Smart Structures and Systems, *Vol. 29, No. 1 (2022) 129-140*
https://doi.org/10.12989/sss.2022.29.1.129

Detection of multi-type data anomaly for structural health monitoring using pattern recognition neural network

Ke Gao [a], Zhi-Dan Chen [b], Shun Weng*, Hong-Ping Zhu [c] and Li-Ying Wu [d]

School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, 1037 Luoyu-road, Wuhan, China

(*Received April 14, 2021, Revised August 11, 2021, Accepted August 12, 2021*)

Abstract. The effectiveness of system identification, damage detection, condition assessment and other structural analyses relies heavily on the accuracy and reliability of the measured data in structural health monitoring (SHM) systems. However, data anomalies often occur in SHM systems, leading to inaccurate and untrustworthy analysis results. Therefore, anomalies in the raw data should be detected and cleansed before further analysis. Previous studies on data anomaly detection mainly focused on just single type of data anomaly for denoising or removing outliers, meanwhile, the existing methods of detecting multiple data anomalies are usually time consuming. For these reasons, recognising multiple anomaly patterns for real-time alarm and analysis in field monitoring remains a challenge. Aiming to achieve an efficient and accurate detection for multi-type data anomalies for field SHM, this study proposes a pattern-recognition-based data anomaly detection method that mainly consists of three steps: the feature extraction from the long time-series data samples, the training of a pattern recognition neural network (PRNN) using the features and finally the detection of data anomalies. The feature extraction step remarkably reduces the time cost of the network training, making the detection process very fast. The performance of the proposed method is verified on the basis of the SHM data of two practical long-span bridges. Results indicate that the proposed method recognises multiple data anomalies with very high accuracy and low calculation cost, demonstrating its applicability in field monitoring.

Keywords: data anomaly detection; feature extraction; pattern recognition neural network; structural health monitoring

1. Introduction

Structural health monitoring (SHM) systems provide detailed information about events on structures and therefore produce large volumes of monitoring data. The monitored data are usually used to detect structural damage and assess the structure's condition (Tian *et al*. 2019, Zhu *et al*. 2020). To make a correct evaluation and meaningful conclusion regarding structural safety and performance, the quality of the received data must be ensured (Bao *et al*. 2019b). Given that SHM systems are usually operated in harsh and noisy environments, abnormal data are inevitable. In particular, outliers, trends, missing and over-range signals have frequently been observed in the monitored data (Smarsly and Law 2014, Luo *et al*. 2015). This situation presents an obstacle to the automatic warning and evaluation of structural damages or accidents, because it is difficult to distinguish which abnormal data are caused by sensor faults or by structural damage (Gul and Catbas 2009). According to field experience, data quality is influenced by factors such as data loss in wireless transmission, electromagnetic interference from construction machinery, noise from the environment or transmission circuit and sensor failure (Ni *et al*. 2009). Consequently, the identification of data anomalies is an important pre-processing step for SHM.

Extensive research has been carried out for the detection and cleansing of different data anomalies. Data loss is the most common anomaly in the SHM systems of large-scale structures (Yang and Nagarajaiah 2016). The missing anomaly is easy to detect but difficult to compensate; therefore, works on the data loss mainly focus on the recovery of missing data. In general, the missing data are restored by the data of correlated sensors (Zhang and Luo 2017) or by other types of data based on the physical relations between the data (Chen *et al*. 2019). As for another common data fault, outliers in the raw data may cause false alarms or misleading analysis results in SHM (Zhang *et al*. 2018). Approaches on outlier detection mainly consist of statistical-based method (Yuen and Mu 2012, Cai *et al*. 2020), distance-based method (Yuen and Ortiz 2017) and clustering-based method (Alessandra *et al*. 2015, Titouna *et al*. 2019). In terms of the measurement noise (Jiang *et al*. 2007, Zvokelj *et al*. 2011, Katicha *et al*. 2014), bias (Calabrese *et al*. 2012, Yang and Nagarajaiah 2014) and other errors in the raw data, such as trend and drift (Wang *et al*. 2017, Peng *et al*. 2018), plenty of studies have been carried out. In summary, the above techniques are all focused on handling single-type anomaly.

Compared with single-type anomaly detection, the capability of detecting multiple anomalies is more suitable and valuable for practical SHM. Kullaa (2013) provided a

*Corresponding author, Ph.D., Professor,
E-mail: wengshun@hust.edu.cn
[a] Ph.D. Candidate
[b] Ph.D. Candidate
[c] Ph.D., Professor
[d] Master Student

http://www.techno-press.org/?journal=sss&subpage=7
ISSN: 1738-1584 (Print), 1738-1991 (Online)

generalised likelihood ratio approach to identify seven fault types of abnormal data in a sensor network based on a multiple hypothesis test. Chang *et al*. (2017) developed an autoregressive modelling and Kalman estimator-based sensor fault recognition method that can detect three types of data anomalies. Fu *et al*. (2019) integrated a distributed similarity test and an artificial neural network to identify drift, spikes and bias anomalies in wireless sensor networks. These mentioned techniques can detect multiple data anomalies, but all of them are verified by numerical simulations or laboratory experiments with artificial sensor faults, which may be very different from the field monitoring.

Recently, researchers have attempted to handle the data anomalies in field monitoring. Huang *et al*. (2020) applied the dynamic independent component analysis to identify two types of data anomalies in the SHM system of a cable-stayed bridge and then infer the structural damage. Bao *et al*. (2019a) applied data visualisation and deep learning network to detect seven types of data anomalies in the SHM system of a long-span bridge in China. Despite these studies, further works are still required to achieve accurate and efficient data anomaly detection.

Pattern recognition of anomalies is the fundamental theory in the aforementioned studies. There are many classification algorithms for pattern recognition, including linear discriminate analysis (LDA) (2004), support vector machine (SVM) (Widodo and Yang 2007) and neural network (Bishop 2006). Unlike LDA and SVM, which use the fixed basis function, neural network is more adjustable to complex problems. In general, neural networks are often used in the research field of image classification (2009). Later, more and more researchers applied this classification algorithm in the field of civil engineering. Lejla *et al*. (2017) used the neural network method to estimate the state and detect the anomaly in a thermal power plant via a health monitoring system with multilayer perception. Nick and Aziminejad (2021) used a modal strain energy-based damage index as the input layer to the artificial neural network to estimate single or multiple damage states in bridges. The artificial neural network has a simpler structure and is more suitable for the situation where the input and output have obvious characteristics in data anomaly detection. And the artificial neural network is more computationally efficient and is more suitable for online SHM. However, the challenge is still existing that the types of anomally in different real structures will be different, and the extracted features are difficult to apply to different data. Therefore, how to propose a general method for identifying anomaly data is of great significance to the health monitoring of the real structure.

In order to achieve a fast and accurate detection of multiple data anomalies, this study provides an approach based on feature extraction and pattern recognition neural network (PRNN). Firstly, nine features are extracted from a time-series data sample and then the long time-series sample is converted into a very short feature vector sample. Next, a small set of randomly selected and manually labelled feature vector samples are used to train a feedforward PRNN. Owing to the short feature sample and small training set, the detection and classification of data anomalies become very efficient. A database of acceleration data from the SHM system of a real long-span highway bridge is used to verify the accuracy of the proposed method. Encouraging results show that the trained PRNN is capable of detecting anomaly occurrences and identifying anomaly patterns from large amounts of monitoring data with high accuracy and low time cost. Besides, the influences of optimal feature selection and label modification on the detection accuracy are also investigated for deeper understanding and better detection accuracy. Furthermore, the proposed method is applied on another practical long-span railway bridge to demonstrate its general applicability.

