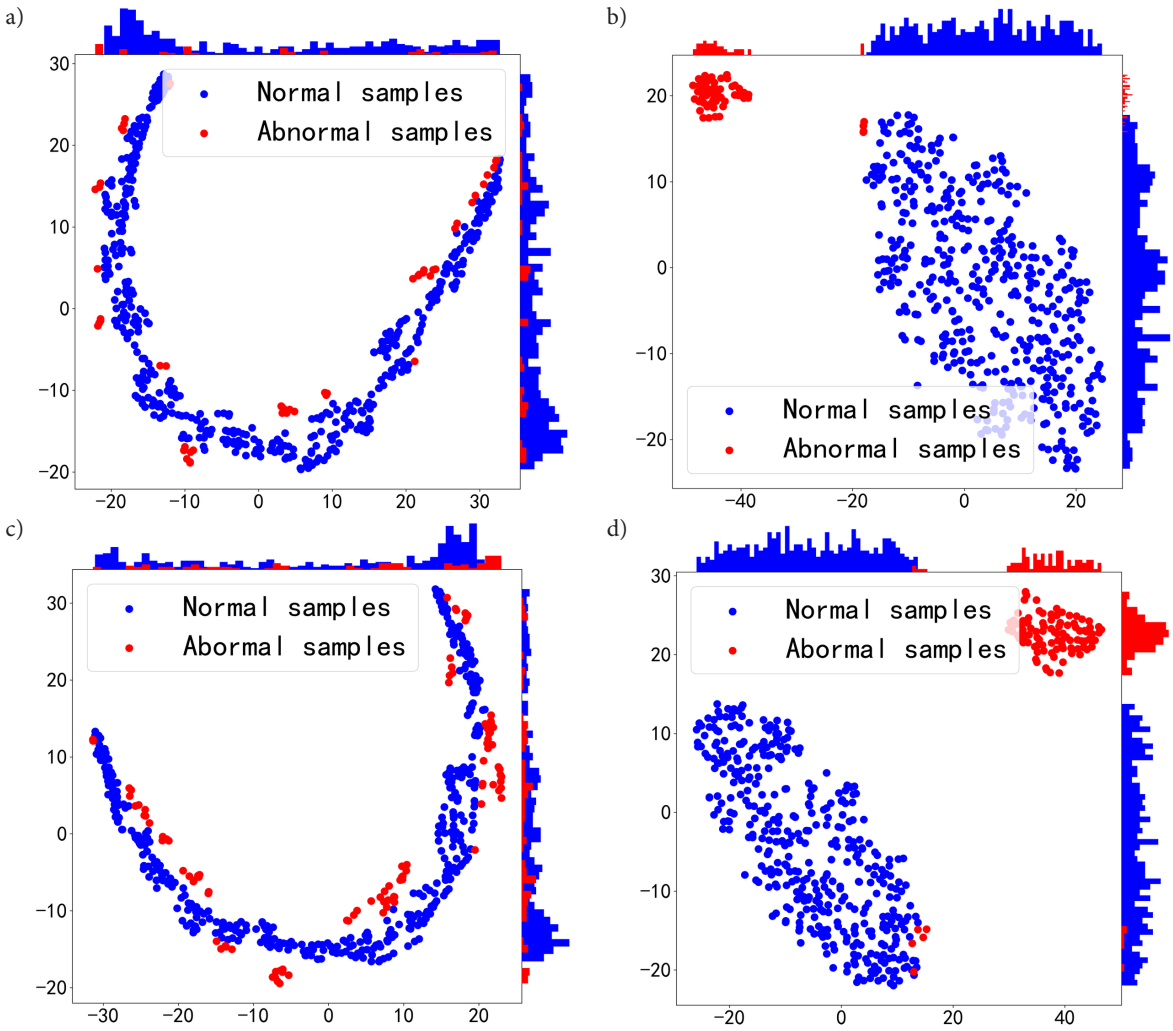


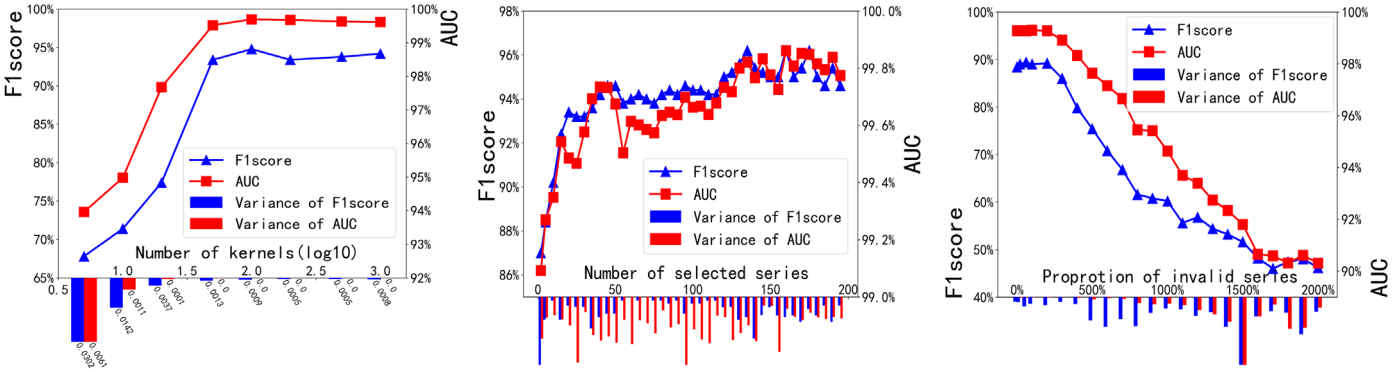
Fig. 11. Comparison on low dimensional distribution of original data with selected generated series data: a) Dataset5 original distribution, b) Dataset5 selected series distribution, c) Dataset7 original distribution, d) Dataset7 selected series distribution

in Fig. 12. In the test to analyze the impact of the number of convolutional kernels on the detection results, only the top 5 feature series are selected as the detection object and the reason is that using fewer features makes it more evident that the change in the detection result comes from the anomalous expressing ability of the generated series. In Fig. 12(a), as the number of convolutional kernels increases, the values of and gradually raise and the variance gradually decreases until reaching a steady state after kernels exceed 100. From this, it can be concluded that increasing the number of kernels can increase the probability of obtaining anomaly-sensitive series and improve the stability of detection result, but the boosting effect will touch a ceiling after a specific number of kernels. Fig. 12(b)(c) provide the analysis of the impact of the number of selected series used in isolation forest on the results and the impact of the proportion of invalid series among the selected series on the results, respectively.

