IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 16, 2018, accepted May 20, 2018, date of publication May 24, 2018, date of current version June 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2840086

Detecting Anomalies in Time Series Data via a Meta-Feature Based Approach

MIN HU[1], ZHIWEI JI[2], KE YAN[3], YE GUO[1], XIAOWEI FENG[1], JIAHENG GONG[2], XIN ZHAO[4], AND LIGANG DONG[2]

[1]SHU-UTS SILC Business School, Shanghai University, Shanghai 201800, China
[2]School of Information & Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China
[3]College of Information Engineering, China Jiliang University, Hangzhou 310018, China
[4]Beijing Chaoyang Hospital Affiliated to Capital Medical University, Beijing 100001, China

Corresponding author: Zhiwei Ji (jzw18@hotmail.com)

This work was supported in part by the Foundation of Zhejiang Provincial Department of Education under Grant 1120KZ0416255, in part by the Foundation of Talent's Start-Up Project in Zhejiang Gongshang University under Grant 1120XJ2116016, and in part by the Shanghai Science and Technology Commission under Grant 16DZ1201704.

ABSTRACT Anomaly detection of time series is an important topic that has been widely studied in many application areas. A number of computational methods were developed for this task in the past few years. However, the existing approaches still have many drawbacks when they were applied to specific questions. In this paper, we proposed a meta-feature-based anomaly detection approach (MFAD) to identify the abnormal states of a univariate or multivariate time series based on local dynamics. Differing from the traditional strategies of "sliding window" in anomaly detection, our method first defined six meta-features to statistically describe the local dynamics of a 1-D sequence with arbitrary length. Second, multivariate time series was converted to a new 1-D sequence, so that each of its segmented subsequence was represented as one sample with six meta-features. Finally, the anomaly detection of univariate/multivariate time series was implemented by identifying the outliers from the samples in a 6-D transformed space. In order to validate the effectiveness of MFAD, we applied our method on various univariate and multivariate time series datasets, including six well-known standard datasets (e.g. ECG and Air Quality) and eight real-world datasets in shield tunneling construction. The simulation results show that the proposed method MFAD not only identifies the local abnormal states in the original time series but also drastically reduces the computational complexity. In summary, the proposed method effectively identified the abnormal states of dynamical parameters in various application fields.

INDEX TERMS Anomaly detection, meta-feature, one-class SVM, time series, shield tunneling.

I. INTRODUCTION

A time series is a collection of observations recorded sequentially following time stamps, which makes the time series data have a natural data organization form. As an important class of temporal data, time series receives increasing attention from researchers, and is valuable for data comprises pattern discoveries, including anomaly detection [1], trend analysis [2], periodic pattern detection [3], and short-term prediction [4], etc. In particular, anomaly detection, as an important topic in the area of time series analysis, aiming at finding abnormal or unexpected sequences [5], was widely applied in many areas, such as fault diagnosis [6], healthcare [7], weather data analysis [8], finance [9], *etc*. In Engineering or mechanical field, the availability of mechanisms for early and reliable fault detections greatly reduces the risks of malfunction or unexpected shutdowns of the system [10].

A subway tunnel is an underground passageway. Shield machine is a tunnel boring device, which works in a narrow workspace with high temperature and large gravels [11]. In the process of tunneling, the rotary cutting wheel, which is located at the front of the shield machine, can effectively excavate soil. In order to generate the tunnel lining, a new tunnel ring will be built using the erector once a certain distance has been excavated (roughly 1.5–2 meters). In the process of excavation and propulsion for shield tunneling, various failures or anomalies occur [12], [13]. If failures are not detected promptly, they may influence the progress of the construction, as well as the safety of workers and surrounding environment [4]. Indeed, accurate and early detection of anomalies or faults in a shield system is crucial to prevent the spread of faults and to reduce considerably losses. Currently, many shield tunneling machines can send back dynamical states of parameters in real time; however, there is no anomaly detection method that is able to effectively identify the abnormal states using recent historic data [14]. The high dimension and noises of the time series in shield tunnel construction raise great challenges to current existing computational methods.

Recently, a number of computational approaches have been introduced to find anomalies in univariate/multivariate time series. These methods can be roughly grouped into four categories: (1) statistics-based methods [15], [16], (2) intelligent computing methods [6], [17], (3) Bayesian networks and other Bayesian reasoning extensions [10], [18], and (4) model-based approaches [19]. Statistical approaches belong to data-based technique, which can detect abnormal changes. Principal component analysis (PCA) [20] and partial east squares (PLS) [16] are two basic methods of fault detection in multivariate analysis [16]. In more recent years, several intelligent computing methods were developed for anomaly detection, such as neural networks (or deep learning) [21]–[23], support vector machines (SVM) [4], fuzzy theory [24], and rough sets theory [25], *etc*. But they still have obvious limitations: 1) the mechanical equipment usually lacks training fault samples [10]; 2) the training and testing stages of these methods are mutually independent, and they lack continuous learning ability. Bayesian networks, which can be represented as directed graphs, seem useful for fault detection and isolation with abrupt and incipient faults [26], [27]. Nevertheless, the potential limitation of these methods comes from the fact that more instances imply more required time to inference the weights (probability distributions) of the edges [5]. For the model-based methods (e.g. state space models, vector models [28], [29]), their advantages and drawbacks are associated with the input time series. Without experts' prior knowledge of the system (temporal data), it is generally difficult to accurately build the model. According to above descriptions, we realize that anomaly detection algorithms are usually domain driven and should be built on experts' knowledge.

In this study, a meta-feature based anomaly detection approach (MFAD) is proposed for predicting the potential risks in the complex process of shield tunnel construction. To predict and diagnose the anomalies in shield tunneling machine, we firstly defined six meta-features for statistically describing the dynamics of a subsequence from the original series. Each data point in the segmented subsequence is associated with the same ring number of tunnel lining, which indicates that the time-series data in tunneling is not only related to time, but also related to the distance or the ring number [30]. Secondly, multivariate time series was converted to a one-dimensional sequence. Each of its segmented subsequence was characterized by defined meta-features. This step drastically reduces the dimensionality of the processed data. Thirdly, a one-class SVM (OCSVM) was optimized on transformed samples, and outliers can be easily recognized. For validating its effectiveness, our developed MFAD framework was applied on several well-known public datasets and real-world time series datasets. The simulation results show that the proposed approach is able to automatically identify the abnormal states of the key factors (variables) in the construction of shield machine. The comparisons with other algorithms show that our MFAD approach is significantly better.

The rest of the paper is structured as follows: in **Section II** the proposed computational approach is presented. In **section III**, the datasets for validation and the experiment design are described in detail. In **Section IV**, the simulation results are analyzed and discussed, while in **Section V** conclusions are drawn and suggestion for future work are presented.

II. META-FEATURE-BASED APPROACH FOR ANOMALY DETECTION (MFAD)

In previous studies, sliding window-based strategy was widely used for time series analysis [31]–[34]. However, the size of sliding window is always determined manually. In addition, sliding window is just for sub-sequence segmentation, but the prediction performance also depends on the similarity metrics. In order to avoid the direct calculation of the distance between two sub-sequences to represent their similarity, we proposed a novel strategy to distinguish normal or abnormal sub-sequences with several representative meat-features, which reflect the local dynamics of a time series. In order to capture the characterizations of the time series of a variable (feature) in detail, we defined six meta-features to represent the temporal dynamics and its curve shape. We assume that a univariate time series $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N}\}$ is consisted of M subsequences with a sorted index list of subsequences $Z = \{1, 2, \ldots, z, \ldots M\}$. $\boldsymbol{N}$ is the length of original series $\boldsymbol{X}$. The z-th ($1 \leq \mathrm{z} \leq \mathrm{M}$) subsequence $X_z = \{x_{z,1}, x_{z,2}, \ldots, x_{z,q}\}$ in $\boldsymbol{X}$, is associated with the time range of z-th *tunnel ring* [35], will be simply represented by a one-dimensional vector with six elements. Using our meta-features, any two unequal sub-sequences are comparable. The definition of the above six meta-features were detailed described in the following sections (**Figure 1**).

A. THE DEFINITION OF META-FEATURES

1) KURTOSIS

Kurtosis is a measure of whether a time series is heavy-tailed or light-tailed relative to a normal distribution [36]. This measurement is used to effectively detect the abrupt peaks from a time series, such as ECG data [37], [38]. Kurtosis also provides a way to reflect the variability of a sequence. Time series with high kurtosis tends to have heavy tails, or outliers; however, low kurtosis indicates light tails or lack of outliers. In this study, the Kurtosis was defined in Eq. (1):

$$K_z = \frac{1}{n}\sum\nolimits_{i=1}^{n} D_i^4 - 3 \tag{1}$$

Where n is the length of z-th subseries; and the D_i values are the standardized data values using the standard deviation defined using n rather than $n-1$ in the denominator.

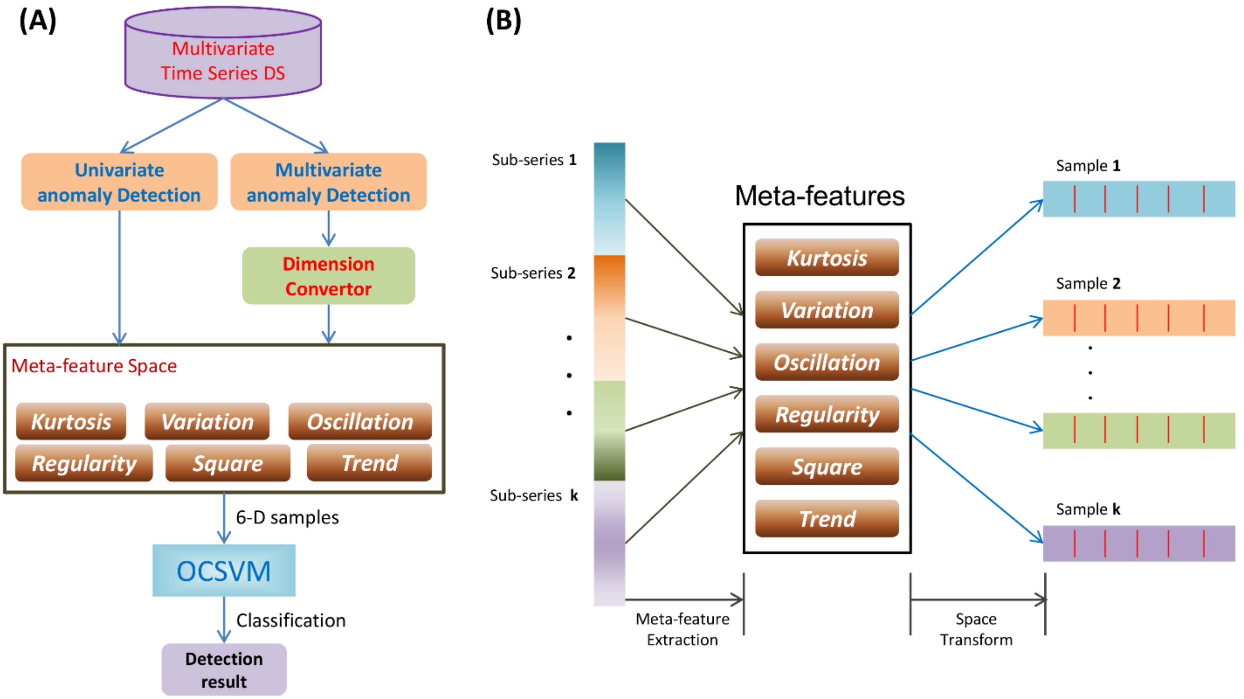


FIGURE 1. (A) The flowchart of the MFAD approach for univariate/multivariate time series; (B) the semantic graph of meta-feature space.