2. Methodology of data anomaly detection

The proposed method consists of two parts. The first part is the feature extraction of the raw data by calculating nine characteristic indices, and the second is the construction and training of a PRNN for data pattern classification.

2.1 Feature extraction from raw data

Inspired by the recognition process of human beings on different objects, a complex object can be described by a small number of essential features rather than a large amount of details. Therefore, feature extraction has the power to make object recognition simple and fast. In terms of data anomaly, we tend to use the simplest and most suitable features to represent various types of anomalies in the time-series data and then bring these features into the pattern recognition algorithm to perform anomaly classification.

According to the common data anomalies in SHM systems (Fig. 1), each anomaly pattern has one or several unique characteristics that can distinguish it from others. Therefore, a series of features are conceived or selected based on our prior knowledge of the data pattern.

Firstly, several parameters are conceived to represent the anomalies with special characteristics. For example, the ‘missing’ anomaly is classified by the empty ratio, which

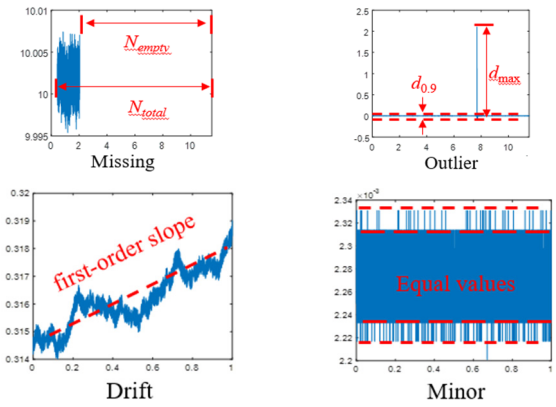


Fig. 1 Regular data anomalies in SHM

reindicates the proportion of empty value in a data sample as

$$\boldsymbol{Empty\_ratio} = \frac{\boldsymbol{N_{empty}}}{\boldsymbol{N_{total}}} \tag{1}$$

where N_{empty} means the number of the empty values and N_{total} means the total number of the data points.

The 'outlier' anomaly can be distinguished from other pattern by the peak intensity as

$$\boldsymbol{Peak_{intensity}} = \frac{\boldsymbol{d_{0.9}}}{\boldsymbol{d_{max}}} \tag{2}$$

where $d_{0.9}$ means the distance between the upper and lower boundaries of 90% of the data points, and d_{max} means the distance between the upper and lower boundaries of all data points.

The 'drift' anomaly often has an overall trend to go upward or downward, thus it is easily to be recognized by the first order slope of the linear fitting line of the data points. The first order slope is calculated by the *polyfit* function in MATLAB. Besides, the drift level is evaluated by the linearity as

$$\boldsymbol{Linearity} = \frac{\boldsymbol{max(\Delta x_i)}}{\boldsymbol{xmin_{max}}} \tag{3}$$

where Δx_i is the difference between the data points and the linear fitting line, x_{max} and x_{min} are the maximum and minimum values in the data vector.

As for the 'minor' anomaly, the data fluctuate among several fixed values rather than random values, hence we use the equal value ratio to distinguish it from others as

$$\boldsymbol{Equal_{ratio}} = \frac{\boldsymbol{N_{equal}}}{\boldsymbol{N_{total}}} \tag{4}$$

where N_{equal} is the number of the fixed values and N_{total} is the number of all the data points.

In addition, since the data samples contain amounts of data points, some popular statistical characters are also considered, namely the standard deviation, the median absolute deviation, the form factor and the over-average ratio as follows.

$$\boldsymbol{SD} = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{\boldsymbol{(x_i - x_a)^2}}{\boldsymbol{n-1}}} \tag{5}$$

where SD means the standard deviation, x_i is the value of each data point, x_a is the mean value of the data vector and n is the number of all the data points in a data vector.

$$MAD = median(|x_i\text{-}median(\mathbf{X})|) \tag{6}$$

where MAD means the median absolute deviation, $median(\mathbf{X})$ indicates the median value of the data vector and $|\ |$ is the operator of absolute value.

$$\boldsymbol{Form_{factor}} = \frac{\boldsymbol{rms(X)}}{\boldsymbol{mean(|X|)}} \tag{7}$$

where $rms(\mathbf{X})$ and $mean(|\mathbf{X}|)$ are the root mean square and the absolute mean value of the data vector, respectively.

$$\boldsymbol{Over\_ave} = \frac{\boldsymbol{N_{over}}}{\boldsymbol{N_{total}}} \tag{8}$$

where N_{over} is the number of the time-series data curve across the average line.

As a result, this study proposes nine features for pattern recognition.

2.2 Training of PRNN

The feedforward network usually has an input layer, an output layer and at least one hidden layer. The neurons of each layer can receive the neuron signals of the previous layer and generate signals to output to the next layer. Fig. 2 illustrates the basic framework of the network. The first layer is called the input layer, the last layer is called the output layer and the middle layer is called the hidden layer.

The hidden layer first performs the summation operation of Eq. (9) on the input parameters X_i of the input layer

$$\boldsymbol{z_j(X)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{w_{ji}^{(m)} X_i + b_j} \tag{9}$$

where w_{ji} and b_j are the weight and the bias of the j-th neuron, respectively. n is the number of neurons of the input layer. The superscript m indicates the quantity of the hidden layers. The network will become more complicated if the quantity of the hidden layer increases.

Next, the input-output mapping relationship of the neuron is established via the activation function

$$h_j = f(z_j(X)) \tag{10}$$

The activation function introduces a non-linear factor to the neuron, so that the neural network can approximate any non-linear function, so that the neural network can be applied to the non-linear model. There are many different types of activation functions. The common sigmoid activation function expressions are as follows

$$f(z) = \frac{1}{1+e^{-z}} \tag{11}$$

The output of the hidden layer will be treated as the input to the output layer. The output Y_i can also be written as

$$Y_i = \sum_{j=1}^{n} v_{ji} z_{ji} + r_{ji} \tag{12}$$

After the connection between neurons is established, it is necessary to determine the weight coefficients and biases between each neuron. The loss function will be used as the evaluation criterion of the predicted value and the true value to determine the weight coefficients and biases. For pattern recognition with n output nodes, the cross-entropy method is often used as the loss function shown as Eq. (13)

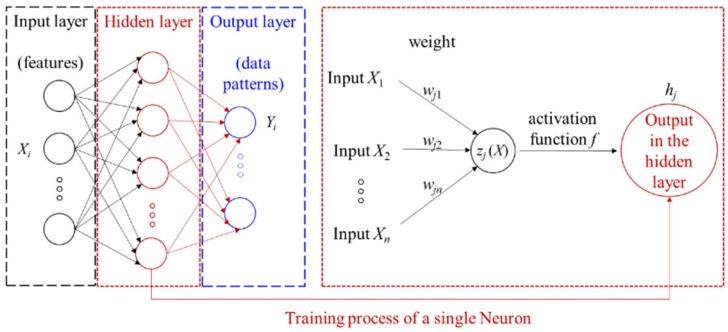


Fig. 2 Framework of the PRNN

$$\mathrm{L}=-\sum_{i=1}^{n}[Y_i\, ln(t_i)+(1-Y_i)\, ln(1-t_i)] \tag{13}$$

where the t_i is the true value of the classification and Y_i is the calculated value of the classification. The weight coefficients and biases of each neuron can be determined by minimising the loss function L with optimisation algorithm.

$$\{w, b, v, r\} = argmin\ \mathrm{L} \tag{14}$$

The Scaled Conjugate Gradient algorithm is used to minimize L. After the parameters $\{w, b, v, r\}$ are determined, the training of the entire neural network is completed. Then bring new inputs into the network, the input can be accurately classified. In this study, one hidden layer with 100 neurons is chosen.

