Smart Structures and Systems, *Vol. 29, No. 1 (2022) 93-103*
https://doi.org/10.12989/sss.2022.29.1.093

Data anomaly detection for structural health monitoring of bridges using shapelet transform

Monica Arul* and Ahsan Kareem [a]

Nathaz Modeling Laboratory, Department of Civil & Environmental Engineering & Earth Sciences, University of Notre Dame, Notre Dame, Indiana 46556, USA

(*Received April 13, 2021, Revised July 20, 2021, Accepted July 29, 2021*)

Abstract. With the wider availability of sensor technology through easily affordable sensor devices, several Structural Health Monitoring (SHM) systems are deployed to monitor vital civil infrastructure. The continuous monitoring provides valuable information about the health of the structure that can help provide a decision support system for retrofits and other structural modifications. However, when the sensors are exposed to harsh environmental conditions, the data measured by the SHM systems tend to be affected by multiple anomalies caused by faulty or broken sensors. Given a deluge of high-dimensional data collected continuously over time, research into using machine learning methods to detect anomalies are a topic of great interest to the SHM community. This paper contributes to this effort by proposing a relatively new time series representation named "Shapelet Transform" in combination with a Random Forest classifier to autonomously identify anomalies in SHM data. The shapelet transform is a unique time series representation based solely on the shape of the time series data. Considering the individual characteristics unique to every anomaly, the application of this transform yields a new shape-based feature representation that can be combined with any standard machine learning algorithm to detect anomalous data with no manual intervention. For the present study, the anomaly detection framework consists of three steps: identifying unique shapes from anomalous data, using these shapes to transform the SHM data into a local-shape space and training machine learning algorithms on this transformed data to identify anomalies. The efficacy of this method is demonstrated by the identification of anomalies in acceleration data from an SHM system installed on a long-span bridge in China. The results show that multiple data anomalies in SHM data can be automatically detected with high accuracy using the proposed method.

Keywords: anomaly detection; long-span bridge; machine learning; shapelet transform; structural health monitoring; time series shapelets

1. Introduction

As the infrastructure demands continue to increase, research into structural health monitoring (SHM) has grown in importance throughout the world. The widespread application of sophisticated SHM systems in civil infrastructure produces a large volume of data. However, the harsh environmental conditions of civil structures cause the data measured by SHM systems to be affected by multiple anomalies caused by faulty or broken sensors. These anomalies pose a significant barrier to assessing the actual structural performance and severely affects the automatic warning system for damage or accidents. Identifying and removing data anomalies due to environmental variations is thus an important preprocessing step in a successful warning system. Several model-based methods have been developed in the past few decades for data anomaly detection in SHM data (Abdelghani and Friswell 2004, Thiyagarajan *et al*. 2017, Wan and Ni 2018, Wang *et al*. 2019). In these methods, several statistical models are initially constructed to predict the measurements. Using appropriate thresholds, measurements that show significant differences between predicted and measured values are identified and treated as anomalies.

Faced with a massive amount of data due to the continuous monitoring of structures, researchers have recently resorted to advanced approaches such as data mining and machine learning techniques for anomaly detection. Bao *et al*. (2019) proposed a computer vision and deep learning-based data anomaly detection method in which the raw time series measurements are first transformed into image vectors which are then fed into the Deep Neural Networks (DNN) to train them to identify various anomalies from SHM data. Fu *et al*. (2019) used a similarity test based on power spectral density to detect anomalies and then trained an artificial neural network to identify the different sensor anomalies. Tang *et al*. (2019) proposed using a Convolutional Neural Network (CNN) for anomaly detection that learned from multiple graphical information. The visualizations of the time series measurements in the time and frequency domain are fed to the neural networks, which then learned the characteristics of each of the anomalies during training. The trained network is then used to identify and classify various anomalies. Mao *et al*. (2020) used Generative Adversarial

*Corresponding author, Ph.D.,
E-mail: maruljay@nd.edu
[a] Ph.D., Professor

ISSN: 1738-1584 (Print), 1738-1991 (Online)

Networks (GAN) combined with autoencoders to identify anomalies. The raw time series from the SHM system is first transformed into Gramian Angular Field (GAF) images which are then used to train the GAN and autoencoders to identify anomalies.

This paper contributes to this effort by proposing the use of a relatively new time series representation named "Shapelet Transform" in combination with Random Forest classifiers for anomaly detection in SHM data. The shapelet transform is a unique time series representation technique that is solely based on the shape of the time series. The raw measurements of every sensor anomaly have a unique time series shape. The shapelet transform utilizes this feature to capture these distinct shapes easily. Random Forest classifier then uses these shapes to identify and classify the different anomalous data patterns from an extensive SHM system database.

Analysis methods based on the global attributes of time series are unintuitive and reduce comprehensibility. By examining local-shape-based features, it is ensured that these small discriminatory shapes are not averaged out but rather used to distinguish the time series, exactly as they are under intuitive visual inspection. The primary advantage of shapelets over the above mentioned competing methods is the interpretability and insight it offers. Shape-based approaches are intuitive, visually meaningful and offer immediate insight into the problem domain that goes beyond their use in accurate detection. Thus the method is a "white-box" machine learning model involving understandable and easily visualizable features that makes the entire process transparent and completely open to inspection. This, in turn, increases the interpretability of the model compared to the other state-of-the-art black-box machine learning models, like neural networks, and helps domain practitioners gain better insights from their data. In terms of applicability, shapelets have been utilized in a wide variety of domains, including motion-capture (Ye and Keogh 2009, 2011, Lines *et al*. 2012, Hartmann and Link 2010), spectrographs (Ye and Keogh 2009, 2011), tornado prediction (McGovern *et al*. 2011), detection of natural hazards (Arul and Kareem 2020), medical and health informatics (Ghalwash *et al*. 2013, Xing *et al*. 2011, 2012) among others. In the present study, the efficacy of this method is demonstrated by the identification of anomalies in SHM data obtained from a long-span bridge in China.

The article is organized as follows. A general overview of the shapelet transform is provided in section 2. A brief description of the SHM data used for this study is given in section 3. The methodology for the proposed anomaly detection process is elaborated in section 4. In this section, the different stages involved in shapelet transform are explained in great detail, along with illustrative examples. The section also explains the step-by-step procedure for detecting anomalous patterns in SHM data obtained from the long-span bridge. Finally, a comprehensive summary of the anomalies detected using shapelet transform is provided in sections 5 and 6.