2) COEFFICIENT OF VARIATION

We also defined "coefficient of variation" to calculate the local variability relatively to the whole sequence [39]. Our *hypothesis* is that the local variability of a subsequence can be significant increased if there is an abrupt peak occurred within this interval. Therefore, this meta-feature indicates if a subsequence sharply changed its curve values. The definition was shown in Eq. (2):

$$C_z = \frac{\sigma_z}{\mu} \tag{2}$$

In Eq. (2), variable σ_z denotes the standard deviation of z-th subseries; and μ is the mean value of variable X on the whole series Z.

3) OSCILLATION

Oscillation is a periodic fluctuation between two things, which is a very common problem in industrial area [40]. The presence of oscillations leads to undesirable increased variability of processed variables, and also provides a new insight to identify the local dynamics from time series [41], [42]. For example, Jäncke *et al.* [43] studied how to identify EEG oscillations, which might be associated with dynamic changes in the acoustic features for the musical stimulus. Particularly, Wang *et al.* [44] proposed a discrete cosine transformation (DCT)-based method to effectively detect oscillations from univariate time series. In this study, we also calculated the oscillation of a subseries by using Wang *et al.*'s approach.

4) REGULARITY

For assessing the complexity of sequence, we applied sample entropy [45] to calculate the regularity of time series. Sample entropy is widely used for diagnosing diseased states from physiological time-series signals [46]. The less number of abrupt peek in the sequence, the larger the value of regularity is.

5) SQUARE WAVE

Square waves are universally encountered in digital switching circuits and are naturally generated by binary logic devices. According to the sampled timeline within a ring, a shield tunneling machine firstly boosts forward for a while, and then the segments are installed [47]. Considering the specialty of our application, the states of some variables within a ring represent obvious square wave: a sequence starts and maintains the signal with significant high values in the first half of the period, and sharply reduces the signal for the second half way. We assumed that the curve of a variable within a ring is normal if square wave is represented and consistent with expectation; otherwise, an abnormal is detected. Given the vector of z-th subseries as $X_z = \{x_{z,1}, \ldots, x_{z,i}, x_{z,i+1}, \ldots, x_{z,N}\}$ and $i = \lfloor N/2 \rfloor$, the binarized sequence of X_z can be easily obtained, which is denoted as TX_z. The calculation of TX_z was defined in Eq. (3):

$$TX_z = X_z > 0.5 * \max(X_z) \tag{3}$$

We tested if z-th subseries is square wave with the following Eq. (4):

$$S_z = 0.5 - rs * TX_z / LEN(TX_z) \tag{4}$$

Where function $LEN(\cdot)$ denotes the length of the sequence TX_z, and vector $rs = \{1, 1, \ldots, 1, 0, 0, \ldots, 0\}$ filters the signal with high values in TX_z. In the vector rs, the number of "1" is i. According to above formulations, the sub-series within the time of a ring corresponds to a lower value if it represents square wave in the sequence.

6) VARIATION OF TREND

The above five meta-features extract the various dynamics from the original time series, however, trend analysis provides a new way to represent the difference between two series [48]. For evaluating the general trend of the subseries related with a ring, we firstly smoothed the original sequence and then calculate the variation on it. The variation of trend of z-th subseries was defined in Eq. (5):

$$T_z = std(smooth(X_z)) \tag{5}$$

Where function $smooth(\cdot)$ and $std(\cdot)$ are used to obtain the smoothed sequence of the original series X_z and its standard deviation, respectively. For a series with random trend, the value T_z will be small if there is no abrupt peak. If a series represent abrupt peak, the value of T_z will be large (see **Figure 1A**).

B. ANOMALY DETECTION BASED ON META-FEATURES

1) DIMENSION CONVERTOR

Often, the anomalies in a multivariate time series can be detected only by analyzing sequence of all variables. The key underlying idea is to reduce a multivariate time series into a univariate time series by exploring the correlation structure of the original variables [5]. In this study, the *dimension convertor* is designed to extract the first component vector of a multivariate time series with PCA or SVD; and the multivariate series was converted to a univariate series. The extracted first component represents the largest amount of information from the original series. After obtained the converted univariate series, six meta-features were calculated from this new generated sequence.

2) SPACE TRANSFORMATION

After defining the six meta-features, each subseries related with a *tunnel ring* will be represented as a sample with six elements, regardless of the length of the original subsequence. For identifying the abnormal states from the whole sequence of variable $\boldsymbol{X}$, each subsequence X_z $(1 \le z \le M)$ was represented as $NS_z = \{K_z, C_z, O_z, R_z, S_z, T_z\}$. Therefore, all the M subsequences of variable $\boldsymbol{X}$ will be transformed into a new 6-dimentional space as M samples, and the abnormal state detection is just to identify the outliers from these samples (**Figure 1B**).

3) ONE CLASS SVM (OCSVM)

The support vector machines (SVM), firstly proposed by Cortes and Vapnik [72], was initially developed to solve the two-class classification problem [4]. The one-class support vector machines (OCSVM), proposed by Scholkopf *et al.*, aims at detecting samples that do not resemble the majority of the dataset [73]. As a method of distribution estimation, OCSVM was considered as a novel tool in outlier detection. Nowadays, OCSVM has been widely adopted in many one-class classification application fields, such as, fault detection and diagnosis [1], [50].

4) META-FEATURE BASED ANOMALY DETECTION USING OCSVM

According to the above description, we identified the abnormal states using OCSVM on the meta-feature-based data space. Each transformed sample for OCSVM corresponds to a subseries within a *tunnel ring*. Therefore, the detecting resolution of our method is one ring, and could not go to intra-ring. Given a small number of transformed samples to OCSVM, the significant outliners can be easily detected. According to the special design of above meta-features, the sampled time series only requires to be calculated one time, which is suitable for online learning.

III. SIMULATION EXPERIMENT

A. DATA PREPROCESSING

For evaluating the performance of the proposed method MFAD, we designed simulation experiments to test a bunch of univariate and multivariate time series datasets, including **1)** four well-known time series datasets, e.g. Nprs44 [51]–[53], chfdb/chf01 [53], [54], mitdbx_108 [53], and Air quality (UCI) [5]; The first three datasets can be download from the link: http://www.cs.ucr.edu/∼eamonn/discords/, and the last one is from UCI public database. **2)** a real-world time series dataset of shield construction, which was collected from a tunneling company located in Shanghai, China. The purpose of this study is to develop an efficient computational approach to identify the potential abnormal states of dynamical parameters in shield tunneling machines. The details of above datasets were described as following.

The time series *Nprs44* [51], [52], showing a patient's respiration, as they wake up. This data series was manually segmented as three stages by a medical expert Dr. J. Rittweger: 1) state II sleep; 2) eyes closed, awake or stage I sleep; and 3) Eyes open, awake. The series was visualized as shown in **Supplementary Fig. S1**. There are two obvious discords occurred in this series. The first discord is very obvious deep breath taken as the patient opened their eyes (stage III). The second discord is much more subtle and impossible to see at this scale (stage I). The details of this dataset was described in [53].

Chfdb/chf01 [54], an ECG (Electrocardiography) time series, is come from the MIDMC congestive Heart Failure Database. This dataset with length 3751, includes two traces (variables), and there is an obvious discord occurred in each trace (See **Supplementary Fig. 2**).

TABLE 1. The univariate time series datasets tested on the MFAD method.

Dataset	Ground Truth	Anomalies predicted by MFAD
Nprs44 (Patient respiration)	Subsequence 3, 9	Subsequence 3, 9
chfdb/**chf01** (ECG, variable 1)	Subsequence 6	Subsequence 6
mitdbx_108(ECG, variable 1)	Subsequence 4, 8, 9	Subsequence 8, 9
Cutter Head: **TS_C01**. (Ring No. 258-290)	Ring 267 Cutter head tripping	260, 266, 267, 287
Cutter Head: **TS_C02**. (Ring No. 551-580)	Ring 568 Cutter head startup failed	560, 564, 567, 568, 569, 573, 579
Grouting system Pump 2#: **TS_G01**. (Ring No. 1011-1040)	Ring 1036 Grouting pump was clogged	**Grouting flow**: 1018, 1024, 1036, 1037, 1038. **Grouting pressure**: 1012, 1031, 1033, 1036, 1037.
Grouting system Pump 4#: **TS_G02**. (Ring No. 1061-1090)	Ring 1079 Piston damage of grouting pump	**Grouting pressure**: 1070, 1073, 1075, 1079, 1080, 1081, 1082, 1087
Hydraulic thrust system: **TS_H01**. (Ring No. 165-200)	Ring 200	**Engine oil pressure**: 172, 175, 177, 186, 194, 197, 200. **Total thrust**: 185, 190, 200.
Hydraulic thrust system: **TS_H02**. (Ring No. 461-490)	Ring 473-475 Oil pressure cannot be controlled	**Engine oil pressure**: 461, 466, 469, 474, 490

TABLE 2. The multivariate time series datasets tested on the MFAD method.

Dataset	Dimension	Ground Truth	Anomalies predicted
chfdb/**chf01**	2	Subsequence 6	**PCA**: Subsequence 6. **SVD**: Subsequence 6.
mitdbx_108_2	2	Subsequence 5	**PCA**: Subsequence 5. **SVD**: Subsequence 5.
Air Quality (UCI)	5	Subsequence 2, 8, 9, 11	**PCA**: Subsequence 8, 9, 11. **SVD**: Subsequence 2, 8, 9, 11.
Shield tunnel construction **MDS1**	6	Ring 568, 569	**PCA**: 560, 561, 565, 566, 567, 568, 569, 573. **SVD**: 560, 561, 565, 567, 568, 569, 573.
Shield tunnel construction **MDS2**	9	Ring 323, 329, 336, 356, 364	**PCA**: 323, 329, 340, 352, 353, 356, 364, 365, 366. **SVD**: 323, 329, 336, 338, 349, 352, 353, 356, 364, 365, 366.

The dataset *mitdbx_108* is recorded from the PhysioNet Web server; and its length is 21600 [53]. There are two variables (features) included in this time series. The first variable (**Supplementary Fig. S3**) was widely studied by other works and the top 3 discords were recognized. Cardiologists from MIT have annotated the discords in this time series, and Keogh *et al*. [53] have added colored markers to draw attentions to those annotations.

The dataset *Air quality* [55] contains 9358 instances of hourly averaged responses that are collected from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multi-sensor Device. All the data was recorded from March 2004 to February 2005 representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. For each of the above three well-known time series datasets, we randomly picked up several non-overlapping sub-series from the original series to validate the effectiveness of our approach.

The real-world time series data of shield construction includes over 400 variables, and each variable is observed once every 10 seconds. After removing out the irrelevant and redundant features [56], [57], we finally obtained 157 features, which related with detailed states on six aspects of shield advance. In our experiments, we mainly focused on the top three most important aspects: 1) the states of cutter [58]; 2) the states of grouting system [59]; and 3) the hydraulic thrust system [60], [61]. We mainly focused on the variable "***rotating speed of cutter head***" to check the states of cutter, "***grouting flow***" and "***grouting pressure***" for the states of grouting system, and "***engine oil pressure***" and "***total thrust***" for the hydraulic thrust system. In our simulation, we randomly chose univariate or multivariate subsequences related with above three aspects and constructed 8 subsets in total; and each one covered at least 30 *tunnel rings*. The details of the selected dataset were described in **Table 1-2**.

B. EXPERIMENT DESIGN

To validate the effectiveness, we tested our developed MFAD approach on various univariate and multivariate time series datasets, including square waves or random sequences.