2.3 Terminologies and evaluation indices

In this study, several common indices are used to evaluate the performance of the proposed PRNN-based classifier, namely the 'Recall', 'Precision', and the receiver operating characteristic (ROC) curve. These indices are defined as follows.

The 'Recall' of one data pattern is the ratio of the recognised true samples to the total true samples of that pattern as

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \times \mathbf{100\%} \tag{15}$$

where TP is true positive, and FN is false negative. This metric is affected by the misclassification of one pattern itself.

The 'Precision' is the ratio of the recognised true samples of one data pattern to all the samples classified as that pattern.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \times \mathbf{100\%} \tag{16}$$

where FP means false positive. The precision of one pattern is affected by the misclassification from other patterns to that pattern.

The abscissa of the ROC curve is the false positive rate and the ordinate is true positive rate. The closer the ROC curve is to the upper left corner, the better the performance of the classifier. The quality of the classifier can be quantified by the area under the curve (AUC). The closer the area is to 1, the better the classifier.

3. Field validation on Su-Tong Yangtze River Highway Bridge

To verify the effectiveness of the proposed method, the monitored data of a long-span cable-stayed bridge in China are used as the database for training and testing.

3.1 Engineering background

In this case, the acceleration data of the Su-Tong Yangtze River Highway Bridge (SYRHB) from January 1 to February 30, 2012, are considered. The data are provided by the committee of the 1st International Project Competition for Structural Health Monitoring (IPC-SHM), 2020 (Bao *et al*. 2021). The bridge has a main span of 1088 m, two side spans of 300 m each and two 306 m-high towers. The deployment of the acceleration sensors is shown in Fig. 3, including 14 two-channel accelerometers on the deck, 2 two-channel accelerometers at the tower top and 2 three-channel accelerometers at the tower bottom. The data anomalies that occurred in the database are divided into six patterns: missing, minor, outlier, square, trend and drift. Table 1 provides a brief description of the characteristics of the normal data and the six patterns of anomalies. Moreover, the quantity ratio of each pattern indicates that there is significant class imbalance among different pattern.

Fig. 4 shows some examples of each anomaly pattern. For instance, the 'missing' anomaly has three different forms, which are empty values, constant value and constant value for most of the time. Besides, the first-order slope of the 'trend' and 'drift' patterns is much larger than that of the other five patterns. On the other hand, it is observed that there are some ambiguities between the 'outlier' and the 'normal' and between the 'trend' and the 'drift'.

The acceleration data are sampled at 20 Hz; hence, the hourly data of each sensor result in a 72,000 × 1 vector. The raw data are split into 1 h-long segments, and each segment is regarded as a sample for anomaly detection. Thus, there are a total of 54,720 samples (38 channels × 24 hours × 60 days). These samples are all pre-labelled in the dataset. A single-label classification criterion is employed for the labelling, that is, a sample with multiple anomalies is

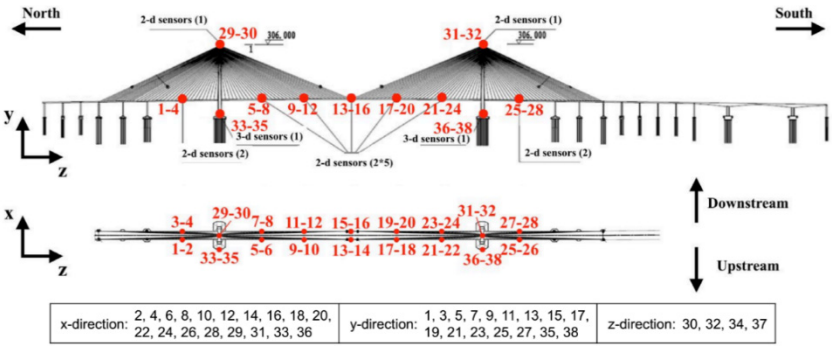


Fig. 3 Deployment of accelerometers on the bridge

Table 1 Description of normal data of the six patterns of data anomaly

No.	Data pattern	Description	Quantity in the database
1	Normal	The time response is normal oscillation curve; frequency response is peak-like (may differ between bridges).	26472 (48.4%)
2	Missing	Most/all of the time response is missing, which makes the time and frequency response zero.	5909 (10.8%)
3	Minor	Relative to normal sensor data, the amplitude is very small in the time domain.	3425 (6.3%)
4	Outlier	One or more outliers appear in the time response.	858 (1.6%)
5	Square	The time response is like a square wave.	6210 (11.3%)
6	Trend	The data has an obvious trend in the time domain and has an obvious peak value in the frequency domain.	10336 (18.9%)
7	Drift	The vibration response is non-stationary, with random drift.	1510 (2.8%)