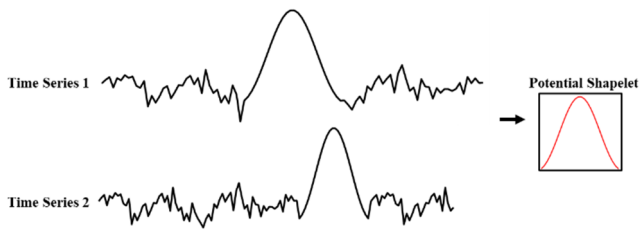

Fig. 1 Time series shapelets

2. Overview of shapelet transform

Consider time series 1 and 2 generated as a result of an event, as shown in Fig. 1. Both the time series have long stretches of aperiodic waveforms. However, a local shape appears for a short duration in the time series that differs substantially from the rest of the time series. These localized shapes are called shapelets. These discriminatory shapes, which are phase independent, serve as a powerful feature for identifying anomalous patterns or classifying events from a large database containing continuous records. Time series shapelets stem from the desire to reify human's innate capacity to visualize the shape of data and identify almost instantly similarities and differences between patterns. Shapelets help computers perform this complex task by identifying the local or global similarity of shape that can offer an intuitively comprehensible way of understanding continuous time series. The shapelets, once discovered, can then be used to transform data into a local-shape space where each feature is the distance between a shapelet and a time series (Lines *et al*. 2012). The result of this transform is that the new representation can be applied to any standard machine learning algorithm to identify anomalous patterns. Shapelet transform has five major stages: generation of shapelet candidates, distance calculation between a shapelet and a time series, assessment of the quality of shapelets, discovery of shapelets, and data transformation. Each of these stages will be elaborated on in detail in the following sections.

3. Data description

In this paper, the SHM dataset from a long-span cable-stayed bridge in China is used. The data is provided by the Asis Pacific Network of Centers for Research in Smart Structures Technology (ANCRiSST), Harbin Institute of Technology (China), and the University of Illinois at Urbana-Champaign (USA) as a part of the 1st International Project Competition for Structural Health Monitoring (IPC-SHM 2020) (Bao *et al*. 2021). The main span of the bridge is 1088 m, two side spans are 300m each, and it consists of two towers that are 306 m high. The structural health monitoring system of the bridge consists of 38 sensors, whose locations are illustrated in Fig. 2. The sensors include accelerometers, anemometers, strain gauges, global positioning systems (GPS), and thermometers. For the present case, one month (2012-01-01 – 2012-01-31) of acceleration data for all 38 sensors from the SHM system is considered for anomaly detection. The sampling frequency of the accelerometers is 20 Hz. The continuous raw measurements are broken down into 1-hour segments and

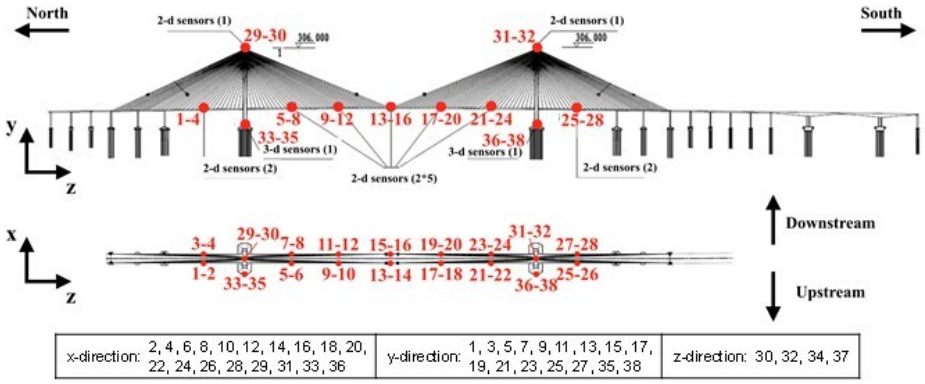

Fig. 2 The bridge and the placement of accelerometers on the deck and towers

Table 1 Description of anomalous data patterns

No	Anomalous patterns	Description	Quantity in 1-month dataset
1	Normal	The time response is normal oscillation curve; frequency response is peak-like (may differ between bridges)	13575 (48%)
2	Missing	Most or all of the time response is missing, which makes the time and frequency response zero	2942 (10.4%)
3	Minor	Relative to normal sensor data, the amplitude is very small in the time domain	1775 (6.3%)
4	Outlier	One or more outliers appear in the time response	527 (1.9%)
5	Square	The time response is like a square wave	2996 (10.6%)
6	Trend	The data has an obvious trend in the time domain and has an obvious peak value in the frequency domain	5778 (20.4%)
7	Drift	The vibration response is non-stationary, with random drift	679 (2.4%)

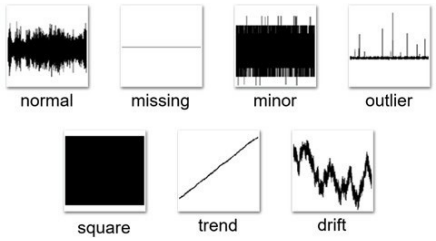

Fig. 3 Examples for each of the anomaly patterns in the SHM data

744 time series measurements for each sensor are obtained for one month resulting in 744 × 38 datasets. The characteristics of the normal data and the six classes of anomalies found in the dataset is described in Table 1. Examples for each data pattern is shown in Fig. 3. The normal time series measurement is labeled as 1, and the other six data anomaly patterns are labeled from 2 – 7. From Table 1, it can be seen that nearly 52% of the data are anomalous. The "trend" is the major anomalous pattern constituting 20% of the dataset, followed by "missing" and "square", each accounting for around 10%. On the other hand, the "outlier" pattern accounts for only 1.9% of the dataset, followed by "drift", which constitutes 2.4% of the data.

4. Methodology for anomaly detection

The methodology for anomaly detection in SHM data involves three major steps, as shown in Fig. 4. In the first step, the raw time series measurements are broken down into 1-hour segments, as mentioned before. The peak envelopes of the time series are extracted to visualize the shape of the time series easily. A time series learning set is constructed along with class labels using these envelopes as shapes. Once the learning set is ready, the process of transforming it into a local-shape space begins. Shapelet transform has five major stages: generation of shapelet candidates, distance calculation between a shapelet and a time series, assessment of the quality of shapelets, discovery of shapelets, and data transformation. In the second step, the original time series-based learning set is transformed into a local-shape space where each element is the distance between a shapelet and a time series. In this transformed learning set, the features are the discovered shapelets, and the instances are the individual time series envelopes. This is fed to a Random Forest classifier for training. Once the training is complete, the trained classifier is used to classify normal and anomalous data from the new incoming time series from the SHM system in the third step.