1) VALIDATION ON UNIVARIATE SEQUENCES

Here, we selected three subsets from the well-known datasets Nprs44, chf01, and mitdbx_108. Considering the original series chf01 [54] and mitdbx_108 [53] are 2D, we selected the first variable of both datasets for validation. All the instances of testing variables were used in the simulation. For each of above three univariate sequences, we randomly selected 9 or 12 non-overlapping subsequences (including proved discords), and transferred them to 6-dimentional data space for outlier detection via OCSVM. Most importantly, we also constructed six subsets of univariate sequences, which were related to the dynamics of shield tunneling machines. The details of above datasets were shown in **Table 1**.

2) VALIDATION ON MULTIVARIATE SEQUENCES

Our computational framework also addresses multivariate time series that represent multiple discords in different variables at different time points (**Table 2**). In our experiments, we selected three multi-dimensional time series to test our MFAD approach: chf01 (2D), mitdbx_108_2 (2D), and Air Quality (5D). All instances (3751) of two variables in chf01 were selected for validation. The number of instances from Air quality dataset is 3500-9000. In addition,mitdbx_108_2 [62], with length 5400, is a subset of mitdbx_108. In addition, we also collected two multivariate time series of shield construction: MDS1 (5D), and MDS2 (9D). Each multivariate time series was converted to one-dimensional sequence via PCA [63] or SVD [64]. The first component vector was further analyzed with our developed meta-features. We simply named the above two strategies as: PCA+MFAD, and SVD+MFAD. The discords occurred in different variables of an original dataset might be reproduced with the first component vector.

3) COMPARISON WITH OTHER ALGORITHMS

To validate the effectiveness, we compared our proposed approach with three typical algorithms on four of above datasets. The algorithms that we selected to be compared are Brute force [65], SAX [53], and K-means based clustering [66], [67]. Four datasets were calculated, including npr44, TS_01, Air Quality, and MDS2 (See the details in **Table 1-2**). The multivariate time series datasets need to be firstly converted to univariate sequence with PCA or SVD. Based on the information of ground truth of each dataset show in **Table 1-2**, accuracy, sensitivity, and specificity were calculated to represent the differences between all of four computational approaches.

C. EXPERIMENTAL PARAMETERS

The simulating experiments were performed under the environment of MATLAB 2017a and LIBSVM 3.22 [68] with Intel Core i7-6600U Processor, 8G RAM (1600MHZ). As a distribution estimation method, OCSVM, has been implemented to identify the outliers (abnormal states) from the whole data samples. After assigning the label of each sample as "1", all the samples were used to construct a one-class SVM. After obtained the optimal trained model, the same samples were input to OCSVM again for the label prediction. The samples which predicted as "-1" were finally determined as abnormal states (outliers). In OCSVM model, Gaussian RBF kernel was employed, and the kernel parameters C and γ were optimized by grid search [56]. In the grid search, we set $C = 2^a$ and $\gamma = 2^b$. Variable a changes from -5 to 15 with step 0.1, and variable b changes from -15 to 10 with step 0.25. Therefore, we have the range of [0.0313, 32768] for C and the range of [0, 1024] for γ. In order to receive the better performance from OCSVM, the data samples were normalized before classification.

To evaluate the accuracy of prediction, three statistical metrics were employed here: (1) accuracy, (2) sensitivity, and (3) specificity. The definition of these metrics were described as shown in Eq. (6-8).

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FP + FN} \tag{6}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

Where variables TP, FP, TN, and FN denote the number of normal subsequences correctly detected as normal (True positives), the number of abnormal subsequences that are detected as normal (False Positives), the number of abnormal subsequences that are recognized as abnormal (True negatives), and the number of normal states that are recognized as abnormal (False negatives).

IV. RESULTS

A. VALIDATION ON UNIVARIATE TIME SERIES

Firstly, our proposed approach was tested on 9 subsequences, which were randomly and manually segmented from the original sequence Nprs44. The simulation result is consistent with the observation. From **Figure 2**, we can clearly observe that the 3-th and 9-th subsequences were detected as abnormal, which were corresponding to the 2-discord and 1-discord shown in **Supplementary Fig. S1**.

Secondly, our approach was tested on 9 subsequences that were extracted from the 1st variable in ECG time series chf01. From **Supplementary Fig. S4**, we found that one obvious discord is detected automatically, which is consistent with **Supplementary Fig. S2**.

Thirdly, our approach was tested on 12 subsequences that were randomly selected from the 1st variable in time series mitdbx_108 [53]. **Supplementary Fig. S5** shows that our

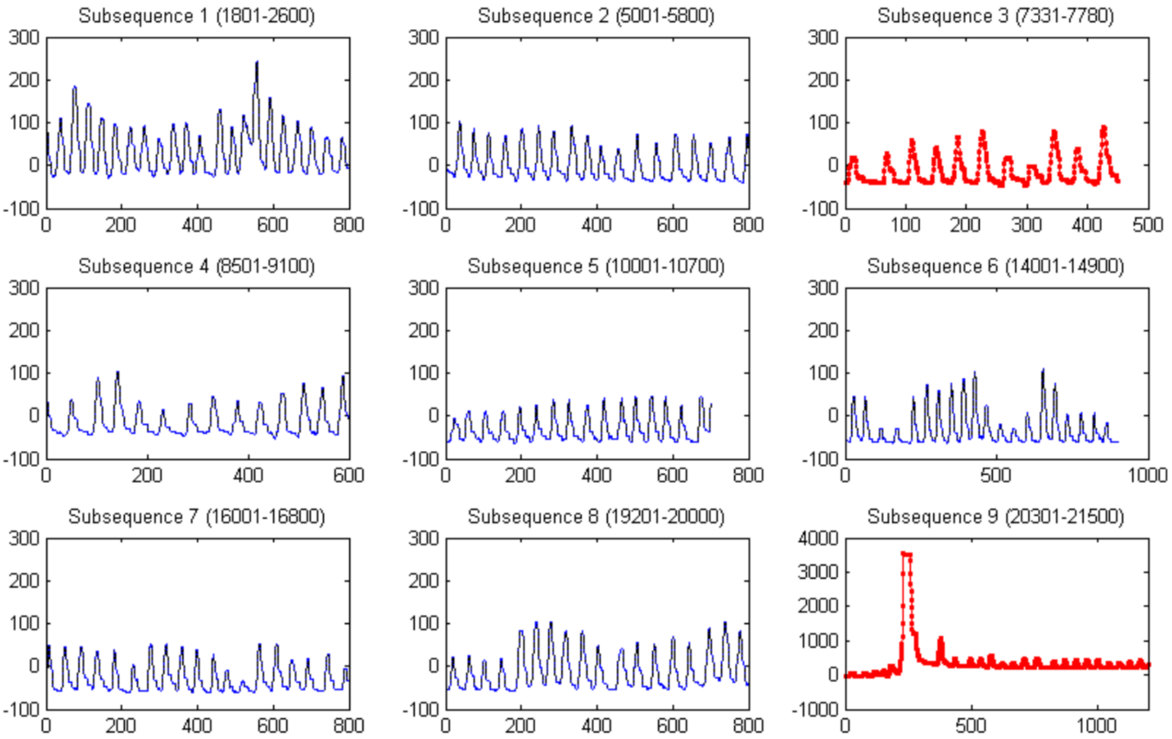


FIGURE 2. Nine subsequences were manually selected from time series Nprs44. Two subsequences were detected as abnormal by MFAD and marked in red.

approach detected 8-th and 9-th subsequences as anomalies, which correspond to discord 1 and 2 reported in [53]. Subsequence 4 was missed by MFAD after detection (**Supplementary Fig. S3**). Compared to the sustaining changes in top 2 discords, the third discord represents a sudden peak so that our method wasn't able to identify it (**Supplementary Fig. S5**). According to the results on the above two datasets, we consider that our method can distinguish the abnormal states from most of the normal states.

The proposed method was also tested on six univariate time series of shield construction.

I) In shield tunnel construction, cutter is used for cutting of frontal soils under the thrust force [69]. The key variable "*rotating speed of cutter head*" is an indicator to reflect if the cutter is normally running [70]. In this study, we selected two univariate series (TS_C01, and TS_C02) related with the state of cutter, which were observed in the tunnel ring 258-290, and 551-580, respectively. In addition, the fault log shows that there is a fault occurred in each sequence (see the details in **Table 1**). **Figure 3** shows the curves of 30 subsequences during the ring 258-290 in series TS_C01. From **Figure 3**, we can clearly observe that the normality of cutter should present an obvious square wave, e. g, the ring 280, or 286, *etc*. However, "the tripping of cutter" indicates that the cutter could not work stably at the first half stage and is always characterized by frequent shock, such as ring 266, or 267. After the simulation with MFAD approach, four subsequences (nos. 260, 266, 267, and 287) were diagnosed as abnormal states. The fault log reported that the cutter met an abnormity ("cutter was turn down") in the ring 267, and induced the tripping of cutter. In fact, this fault was recognized by our approach in advance (see the curve of ring 266). Particularly, our **trend analysis** of this variable shows more clearly that the detected abnormal states are significant different with others (**Supplementary Fig. S6**). In **Supplementary Fig. S6**, the dynamic of a normal cutter shows as a square wave; otherwise, a straight line or random curve indicates cutter cannot be started. Also, **Supplementary Fig. S7** presents the curves of 30 subsequences in the period 551-580 (**TS_C02**), in which the curves of predicted abnormal states are significantly different from others. Our prediction shows that seven subsequences (560, 564, 567, 568, 569, 573, and 579) were recognized as abnormal states. Particularly, ring "568" was reported as a fault state in the fault log: the cutter cannot be started. Actually, the rings from 567-569 were all abnormal, and the rotating speed of cutter is hard to maintain on a normal level. Taken together, above simulation results indicate that our MFAD method effectively identify the significant abnormal states from the original series, and is able to find the occurrence of fault in advance.

II) Secondly, we applied our method to two sets related with states of grouting system (Dataset TS_G01, and TS_G02). These two datasets were collected from different *grouting pumps* during the periods with ring number 1011-1040, 1061-1090, respectively. In our fault log, the grouting pump #2 was diagnosed as fault in the ring 1036, and

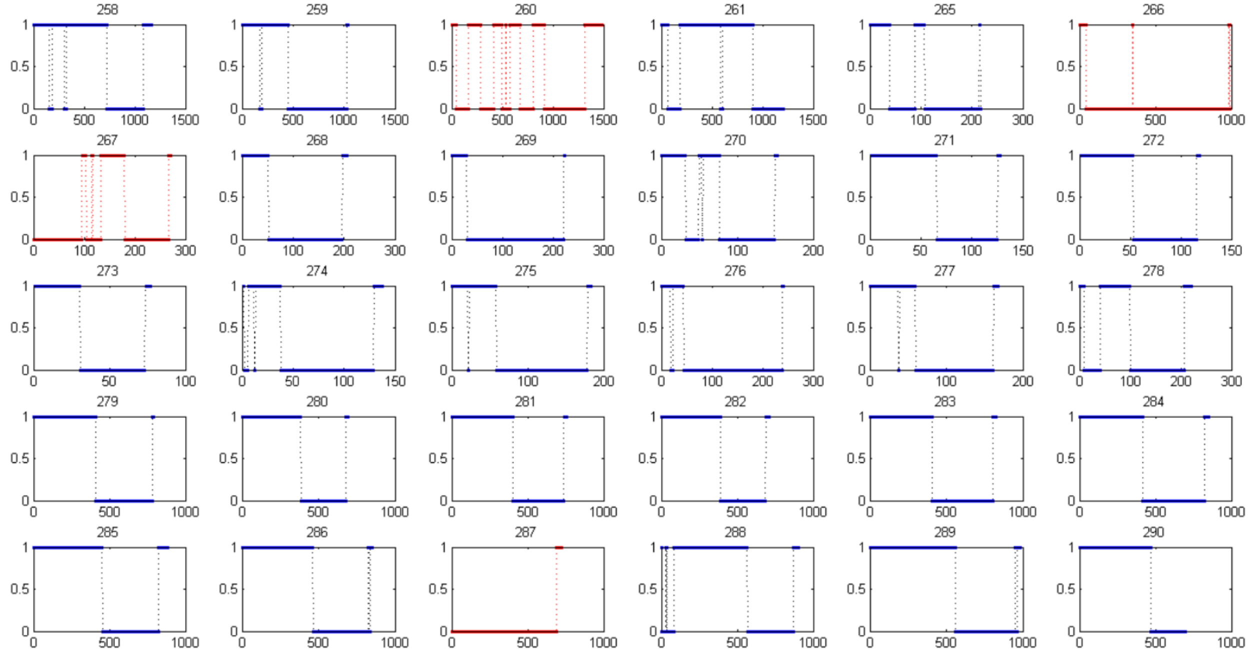


FIGURE 3. Univariate time series TS_C01 (rotating speed of cutter head) includes 30 subsequences, which were observed from ring no. 258 to 290. Four subsequences were detected as abnormal by MFAD and marked in red.