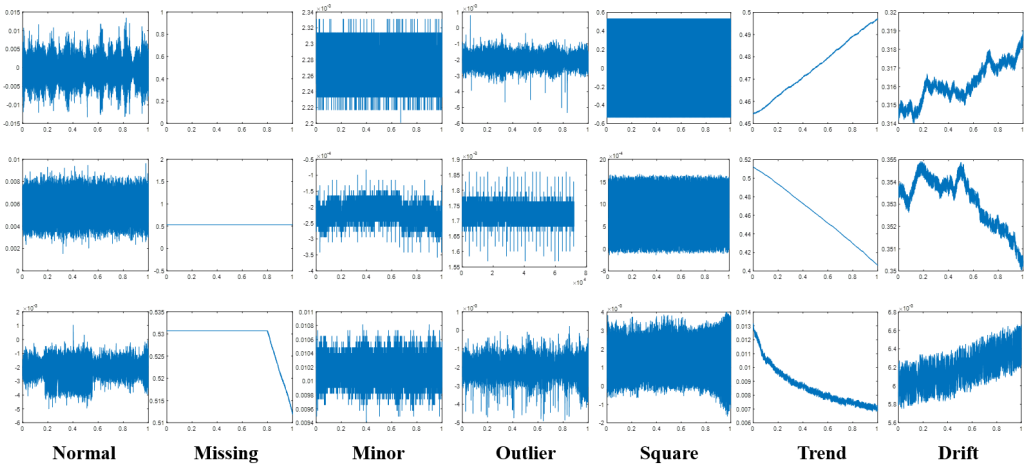


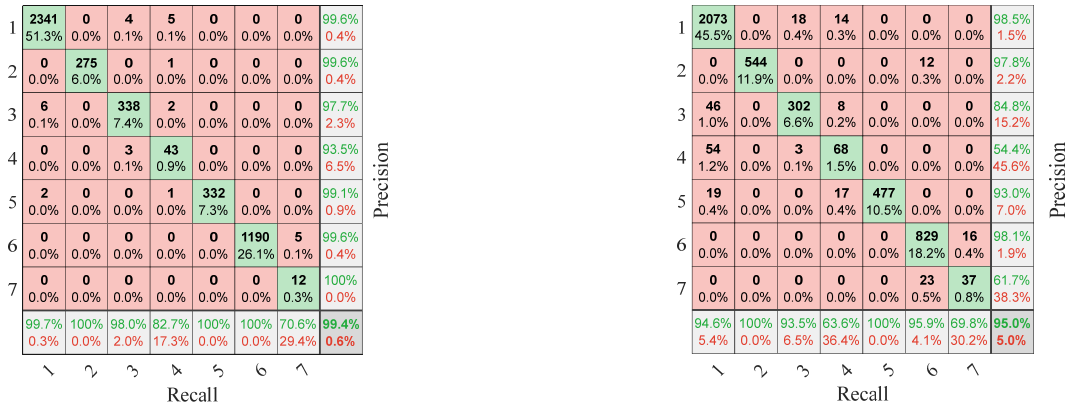
Fig. 4 Examples of each data pattern (normal and six kinds of anomalies)

labelled as the principal anomaly pattern. The third picture of the 'missing' anomaly in Fig. 4 can illustrate the problem of multiple anomalies. Most of the data is 'missing' and a small part is 'trend', so label it as 'missing'.

Consequently, the samples and labels are used to establish training and testing sets to verify the performance of the proposed method. Specifically, in the training set, samples with labels are used to train the PRNN which has a single hidden layer with 150 neurons, and in the testing set, samples without labels are recognised and given labels using the trained network. The predefined labels of the samples in the testing set are regarded as the reference and compared with the recognised labels.

3.2 Detection of data anomalies

Firstly, data of 10 days (9120 samples) were randomly chosen, half of which generated the training set while the



(a) Training set (b) Testing set

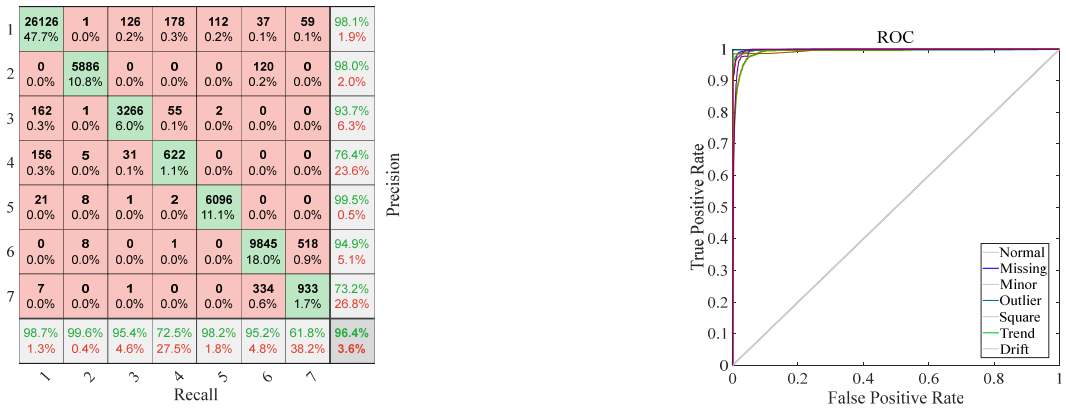
Fig. 5 Confusion matrix (Data patterns: 1-normal, 2-missing, 3-minor, 4-outlier, 5-square, 6-trend and 7-drift)

other half was used as the testing set. Fig. 5 shows the confusion matrices of the detection results, which contains the 'Recall' at the bottom row and the 'Precision' at the rightmost column. The training set and testing set give global accuracy levels of 99.4% and 95.0%, respectively.

The 'Recall' means the reliability of the pattern classification. Notes that the recalls of the 'normal', 'missing', 'minor', 'square' and 'trend' patterns are all above 93%. In particular, the recalls of the 'missing' and 'square' are both 100%. On the other hand, the number of 'outlier' sample in the training set is too small to obtain a well-trained PRNN for detecting the 'outlier' anomaly. As a result, in the testing, just 63.6% of the true 'outlier' samples are correctly classified. Similarly, the recall of the 'drift' anomaly is 69.8%, which is caused by the rare 'drift' samples in the training set.

The 'Precision' of one data pattern means the probability that a data sample is correctly recognised. Note that the precision of the 'outlier' pattern is 55.4% due to just 54 misclassified 'normal' samples. Meanwhile, the precision of the 'drift' pattern is 61.7% due to just 23 misclassified 'trend' samples. As a result, the precision is highly affected by the class imbalance, that is, when the sample number of one pattern is small, its precision will be significantly affected by the misclassified samples of other patterns.

Next, the performance of the PRNN is validated via a much larger testing set. In this case, the data of 5 days (January 5, 9, 15, 24 and 31) were selected as the training set. The specified selection aims to get as many samples of each anomaly pattern as possible to reduce the negative effect of the class imbalance in the training set. Subsequently, data of the other 55 days were used as the testing set. Fig. 6 shows the confusion matrix and the ROC curve of the detection result. The global accuracy of anomaly detection for the testing set is 96.4%. As for the 'normal', 'missing', 'minor', 'square' and 'trend' patterns, both the recall and precision keep high levels over 93%, demonstrating that the PRNN can accurately recognise these five types of anomaly. Besides, the recall of the 'outlier' is 72.5% because 20% of the true 'outlier' samples are misclassified as 'normal'. Meanwhile, just 61.8% of the true 'drift' samples are correctly recognised while 34.3% of that are misclassified as the 'trend'. Consequently, the ambiguities between 'outlier' and 'normal' and between 'trend' and 'drift' significantly drag down the detection accuracy of the 'outlier' and 'drift'. In addition, the ROC curve of each data pattern is very close to the upper left axis of the coordinate system which means that the detection accuracy for each pattern is high and the misjudgement rate



(a) Confusion matrix (b) ROC curve

Fig. 6 Detection results (Data patterns: 1-normal, 2-missing, 3-minor, 4-outlier, 5-square, 6-trend and 7-drift)

