Received November 11, 2019, accepted November 23, 2019, date of publication November 28, 2019,
date of current version December 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2956494

Outlier Detection for Monitoring Data Using Stacked Autoencoder

FANGYI WAN[1], GAODENG GUO[1], CHUNLIN ZHANG[1], QING GUO[1], AND JIE LIU[1,2], (Senior Member, IEEE)

[1]School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China
[2]Department of Mechanical and Aerospace Engineering, Carleton University, Ottawa, ON K1S5B6, Canada

Corresponding author: Chunlin Zhang (zchunlin@nwpu.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 10402035 and Grant 11702216, in part by the Fundamental Research Funds for the Central Universities, and in part by the China Postdoctoral Science Foundation under Grant 2018M631196.

ABSTRACT Monitoring data contain the important status information of the monitored object, and are the basis for following data mining and analysis. However, the monitoring data usually suffer the pollution of the outliers, leading to negative effect on the subsequent data processing. To address the problem, this paper proposed an outlier detection method based on stacked autoencoder (SAE). SAE has a powerful capability of feature extraction and greatly preserves the original information of the data. The trained SAE by normal data can learn the characteristics of normal data. When a set of data with outliers are inputted to the trained network, there are larger reconstruction errors at the outliers between the original input data and the reconstructed data obtained by using the encoding parameters and the decoding parameter mapping, which provides a basis for locating outliers. Meanwhile, this paper introduced the Grubbs criterion and the PauTa criterion to identify the reconstruction errors corresponding to the outliers based on the traditional threshold method. The method can quickly isolate the abnormal data from the normal data according to the reconstruction error and the identification criterion. The effectiveness and superiority of the proposed method have been validated by experiment on real data and comparisons with traditional outlier detection algorithms.

INDEX TERMS Condition monitoring, outlier detection, stacked autoencoder, monitoring data.

I. INTRODUCTION

With the continuous advancement of information technology and the advent of the era of big data, all kinds of information are presented in the form of monitoring data, which are characterized by large scale, diversity, complicated information and sparse value. At the same time, these data contain important information, which is the basis for analyzing the data and extracting value of the monitored object [1], [2]. However, the original monitoring data probably are contaminated by outliers due to the influences of environmental interference (noise, shock and vibration), communication obstacles, sensor fault and other factors [3]–[5]. On the one hand, the data with outliers contain valuable information. On the other hand, they cannot reflect the real situation completely. If these data are directly used to the scenarios of modeling, state analysis, fault diagnosis and prediction, which will affect the accuracy of the model, increase the probability of false-alarm and false negatives and make wrong decisions [6]–[8]. Therefore, it is necessary to adopt an efficient method to detect the outliers of the monitoring data, thereby ensuring the quality of the data and providing guarantee for the reliable process of the subsequent analysis.

Outlier detection is a crucial step that must take precedence over data analysis, and it seeks to separate abnormal data from normal data in the dataset. Outliers usually exist in the data in an isolated or continuous manner. In different fields, the characteristics and definitions of outliers are not the same, and there are detection methods suitable for their own fields. Currently, the common outlier detection methods can be roughly divided into five categories [9]–[12]: statistics-based, distance-based, density-based, clustering-based approaches and computational intelligence algorithms.

The associate editor coordinating the review of this manuscript and approving it for publication was Gongbo Zhou.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