Table 2. Comparison between proposed method and baselines (AUC)

	ECOD	COPOD	ABOD	CBLOF	HBOS	KNN	PCA	OCSVM	Auto Encoder	Ablation Experiment	
										Isolation forest	Proposed method
Dataset1	0.9265	0.9830	0.9119	0.7881	0.9400	0.9875	0.9850	1.0	0.7865	0.9751	1.0
Dataset2	0.7625	0.8465	0.8525	0.9890	0.8015	0.9395	0.8285	0.9450	0.7780	0.9159	0.9991
Dataset3	0.6715	0.6930	0.8950	0.9919	0.6825	0.9875	0.7160	0.9625	0.8007	0.9239	0.9999
Dataset4	0.5480	0.5300	0.9500	0.9329	0.5675	0.9565	0.5935	0.8830	0.8513	0.8620	0.9807
Dataset5	0.5235	0.5540	0.8985	0.7510	0.5370	0.9420	0.5435	0.7975	0.9768	0.7792	0.9914
Dataset6	0.5369	0.5609	0.8215	0.8880	0.5400	0.8887	0.5683	0.7314	0.6004	0.7884	0.9630
Dataset7	0.5226	0.5078	0.8179	0.7532	0.5167	0.9010	0.5374	0.7375	0.8422	0.7150	0.9687
Dataset8	0.9137	0.8898	-	0.8005	0.8041	0.9063	0.6039	0.9156	-	0.9140	0.9957
Dataset9	0.7262	0.7206	-	0.6566	0.7210	0.7056	0.6634	0.7171	-	0.7246	0.9339
Dataset10	0.9056	0.8532	-	0.8820	0.6502	0.8966	0.6404	0.9052	-	0.9053	0.9983
Dataset11	0.6274	0.7764	0.8862	0.8005	0.7597	0.9178	0.5027	0.5334	0.5768	0.9036	0.9683

Note: The ABOD and Autoencoder can only be used for multi-dimensional data, thus, there is no result in the Dataset8~ Dataset10 which are one-dimensional dataset



a) Impact of the number of kernels *b) Impact of the number of selected series* *c) Impact of the proportion of invalid series*