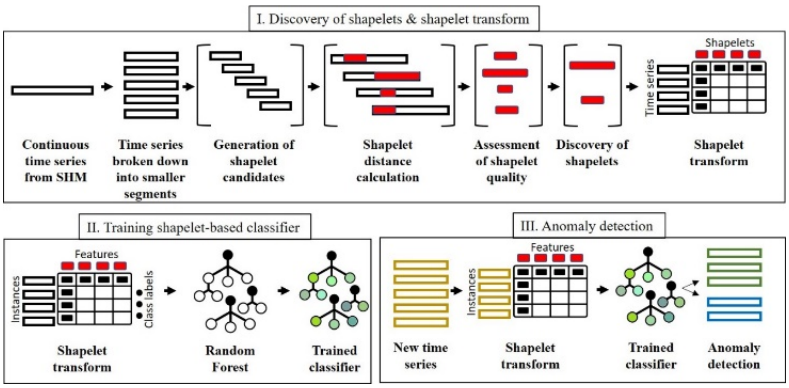

Fig. 4 Methodology for anomaly detection in SHM data

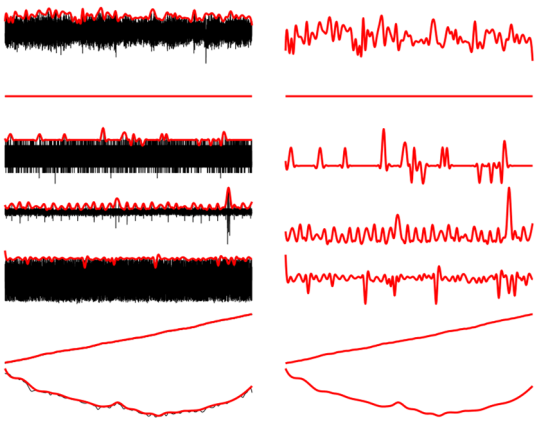

Fig. 5 Extraction of peak envelopes from anomalous patterns

4.1 Step 1: Discovery of shapelets and shapelet transform

4.1.1 Preprocessing of raw data

Based on the visual inspection of Fig. 3, it is easy to differentiate between the different anomalies. However, it is quite challenging to use the raw time histories for shapelet detection due to the long periods of periodic waveforms present in the vibrations. This can be overcome by extracting the acceleration time history envelopes, which gives an overall shape to the vibration time series. The envelopes can then be easily used as input for the discovery of shapelets. Fig. 5 shows the extraction of peak envelopes of the bridge acceleration time history calculated using a moving window. A peak envelope is used here instead of a root-mean-square (RMS) envelope, as a peak provides better differentiation between anomalous patterns when noisy or spurious signals are present. By looking at the peak envelopes of anomalies, the classification of anomalous data has become a much easier task now. Considering the computational demand of the algorithm, the envelopes obtained from the raw time series is down sampled to 1 Hz to improve the efficiency of the algorithm. Down sampling, the data did not affect the shapes of the envelopes, and hence the reliability of the method remains unchanged.

4.1.2 Generation of shapelet candidates

Consider a time series dataset TS. Let C be the set of corresponding class labels for each time series. A time series learning set $\varphi\{TS, C\}$ is first created by a vector of instance input-output pairs $\varphi_i\{TS_i, C_i\}$. Each subsequence in each time series in φ is considered as a potential shapelet candidate. So, there are $(m-l)+1$ discrete subsequences of length l between a subsequence X of length l of a time series TS of length m. If W_1 is the set of all candidate shapelets of length l in a time series TS_1, then

$$W_1 = \{w\,min + 1_{max_{min}}\} \tag{1}$$

where $min \geq 3$ as it is the minimum meaningful length for a time series and $max \leq m$. It should be noted that the shapelet algorithm independently normalizes all candidate shapelets so as to be invariant to scale and offset. For the present case, the time series learning set consists of 700 labeled sets of time series envelopes extracted from the raw measurements, as shown in Fig. 6. It should be noted that the learning set contains equal samples of patterns obtained from the SHM data, i.e., the set contains 100 samples of the normal pattern, 100 samples of missing pattern, 100 samples of minor pattern and so on. This is done to achieve a balanced training set to avoid classifier bias during the detection of anomalies. The reason for choosing 100 as the sample number is as follows. The data from 2012-01-01 to 2012-01-16 is used for training the algorithm, and the data from the other fifteen days (2012-01-17 to 2012-01-31) is used for testing. In the training dataset, the "outlier" pattern had the lowest quantity of about 100 datasets. Hence this number has been established as a baseline for choosing the number of samples for each pattern.

Thus, the time-series learning set, φ has a total of 700 datasets, as shown in Fig. 6. Each time series in the training set has 3600 data points as the sampling frequency is 1 Hz. Let us take the first time series in the training set for illustration. As per Eq. (1)

$$W_1 = \{w_3, w_4, \ldots, w_{3599}, w_{3600}\} \tag{2}$$

where w_3 (first 3 data points) is the shortest shapelet length and w_{3600} (entire time series) is the longest shapelet length. Thus, the set W_1 contains 3598 different lengths of shapelets obtained from the first time series. Similarly,

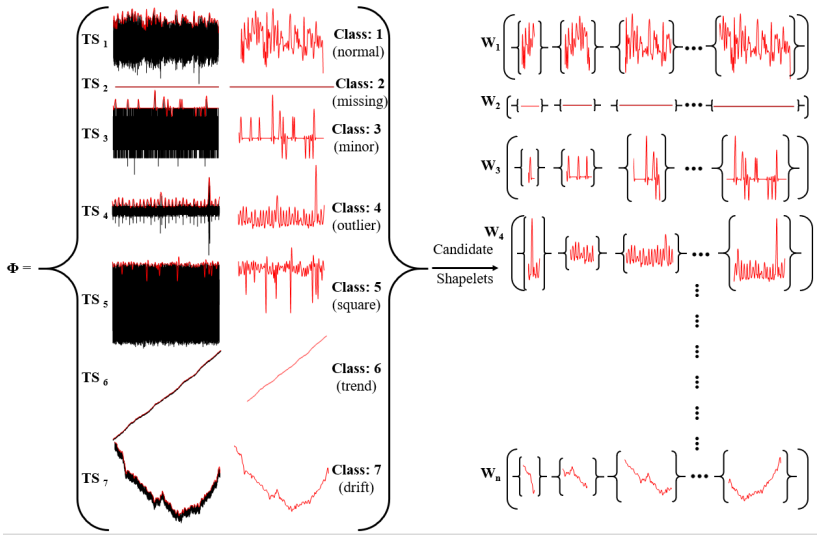

Fig. 6 Illustration of generation of shapelet candidates for each time series in the time series learning set

shapelet candidates are generated from all the time series in the learning set.