1) The statistical-based method is the most traditional method in outlier detection. The foothold of this methods is that outlier is usually an observation that significantly deviates from other observations, so the outliers can be identified by the statistical models of the data [13]. Although statistical-based methods are more difficult to adapt as increase of data's complexity and dimension, they are still the best algorithms for certain problems in reality, and related research has continued. For instance, Ahsan *et al.* [14] proposes PCA Mix based control chart for outlier detection of mixed continuous and categorical data. Hu *et al.* [15] proposes a meta-feature-based anomaly detection approach (MFAD) to identify the abnormal states of a univariate or multivariate time series based on local dynamics. Huan *et al.* [9] uses model selection-based support vector data description (SVDD) to detect outlier in Wireless Sensor Networks.
2) The distance-based method mainly quantifies the degree of deviation of outliers by distance, such as Euclidean distance, Mahalanobis distance and Maximal Data Piling distance [16]. Among these methods, the K-Nearest Neighbor (KNN) is the most common method for distinguishing data by finding the nearest K neighbors from the data to be identified. The traditional KNN is often difficult to adapt to the situations that there are uneven distribution or abnormal clusters in the dataset. For this reason, many researchers have improved KNN and achieved certain results [17]–[20].
3) The density-based approach is based on the idea of neighborhood as well as the distance-based approach. In most of the density-based approaches, they assume that the density around a normal data object is similar to the density around its neighbors, whereas the density is considerably low than that of its neighbors in case of an outlier. Among them, local outlier factor (LOF), connectivity-based outlier factor (COF) and influenced outlier (INFLO) are examples of some well-known density-based approaches for outlier detection [21]. The isolated forest algorithm is a recently developed outlier detection method based on density, it is widely used in industry because of linear time complexity, high precision and the ability of dealing with the big data [22]–[24].
4) The cluster-based method identifies the outliers by the distance between the data to be identified and the cluster center. Many different types of clustering methods have emerged in the past several years, including several algorithms based on partitioning, density, fuzziness, grids and hierarchies [25], [26]. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is the most distinguished density-based clustering algorithm among them, and many clustering methods for outlier detection are inspired by it. Some clustering methods accommodate outliers by introducing an additional cluster, which consists of the data far away from all cluster centers. In this way, outliers will be isolated from normal data [27].
5) Most methods based on computational intelligence are inspired by the natural or living nature to imitate and solve problems with their principles and ideas. Some of these methods have also been applied on the field of anomaly data detection, such as artificial neural networks [28], [29], genetic algorithms (GA) [30]–[33], simulated annealing [34], [35], ant colony algorithm [36], artificial bee colony algorithm [37]. The methods based on computational intelligence usually try to find the optimal solution satisfying certain conditions by simulating the biological or natural principles when detecting outliers, and those data that do not satisfy the particular condition will be regarded as outliers.

Although fruitful achievements have been reported by these methods, most of these existing outlier detection approaches still have the following drawbacks. Statistical-based methods are often difficult to deal with the data with a high degree of nonlinearity. It takes time and effort to detect outliers with distance-based and density-based algorithms because it is necessary to compare with a large amount of historical data for the methods, so their real-time is poor. The performance of clustering-based algorithms mainly depends on the selection of parameters, but the optimal parameters are difficult to estimate in high dimensional data. For different problems, methods based on computational intelligence often need to find the optimal solution to the problem according to different conditions. If researchers can't choose the appropriate screening conditions or parameters, this will cause long-term operation of the algorithm, fall into local optimum and premature, etc. Therefore, the use of these methods based on computational intelligence requires the researchers to have good professional knowledge.

In recent years, deep learning has received extensive attention due to its strong ability of exploiting the high dimensional and large-scale data. Based on the significant advantages of deep learning algorithms and combined with the characteristics of outlier detection, this paper proposes an outlier detection algorithm based on stacked autoencoder (SAE) for monitoring data. SAE only needs to use normal data to train the model, which avoids the facts that there are fewer outliers in real cases and it is difficult to extract the features of outliers. At the same time, the SAE has the ability to handle non-linear, high dimensional and large-scale data, and also adapts to online detection. When a set of data with outliers are input to the trained SAE model which has automatically learnt the characteristics of normal data, larger reconstruction errors will occur on outliers. The outliers in monitoring data can be quickly detected based on the reconstruction errors.

II. THE PROPOSED METHOD

A. BASIC CONCEPT OF SAE ALGORITHM

A stacked autoencoder model is usually constructed by stacking several autoencoders that are the most typical feed-forward neural networks [38]. The autoencoder consists of encoder and decoder, the structure is shown in Fig. 1.

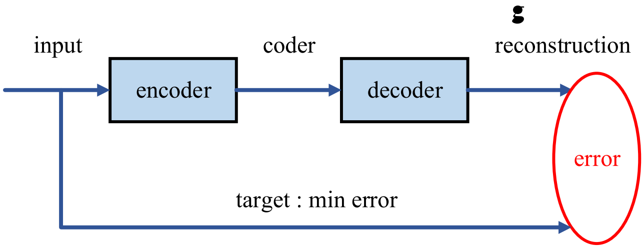


FIGURE 1. The structure of autoencoder.