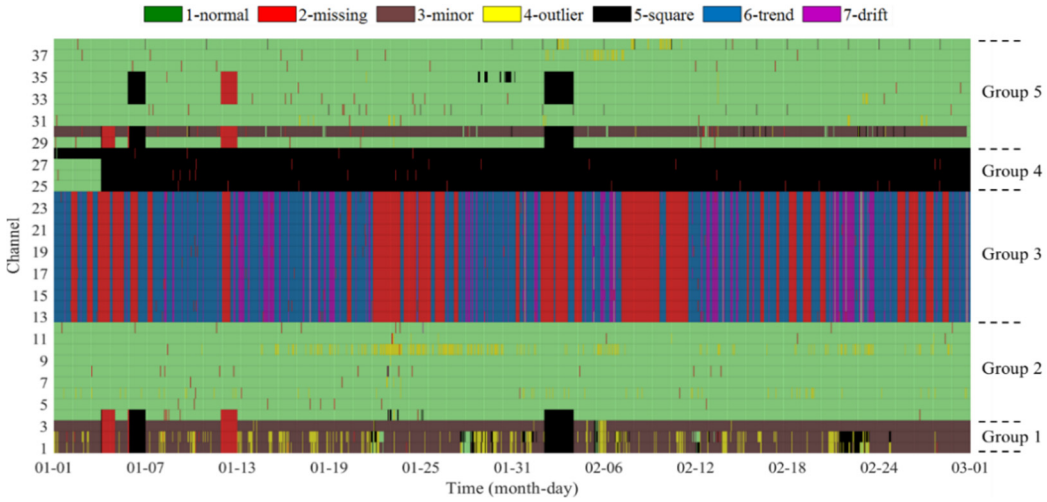


Fig. 7 Detection of the data anomalies from 2012-01-01 to 2012-02-29

is low. The area under the curve (AUC) is usually adopted to quantify the evaluation index of the accuracy of the ROC curve. The value range of AUC is between 0 and 1. The closer to 1, the higher the detection accuracy. The value of AUC of each pattern is 0.9985, 0.9978, 0.9981, 0.9918, 0.9967, 0.9985 and 0.9950 respectively. The confusion matrix, the ROC curve and AUC values show the outstanding ability of the proposed method for data anomaly detection.

As for the calculation efficiency of the proposed PRNN, the feature extraction transforms a 72,000 × 1 time-series data vector into a 9 × 1 feature vector, leading to a significant reduction of the computation in both the training and testing processes. In this case, it took about 1432 s for the feature extraction of the 60 days data, 4 s for the training of the PRNN and 4 s for classifying the data anomaly patterns (CPU: Intel i7-6700K, RAM: 16 GB, 7200 r/min HDD), which is extraordinary time-saving for data anomaly detection.

Fig. 7 shows the detected occurrence time and corresponding pattern of the data anomalies within two months. The 38 channels are roughly divided into five groups according to the classification results: channels 1–3, 4–12, 13–24, 25–28 and 29–38. Channels in a group are observed to have similar anomaly distributions and temporal trend. The data quality of groups 2 and 5 is relatively good while the other three groups are almost dominated by the six patterns of data anomalies. For group 1, nearly 70.5% of the data are 'minor' anomalies. Note that almost all the data in group 3 are anomalies, which consist of 66.2% 'trend', 26.7% 'missing' and 7.1% 'drift'. Moreover, 87.6% of the data in group 4 are 'square'. Besides, the 'missing' and 'outlier' anomalies occur occasionally in almost all channels. According to the quantities and types of anomalies, the anomalies in groups 1, 3 and 4 are mainly due to the failures of sensors or acquisition instrument. Additionally, the occasional anomalies in groups 2 and 5 are hitches caused by power cut, circuit interference and so on.

In summary, the proposed method is able to detect the data anomalies in the dataset with high global accuracy and low time cost. Nevertheless, the detection accuracy of the 'outlier' and 'drift' still needs further improvement.

4. Discussion

In this section, influences of the optimal feature selection and the accuracy of the artificial label are investigated for deeper understanding of the proposed method and further improvement of the detection accuracy based on the data of the SYRHB. Also, the proposed method is likewise applied on Ganjiang Railway cable-stayed bridge to verify the generality in engineering field.

4.1 Optimisation of feature selection

As mentioned above, most of the time cost is spent on feature extraction. Therefore, using lesser features can improve the operation efficiency of the proposed method. A question arises: How can we get accurate and reliable detection results with the least features? To answer the question, it is necessary to investigate the contribution and importance of each feature in the data anomaly detection. In this section, the investigation is based on the 5-day training set and 55-day testing set.

Firstly, the PRNN is trained using only one feature at a time and then the data samples in the testing set are recognised based on this network. Table 2 shows the recalls of each data pattern using just one feature. For instance, if we train the network using only the 'linearity', then the network can distinguish the 'normal', 'missing', 'square' and 'trend' patterns with high accuracy of over 93%. Note that even though the recall of the 'drift' is just 67.1%, it is the maximum value and much larger than others in the same row. This means that 'linearity' is the most important feature to distinguish the 'drift' pattern. Specifically, the 'linearity' feature is very helpful in detecting five data patterns. Similarly, the 'peak intensity' feature has great contributions to five types of data, especially for the 'outlier'. This is very logical because the 'peak intensity' feature is specially conceived to distinguish the 'outlier'

Table 2 Recall of each anomaly pattern with only one feature

	Linearity	Peak intensity	Equal value ratio	Over-average ratio	Empty ratio	First-order slope	Standard deviation	Median absolute deviation	Form factor
Normal	93.9	99.0	75.8	96.0	100	100	93.0	98.0	87.0
Missing	99.6	99.6	95.1	99.8	0	0	0	99.6	99.7
Minor	32.7	0	82.9	25.4	0	0	0	0	0
Outlier	14.5	68.4	0	0	0	0	0	0	0.6
Square	98.0	89.6	91.3	87.5	0	0	94.9	43.7	94.5
Trend	92.9	98.4	92.1	90.6	0	82.1	74.7	77.0	77.4
Drift	67.1	29.6	36.4	20.4	0	0	0	0	0
Global accuracy	89.0	89.3	80.9	86.4	48.4	63.9	69.9	77.7	78.2

anomaly, as presented in Section 2.1.

Additionally, both ‘equal value ratio’ and ‘over-average ratio’ features are shown to be capable of detecting four types of data with agreeable accuracy. Note that the ‘equal value ratio’ feature is very valuable for the ‘minor’ pattern. As for the rest of the features, they are important to just two or three types of data anomalies. The worst case is that the ‘null ratio’ is not capable of detecting any type of data anomaly, though this does not mean it is useless. On the other hand, the ‘null ratio’ is still required to recognise some situations in the ‘missing’ pattern.

As a result, the first four features in Table 2, namely the ‘linearity’, ‘peak intensity’, ‘equal value ratio’ and ‘over-average ratio’, are clearly essential features for data anomaly detection. Therefore, the PRNN is trained based on these four features and then used to detect the anomalies in the testing set. As shown in Fig. 8, the recalls and precisions of each data pattern is close to those shown in Fig. 6, which is based on the network trained by nine features. Consequently, using four appropriate features can obtain a global detection accuracy of 95.9%. Compared with the nine features method, it only took 672 s to extract features and 4 s for the training of the PRNN and 4 s for classifying the data anomaly patterns. In this study, we still use nine features not only to guarantee high global accuracy but also to keep a dependable network with the best recognition of each data pattern. It is worth mentioning that when the amount of data that needs to be extracted feature is huge, reducing the number of features will greatly improve the computational efficiency.