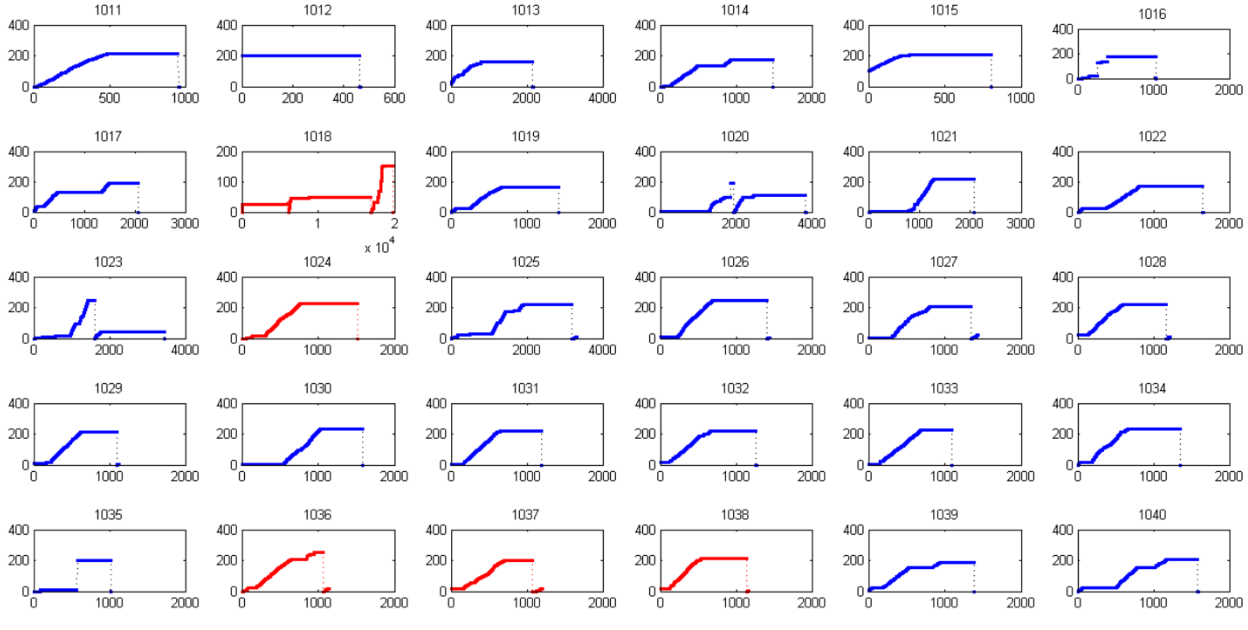


FIGURE 4. Univariate time series TS_G01 (grouting flow in pump #2) includes 30 subsequences, which were observed from ring no. 1011-1040. Five abnormal subsequences were detected as abnormal by MFAD and marked in red.

pump #4 has fault in the ring 1079 (**Table 1**). **Figure 4** shows the curves of variable "*grouting flow*" related to the pump #2 from ring 1011-1040 (TS_G01). In **Figure 4**, most of the curves indicate that the quantity of grouting gradually increase and will maintain at a high level. Our approach predicted five abnormal states of quantity of routing in the period 1011-1040, including 1018, 1024, 1036, 1037, and 1038. Obviously, the subsequence at ring 1036 presents an abnormity: when the signal grew from zero to high level, it suddenly dropped down. **Supplementary Fig. S8** presents all the curves of variable "*grouting pressure*" for pump #2 from ring 1011-1040. Although the dynamic trend of grouting pressure is complicated, there are five subsequences were detected as anomalies by our approach, including 1012,

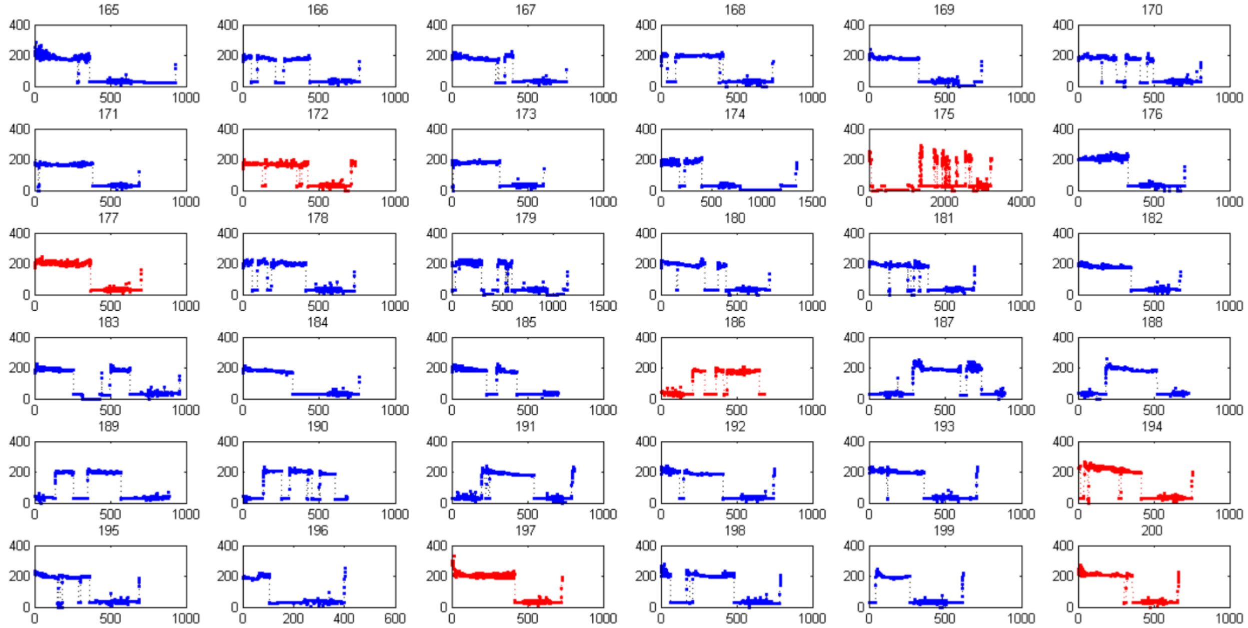


FIGURE 5. Univariate time series TS_H01 (engine oil pressure) includes 30 subsequences, which were observed from ring no. 165-200. Seven abnormal subsequences were detected as abnormal by MFAD and marked in red.

1031, 1033, 1036, and 1037. Therefore, the series of *grouting flow*and *grouting pressure* have abnormal states in the ring 1036, which is consistent with the information in fault log. We also applied our method to the variable of *grouting pressure,*associated with pump #4 for ring 1061-1090 in time series TS_G02 (**Supplementary Fig. S9**). There are eight subsequences predicted as abnormal states (see the details in **Table 1**). According to the above simulation results, we can conclude that our computational approach is capable of identifying the potential abnormal states from complex series.

III) Finally, our MFAD approach was applied to the time series of variables, e.g. *engine oil pressure*, and *total thrust*, which were considered as key characterizations of hydraulic thrust system in a shield tunneling machine [60], [61]. In total, we selected two univariate time series (TS_H01, and TS_H02) of the hydraulic thrust system, which were collected in the periods 165-200 and 461-490, respectively (see the details in **Table 1**).

Figure. 5 shows the curves of 30 subsequences about *engine oil pressure* from the ring 165-200 in TS_H01. In general, most of subsequences in **Figure. 5** indicate that *engine oil pressure* should be maintained on high level in the first half of a ring, and then be turn down in the last half. After calculation, our approach found there were seven abnormal states, including ring 172, 175, 177, 186, 194, 197, and 200. Comparing with other cases, the variable "*engine oil pressure*" in some certain stages (e.g. ring nos 175, or 186) presents opposite trend. In addition, we further tested the series of "*total thrust"* in the period 165-200, and found that three subsequences (185, 190, and 200) might be potential abnormal states (**Supplementary Fig. S10**). Finally, the prediction related with *engine in pressure* in the period from ring 461 to 490 was presented (**Supplementary Fig. S11**). The prediction results show that five subsequences (nos. 461, 466, 469, 474, and 490) were detected as abnormal states (**Table 1**). Specially, we discussed the predicted subseries in ring 474. Our fault log shows that the variable "*engine oil pressure*" includes an abnormal state in the ring 473; however, the fault record claimed that there was something wrong with the hydraulic in the process of shield advance in the ring 475 so that the speed of the advance was delayed. The performance of MFAD on all the univariate time series was detailed presented in **Table 3**.

TABLE 3. The performance of MFAD on nine univariate time series.

Dataset (univariate)	Accuracy	Sensitivity	Specificity
Nprs44	1	1	1
Chfdb/ch01 (Variable 1)	1	1	1
Mitdbx_108 (Variable 1)	0.9167	1	0.90
TS_C01	0.90	0.897	1
TS_C02	0.80	0.793	1
TS_G01	0.867	0.862	1
TS_G02	0.767	0.759	1
TS_H01	0.867	0.862	1
TS_H02	0.867	0.862	1

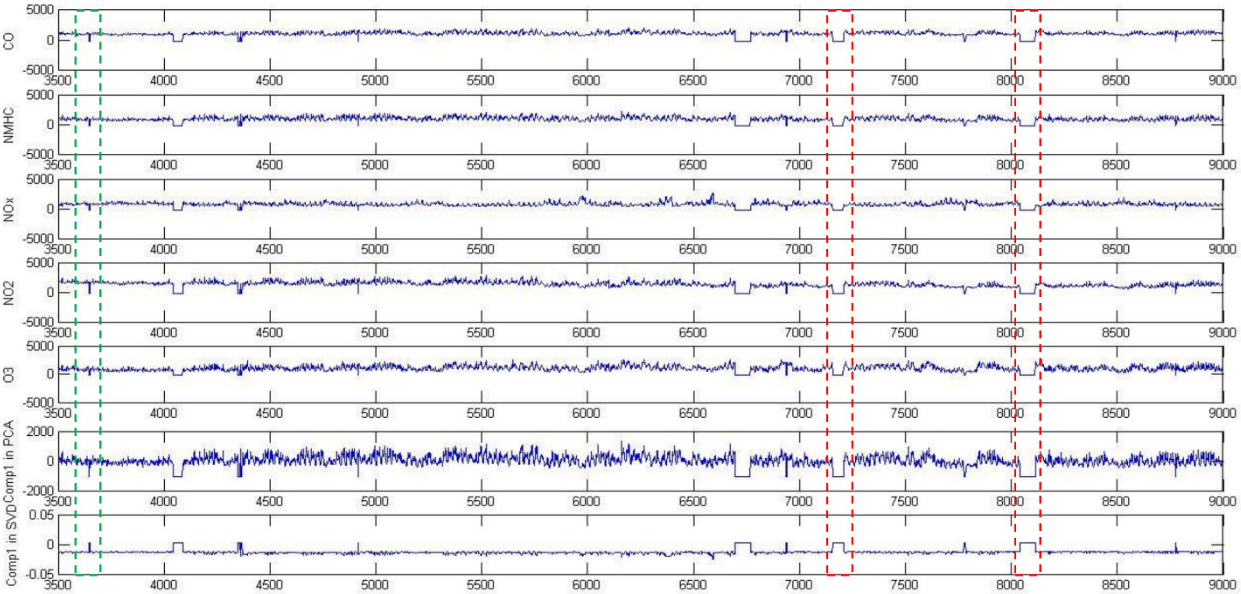


FIGURE 6. Multivariate time series Air Quality includes 5 variables, which corresponds to 5 sensors. The last two series are the first component vectors of original series extracted by PCA and SVD, respectively. The abnormal subsequences highlighted with red color needed to be detected.