The encoder maps input data into hidden layer with (1)

$$y = s(Wx + b) \tag{1}$$

where x is input vector, W is weight matrix connecting the input layer to hidden layer, b is bias vector belonging to the nodes of hidden layer, s represents the sigmoid activate function.

The decoder maps y into reconstruction vector z according to

$$z = s(W'y + b') \tag{2}$$

where W' is weights connecting the hidden layer to output layer, b' is the bias vector belonging to nodes of output layer.

The reliability of the auto-encoder is estimated by its reconstruction capability. To recover the input data from the output layer as far as possible, the optimization of the model parameters is to minimize the reconstruction error [39], [40]:

$$L(X, Z) = \begin{cases} H(B(x) \ |B(z)) & x \in \{0,\ 1\} \\ \|x - z\|^2 & x \in R \end{cases} \tag{3}$$

where X is a set of input vectors x, Z is the corresponding set of reconstructed vector z, H represents the Bernoulli cross entropy, $B(x)$ and $B(z)$ are the mean values of x and z.

SAE can be composed by stacking the encoder parts of AE in each layer in the form of putting the hidden coders of the upper layer as the input of the next layer, as presented as Fig. 2.

SAE first performs pre-training layer-by-layer, which is unsupervised, so the samples do not require labels in this process. After pre-training, the network only needs to use a small number of labeled samples for fine-tuning to get better performance. This means can effectively avoid that the SAE model falls into the local optimum [41]. The essence of SAE is to encode the input layer by layer without losing key information, so the decoder can be reconstructed back into the input with a sufficiently small error [42].

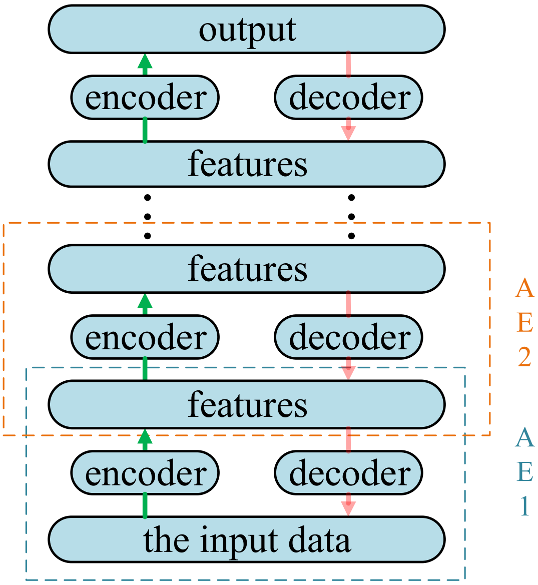


FIGURE 2. The structure of SAE.

B. OUTLIER DETECTION BASED ON THE SAE

When the SAE algorithm is used for outlier detection, normal data that are not contaminated by abnormal data constitute the training set for training the SAE model. The training set contain the basic characteristics of normal data. The SAE model extracts and learns the distribution characteristics of normal data through its deep structure.

The resulting SAE model has the ability to extract features of normal data by tuning the parameters of the model, and fully preserves the key information of the input data and maintains an optimal reconstruction error. The training algorithm is summarized as shown in Algorithm 1.

When constructing a SAE model for anomaly data detection, the key parameters include the number of network layers, the number of neurons in the hidden layer, the learning rate, iterations, and the batch size. These parameters are super-parameters. There is currently no a great way to explain how to set them up.

For the number of network layers, the parameter is related to the dimension of input data. When the data dimension is large, more layers could be tried. At the same time, a model with three-layer can often achieve a good detection effect for most data. The number of neurons in the hidden layer is also related to the dimension of the input data. The number of neurons in each layer can be set according to the strategy of step-down, realizing layer-by-layer compression and feature extraction of data. Meanwhile, the compression of neurons in each layer should not be too large, and excessive compression may result in more information loss.