4.1.3 Shapelet distance calculation

Euclidean distance is used as a similarity measure in shapelets, and the squared Euclidean distance between a subsequence X of length l and another subsequence Y of the same length is defined as

$$d(X,Y) = \sum_{i=1}^{l} (x_i - y_i)^2 \tag{3}$$

The distance between a potential shapelet candidate and all normalized series in TS is computed to create a list of n distances called an orderline DS. An orderline consists of distance values and the class label corresponding to the time series for which the distance value is calculated. The orderline is then sorted in increasing order of the distance value. In the present study, each time series leads to the generation of 3598 shapelet candidates. Each of these 3598 shapelets is then compared with other time series using Euclidean distance. For illustration purposes, consider a shapelet candidate S_1 as shown in Fig. 7. The shapelet candidate moves over every time series, and the minimum distance between the candidate and all the normalized lengths of subsequences in the time series set is calculated. It should be noted that Fig. 7 is an exaggerated illustration to facilitate easy understanding and not an accurate depiction of the normalization process. If the shapelet candidate is generated from a pattern that is different from the time series being compared to, then it will lead to a large normalized Euclidean distance. However, if the shapelet is similar to the one being compared to, then it will have a minimum normalized Euclidean distance, as seen in $d_{S_1,7}$ in Fig. 7. Thus, the distance between a shapelet candidate S_1 and all the normalized lengths of subsequences in time series set TS is given by

$$DS = \langle d_{S_1,1}, d_{S_1,2}, \ldots, d_{S_1,n} \rangle \tag{4}$$

It is a time-consuming task to calculate DS, and hence several speed-up techniques have been proposed in the literature to handle the large volume of calculations (Ye and Keogh 2009, 2011, Mueen *et al*. 2011, Rakthanmanon and Keogh 2013, Hills *et al*. 2014).

4.1.4 Assessment of shapelet quality

Information Gain (IG) (Shannon and Weaver 1949) is used as the standard approach to calculating the quality of

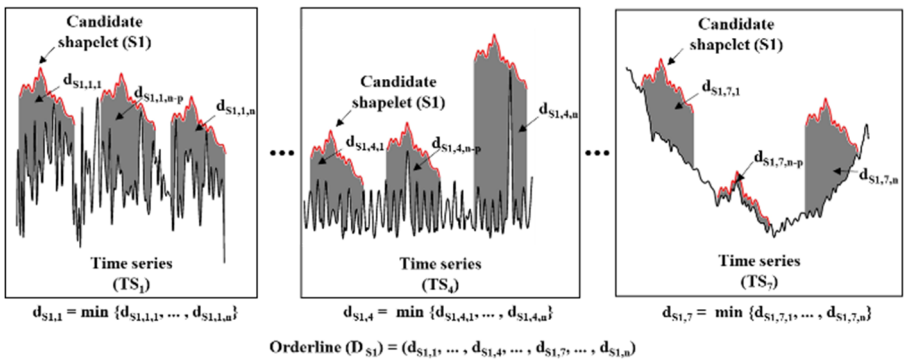

Fig. 7 Illustration of generation of shapelet candidates for each time series in the time series learning set

a shapelet (Ye and Keogh 2009, 2011, Mueen *et al*. 2011). If a time series dataset T can be split into two classes, 1 and 2, then the entropy of T is

$$H(T) = -p(1)\,log\big(p(1)\big) - p(2)\,log\big(p(2)\big) \qquad (5)$$

where $p(1)$ and $p(2)$ are the proportion of objects in classes 1 and 2, respectively. Thus, every splitting strategy partitions the dataset T into two sub-datasets T_I and T_{II}. The Information Gain of this split is the difference between the entropy of the entire dataset and the sum of the weighted average of entropies for each split. In the present case, the splitting rule is based on the distance from the shapelet candidate S to every series in the dataset. The best possible shapelet will generate small distance values compared to a time series of its own class and large distance values for time series from the other class. Thus, the best arrangement for the orderline is to have all the distance values corresponding to the class of the shapelet in T_I and the other in T_{II}. Thus, the information gain for each split is calculated as

$$IG = H(T) - \left(\frac{|T_I|}{|T|}H(T_I) + \frac{|T_{II}|}{|T|}H(T_{II})\right) \qquad (6)$$

where $0 \le IG \le 1$.

The same procedure is extended to the 7-class problem as in the present study. For illustration purposes, consider the shapelet candidate S_1, mentioned in the previous section. S_1 is compared with 699 other time series in the learning set, and thus 699 distances are obtained. These distances values are ordered in increasing value in the orderline, and the information gain is calculated as shown in Fig. 8. The same procedure is extended to all the shapelets candidates that are generated. Whichever length of the shapelet surpasses the provided information gain threshold (0.05 in the present case) is retained, and the other shapelet lengths are discarded. This makes sure that the selected shapelets are meaningful and have discriminatory power. Predetermining the optimal length of the shapelet is impossible and unnecessary as it hinders the detection accuracy of the algorithm. It is also challenging to interpret the variety of shapelet lengths obtained from the algorithm.

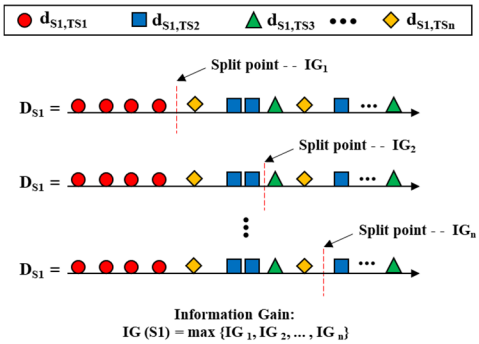

Fig. 8 One-dimensional representation of the arrangement of time series objects by the distance to the candidate shapelet. Information Gain is calculated for each possible split point

These lengths have been chosen from several thousands of shapelet lengths compared with several other time series. However, there is a provision in the shapelet algorithm to set the maximum and minimum shapelet length to achieve speedup. This provision should be used with care and should only be utilized when only a certain length of shapelets are of interest.