B. VALIDATION ON THE MULTIVARIATE TIME SERIES

To test the effectiveness of our strategy, five multivariate time series datasets were selected, including **1)** three well-known time series datasets (chf01, mitdb_108, and Air quality); **2)** two multivariate time series of shield construction. The details of these datasets were introduced in **Section III** (Also see **Table 2**). For each given multivariate series, we applied PCA and SVD to extract the first component and therefore convert the original series to a univariate series. Finally, MFAD will be performed on the one-dimensional component vector to identify the outliers.

First, let us look at the **Supplementary Fig. S12**. The first two curves represent two variables in dataset chf01 [54]. Obviously, each variable has a discord, which is marked in red color. After transforming using PCA or SVD, the corresponding first component was shown in the bottom. From **Supplementary Fig. S12**, we found that both PCA and SVD can reserve the characteristics of the original multivariate series, with the abnormal stages revealed. Since two discords in the original series are overlapping in the time line, the component vectors obtained from PCA or SVD merged two abnormal regions that become a new wider discord. Using the same segmented regions in the time line (**Supplementary Fig. S4**), 9 subsequences of the first component vector derived from PCA or SVD were obtained. MFAD were then used on these two converted one-dimensional vectors; and the abnormal subsequences were identified (**Supplementary Fig. S13-14**). We found that both PCA+MFAD and SVD+MFAD is capable of identifying the unique discord locating in the region 2301-2600.

Second, **Supplementary Fig. S15** shows the original series of mitdbx_108_2 and the extracted first component from PCA or SVD. Also, our results show that PCA+MFAD and SVD+MFAD can recognize the same discord locating in the region 3911-4300 (**Supplementary Fig. S16-17**), which corresponds to the 3rd discord in the whole dataset mitdbx_108 shown in **Supplementary Fig. S3**.

Third, our approach was tested on the Air quality dataset. **Figure 6** shows that the first five curves correspond to 5 variables, which are observed by 5 chemical sensors. From **Figure 6**, it is obvious that one discord is represented as a sudden decreased wave (marked with red color). In addition, the markers with green color in **Figure 6** denote the noise signals. Comparing with PCA+MFAD, we found that SVD+MFAD provide a much distinguished component vector, which significantly emphasized on the potential abnormal subsequences and weakened the normal signals (**Figure 6**). We manually selected 12 subsequences from the first component derived from PCA+MFAD or SVD+MFAD, and identified the outliers. **Figure 7** shows that PCA+MFAD can recognize three abnormal subsequences (8th, 9th, and 11th), however, SVD+MFAD has a higher accuracy with four diagnosed discords: 2nd, 8th, 9th, and 11th subsequences in **Figure 8**.

Finally, our approach was tested on two multivariate time series of shield construction. We manually constructed two datasets, and there were multiple anomalies occurred in different original variables. We want to verify PCA+MFAD or SVD+MFAD can recognize the abnormal states in the converted space, which might correspond to the discord in certain original variables. The first multivariate time series is MSD1, which includes 5 variables, including "*grouting flow (pump 4#)*", "*grouting pressure(pump 4#)*", "*engine oil pressure*","*total thrust*", *and* "*rotating speed of*

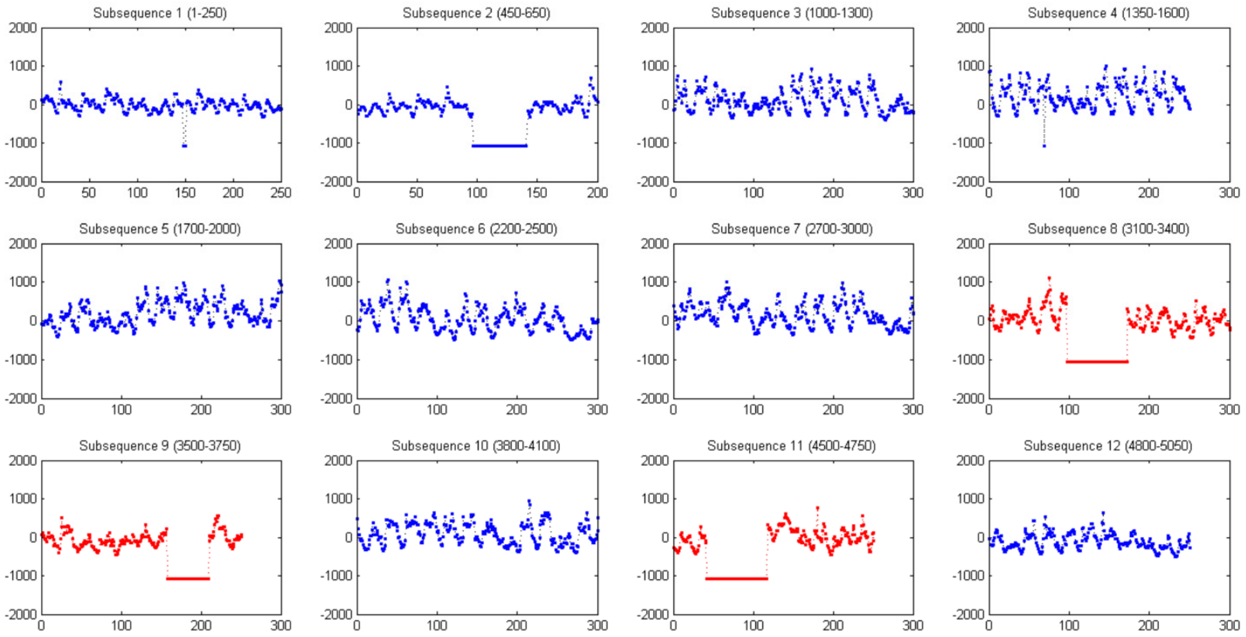


FIGURE 7. 12 subsequences were manually selected from PCA-extracted first component vector of Air quality. Three subsequences were detected as abnormal by MFAD and marked in red.

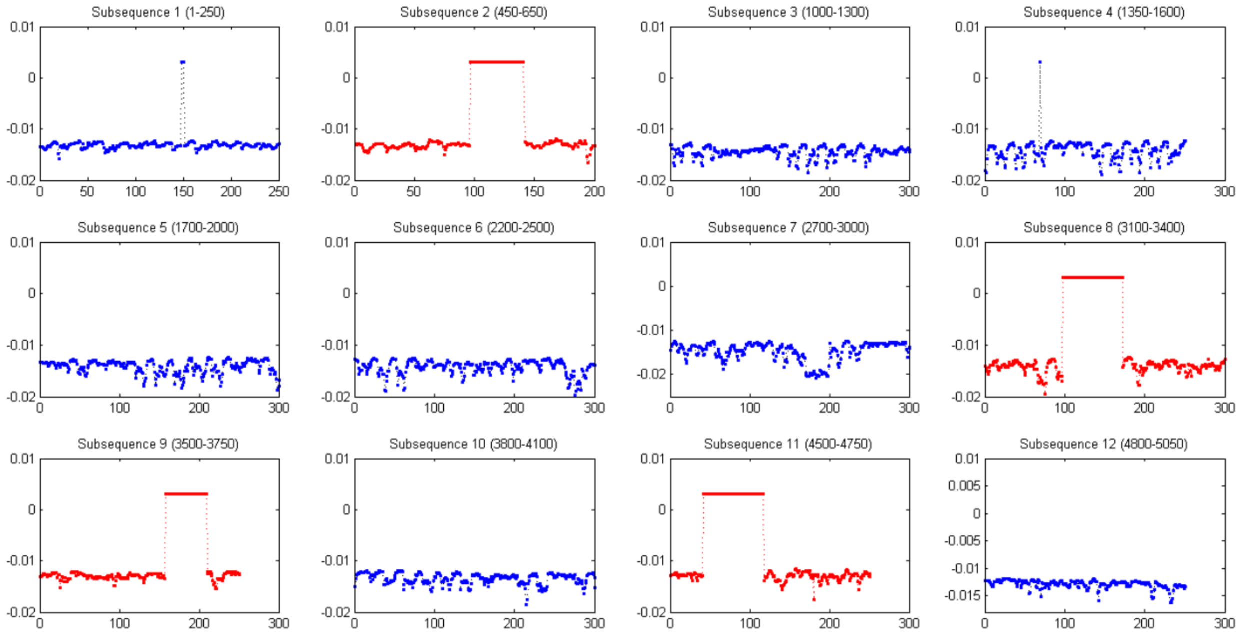


FIGURE 8. 12 subsequences were manually selected from SVD-extracted first component vector of Air quality. Four subsequences were detected as abnormal by MFAD and marked in red.

cutter head". **Figure 9** shows the five original 5 variables and the first components calculated from PCA and SVD. In **Figure 10**, there are 8 abnormal subsequences (ring nos. 560, 561, 565, 566, 567, 568, 569, 573) were identified by PCA+MFAD. SVD+MFAD found 7 abnormal subsequences (ring nos. 560, 561, 565, 567, 568, 569, and 573) in **Figure 11**. The fault log reported that, 1) the cutter can't be started in the ring 568; and 2) grouting pump 4# were damaged in the ring 569. Similarly, another multivariate time series is MSD2 (figure not shown), which includes 9 variables (see the details in **Table S1**). We also extracted the first component vectors from MSD2 with PCA and SVD. **Supplementary Fig. S18** shows that 9 abnormal subsequences are identified (PCA+MFAD), which represent significant