Learning rate is a crucial parameter in the SAE model with values between 0 and 1. A too large learning rata may produce loss explosion and shock; if the learning rate is too small, the model will converge too slowly or over-fitting. As setting the learning rate, experiment can try 1, 0.1, 0.01,

Algorithm 1 Training SAE model

```
Input: the train dataset X={x};
Output: the trained SAE model
Step 1: Set training parameters: the number of network
        layers L, the numbers of every layers' neuron,
        pretraining iteration Pi, fine-tuning iteration Fi,
        the learning rata $\alpha$, momentum m, batch size,
        activation function.
Step 2: Pretraining
        for i = 1 to Pi do
          for j = 1 to L do
            forward propagation to compute Z
            $y_j$ is mapped to (j+1)$th$ SAE
          end for
        end for
Step 3: Fine-tuning
        for i = 1 to Fi do
          Compute the reconstruction error of output of
          layer L
          for j = L-1 to 1 do
            Compute the reconstruction error $\delta_j$
          end for
          for j = 1 to L do
            Update model parameter
            $\nabla W_j = \delta_j (f(x))^T$ , $\nabla b_j = \delta_j$
            $W_j = W_j - \alpha \nabla W_j - mW_i$, $b_j = b_j - \alpha \nabla b_j$
          end for
        end for
```

0.001, 0.0001 as the test value one by one, and then continue to approach the optimal learning rate by combining the dichotomy strategy.

Iterations can be determined by directly observing the loss of each iteration. When the minimum loss is reached in a certain iteration or the loss is substantially constant after certain iteration, the iterations can be set to the optimal iterations. Increasing the number of iteration with the same other parameters tends to improve the accuracy of the model, but the model should prevent over-fitting and considers of training costs in the iteration.

Batch size has relatively less impact on the model, and it is a selected subset of all data in each step of the training. When setting up data batches, it need only to ensure that the subset of data under each batch can reflect the characteristics of most data. It should be avoided that the subsets can only reflect very local data features compared with the entire dataset.

At this time, when a group of data with random outliers are input into the trained model, there is a bigger reconstruction error will present between input data and the reconstructed data obtained by the SAE parameters. In the paper, the reconstruction error is defined as follow:

$$error = (x - z)^2 \tag{4}$$

where x is input data, z represents corresponding reconstructed vector, x and z have same dimension.

The size of the reconstruction error is the basic criterion for evaluating whether the data point is abnormal. In order to obtain the exact location of the outlier more accurately, the reconstruction error will be further analyzed here. Firstly, the upper threshold T of the reconstruction error is determined. The data points will be classified as normal values if the corresponding reconstruction errors are less than the upper threshold. For suspicious data points that the reconstruction errors are greater than or equal to the threshold, further analysis is processed with Grubbs or variant of the PauTa criterion (3σ criterion).

When the dimension of the input vector is less than or equal to 100, the suspect data points will be determined with the Grubbs criterion. First, the statistic g_n is calculated as:

$$G_n = \frac{e_n - \bar{e}}{s} \tag{5}$$

where e_n is reconstruction error, $\bar{e}$ is the mean of the reconstruction error, s is the standard deviation of the reconstruction error, and the dimension of G_n is consistent with the input vector. Then the significance level α need to be determined, and the corresponding critical value is found. If one suspect data point satisfies with:

$$G_n > G_{1-a}(n) \tag{6}$$

the point is an abnormal value, otherwise it is a normal value.

When the dimension of the input vector is greater than 100, there is variant of the PauTa criterion for judging the suspect data points. The standard form of the PauTa criterion that is used to recognize coarse error is:

$$P_n > 3\sigma \tag{7}$$

where $P_n = abs(e_n - \bar{e})$, σ is standard deviation of the reconstruction error, and the dimension of P_n is consistent with the input vector. In order to make the PauTa criterion more suitable for the reconstruction error of the SAE model, the PauTa criterion is modified as follows:

$$P_n > 3M\sigma \tag{8}$$

where M is a constant that is greater than 0. if P_n corresponding to a certain reconstruction error satisfies the above formula, the corresponding input data point is recognized as an abnormal value. Otherwise, the data point is normal.