4.1.5 Discovery of shapelets and shapelet transform

An algorithm combining all of the above-mentioned components of shapelet discovery was developed by Bostrom and Bagnall (2017a), Lines *et al*. (2012), Hills *et al*. (2014) and is available at www.sktime.org. The same algorithm has been adopted and modified to suit the datasets under consideration for the present study. The input to the algorithm is the time series leaning set φ. As mentioned in the previous sections, the default minimum length of the shapelets is set to 3, and the maximum length is equal to the length of each time series. The number of shapelets to store (r) is set to a default of 10 times the number of time series in the training set. Moreover, based on the number of classes $(numC)$ in the training set, a limit of $r/numC$ shapelets for each class is set as the maximum number of shapelets to store per class. The minimum information gain threshold is set to a default value of 0.05. This makes sure that poor quality shapelets below this threshold are removed during the shapelet finding process. Using the provided parameters, the algorithm then makes a single pass through the time series data in φ taking each subsequence of every time series as a potential shapelet candidate. The generated shapelet candidates are also normalized to make them independent of scale and offset. The distance between each shapelet candidate and time series in the training dataset is calculated, and the order list D_S is formed to assess the quality of shapelets using Information Gain. Once all the shapelets in a time series have been assessed, the poor quality shapelets are removed, and the rest is added to the shapelet set. After all the time series in the training set have been evaluated this way, the algorithm returns the discovered shapelets. For the present study, the shapelet algorithm is implemented in Python as a single-core serial job on an Intel Xeon Processor E5-2620 (2.6-GHz CPU) for 1 hour, and the algorithm discovered a total of 68 shapelets. Various shapes were discovered for each of the seven anomalous patterns, as shown in Figs. 9-11. Shapelets corresponding to the "missing" and "normal" patterns have the highest information gain as these shapes separate the classes easily.

Once the shapelets are discovered, the next step is to transform the learning set φ into a local-shape space where each feature is the distance between a shapelet and a time series. So, given a set of a time series dataset TS containing n time series and a set of r discovered shapelets, the shapelet transform algorithm calculates the minimum distance between each discovered shapelet and each time series in the dataset. This transformation creates a matrix G that contains n rows and r columns matrix as illustrated in Fig. 12, where each element is the minimum Euclidean distance between each shapelet and time series,

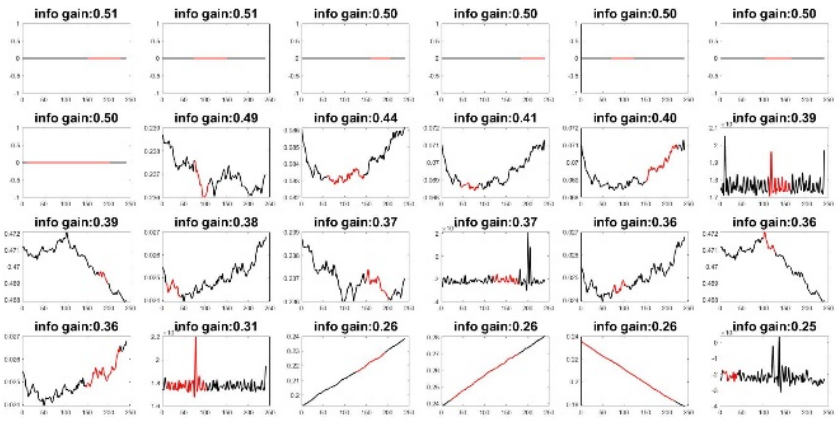

Fig. 9 Shapelets (1-24) discovered for anomaly detection

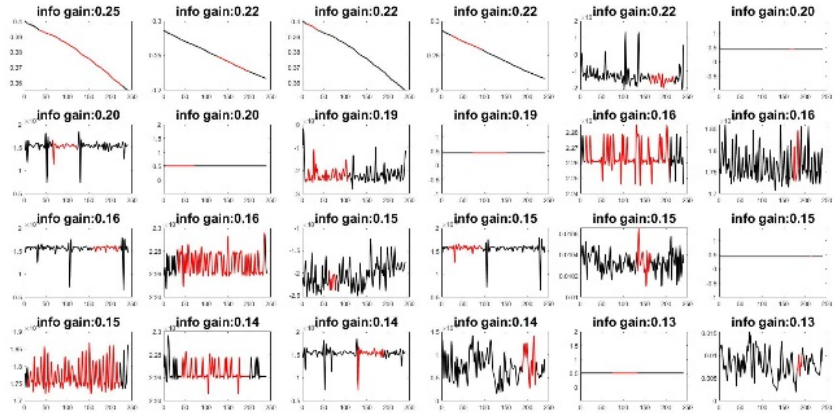

Fig. 10 Shapelets (25-48) discovered for anomaly detection

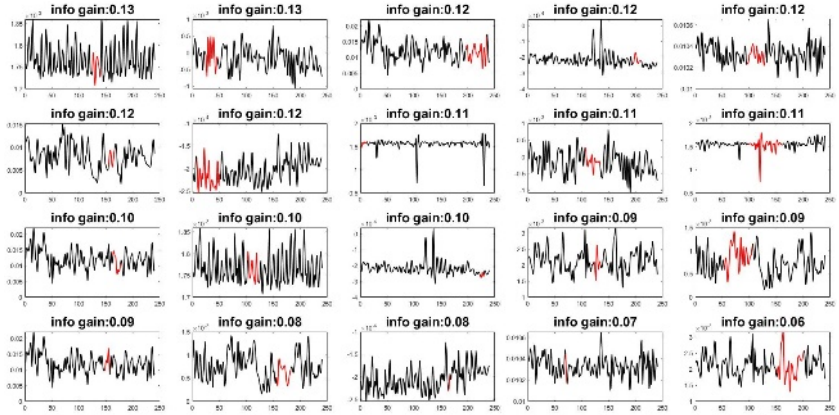

Fig. 11 Shapelets (49-68) discovered for anomaly detection

with the class values appended to the end of each row. It should be noted that S_1 is the first shapelet in the set of all shapelets and does not denote that it is obtained from TS_1. The matrix G now serves as the standard instance-attribute dataset that is used in machine learning tasks that can be used with any supervised or unsupervised algorithm. In the present study, shapelet transform constructs a 700 × 68 matrix where each element corresponds to the minimum Euclidean distance between each shapelet and the time series.