Fig. 12. Impact of key parameters on the result

From the Fig. 12(b), it can be found that the detection accuracy raises with the increase of the number of selected series. When the number of selected series exceeds 5, the change rate of and are less than 3.17% and 0.4% respectively, and the variance of both are slightly reduced. It indicates that using only a few series in the top of the anomaly-sensitive rank can significantly distinguish normal data from abnormal, meanwhile the increase in the number of selected series can improve the stabilization of the detection. Therefore, the number of the selected feature series in the study is set to 10 to avoid excessive information loss and feature series redundancy in detection object. Fig. 12(c) record the impact of proportion of invalid series, and the experiment is conducted by combining top 5 series and some invalid series at the bottom of anomaly sensitivity rank to be the detecting object of the isolation forest. As it depicted, the detection result is relatively stable when the proportion of invalid series is in the range of 0%-100%, which is mainly contributed by the robustness of the isolation forest. As the proportion continually raise, drops to the unacceptable points. The main reason is that the massive invalid series cover the abnormally sensitive series, distract the tree-building process of the isolation forest, leading to the almost useless result. Combining the above analysis can conclude that the proposed method has strong robustness.

3.3.3. Comparing with other methods

We compare the proposed method with the commonly used anomaly detection algorithms, and the ablation experiment is also conducted to verify the effect of the random convolution kernels. The programs of common anomaly detection algorithms are provided by the anomaly detection package named pyod (https://github.com/yzhao062/pyod). Among them, we fine-tune the structure and parameters of the Autoencoder to ensure the net can fit the training data and reach a steady state.

- Table 2 records the of methods applied in each dataset. The best records are marked in bold.
- Fig. 13 records ROC curve of each dataset, which are plotted base on the anomaly scores provided by each method.
- Fig. 14 record the impact of the rate of the degradation on the result. The degradation rate of Dataset1~Dataset5 varying from 4% to 0.25% with the difficulty of detecting abnormal gradually increase.

As shown in Table 2, the ABOD, CBLOF and KNN achieve relatively high on Datasets1~7, exceeding the isolation forest, while on Dataset8~11, the result of ECOD, OCSVM and isolation forest are close. The performance gap indicates that different methods have different applicability on different data types. But the proposed method benefits from the cross-dimensional and multi-scale feature extraction ability of the random convolution kernel, which can effectively process data of different frequencies, and achieve superior detection results on both data types. It can also be found in the ablation experiment that the combination of the random convolution kernel greatly boosts the detection effect of the isolation forest.

The ROC curves depicted in Fig. 13 visualizes the performance in separating the abnormal from the normal by using the anomaly scores of each method. It can be found that under almost all of the false positive rates the accuracy of proposed method surpasses the others. (a) Impact on precision

Fig. 14 shows that most anomaly detection algorithms gradually lose its ability to identify outlies as the degradation rate decreases. Among them, the precision of KNN on some datasets is close to the proposed method but the recall is inferior, indicating that using Euclidean distance to measure the difference between normal and abnormal of these datasets is not accurate enough. As the dimension increased, the KNN has the risk of distance failure, which will lead to a further decline in detection effect. The performance of ABOD suggests that the perspective of angle is also inappropriate to distinguish outliers. The CBLOF detects abnormalities from data density, and its performance is close to the proposed method when abnormalities are more apparent in Dataset2 and Dataset3, but declines significantly as the degradation rate decreases. The stability of precision and recall of the proposed method outperforms others, and the decline is much smaller than isolation forest used alone, which can prove that the feature series calculated by random convolution kernel and feature selection are more stable in anomaly expression.

4. Discussion

The methods performance by training using normal samples only is discussed in this section. It can be found in Table 2 that Autoencoder of DNN method performs poorly among these datasets, which probably caused by the mixture of anomalies in the training set. Therefore, we choose the Autoencoder to the comparison with the proposed method. The validation strategy is as follow: Firstly, training the both methods with pure normal data in training set of Dataset1~7, then testing the model using the complete training dataset to set the anomaly threshold according to its percentage of anomalies, and finally evaluating it on the test dataset. The of the Autoencoder and the proposed method are shown in Table 3, and the best records are marked in bold.