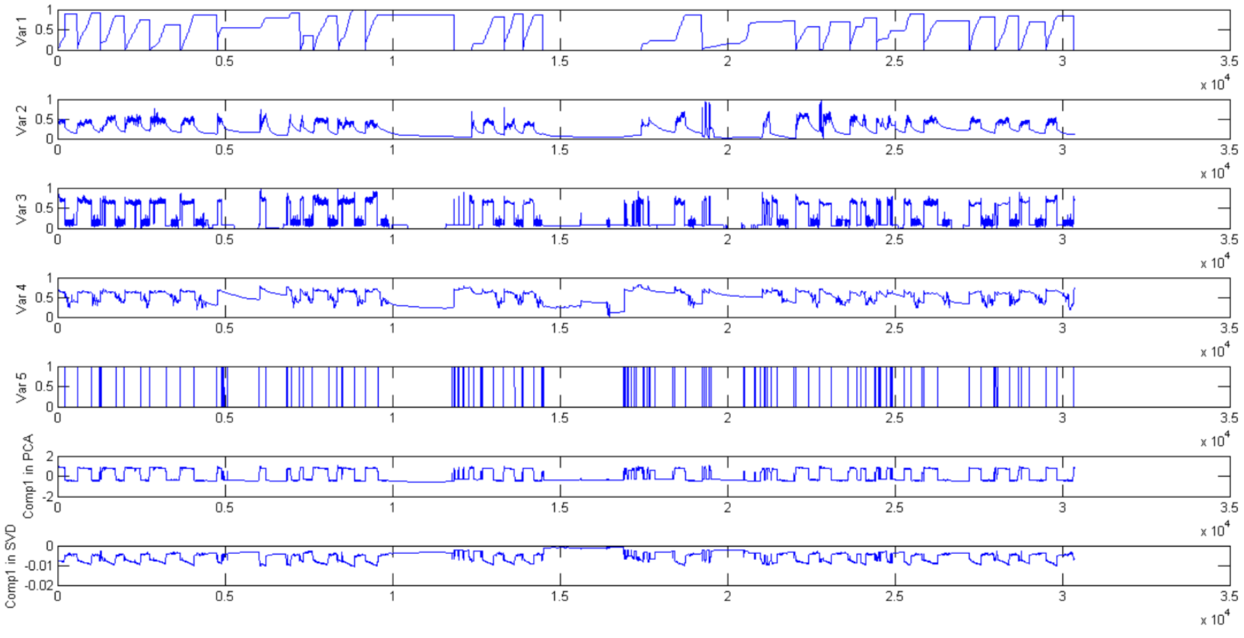


FIGURE 9. Multivariate time series MDS1 includes 5 variables, which corresponds to 5 state parameters in shield tunneling construction. The last two series are the first component vectors of original series extracted from PCA and SVD, respectively.

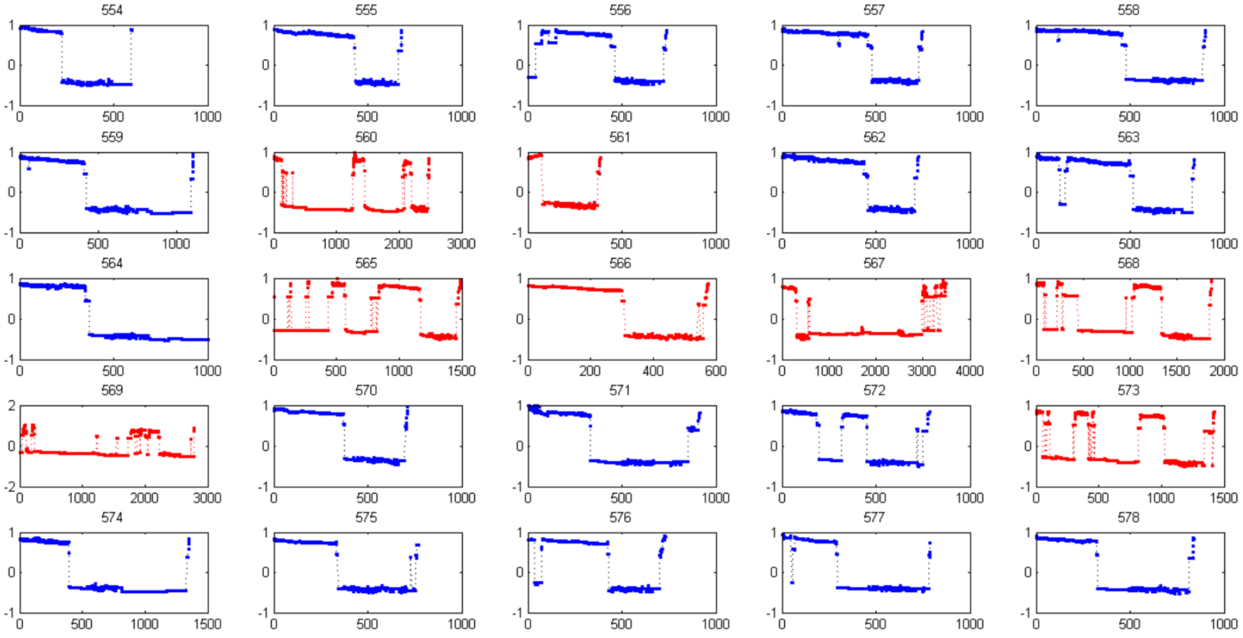


FIGURE 10. The first component vector of MDS1 via PCA includes 30 subsequences, which correspond to ring no. 554-578. Eight abnormal subsequences were detected as abnormal by MFAD and marked in red.

different curves rather than others. SVD+MFAD found 11 abnormal subsequences (**Supplementary Fig. S19**). Based on above simulation, we concluded that the proposed approach can be used for multivariate time series anomaly test. The key step is to convert the multi-dimensional time series to one-dimensional series so that the meta-features can be extracted and calculated for determining the outliers. The performance of MFAD on all the multivariate time series was detailed described in **Table 4**.

C. COMPARISON WITH OTHER ALGORITHMS

To verify the performance of the proposed approach MFAD, we tested four of above datasets on three typical discord detection methods: Brute force [65], Hot SAX [53], and

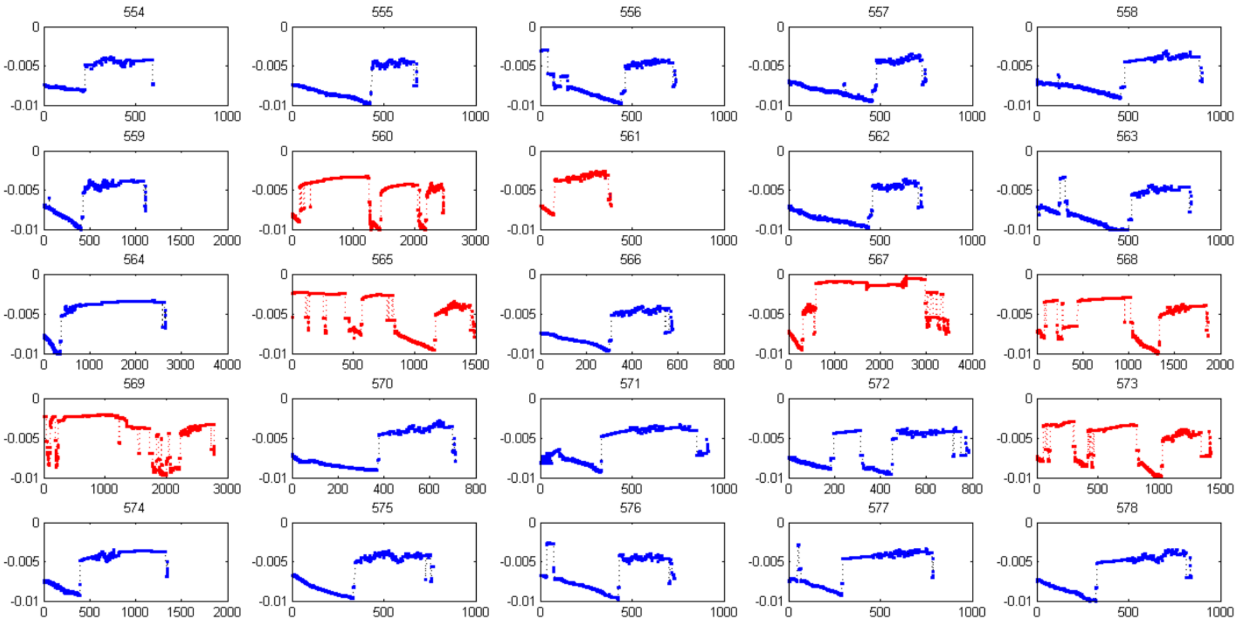


FIGURE 11. The first component vector of MDS1 via SVD includes 30 subsequences, which correspond to ring no. 554-578. Seven abnormal subsequences were detected as abnormal by MFAD and marked in red.

TABLE 4. The performances of MFAD+PCA and MFAD+SVD on five multivariate series.

MFAD+PCA	Dimension	Accuracy	Sensitivity	Specificity
Chfdb/chf01	2	1	1	1
Mitdbx_108_2	2	1	1	1
Air Quality (UCI)	5	0.9167	1	0.75
MDS1	6	0.760	0.739	1
MDS2	9	0.867	0.875	0.80
MFAD+SVD	**Dimension**	**Accuracy**	**Sensitivity**	**Specificity**
Chfdb/chf01	2	1	1	1
Mitdbx_108_2	2	1	1	1
Air Quality (UCI)	5	1	1	1
MDS1	6	0.80	0.783	1
MDS2	9	0.867	0.850	1

k-means based clustering [66] with sliding window 100 and 200. The performance of these approaches were calculated as the average of two different sliding window. Due to the fact that MFAD works on a set of sub-sequences segmented from the original series, we thus used the same coordinates to count the discords within the range of each sub-sequence. We consider a sub-sequence as an anomaly if it includes a discord identify by one of above three methods.

For the univariate series Nprs44, all of three methods can detect the most obvious discord but lost another one. **Table 5** shows that the performance of these methods on Nprs44 were close to MFAD. As to the series TS_C01, Brute Force and SAX shows quite good accuracy, but are hard to identify the discords. For the Air Quality dataset, clustering method works better than Brute Force and SAX. And the results on MDS2 indicates that Brute Force and SAX are much better to identify normal states rather than clustering. Observing the local dynamics of these four time series, we found that the anomalies were difficult to be identified from **shock signals**, such as TS_C01 and MDS2. For the non-shock signals (e.g. peaks on steady signals), it is easier to detect discords with clustering strategy. Based on the results shown in **Table 5**, we concluded that MFAD represented a significant performance rather than other three approaches.

TABLE 5. The comparison of NFAD with other three algorithms. The performance of brute force, SAX, and clustering were expressed as the average of two different sliding window: n = 100 and n = 200.

Approaches	Datasets	Ground Truth	Accuracy	Sensitivity	Specificity
MFAD	Nprs44	2	1	1	1
Brute Force	Nprs44	2	0.889	1	0.5
SAX	Nprs44	2	0.889	1	0.5
Clustering	Nprs44	2	0.889	1	0.5
MFAD	TS_C01	1	0.90	0.897	1
Brute Force	TS_C01	1	0.933	0.966	0
SAX	TS_C01	1	0.767	0.793	0.5
Clustering	TS_C01	1	0.483	0.5	0
MFAD	Air Quality (PCA)	4	0.9167	1	0.75
Brute Force	Air Quality (PCA)	4	0.708	1	0.125
SAX	Air Quality (PCA)	4	0.667	1	0
Clustering	Air Quality (PCA)	4	0.708	0.562	1
MFAD	Air Quality (SVD)	4	1	1	1
Brute Force	Air Quality (SVD)	4	0.75	1	0.25
SAX	Air Quality (SVD)	4	0.667	1	0
Clustering	Air Quality (SVD)	4	0.833	0.75	1
MFAD	MDS2 (PCA)	5	0.867	0.875	0.8
Brute Force	MDS2 (PCA)	5	0.867	0.975	0
SAX	MDS2 (PCA)	5	0.844	0.95	0
Clustering	MDS2 (PCA)	5	0	0	0
MFAD	MDS2 (SVD)	5	0.867	0.850	1
Brute Force	MDS2 (SVD)	5	0.867	0.975	0
SAX	MDS2 (SVD)	5	0.867	0.975	0
Clustering	MDS2 (SVD)	5	0.044	0.025	0.2

The details of the comparisons on each dataset were represented in **Fig. S20-S37**.