4.2 Step 2: Training of shapelet-based classifier

The shapelet based classifier originally developed by Ye and Keogh (2009) embeds shapelet finding in a decision tree classifier where shapelets are found at every node. Many researchers ever since have demonstrated that higher accuracy can be achieved by using shapelets with more complex classifiers or ensemble of classifiers than with decision trees, where overfitting is a major issue (Lines *et al*. 2012, Hills *et al*. 2014, Bagnall *et al*. 2017, Bostrom and

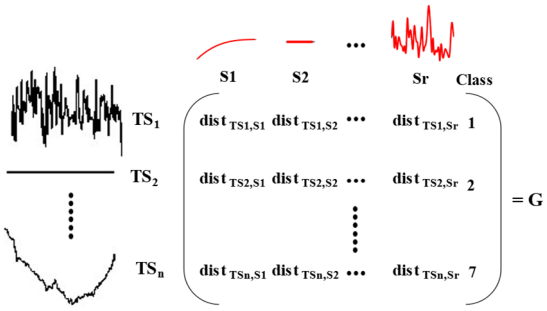

Fig. 12 Shapelet Transform containing a matrix of Euclidean distance between the discovered shapelets and the other time series in the learning set

Bagnall 2017b). For the present study, Random Forest (Breiman 2001) is used as a classifier for time series classification. The Random Forest algorithm seeks to solve the issues with decision trees by classifying examples using many decision trees and predicting the class of a sample based on the mean probability estimate across all the trees. Thus, a Random Forest classifier with 500 trees is used for training on the shapelet-transformed dataset. The training time is 1 hour. Hills *et al*. (2014), Bagnall *et al*. (2017) compared the performance of shapelet transform using several standard classifiers and ensemble classifiers on various datasets from the UCR time-series archive (Chen *et al*. 2015). According to their study, a shapelet-based random forest classifier with 500 trees is found to be optimal. Hence the same has been adopted in the present study. It is also found that increasing the number of trees beyond 500 did not significantly increase accuracy.

4.3 Step 3: Anomaly detection

As mentioned in section 4.1.1, the data from 2012-01-01 to 2012-01-16 is used for training the algorithm, and the data from the other fifteen days (2012-01-17 to 2012-01-31) is used for testing. The raw measurements are broken down into 1-hour segments resulting in a total of 13679 datasets. The peak envelopes of the time series are extracted and down sampled to 1 Hz. The shapelet transform algorithm is used on the test set to transform the data onto shape-space where each element is the minimum Euclidean distance between the discovered shapelets and the time series in the test set. Thus a 13679×68 matrix is obtained where 13679 are the time series instances, and 68 are the shapelet-based features. The trained Random Forest classifier is then tested on this transformed test set.

5. Results and discussion

The detection of anomalies using the shapelet-based classifier was implemented as a single-core serial job on an Intel Xeon Processor E5-2620 (2.6-GHz CPU), and the algorithm took 2.5 hours to output the results. The run time will drastically increase if event detections are made for months or years of continuous data. The run time can be reduced in two ways. Incorporating parallelism in the algorithm so that distance calculations can be executed in parallel on multicore machines. Another way is to redesign the algorithm to make it suitable for parallel Graphics Process Units (GPUs). Chang *et al*. (2012) improved the shapelet algorithm for GPU implementation and achieved speedups nearly two orders of magnitude faster than CPU implementation. This means that a 1.7-hour CPU implementation of shapelets will only take 2 minutes using GPUs. Such an algorithm redesign will be explored in future studies to render this method efficient for processing large volumes of data.

The detection results are shown in Table 2. The definitions in the following section will help understand the performance metrics of the classifier better.

5.1 Terminologies used in assessing the performance of the classifier

- True Negative (TN) - The actual value is False, and the classifier also predicted False.
- False Positive (FP) - The actual value is False, and the classifier predicted True.
- False Negative (FN) - The actual value is True, and the classifier predicted False.
- True Positive (FP) - The actual value is True, and the classifier also predicted True.
- Accuracy - Accuracy is the sum of true positives and true negatives divided by the total number of instances. From the confusion matrix, accuracy is the sum of the diagonal elements divided by the total number of predictions made. Accuracy is calculated as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

- Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It is a measure of the correctness of a positive prediction. Precision will have a value of 1 for an ideal classifier with no false positives. Precision is given by:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

- Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to all actual class observations. It is the measure of how many true positives get predicted from all the positives in the dataset. Recall will have a value of 1 for an ideal classifier with no false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

- F1 Score - F1 score is the harmonic mean of precision and recall and is a combined measure of the two. F1 score is high when both precision and recall are high.

$$F1 = 2 \times \left(\frac{precision \times recall}{precision + recall}\right) \tag{10}$$

Table 2 Performance metrics of the shapelet-based Random Forest classifier

Class	Anomalous patterns	Accuracy (%)	Precision	Recall	F1 score
1	Normal	95.26	0.92	0.98	0.95
2	Missing	100	1.0	1.0	1.0
3	Minor	98.82	0.87	0.96	0.91
4	Outlier	96.94	0.87	0.54	0.67
5	Square	99.06	0.99	0.92	0.96
6	Trend	98.06	0.90	1.0	0.94
7	Drift	97.80	1.0	0.50	0.66
Overall accuracy		**92.97%**			

5.2 Discussion of results

The performance metrics are also shown visually in the form of a confusion matrix in Fig. 13. In the confusion matrix, the diagonal elements are the correctly classified instances, and their corresponding precision is provided underneath within brackets. An overall accuracy of 93% is obtained using the shapelet-based classifier. The individual accuracy of all the classes is above 95%, with classes 2 and 5 having an accuracy of about 100%. In terms of precision and recall, classes "normal", "square", and "trend" have a high value of over 90%, with class "missing" having a maximum of 100%. For classes "outlier" and "drift", the recall value is very low even though the precision is high. This is because a small number of instances in class "normal" were predicted as belonging to class "outlier" due to the presence of significant outliers. Also, some instances in class "normal" were predicted as "class square" as the time series has a very close resemblance to a square shape.