As shown in Table 3, the Autoencoder obtains the priority on six datasets, meanwhile the proposed method is close to its with a maximum fallback of 0.09%. Comparing to the condition that the models trained by abnormal mixed datasets, the performance of Autoencoder decreases significantly while the proposed method only decreases by 3.61% at the maximum. This result confirms the aforementioned analysis that when the training samples are mixed with abnormal data, the DNN models will indiscriminately learn and reconstruct the normal and abnormal data due to its powerful fitting ability, thus causing deviation in representation of the normal samples and leading to the poor performance. From the comparison, we can conclude that the proposed method is less sensitive to the purity of the training data. Considering that the desirable pure training dataset is difficult to ob-

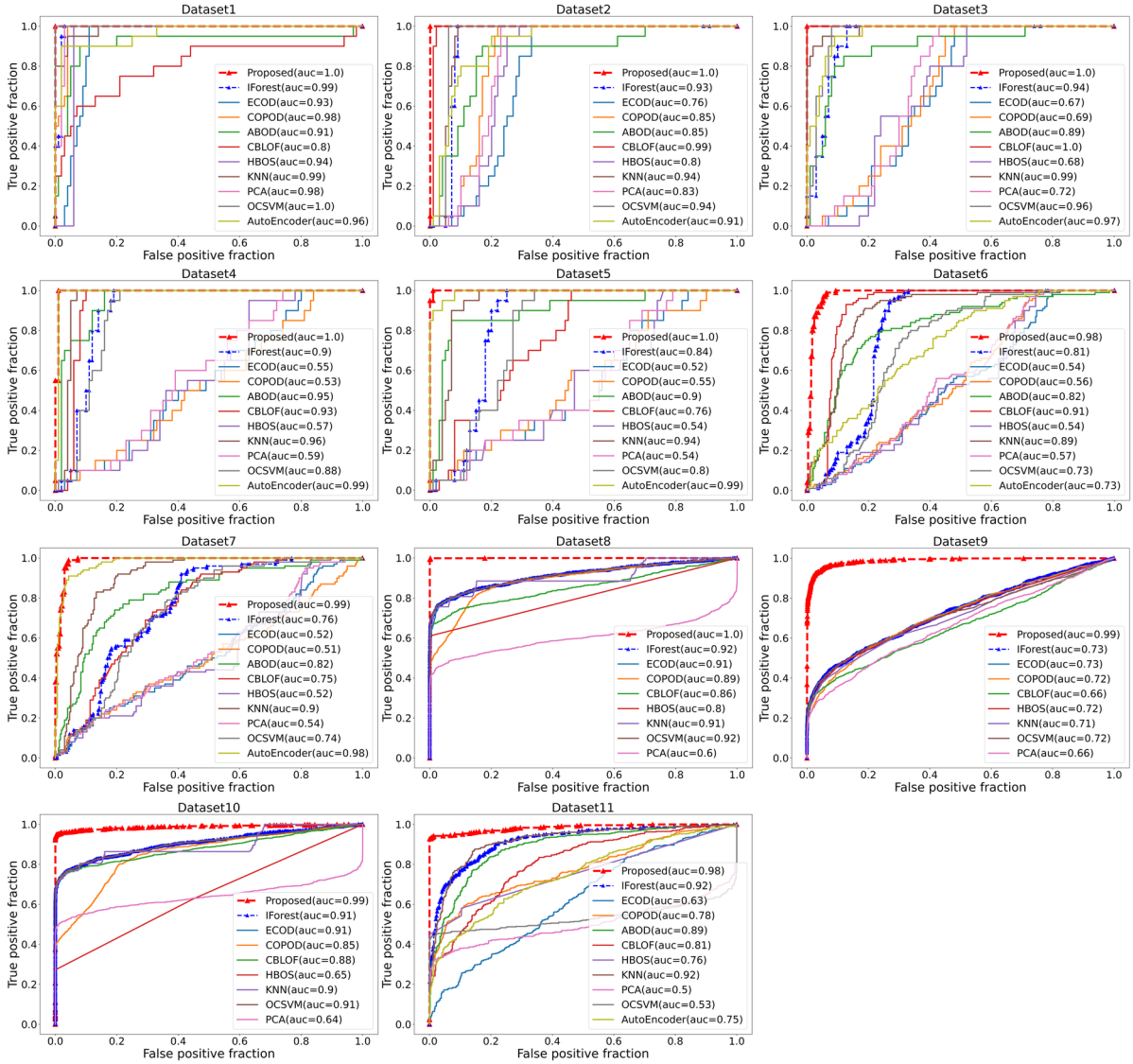


Fig. 13. ROC curve of all methos in each dataset

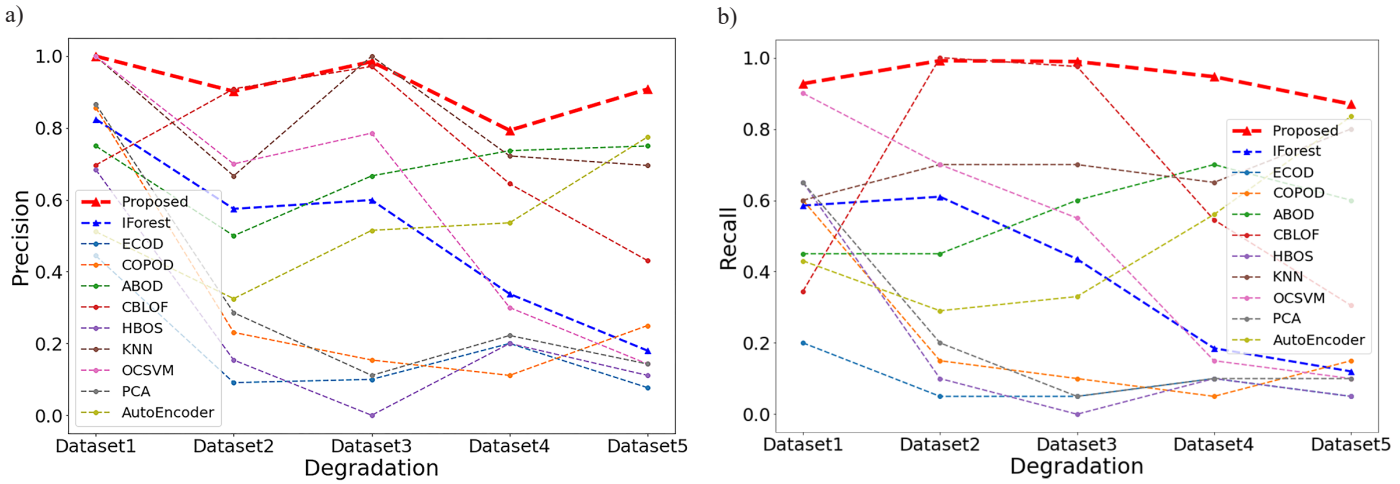


Fig. 14 Impact of the rate of the degradation: a) Impact on precision, b) Impact on recall

tain in practical, the proposed method is more suitable for realistic scenario.

4. Conclusions

In this paper, we propose an anomaly detection method combining random convolution kernel and isolation forest. The method consists of three parts: feature series generation with random convolution kernel, feature filtering based on time series decomposing and anomaly