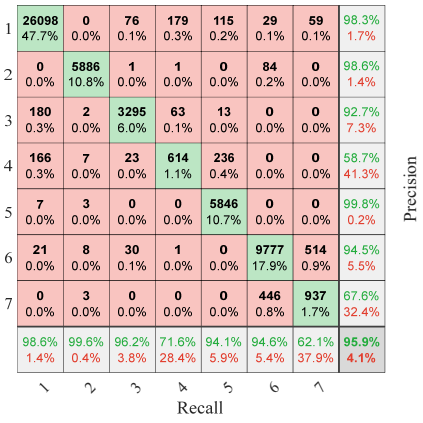


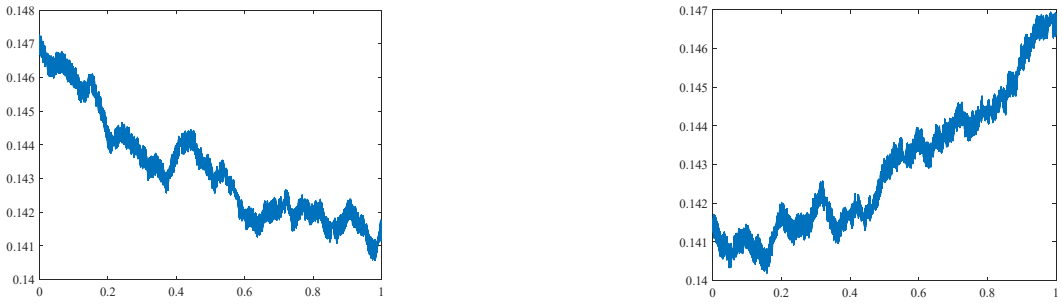
Fig. 8 Confusion matrix based on the optimal four features

4.2 Label modification

The accuracy of the data labels used for training affects the accuracy of the PRNN and, therefore, the accuracy of anomaly detection. In practice, one category of data anomaly may have different forms and characters according to the sensor type and location. Besides, the ambiguity between different anomaly patterns, such as the ‘trend’ and ‘drift’, makes the recognition more difficult as shown in Fig. 9. These uncertainties affect the accuracy of the manual labelling process, especially for large amounts of data. The two pieces of data should belong to the same type, but they are incorrectly labelled as drift and trend, then the wrong trend label is changed as the drift label. Other similar error labels have been also modified.

To eliminate the adverse impact of the label error, the original labels of the two months’ database are carefully re-checked. Only obviously wrong labels in the dataset are modified to avoid misunderstandings. Table 3 compares the quantity and ratio of each pattern in the original label and modified label sets. In the modified label set, the ratio of ‘normal’ and ‘trend’ is reduced by 2.4% and 2.1%, respectively, while the ratio of ‘minor’ and ‘drift’ is increased accordingly. This outcome indicates that mislabelling mainly exists between ‘normal’ and ‘minor’ and between ‘trend’ and ‘drift’.

Table 4 compares the detection accuracy using the original label and modified label sets. The recall and precision of almost all data patterns increase moderately or slightly. In particular, the recall of the ‘drift’ sharply climbs from 61.8% to 90.7%, which also brings a conspicuous improvement of the precision of the ‘drift’, climbing from 73.2% to 82.8%. Finally, the global accuracy for the modified label set has a high level of 97%. Consequently, the modified label set helps achieve a more accurate and dependable anomaly detection. In fact, the ‘trend’ anomaly can be seen as a ‘drift’ anomaly that only drifts in one direction. Hence, these two anomaly patterns should be regarded as the same pattern to obtain better detection accuracy.



(a) Drift (b) Trend

Fig. 9 Raw sample with wrong label

Table 3 Quantity and ratio of each pattern of the original label and the modified label

No.	Anomaly pattern	Quantity		Ratio of total samples (%)	
		Original label	Modified label	Original label	Modified label
1	Normal	99.6	95.1	99.8	99.6
2	Missing	0	82.9	25.4	0
3	Minor	68.4	0	0	0
4	Outlier	89.6	91.3	87.5	43.7
5	Square	98.4	92.1	90.6	77.0
6	Trend	29.6	36.4	20.4	0
7	Drift	89.3	80.9	86.4	77.7

Table 4 Comparison of the detection accuracy using the original label set and the modified label set

No.	Anomaly pattern	Original label set		Modified label set	
		Recall (%)	Precision (%)	Recall (%)	Precision (%)
1	Normal	98.7	98.1	98.7	98.1
2	Missing	99.6	98.0	99.6	98.0
3	Minor	95.4	93.7	95.4	93.7
4	Outlier	72.5	76.4	72.5	76.4
5	Square	98.2	99.5	98.2	99.5
6	Trend	95.2	94.9	95.2	94.9
7	Drift	61.8	73.2	61.8	73.2
Global accuracy		96.4%		97%	

4.3 Extended field application

The Ganjiang Railway cable-stayed bridge (GRCSB) is currently the longest ballast-less track cable-stayed bridge for high-speed railways in the world, with a main span of 300 m and a design speed of 350 km/h. Fig. 10 shows the deployment of the sensor system on the bridge. There are 4 two-channel accelerometers and 1 three-channel accelerometer on the deck of the main span (i.e., 11 channels in total). In this study, 28 days of data of these 11 channels within September 2019 are used to establish a database. Similar to the SYRHB, the data samples in this application are also hourly data; hence, there are 7392 samples in total.

The data anomalies that occurred in the database mainly include four patterns: missing, step, outlier and drift (Fig. 11). Note that there is a new type of data anomaly called ‘step’, wherein the piecewise signal jumps to a big fixed value and then jumps back to the normal range. Additionally, different from the SYRHB, the GRCSB is excited by the high-speed train, leading to a much larger acceleration response in the ‘normal’ pattern than that of the SYRHB. In this situation, it is difficult to distinguish the ‘normal’ sample and the ‘outlier’ sample in GRCSB because both may have obvious extra peaks. Moreover, the ‘normal’ data in the situation of ambient excitation is similar to the ‘drift’.

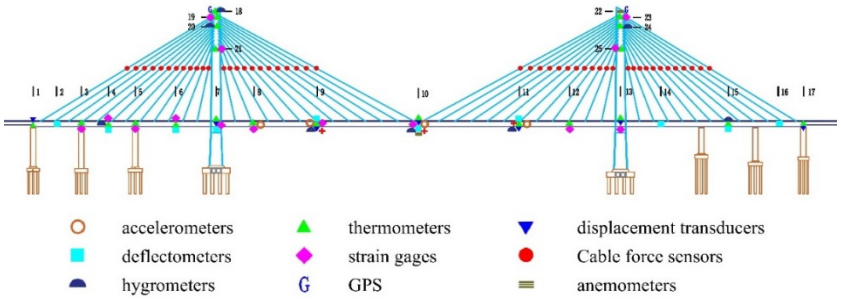


Fig. 10 Deployment of the sensor system

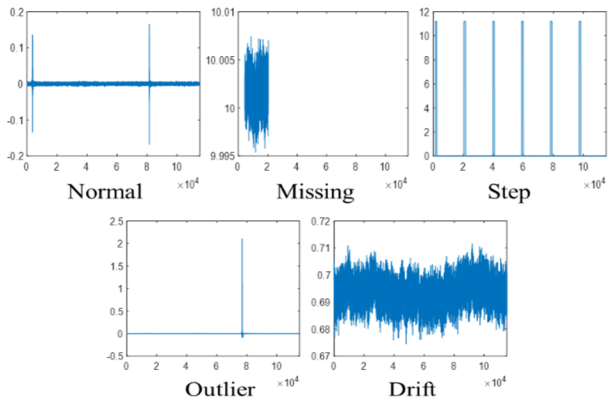


Fig. 11 Examples of different data patterns in the database

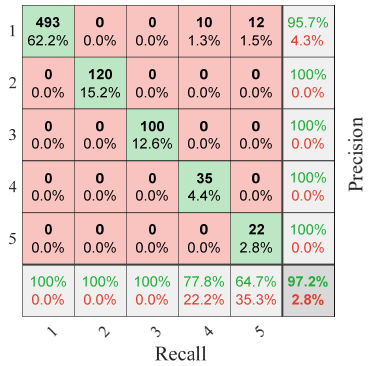


Fig. 12 Confusion matrix of the training set (Data pattern: 1-normal, 2-missing, 3-step, 4-outlier, 5-drift)