When a group of data with random outliers are input into the trained model, the detection process is summarized as shown in Algorithm 2.

The trained SAE model can quickly identify outlier and is still effective for high-dimensional and nonlinear data. On the one hand, the trained SAE parameters can remove abnormal data from the dataset to ensure reliability in subsequent process of data modeling or analysis with an offline fashion. On the other hand, it can isolate the abnormal data in time and feed back to the corresponding processing mechanism for monitoring data in an online way, which avoids false-alarm and false negative due to abnormal data.

Algorithm 2 The Detection Algorithm With Trained SAE

```
Input: the data x may be contaminated by outliers
Output: the location of outlier in x
Step 1: Encoder process
    $y_1 = f\left(W_1 x + b_1\right)$
    for i = 2 to L do
        $y_i = f\left(W_i y_{i-1} + b_i\right)$
    end for
Step 2: Decoder process
    $z_L = f\left(W'_L y_L + b'_L\right)$
    for i = L-1 to 1 do
        $z_i = f\left(W'_i z_{i+1} + b'_i\right)$
    end
Step 3: Locate the location of the outliers
    Calculate the $error = (x - z)^2$
    if error < T do
        the data point is normal data
    else
        analysis the error with Grubbs or PauTa
    end if
```

TABLE 1. Confusion matrix.

Actual Data Status	Tested Results	
	Normal	Outlier
Normal	TP	FN
Outlier	FP	TN

III. EXPERIMENTS

To test the performance of the proposed method, experiments are designed on real dataset and simulation dataset. In order to clarify the effect of SAE algorithm on outlier detection, the detection results are divided into four categories: true positive (TP), false positive (FP), true negative (TN) and false negative (FN), as listed in Table 1.

The combination of the true attributes of the data and the detection results of the algorithm can be presented in the form of a confusion matrix [43].

With the confusion matrix, the detection effect of the SAE algorithm is measured by accuracy, which is defined as:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

A. THE EXPERIMENT ON REAL DATASET

Four public datasets which are available for free from the UCR public database [44] are selected to test the practicability of the proposed method. The four datasets are ADIAC, Chlorine concentration (Chl), FordA, Mallat, respectively. Since the original dataset does not have abnormal data, we artificially add the abnormal data in the dataset to test the proposed algorithm. The experiments were performed under the environment of MATLAB 2015b with Intel (R) Core (TM)2 Quad CPU, 4G RAM (2.40GHz). Table 2 shows the basic information and crucial parameters for training a SAE model using a dataset.

First, the abnormal data were built by adding an isolated outlier to the data sequence randomly. Half of the test data are normal data; the others are abnormal data. When there is a single outlier in the data, the abnormal data point will generate a large reconstruction error when the reconstruction errors of normal data are smaller after the data are reconstructed, as shown in Fig. 3. After the reconstruction error is obtained, the position of the abnormal data point can be accurately located by (5)-(8).

The detection effects of proposed method are presented in Table 3 over the entire test set. From the table, it can be seen that the data can be recognized well whatever normal data or abnormal data, and false-alarm and false negative were rare.