Similarly, several instances in class "trend" were predicted as belonging to class "drift" as the time series closely resembled class "drift". Each of these cases is examined in detail, and remedial measures are proposed in the following sections. Meanwhile, in the present study, since the learning set is constructed as a well-balanced dataset, of all the performance metrics, accuracy measures can be used as a useful indicator to comment on the performance of the classifier. Based on high individual and overall accuracy, the proposed shapelet-based classifier has an excellent ability to identify anomalies in SHM data.

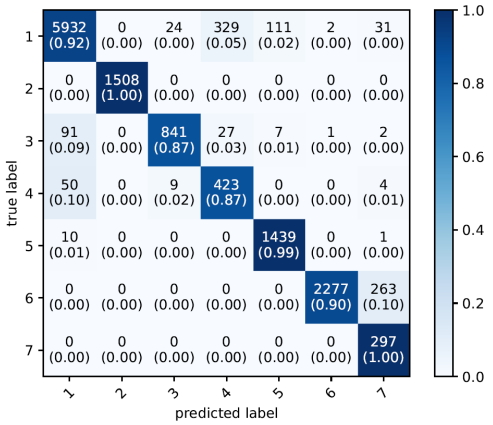

Fig. 13 Confusion matrix for detected anomalies

6. Remedial measures for increasing the performance of the classifier

4.1 Removal of outliers during preprocessing

From the confusion matrix, it can be seen that 329 instances in class "normal" are predicted as class "outlier". On closer inspection, it is found that the outliers not only affect the instances in class "outlier" but also the instances in class "normal" as 50 instances in class "outlier" is predicted as "normal". One such example of a class "normal" instance with outliers is shown in Fig. 12 in the upper left corner. This confuses the machine learning algorithm as it has learned that an "outlier" is the only class with large outliers. Hence it is wise to remove all the predominant outliers in the preprocessing step so that class "outlier" becomes a pure class containing datasets with significant outliers. This can be easily done using the 'rmoutliers' command in MATLAB that detects and removes predominant outliers according to a user-specified window. It should be noted that this command not only removes outliers from class "normal, but it also removes significant outliers in class "outlier". From Fig. 14, it can be seen that in the first column, after removal of outliers, class "normal" appears clean. This will increase the accuracy of the classifier as it will not be confused over the presence of outliers in class "1". In the second column, a single outlier in an instance in class "outlier" is removed which transforms the time series to class "minor". In the third column, even after the removal of predominant outliers, certain pesky outliers remain and hence this instance still belongs to class "outlier". In this way, relabeling datasets will lead to pure classes, which leads to better classifier performance after outlier removal.

4.2 Detrending time series during preprocessing

It can be seen from the confusion matrix that 263 instances in class "trend" are labeled as "class "drift". On

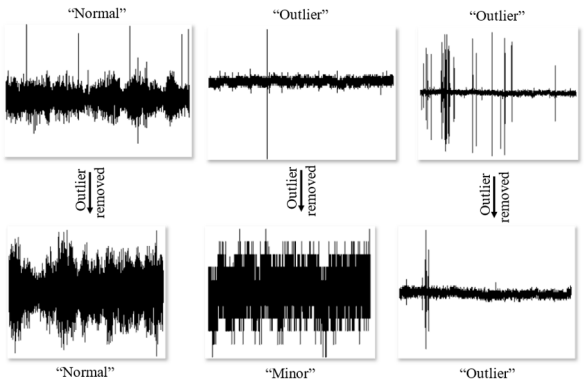

Fig. 14 Removal of outliers during preprocessing

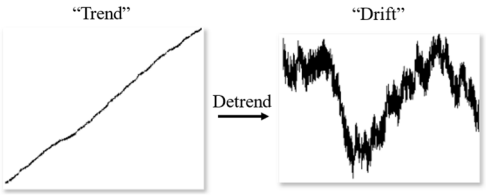

Fig. 15 Detrending during preprocessing

close inspection, it is noted that the instances in class "trend" is nothing but the instances in "drift" with a trend in time series. Since the algorithm is trained on time series envelopes, the classifier finds many similarities between the two classes. Moreover, the class "trend" contains varieties of time series trends increasing from left to right and vice versa. This introduces difficulty during the learning process. In Fig. 15, a time series instance in class "trend" is detrended and it transforms to class "drift". After thorough inspection, it is found that this is the case for all of the instances in class "trend". So, if all the time series instances are detrended this way, during the preprocessing step, this will transform the 7-class problem into a 6-class problem. This is a considerable advantage in terms of computational efficiency and classifier performance. Detrending can be applied together with the removal of outliers in the preprocessing step, and datasets need to be relabeled before feeding them to the machine learning algorithm. These simple preprocessing steps will drastically increase the performance of the classifier.

7. Conclusions

Anomaly detection is a long standing problem in the SHM community. In this paper, this fundamental problem is addressed by autonomously identifying anomalous data patterns in 1-month of acceleration data from an SHM system installed on a long-span bridge in China. This is achieved using a relatively new and efficient time series representation named "Shapelet transform" combined with a machine learning algorithm (Random Forest classifier) to identify anomalies in SHM data. Shapelet transform is a unique time series representation technique that is solely based on the shape of the time series and provides a universal standard feature for detection based on the distance between a shapelet and a time series. The raw measurements of every sensor anomaly have a unique time series shape, and the shapelet transform utilizes this feature to capture these distinct shapes easily. These shapes are used to transform the SHM data into a local-shape space, and the Random Forest classifier is then trained on this transformed dataset to identify and classify the different anomalous data patterns.

The data used in the current study has six different anomalous patterns of time series. From the 1-month acceleration data, the first sixteen days is used for training the algorithm, and the data from the other fifteen days is used for testing. A balanced dataset is created that contains equal samples from all classes of anomalies. The shapelet algorithm discovered 68 shapes from the training dataset. These shapes are used to transform the dataset into a local shape-space. The transformed dataset is then used to train a Random Forest classifier for anomaly detection. The classifier has an overall accuracy of 93%, which indicates that the proposed shapelet-based classifier has an excellent ability to identify anomalies in SHM data. The individual accuracies of all the classes are also above 95%. Various preprocessing measures are also proposed in this paper to increase the classifier performance even further, and this will be pursued in future studies.

Acknowledgments

This work was supported in part by the Robert M. Moran Professorship and National Science Foundation Grant (CMMI 1612843). The authors would like to thank the organizers of the International Project Competition for Structural Health Monitoring (IPC-SHM 2020), Asia-Pacific Network of Centers for Research in Smart Structures Technology (ANCRiSST), Harbin Institute of Technology (China), and the University of Illinois at Urbana-Champaign (USA) for providing the structural health monitoring data of the long-span bridge. The authors also would like to thank the chairs of IPC-SHM 2020, Prof. Hui Li and Prof. Billie F. Spencer Jr, for their leadership in the competition.