V. DISCUSSION AND CONCLUSION

This paper presents a novel computational framework for recognition of abnormal states in the complex process of shield tunneling construction using a meta-feature based approach that is called MFAD. In MFAD, we firstly defined six meta-features for describing the dynamics of a time series. Secondly, all the subsequences extracted from the same variable can be represented by our defined meta-features. Thirdly, segmented subsequences with different length can be compared in a transformed feature space with these meta-features. As a result, the abnormal subsequences can be easily identified. Different from most of existing approaches, MFAD does not directly detect the discords on the original time series, but in a simplified data space. Moreover, it is also suitable for online adaptive learning, since all the previous subsequences only need to be calculated one time.

MFAD approach is suitable for both univariate and multivariate time series. For the univariate sequence, it can be directly addressed by MFAD framework. For the multivariate time series, a conversion from multivariate time series to one-dimensional series is performed so that the meta-features can be further extracted and calculated. In this work, we found that MFAD+SVD represents better performance than MFAD+PCA. Except the PCA or SVD mentioned in this study, other strategies (e.g. clustering-based methods) also can be used as dimension converters in our MFAD framework.

We carried out the simulation experiments on a large number of datasets, including several well-known time series datasets and real datasets collected from shield tunneling construction. The results show that the proposed approach is not only capable of identifying the real faults, but also recognizing several abnormal states that are significantly different from most cases. Our developed approach is also suitable to detect the stage-based abnormal patterns from a univariate/multivariate time series.

Moreover, we compared our MFAD with three traditional anomaly detection methods (Brute Force, SAX, K-means based clustering) on these four time series datasets.

The computational results show that the performance of MFAD is outstanding (**Table 5**). Firstly, sliding window based strategies evenly segment an original time series as a set of redundant sub-sequence, which is the reason for the increase of computation cost. However, the sub-sequences addressed by MFAD are non-overlapping, so that the number of segments calculated in MFAD is much fewer. Secondly, many sliding window based algorithms, including Brute Force or Clustering, represent the similarity of a segment pair based on distance measure, which is likely to be unreliable [71]. The contribution of our MFAD is that it avoids the direct calculation of the distance between two sequences to represent their similarity, but uses several representative meta-features to distinguish normal or abnormal sub-sequences. Therefore, MFAD not only simplifies the computational complexity, but also makes any two unequal sub-sequences comparable. Finally, we found that the time series collected from shield construction are much complicated rather than EEG or ECG datasets. Identifying the abnormal states in ring is in line with the actual application requirements. In conclusion, our study provides a novel strategy of comparing the differences between stages within a long series.

Due to the local dynamics, periodicity, or randomness in different time series, the differences between the abnormal subsequences and other normal subsequences from the same series might be reflected on a subset of meta-features. In other word, the feature ranking is varied in different time series dataset. Which meta-feature is the most important one depends on which dataset was tested. Therefore, we combined all the six meta-features to construct one-class SVM for outlier detection in our experiments. Our simulation results show the combination of six meta-features in MFAD works well in most cases.

Limitations exist in the proposed MFAD approach. MFAD recognizes the minority of outliers as anomalies. The number of identified discords is usually more than the detected outliers in the real-world situations. For example, our fault log in the real shield tunneling construction only report a few of anomalies; however, MFAD actually can detect more cases. To improve the accuracy, there were two ways need to be considered in the near future: 1) record the fault events timely and seriously (ground truth); 2) refine the current meta-features to better summarize local dynamics of time series. In the next step, we will plan to release this MFAD framework and further test its effectiveness in the real-world shield tunneling construction.

CONFLICT INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

AUTHOR CONTRIBUTIONS

Conceived and designed the algorithms: M.H., Z.J. Performed the simulations: K.Y., Z.J. Processed and Analyzed the data: X.F., Y.G., J.G. Wrote the paper: Z.J., M.H. Provide ideas to improve the computational approach: Z.J., M.H., X.Z., L.D.

REFERENCES

[1] K. Yan, Z. Ji, and W. Shen, "Online fault detection methods for chillers combining extended Kalman filter and recursive one-class SVM," *Neurocomputing*, vol. 228, pp. 205–212, Mar. 2017.

[2] C. S. Sharma, S. N. Panda, R. P. Pradhan, A. Singh, and A. Kawamura, "Precipitation and temperature changes in eastern India by multiple trend detection methods," *Atmos. Res.*, vol. 180, pp. 211–225, Nov. 2016.

[3] A. K. Chanda, C. F. Ahmed, M. Samiullah, and C. K. Leung, "A new framework for mining weighted periodic patterns in time series databases," *Expert Syst. Appl. Int. J.*, vol. 79, pp. 207–224, Aug. 2017.

[4] Z. Ji, B. Wang, S. P. Deng, and Z. You, "Predicting dynamic deformation of retaining structure by LSSVR-based time series method," *Neurocomputing*, vol. 137, pp. 165–172, Aug. 2014.

[5] J. Li, W. Pedrycz, and I. Jamal, "Multivariate time series anomaly detection: A framework of hidden Markov models," *Appl. Soft Comput.*, vol. 60, pp. 229–240, Nov. 2017.

[6] L. I. Dong, S. Liu, and H. Zhang, "A method of anomaly detection and fault diagnosis with online adaptive learning under small training samples," *Pattern Recognit.*, vol. 64, pp. 374–385, Apr. 2017.

[7] R. B. Penfold and F. Zhang, "Use of interrupted time series analysis in evaluating health care quality improvements," *Acad. Pediatrics*, vol. 13, no. 6, pp. 38–44, 2013.

[8] A. Rahmani *et al.*, "Graph-based approach for outlier detection in sequential data and its application on stock market and weather data," *Knowl.-Based Syst.*, vol. 61, pp. 89–97, May 2014.

[9] M. Podsiadlo and H. Rybinski, "Financial time series forecasting using rough sets with time-weighted rule voting," *Expert Syst. Appl.*, vol. 66, pp. 219–233, Dec. 2016.

[10] J. C. M. Oliveira, K. V. Pontes, I. Sartori, and M. Embiruçu, "Fault detection and diagnosis in dynamic systems using weightless neural networks," *Expert Syst. Appl.*, vol. 84, pp. 200–219, Oct. 2017.

[11] Y. Zhao, H. Pan, H. Wang, and H. Yu, "Dynamics research on grouping characteristics of a shield tunneling machine's thrust system," *Automat. Construct.*, vol. 76, pp. 97–107, Apr. 2017.

[12] G. Zheng *et al.*, "Study of the collapse mechanism of shield tunnels due to the failure of segments in sandy ground," *Eng. Failure Anal.*, vol. 79, pp. 464–490, Sep. 2017.

[13] M. Bayati and J. K. Hamidi, "A case study on TBM tunnelling in fault zones and lessons learned from ground improvement," *Tunnelling Underground Space Technol.*, vol. 63, pp. 162–170, May 2017.

[14] S. Li *et al.*, "An overview of ahead geological prospecting in tunneling," *Tunnelling Underground Space Technol.*, vol. 63, pp. 69–94, Mar. 2017.

[15] E. Garoudja, F. Harrou, Y. Sun, K. Kara, A. Chouder, and S. Silvestre, "Statistical fault detection in photovoltaic systems," *Solar Energy*, vol. 15, pp. 485–499, Jul. 2017.

[16] M. Madakyaru, F. Harrou, and Y. Sun, "Improved data-based fault detection strategy and application to distillation columns," *Process Saf. Environ. Protection*, vol. 107, pp. 22–34, Apr. 2017.

[17] H. Mekki, A. Mellit, and H. Salhi, "Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules," *Simul. Model. Pract. Theory*, vol. 67, pp. 1–13, Sep. 2016.

[18] Z. Ji, Q. Xia, and G. Meng, "A review of parameter learning methods in Bayesian network," in *Proc. Int. Conf. Intell. Comput. (ICIC)*, 2015, pp. 3–12.

[19] M. Gan, C. L. P. Chen, H. X. Li, and L. Chen, "Gradient radial basis function based varying-coefficient autoregressive model for nonlinear and nonstationary time series," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 809–812, Jul. 2015.

[20] J. F. Barragan, C. H. Fontes, and M. Embiruçu, "A wavelet-based clustering of multivariate time series using a multiscale SPCA approach," *Comput. Ind. Eng.*, vol. 95, pp. 144–155, May 2016.

[21] C. K. Lau, K. Ghosh, M. A. Hussain, and C. R. C. Hassan, "Fault diagnosis of Tennessee Eastman process with multi-scale PCA and ANFIS," *Chemometrics Intell. Lab. Syst.*, vol. 120, pp. 1–14, Jan. 2013.

[22] J. Zarei, "Induction motors bearing fault detection using pattern recognition techniques," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 68–73, 2012.

[23] S. Kanarachos, S.-R. G. Christopoulos, A. Chroneos, and M. E. Fitzpatrick, "Detecting anomalies in time series data via a deep learning algorithm combining wavelets, neural networks and Hilbert transform," *Expert Syst. Appl.*, vol. 85, pp. 292–304, Nov. 2017.

[24] G. C. Silva, R. M. Palhares, and W. M. Caminhas, "Immune inspired fault detection and diagnosis: A fuzzy-based approach of the negative selection algorithm and participatory clustering," *Expert Syst. Appl.*, vol. 36, no. 16, pp. 12474–12486, 2012.

[25] Z. Geng and Q. Zhu, "Rough set-based heuristic hybrid recognizer and its application in fault diagnosis," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2711–2718, 2009.

[26] S. Mascaro, A. E. Nicholso, and K. B. Korb, "Anomaly detection in vessel tracks using Bayesian networks," *Int. J. Approx. Reason.*, vol. 55, no. 1, pp. 84–98, 2014.

[27] B. Cai, H. Liu, and M. Xie, "A real-time fault diagnosis methodology of complex systems using object-oriented Bayesian networks," *Mech. Syst. Signal Process.*, vol. 80, pp. 31–44, Dec. 2016.

[28] G. Heydari, M. A. Vali, and A. A. Gharaveisi, "Chaotic time series prediction via artificial neural square fuzzy inference system," *Expert Syst. Appl.*, vol. 55, pp. 461–468, Aug. 2016.

[29] K. Bisht and S. Kumar, "Fuzzy time series forecasting method based on hesitant fuzzy sets," *Expert Syst. Appl.*, vol. 64, pp. 557–568, Dec. 2016.

[30] Y. Zhou, H. Luo, and Y. Yang, "Implementation of augmented reality for segment displacement inspection during tunneling construction," *Automat. Construct.*, vol. 82, pp. 112–121, Oct. 2017.

[31] T. Zhang, D. Yue, Y. Gu, Y. Wang, and G. Yu, "Adaptive correlation analysis in stream time series with sliding windows," *Comput., Math. Appl.*, vol. 57, no. 6, pp. 937–948, 2008.

[32] Z. Ding and M. Fei, "An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window," *IFAC Proc.*, vol. 46, no. 20, pp. 12–17, 2013.

[33] H. Ren, Z. Ye, and Z. Li, "Anomaly detection based on a dynamic Markov model," *Inf. Sci.*, vol. 411, pp. 52–65, Oct. 2017.

[34] J. S. Chou and N. T. Ngo, "Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns," *Appl. Energy*, vol. 177, pp. 751–770, Sep. 2016.

[35] H.-N. Wu, S.-L. Shen, S.-M. Liao, and Z.-Y. Yin, "Longitudinal structural modelling of shield tunnels considering shearing dislocation between segmental rings," *Tunnelling Underground Space Technol.*, vol. 50, pp. 317–323, Aug. 2015.