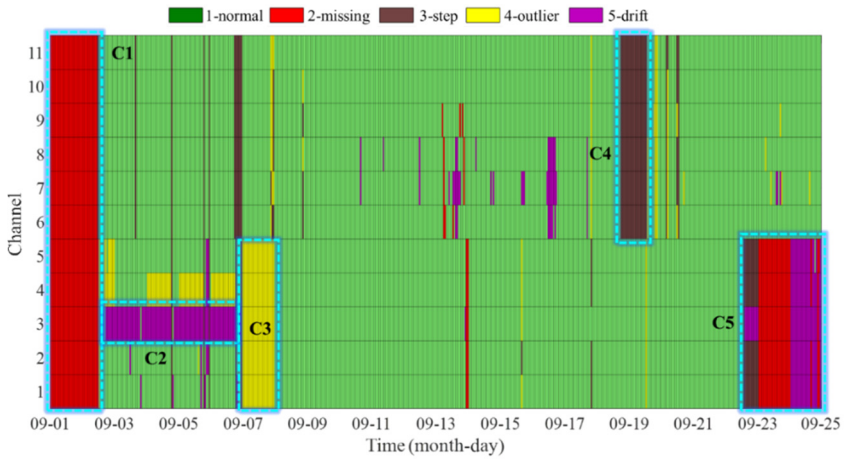


Fig. 13 Anomaly distribution within September 2019 (C means Cluster)

In this application, data samples of 3 days (792 samples) are chosen and manually labelled to build the training set. Fig. 12 shows the confusion matrix of the training set. Clearly, the trained network achieves outstanding classification accuracy, which indicates its ability for data anomaly detection.

Next, the data samples of the other 25 days without label are recognised and classified by the trained network. Fig. 13 illustrates the anomaly distribution within the month. It is obvious that there are five clusters of anomalies, which are marked in chronological order. In cluster 1, the 'missing' occurred in all the channels for the first two days, which may be caused by a power cut. Cluster 2 just consists of the 'drift' anomaly in channel 3. For cluster 3, the 'outlier' anomaly occurred in channels 1–5 within a day. In cluster 4, the 'step' anomaly occurred in channels 6–11 for a day. Channels 1 to 5 constitute cluster 5, which consists of the 'step', 'missing' and 'drift' patterns for nearly three days. Besides these anomaly clusters, all the channels operate normally most of the time.

5. Conclusions

This paper proposes an effective and accurate data anomaly detection method. Nine expertise-based features are extracted from the piecewise acceleration data. The features are then used to train a PRNN for recognising different data patterns. A database containing 60 days of acceleration data of a long-span cable-stayed bridge in China is used to verify the detection accuracy of the proposed method. Results confirm that the seven patterns of data (normal data and six types of anomalies) are recognised with high global accuracy of 96.4% and very low time cost.

The influence of feature selection and label accuracy is investigated as well. It turns out that the detection accuracy based on four appropriate features (linearity, peak intensity, equal value ratio and over-average ratio) is nearly the same as that based on nine features. Additionally, the modified label set further improves the accuracy of the data anomaly detection. Finally, the proposed method is also used to detect the data anomalies on another long-span cable-stayed railway bridge. Favourable results prove that the proposed method is applicable and promising in practical engineering structures.

Acknowledgments

This work was financially supported by the grants from the National Natural Science Foundation of China (NSFC,

contract number: 51922046, 51778258 and 51838006), Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 152621/16E), and the Research Funds from China Railway Eryuan Engineering Group CO.LTD (KYY2019029), and the Research Fund of China Railway Siyuan Survey and Design Group CO.LTD (contract number: 2020K006).

References

Alessandra, D., Salvatore, G., Giuseppe, L., Fabrizio, M. and Marco, O. (2015), "Adaptive distributed outlier detection for WSNs", *IEEE T. Cybernetics*, **45**(5), 888-899. https://doi.org/10.1109/TCYB.2014.2338611

Bao, Y., Tang Z., Li, H. and Zhang, Y. (2019a), "Computer vision and deep learning–based data anomaly detection method for structural health monitoring", *Struct. Health Monitor.*, **18**(2), 401-421. https://doi.org/10.1177/1475921718757405

Bao, Y., Chen Z., Wei, S., Xu, Y., Tang, Z. and Li, H. (2019b), "The state of the art of data science and engineering in structural health monitoring", *Engineering*, **5**(2), 234-242. https://doi.org/10.1016/j.eng.2018.11.027

Bao, Y., Li, J., Nagayama, T., Xu, Y., Spencer Jr, B.F. and Li, H. (2021), "The 1st International Project Competition for Structural Health Monitoring (IPC-SHM, 2020): A summary and benchmark problem. *Struct. Health Monitor.*, **20**(4), 2229-2239. https://doi.org/10.1177/14759217211006485

Banjanovic-Mehmedovic, L., Hajdarevic, A., Kantardzic, M., Mehmedovic, F. and Dzananovic, I. (2017), "Neural network-based data-driven modelling of anomaly detection in thermal power plant", *Automatika*, **58**(1), 69-79. https://doi.org/10.1080/00051144.2017.1343328

Bishop, C. (1995), *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, UK.

Bishop, C. (2006), *Pattern Recognition and Machine Learning*, Springer Press, Cambridge, UK.

Cai, S., Li, L., Li, S., Sun, R. and Yuan, G. (2020), "An efficient approach for outlier detection from uncertain data streams based on maximal frequent patterns", *Expert Syst. Appl.*, **160**, 113646. https://doi.org/10.1016/j.eswa.2020.113646

Calabrese, L., Campanella, G. and Proverbio, E. (2012), "Noise removal by cluster analysis after long time AE corrosion monitoring of steel reinforcement in concrete", *Constr. Build. Mater.*, **34**, 362-371. https://doi.org/10.1016/j.conbuildmat.2012.02.046

Chang, C., Chou, J., Tan, P. a

nd Wang, L. (2017), "A sensor fault detection strategy for structural health monitoring systems", *Smart Struct. Syst.*, *Int. J.*, **20**(1), 43-52. https://doi.org/10.12989/sss.2017.20.1.043

Chen, Z., Li, H. and Bao, Y. (2019), "Analyzing and modeling inter-sensor relationships for strain monitoring data and missing data imputation: a copula and functional data-analytic approach", *Struct. Health Monitor.*, **18**(4), 1168-1188. https://doi.org/10.1177/1475921718788703

Fu, Y., Peng, C., Gomez, F., Narazaki, Y. and Spencer Jr., B. (2019), "Sensor fault management techniques for wireless smart sensor networks in structural health monitoring", *Struct. Control Health.*, **26**(7), e2362. https://doi.org/10.1002/stc.2362

Gul, M. and Catbas, F. (2009), "Statistical pattern recognition for structural health monitoring using time series modeling: Theory and experimental verifications", *Mech. Syst. Signal Process.*, **23**(7), 2192-2204. https://doi.org/10.1016/j.ymssp.2009.02.013