(

Table 3. Performance comparison on pure and mixed training dataset (AUC)

Datasets	Pure normal training dataset		Abnormal mixed training dataset	
	Autoencoder	Proposed method	Autoencoder	Proposed method
Dataset1	1.0	1.0	0.7865	1.0
Dataset2	1.0	1.0	0.7780	0.9980
Dataset3	1.0	1.0	0.8007	0.9997
Dataset4	0.9994	0.9939	0.8513	0.9782
Dataset5	0.9995	0.9993	0.9768	0.9895
Dataset6	1..0	0.9991	0.6004	0.9630
Dataset7	0.9973	0.9988	0.8422	0.9687

detection by isolation forests. The first two part combined as an automatic feature generation method alleviates the reliance of manual feature engineering, and the generated anomaly sensitive feature series enhances the isolation forest performance the on anomaly-mixed data conditions. The main contributions of the research are that: (1) We apply the concept of random convolution kernel to the anomaly detection task and established the initialization strategy of kernel parameters. (2) We propose a feature series selection method based on time series decomposing, and achieve automatic anomaly-sensitive feature series generation by combining it with random convolution kernel. In the experiment the proposed method outperforms the other commonly used methods on different types of data, providing a new solution to the unsupervised anomaly detection problem.

Acknowledgement

We greatly appreciate the support provided by the National Natural Science Foundation of China (Grant No. 62176262).