References

Abdelghani, M. and Friswell, M.I. (2004), "Sensor validation for structural systems with additive sensor faults", *Struct. Health Monitor.*, **3**(3), 265-275. https://doi.org/10.1177/1475921704045627

Arul, M. and Kareem, A. (2021), "Applications of shapelet transform to time series classification of earthquake, wind and wave data", *Eng. Struct.*, **228**, 111564. https://doi.org/10.1016/j.engstruct.2020.111564

Bagnall, A., Lines, J., Bostrom, A., Large, J. and Keogh, E. (2017), "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances", *Data Min. Knowl. Discov.*, **31**(3), 606-660. https://doi.org/10.1007/s10618-016-0483-9

Bao, Y., Tang, Z., Li, H. and Zhang, Y. (2019), "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., 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

Bostrom, A. and Bagnall, A. (2017a), "Binary shapelet transform for multiclass time series classification", In: *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXII*, Springer, pp. 24-46.

Bostrom, A. and Bagnall, A. (2017b), "A shapelet transform for multivariate time series classification", *arXiv preprint* arXiv:1712.06428.

Breiman, L. (2001), "Random forests", *Mach. Learn.*, **45**(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chang, K.W., Deka, B., Hwu, W.M.W. and Roth, D. (2012), "Efficient pattern-based time series classification on GPU",

Proceedings of 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, Belgium, pp. 131-140. https://doi.org/10.1109/ICDM.2012.132

Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A. and Batista, G. (2015), The UCR Time Series Classification Archive. URL: www.cs.ucr.edu/~eamonn/time_series_data/

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

Ghalwash, M.F., Radosavljevic, V. and Obradovic, Z. (2013), "Extraction of interpretable multivariate patterns for early diagnostics", *Proceedings of 2013 IEEE 13th International Conference on Data Mining*, Dallas, TX, USA, December, pp. 201-210. https://doi.org/10.1109/ICDM.2013.19

Hartmann, B. and Link, N. (2010), "Gesture recognition with inertial sensors and optimized dtw prototypes", *Proceedings of 2010 IEEE International Conference on Systems, Man and Cybernetics*, Istanbul, Turkey, October, pp. 2102-2109. 10.1109/ICSMC.2010.5641703

Hills, J., Lines, J., Baranauskas, E., Mapp, J. and Bagnall, A. (2014), "Classification of time series by shapelet transformation", *Data Min. Knowl. Discov.*, **28**(4), 851-881. https://doi.org/10.1007/s10618-013-0322-1

Lines, J., Davis, L.M., Hills, J. and Bagnall, A. (2012), "A shapelet transform for time series classification", *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, China, August, pp. 289-297. https://doi.org/10.1145/2339530.2339579

Mao, J., Wang, H. and Spencer Jr, B.F. (2020), "Toward data anomaly detection for automated structural health monitoring: Exploiting generative adversarial nets and autoencoders", *Struct. Health Monitor.*, **20**(4), 1609-1626. https://doi.org/10.1177/1475921720924601

McGovern, A., Rosendahl, D.H., Brown, R.A. and Droegemeier, K.K. (2011), "Identifying predictive multidimensional time series motifs: an application to severe weather prediction", *Data Min. Knowl. Discov.*, **22**(1-2), 232-258. https://doi.org/10.1007/s10618-010-0193-7

Mueen, A., Keogh, E. and Young, N. (2011), "Logical-shapelets: an expressive primitive for time series classification", *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA, August, pp. 1154-1162. https://doi.org/10.1145/2020408.2020587

Rakthanmanon, T. and Keogh, E. (2013), "Fast shapelets: A scalable algorithm for discovering time series shapelets", *Proceedings of the 2013 SIAM International Conference on Data Mining*, Austin, TX, USA, May, pp. 668-676. https://doi.org/10.1137/1.9781611972832.74

Shannon, C.E. and Weaver, W. (1949), *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL, USA, p. 117.

Tang, Z., Chen, Z., Bao, Y. and Li, H. (2019), "Convolutional neural network-based data anomaly detection method using multiple information for structural health monitoring", *Struct. Control Health Monitor.*, **26**(1), e2296. https://doi.org/10.1002/stc.2296

Thiyagarajan, K., Kodagoda, S. and Van Nguyen, L. (2017), "Predictive analytics for detecting sensor failure using autoregressive integrated moving average model", *Proceedings of 2017 12th IEEE Conference on Industrial Electronics and Applications* (*ICIEA*), Siem Reap, Cambodia, June, pp, 1926-1931. https://doi.org/10.1109/ICIEA.2017.8283153

Wan, H.-P. and Ni, Y.-Q. (2018), "Bayesian modeling approach for forecast of structural stress response using structural health monitoring data", *J. Struct. Eng.*, **144**(9), 04018130. https://doi.org/10.1061/(ASCE)ST.1943-541X.0002085

Wang, H., Zhang, Y.-M., Mao, J.-X., Wan, H.-P., Tao, T.-Y. and Zhu, Q.-X. (2019), "Modeling and forecasting of temperature-induced strain of a long-span bridge using an improved bayesian dynamic linear model", *Eng. Struct.*, **192**, 220-232. https://doi.org/10.1016/j.engstruct.2019.05.006

Xing, Z., Pei, J., Yu, P.S. and Wang, K. (2011), "Extracting interpretable features for early classification on time series", *Proceedings of the 2011 SIAM International Conference on Data Mining*, Mesa, AZ, USA, April, pp. 247-258. https://doi.org/10.1137/1.9781611972818.22

Xing, Z., Pei, J. and Philip, S.Y. (2012), "Early classification on time series", *Knowl. Inform. Syst.*, **31**(1), 105-127. https://doi.org/10.1007/s10115-011-0400-x

Ye, L. and Keogh, E. (2009), "Time series shapelets: a new primitive for data mining", *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, June-July, pp. 947-956. https://doi.org/10.1145/1557019.1557122

Ye, L. and Keogh, E. (2011), "Time series shapelets: a novel technique that allows accurate, interpretable and fast classification", *Data Min. Knowl. Discov.*, **22**(1-2), 149-182. https://doi.org/10.1007/s10618-010-0179-5