Huang, H., Yi, T. and Li, H. (2020), "Anomaly identification of structural health monitoring data using dynamic independent component analysis", *J. Comput. Civil. Eng.*, **34**(5), 04020025. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000905

Jiang, X., Mahadevan, S. and Adeli, H. (2007), "Bayesian wavelet packet denoising for structural system identification", *Struct. Control Health.*, **14**(2), 333-356. https://doi.org/10.1002/stc.161

Karthick, P., Ghosh, D.M. and Ramakrishnan, S. (2018), "Surface electromyography based muscle fatigue detection using high-resolution time-frequency methods and machine learning algorithms", *Comput. Meth. Prog. Bio.*, **154**, 45-56. https://doi.org/10.1016/j.cmpb.2017.10.024

Katicha, S., Flintsch, G., Bryce, J. and Ferne, B. (2014), "Wavelet denoising of TSD deflection slope measurements for improved pavement structural evaluation", *Comput.-Aided Civil Infrastr. Eng.*, **29**(6), 399-415. https://doi.org/10.1111/mice.12052

Kavzoglu, T. (2009), "Increasing the accuracy of neural network classification using refined training data", *Environ. Modell. Softw.*, **24**(7), 850-858. https://doi.org/10.1016/j.envsoft.2008.11.012

Kullaa, J. (2013), "Detection, identification, and quantification of sensor fault in a sensor network", *Mech. Syst. Signal. Pr.*, **40**(1), 208-221. https://doi.org/10.1016/j.ymssp.2013.05.007

Lejla, B., Amel, H., Mehmed, K., Fahrudin, M. and Izet, D. (2017), "Neural network-based data-driven modelling of anomaly detection in thermal power plant", *Automatika*, **58**(1), 69-79. https://doi.org/10.1080/00051144.2017.1343328

Li, M. and Yuan, B. (2005), "2D-LDA: A statistical linear discriminant analysis for image matrix", *Pattern Recogn. Lett.*, **26**(5), 527-532. https://doi.org/10.1016/j.patrec.2004.09.007

Luo, Y., Ye, Z., Guo, X., Qiang, X. and Chen, X. (2015), "Data missing mechanism and missing data real-time processing methods in the construction monitoring of steel structures", *Adv. Struct. Eng.*, **18**(4), 585-601. https://doi.org/10.1260/1369-4332.18.4.585

Ni, K., Ramanathan, N., Chehade, M., Balzano, L., Nair, S., Zahedi, S., Kohler, E., Pottie, G., Hansen, M. and Srivastava, M. (2009), "Sensor network data fault types", *ACM Transct. Sensor Network*, **5**(3), 1-29. https://doi.org/10.1145/1525856.1525863

Nick, H. and Aziminejad, A. (2021), "Vibration-Based Damage Identification in Steel Girder Bridges Using Artificial Neural Network Under Noisy Conditions", *J. Nondestruct. Eval.*, **40**(1), 15. https://doi.org/10.1007/s10921-020-00744-8

Peng, Y., Qiao, W., Qu, L. and Wang, J. (2018), "Sensor fault detection and isolation for a wireless sensor network-based remote wind turbine condition monitoring system", *IEEE Transact. Ind. Appl.*, **54**(2), 1072-1079. https://doi.org/10.1109/IAS.2017.8101845

Smarsly, K. and Law, K. (2014), "Decentralized fault detection and isolation in wireless structural health monitoring systems using analytical redundancy", *Adv. Eng. Softw.*, **73**, 1-10. https://doi.org/10.1016/j.advengsoft.2014.02.005

Tian, W., Weng, S., Xia, Y., Zhu, H., Gao, F., Sun, Y. and Li, J. (2019), "An iterative reduced-order substructuring approach to the calculation of eigensolutions and eigensensitivities", *Mech. Syst. Signal. Pr.*, **130**, 361-377. https://doi.org/10.1016/j.ymssp.2019.05.006

Titouna, C., Naït-Abdesselam, F. and Khokhar, A. (2019), "DODS: A distributed outlier detection scheme for wireless sensor networks", *Comput. Networks*, **161**, 93-101. https://doi.org/10.1016/j.comnet.2019.06.014

Venugopal, G., Navaneethakrishna, M. and Ramakrishnan, S. (2014), "Extraction and analysis of multiple time window features associated with muscle fatigue conditions using sEMG signals", *Expert Syst. Appl.*, **41**(6), 2652-2659. https://doi.org/10.1016/j.eswa.2013.11.009

Wang, Y., Yang, A., Chen, X., Wang, P., Wang, Y. and Yang, H. (2017), "A deep learning approach for blind drift calibration of

sensor networks", *IEEE Sens. J.*, **17**(13), 4158-4171. https://doi.org/10.1109/JSEN.2017.2703885

Widodo, A. and Yang, B. (2007), "Support vector machine in machine condition monitoring and fault diagnosis", *Mech. Syst. Signal. Pr.*, **21**(6), 2560-2574. https://doi.org/10.1016/j.ymssp.2006.12.007

Yang, Y. and Nagarajaiah, S. (2014), "Blind denoising of structural vibration responses with outliers via principal component pursuit", *Struct. Control Health.*, **21**(6), 962-978. https://doi.org/10.1002/stc.1624

Yang, Y. and Nagarajaiah, S. (2016), "Harnessing data structure for recovery of randomly missing structural vibration responses time history: Sparse representation versus low-rank structure", *Mech. Syst. Signal. Pr.*, **74**, 165-182. https://doi.org/10.1016/j.ymssp.2015.11.009

Yuen, K. and Mu, H. (2012), "A novel probabilistic method for robust parametric identification and outlier detection", *Probabilist. Eng. Mech.*, **30**, 48-59. https://doi.org/10.1016/j.probengmech.2012.06.002

Yuen, K. and Ortiz, G. (2017), "Outlier detection and robust regression for correlated data", *Comput. Method Appl. Mech. Eng.*, **313**, 632-646. https://doi.org/10.1016/j.cma.2016.10.004

Zhang, Z. and Luo, Y. (2017), "Restoring method for missing data of spatial structural stress monitoring based on correlation", *Mech. Syst. Signal. Pr.*, **91**, 266-277. https://doi.org/10.1016/j.ymssp.2017.01.018

Zhang, Z., Mehmood, A., Shu, L., Huo, Z., Zhang, Y. and Mukherjee, M. (2018), "A survey on fault diagnosis in wireless sensor networks", *IEEE Access*, **6**, 11349-11364. https://doi.org/10.1109/ACCESS.2018.2794519

Zhu, H., Gao, K., Xia, Y., Gao, F., Weng, S., Sun, Y. and Hu, Q. (2020), "Multi-rate data fusion for dynamic displacement measurement of beam-like supertall structures using acceleration and strain sensors", *Struct. Health Monitor.*, **19**(2), 520-536. https://doi.org/10.1177/1475921719857043

Zvokelj, M., Zupan, S. and Prebil, I. (2011), "Non-linear multivariate and multiscale monitoring and signal denoising strategy using Kernel Principal Component Analysis combined with Ensemble Empirical Mode Decomposition method", *Mech. Syst. Signal. Pr.*, **25**, 2631-2653. https://doi.org/10.1016/j.ymssp.2011.03.002