References

1. Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 2013; 35(8): 1798–1828, https://doi.org/10.1109/TPAMI.2013.50.
2. Calheiros R N, Ramamohanarao K, Buyya R et al. On the effectiveness of isolation-based anomaly detection in cloud data centers: On the effectiveness of isolation-based anomaly detection in cloud data centers. Concurrency and Computation: Practice and Experience 2017; 29(18): e4169, https://doi.org/10.1002/cpe.4169.
3. Chalapathy R, Chawla S. Deep Learning for Anomaly Detection: A Survey. 2019. http://arxiv.org/abs/1901.03407
4. Cheng Z, Wang S, Zhang P et al. Improved autoencoder for unsupervised anomaly detection. International Journal of Intelligent Systems 2021; 36(12): 7103–7125, https://doi.org/10.1002/int.22582.
5. Dempster A, Petitjean F, Webb G I. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery 2020; 34(5): 1454–1495, https://doi.org/10.1007/s10618-020-00701-z.
6. Dempster A, Schmidt D F, Webb G I. MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Virtual Event Singapore, ACM: 2021: 248–257, https://doi.org/10.1145/3447548.3467231.
7. Guo K, Liu D, Peng Y, Peng X. Data-Driven Anomaly Detection Using OCSVM with Boundary Optimzation. 2018 Prognostics and System Health Management Conference (PHM-Chongqing), Chongqing, IEEE: 2018: 244–248, https://doi.org/10.1109/PHM-Chongqing.2018.00048.
8. Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks. Science 2006; 313(5786): 504–507, https://doi.org/10.1126/science.1127647.
9. Jahromi A F, Hajiloei M, Dehghani Y, Lahoninezhad S. Improved subspace-based and angle-based outlier detections for fuzzy datasets with a real case study. Journal of Intelligent & Fuzzy Systems 2022; 42(6): 5471–5481, https://doi.org/10.3233/JIFS-211955.
10. Jimenez A, Raj B. Time Signal Classification Using Random Convolutional Features. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, IEEE: 2019: 3592–3596, https://doi.org/10.1109/ICASSP.2019.8682489.
11. Kingma D P, Welling M. Auto-Encoding Variational Bayes. 2014. http://arxiv.org/abs/1312.6114
12. Lei Z, Zhu L, Fang Y et al. Anomaly detection of bridge health monitoring data based on KNN algorithm. Journal of Intelligent & Fuzzy Systems 2020; 39(4): 5243–5252, https://doi.org/10.3233/JIFS-189009.
13. Li Y, Wang Y, Ma X. Variational autoencoder-based outlier detection for high-dimensional data. Intelligent Data Analysis 2019; 23(5): 991–1002, https://doi.org/10.3233/IDA-184240.
14. Liu F T, Ting K M, Zhou Z-H. Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, IEEE: 2008: 413–422, https://doi.org/10.1109/ICDM.2008.17.
15. Mensi A, Bicego M. Enhanced anomaly scores for isolation forests. Pattern Recognition 2021; 120: 108115, https://doi.org/10.1016/j.patcog.2021.108115.
16. Puggini L, McLoone S. An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data. Engineering Applications of Artificial Intelligence 2018; 67: 126–135, https://doi.org/10.1016/j.engappai.2017.09.021.
17. Saxe A M, Koh P W, Chen Z et al. On Random Weights and Unsupervised Feature Learning. International Conference on Machine Learning (ICML 2011), Bellevue, Washington, USA, 2011.
18. Smith W A, Randall R B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. Mechanical Systems and Signal Processing 2015; 64–65: 100–131, https://doi.org/10.1016/j.ymssp.2015.04.021.
19. Tan C W, Dempster A, Bergmeir C, Webb G I. MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. Data Mining and Knowledge Discovery 2022. doi:10.1007/s10618-022-00844-1, https://doi.org/10.1007/s10618-022-00844-1.
20. Tian H D, Khoa N, Anaissi A et al. Concept Drift Adaption for Online Anomaly Detection in Structural Health Monitoring. PROCEEDINGS

OF THE 28TH ACM INTERNATIONAL CONFERENCE ON INFORMATION & KNOWLEDGE MANAGEMENT (CIKM '19) 2019: 2813–2821, https://doi.org/10.1145/3357384.3357816.
21. Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th international conference on Machine learning - ICML '08, Helsinki, Finland, ACM Press: 2008: 1096–1103, https://doi.org/10.1145/1390156.1390294.
22. Zhang L, Lin J, Karim R. Adaptive kernel density-based anomaly detection for nonlinear systems. Knowledge-Based Systems 2018; 139: 50–63, https://doi.org/10.1016/j.knosys.2017.10.009.
23. Zhao J, Itti L. Decomposing time series with application to temporal segmentation. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, IEEE: 2016: 1–9, https://doi.org/10.1109/WACV.2016.7477722.
24. Zhong S, Fu S, Lin L et al. A novel unsupervised anomaly detection for gas turbine using Isolation Forest. 2019 IEEE International Conference on Prognostics and Health Management (ICPHM), San Francisco, CA, USA, IEEE: 2019: 1–6, https://doi.org/10.1109/ICPHM.2019.8819409.
25. Zong B, Song Q, Min M R et al. Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. ICLR, 2018.