[36] M. K. Cain, Z. Zhang, and K. H. Yuan, "Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation," *Behav. Res. Methods*, vol. 49, no. 5, pp. 1716–1735, 2017.

[37] F. I. Donoso, R. L. Figueroa, E. A. Lecannelier, E. J. Pino, and A. J. Rojas, "Atrial activity selection for atrial fibrillation ECG recordings," *Comput. Biol. Med.*, vol. 43, no. 10, pp. 1628–1636, 2013.

[38] G. R. Iannotti, F. Pittau, C. M. Michel, S. Vulliemoz, and F. Grouiller, "Pulse artifact detection in simultaneous EEG–fMRI recording based on EEG map topography," *Brain Topogr.*, vol. 28, no. 1, pp. 21–32, 2015.

[39] R. Knvps and R. Dhuli, "Classification of ECG heartbeats using nonlinear decomposition methods and support vector machine," *Comput. Biol. Med.*, vol. 87, pp. 271–284, Aug. 2017.

[40] X. Li, J. Wang, B. Huang, and S. Lu, "The DCT-based oscillation detection method for a single time series," *J. Process Control*, vol. 20, no. 5, pp. 609–617, 2010.

[41] M. Kiyama, M. Yamada, and M. Tatsumi, "Quantitative analysis of low-frequency current oscillation in semi-insulating GaAs," *Eur. Phys. J.-Appl. Phys.*, vol. 27, nos. 1–3, pp. 185–188, 2004.

[42] R. Yan, T. M. Antonsen, and G. S. Nusinovich, "Nonlinear analysis of low-frequency oscillations in gyrotrons," *IEEE Trans. Plasma Sci.*, vol. 38, no. 6, pp. 1178–1184, Jun. 2010.

[43] L. Jäncke, J. Kühnis, L. Rogenmoser, and S. Elmer, "Time course of EEG oscillations during repeated listening of a well-known aria," *Frontiers Hum. Neurosci.*, vol. 9, p. 401, Jul. 2015.

[44] J. Wang, B. Huang, and S. Lu, "Improved DCT-based method for online detection of oscillations in univariate time series," *Control Eng. Pract.*, vol. 21, no. 5, pp. 622–630, 2013.

[45] R. K. Tripathy, S. Deb, and S. Dandapat, "Analysis of physiological signals using state space correlation entropy," *Healthcare Technol. Lett.*, vol. 4, no. 1, pp. 30–33, 2017.

[46] P. Marwaha and R. K. Sunkaria, "Complexity quantification of cardiac variability time series using improved sample entropy (I-SampEn)," *Australas. Phys. Eng. Sci. Med.*, vol. 39, no. 3, pp. 755–763, 2016.

[47] O. Purcell, B. M. Di, C. S. Grierson, and N. J. Savery, "A multi-functional synthetic gene network: A frequency multiplier, oscillator and switch," *PLoS ONE*, vol. 6, no. 2, p. e16140, 2011.

[48] W. W. Melek, Z. Lu, A. Kapps, and W. D. Fraser, "Comparison of trend detection algorithms in the analysis of physiological time-series data," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 4, pp. 639–651, Apr. 2005.

[49] X. He *et al.*, "Multi-task learning with one-class SVM," *Neurocomputing*, vol. 133, pp. 416–426, Jun. 2014.

[50] Y. Xiao, H. Wang, W. Xu, and J. Zhou, "Robust one-class SVM for fault detection," *Chemometrics Intell. Lab. Syst.*, vol. 151, pp. 15–25, Feb. 2016.

[51] G. Li, O. Bräysy, L. Jiang, Z. Wu, and Y. Wang, "Finding time series discord based on bit representation clustering," *Knowl.-Based Syst.*, vol. 54, pp. 243–254, Dec. 2013.

[52] W. Luo and M. Gallagher, "Faster and parameter-free discord search in quasi-periodic time series," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining (PAKDD) Lecture Notes Comput. Sci.*, vol. 6635, 2011, pp. 135–148.

[53] E. Keogh, J. Lin, and A. Fu, "HOT SAX: Efficiently finding the most unusual time series subsequence," in *Proc. 5th IEEE Int. Conf. Data Mining*, Nov. 2005, p. 8.

[54] A. M. Vanage, R. H. Khade, and D. B. Shinde, "Classifying five different arrhythmias by analyzing the ECG signals," *Int. J. Comput. Eng. Manage.*, vol. 15, no. 4, pp. 75–80, 2012.

[55] *UCI Machine Learning Repository*. Accessed: 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets.html

[56] Z. Ji, G. Meng, D. Huang, X. Yue, and B. Wang, "NMFBFS: A NMF-based feature selection method in identifying pivotal clinical symptoms of hepatocellular carcinoma," *Comput. Math. Methods Med.*, vol. 2015, Jul. 2015, Art. no. 846942.

[57] Z. Ji and B. Wang, "Identifying potential clinical syndromes of hepatocellular carcinoma using PSO-based hierarchical feature selection algorithm," *Biomed. Res. Int.*, vol. 2014, Mar. 2014, Art. no. 127572.

[58] X. Li and X. Chen, "Using grouting of shield tunneling to reduce settlements of overlying tunnels: Case study in Shenzhen metro construction," *J. Eng. Manage.*, vol. 138, no. 4, pp. 574–584, 2012.

[59] D. G. Aggelis, T. Shiotani, and K. Kasai, "Evaluation of grouting in tunnel lining using impact-echo," *Tunnelling Underground Space Technol.*, vol. 23, no. 6, pp. 629–637, 2008.

[60] K. Deng, X. Tang, L. Wang, and X. Chen, "Force transmission characteristics for the non-equidistant arrangement thrust systems of shield tunneling machines," *Automat. Construct.*, vol. 20, no. 5, pp. 588–595, 2011.

[61] K. Deng, J. Huang, and H. Wang, "Layout optimization of non-equidistant arrangement for thrust systems in shield machines," *Automat. Construct.*, vol. 49, pp. 135–141, Jan. 2015.

[62] *Time Series Datasets Tested by Hot SAX*. Accessed: 2005. [Online]. Available: http://www.cs.ucr.edu/~eamonn/discords/ICDM05_discords.pdf

[63] R. Zarei, J. He, S. Siuly, and Y. Zhang, "A PCA aided cross-covariance scheme for discriminative feature extraction from EEG signals," *Comput. Methods Programs Biomed.*, vol. 146, pp. 47–57, Jul. 2017.

[64] Z. Liu, J. A. de Zwart, P. van Gelderen, L.-W. Kuo, and J. H. Duyn, "Statistical feature extraction for artifact removal from concurrent fMRI-EEG recordings," *NeuroImage*, vol. 59, no. 3, pp. 2073–2087, 2012.

[65] D. Zheng, F. Li, and T. Zhao, "Self-adaptive statistical process control for anomaly detection in time series," *Expert Syst. Appl.*, vol. 57, pp. 324–336, Sep. 2016.

[66] C. T. Yiakopoulos, K. C. Gryllias, and I. A. Antoniadis, "Rolling element bearing fault detection in industrial environments based on a K-means clustering approach," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2888–2911, 2011.

[67] P. S. Brandley and U. M. Fayyad, "Refining initial points for K-means clustering," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 91–99.

[68] X. Xu, Y. Ji, and G. Stormo, "Discovering cis-regulatory RNAs in Shewanella genomes by support vector machines," *PLoS Comput. Biol.*, vol. 5, no. 4, p. e1000338, 2009.

[69] H. Yang, G. Gong, and L. Wang, "Determination of the cutterhead torque for EPB shield tunneling machine," *Automat. Construct.*, vol. 20, no. 8, pp. 1087–1095, 2011.

[70] J. Liao, Z. Chen, and B. Yao, "High-performance adaptive robust control with balanced torque allocation for the over-actuated cutter-head driving system in tunnel boring machine," *Mechatronics*, vol. 46, pp. 168–176, Oct. 2017.

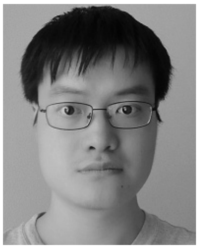
[71] E. Keogh, J. Lin, and W. Truppel, "Clustering of time-series subsequences is meaningless: Implications for previous and future research," *Knowl. Inf. Syst.*, vol. 8, no. 2, pp. 154–177, 2005.

[72] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

[73] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a highdimensional distribution," *Neural Comput.*, vol. 13, pp. 1443–1472, 2001.

MIN HU received the Ph.D. degree in control theory and control engineering from Shanghai University, China, in 2004. She is currently an Associate Professor of information management with the SHU-UTS SILC Business School, Shanghai University, where she is also the Dean of the Department of Information Management and Information System. Her research interests are building information management and artificial intelligence.

ZHIWEI JI received the bachelor's degree from Zhejiang A&F University (ZAFU), Linan, China, in 2003, the master's degree from Shanghai University, Shanghai, China, in 2009, and the Ph.D. degree from the School of Electronics and Information Engineering, Tongji University, Shanghai, in 2016. Between 2016 and 2017, he was a Post-Doctoral Research Fellow with Wake Forest University, USA. Since 2003, he has been with the School of Information Engineering, ZAFU. He is currently an Assistant Professor with Zhejiang Gongshang University, Hangzhou, China. His research interests mainly focus on pattern recognition, machine learning, computational biology, and bioinformatics.



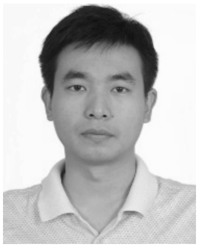
KE YAN received the bachelor's degree and the Ph.D. degree in computer science from the National University of Singapore in 2007 and 2012, respectively, under the supervision of Dr. H.-L. Cheng. Between 2013 and 2014, he was a Post-Doctoral Researcher with the Masdar Institute of Science and Technology, Abu Dhabi, UAE. He is currently a Lecturer with China Jiliang University, Hangzhou, China. His main research field includes computer graphics, computational geometry, data mining, and machine learning.

YE GUO received the B.E. degree in industrial engineering from the School of Management Science and Engineering, Nanjing University of Finances and Economics, China, in 2016. She is currently pursuing the M.B.A. degree in management science and engineering with the SHU-UTS SILC Business School, Shanghai University, China. Her current research interest is information management of the intelligent city, especially anomaly detection.

XIAOWEI FENG received the B.B.M. degree in information management and information system from the SHU-UTS SILC Business School, Shanghai University, China, in 2016, where she is currently pursuing the M.B.A. degree in management science and engineering. Her current research interests are anomaly detection and information construction of urban public facilities.

JIAHENG GONG received the bachelor's degree from the Dalian University of Technology, Dalian, China, in 2017. He is currently pursuing the master's degree with Zhejiang Gongshang University, Hangzhou, China. His current research interests are time series analysis and big data mining.

XIN ZHAO received the bachelor's degree and the M.D. degree in clinical medicine from Capital Medical University in 2004 and 2012, respectively, under the supervision of Dr. D.-Z. Chen. Since 2015, he has been an Associate Chief Doctor with the Hepatic-Biliary Department, Beijing Chaoyang Hospital Affiliated to Capital Medical University. His main research fields include clinical medicine and medical big data analysis.



LIGANG DONG received the bachelor's and master's degrees from Zhejiang University (Mixed Class), China, in 1995 and 1998, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, in 2003. He is currently a Professor with Zhejiang Gongshang University, Hangzhou, China, where he is also the Dean of the School of Information and Electronic Engineering. His research interests include various topics in smart networks and educations.