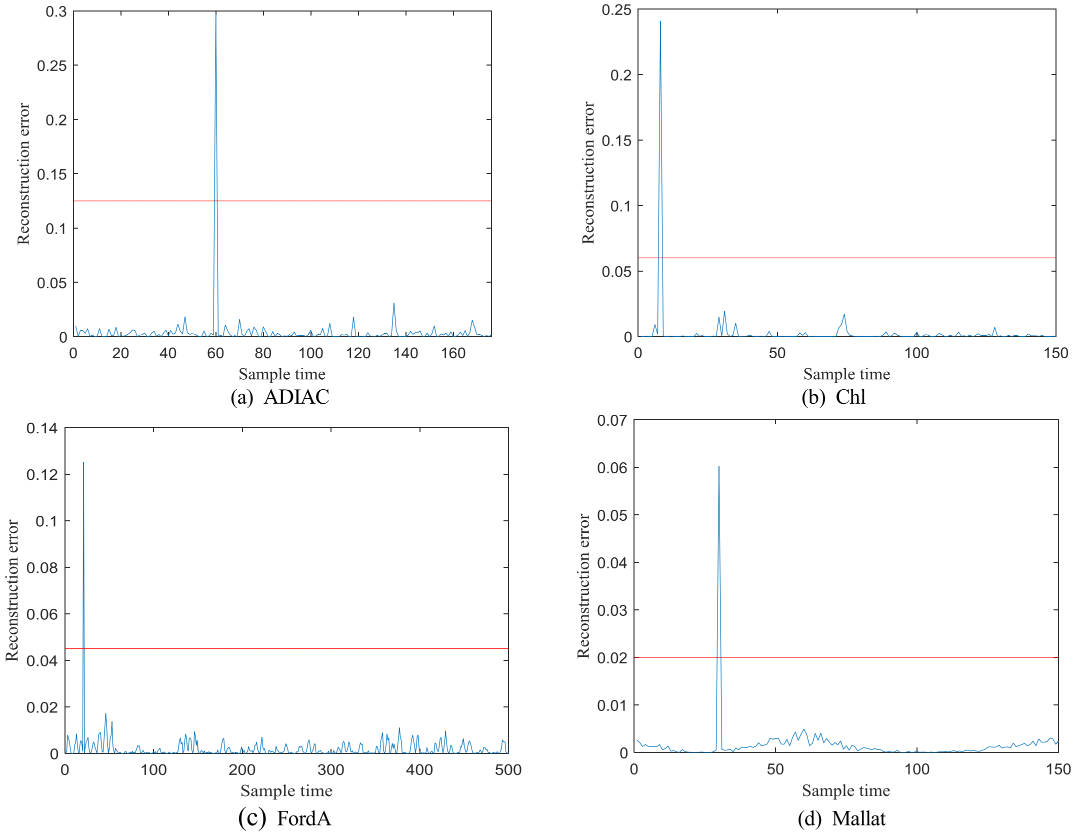
Meanwhile, the experiments were conducted to compare SAE with isolation Forest (iForest), Genetic Algorithm (GA), K-means clustering method (K-means), K-nearest neighbor algorithm (KNN) and principal component analysis (PCA) in detection accuracy. Table 4 reported the detection results of all methods.

In terms of accuracy, the SAE algorithm achieves more than 93% detection accuracy on four datasets, and enjoys the highest average accuracy. The other five algorithms all have the problem of severely degraded detection accuracy on a certain dataset. This shows that the SAE method has wide applicability on detecting abnormal data, and is more adaptable to various scenarios where abnormal data exists.

Next, the trained model was tested with continuous outliers. The experiment built anomalous data by adding 5 consecutive outliers to the original data randomly. The corresponding reconstruction errors were presented in Fig. 4. It can be observed that the continuous outliers will continuously generate large reconstruction errors after reconstruction from Fig. 4, and the outliers in the middle will not be submerged due to the influence of the adjacent outliers. It also shows that the SAE model retains the key information of the data as much as possible, the original basic features are not missing due to the deep extraction of the data characteristics.

TABLE 2. The basic information and crucial parameters of datasets for training SAE.

Datasets	Contents							
	Train size	Test size	Data length	layer	Structure of SAE	Learning rate	Iterations	Batch
ADIAC	500	280	176	2	176-120-50	0.25-0.5	1000	100
Chl	2000	2000	150	3	150-80-40-30	1-1-1	5000	500
FordA	1500	600	500	3	500-250-100-50	0.5-0.25-0.25	10000	500
Mallat	1400	1000	150	3	150-120-50-30	0.25-0.5-0.5	100	200



(a) ADIAC

(b) Chl

(c) FordA

(d) Mallat

FIGURE 3. The reconstruction errors of four datasets when an isolated outlier in the data, where the red line represents the upper threshold and the blue line represents the reconstruction errors.

TABLE 3. The detection results of isolated outlier.

Results	Datasets			
	ADIAC	Chl	FordA	Mallat
TP	139	994	281	494
FP	1	6	19	6
FN	4	28	19	35
TN	136	972	281	465
Accuracy	98.21%	98.30%	93.67%	95.90%

TABLE 4. The detection results of isolated outlier in four datasets.

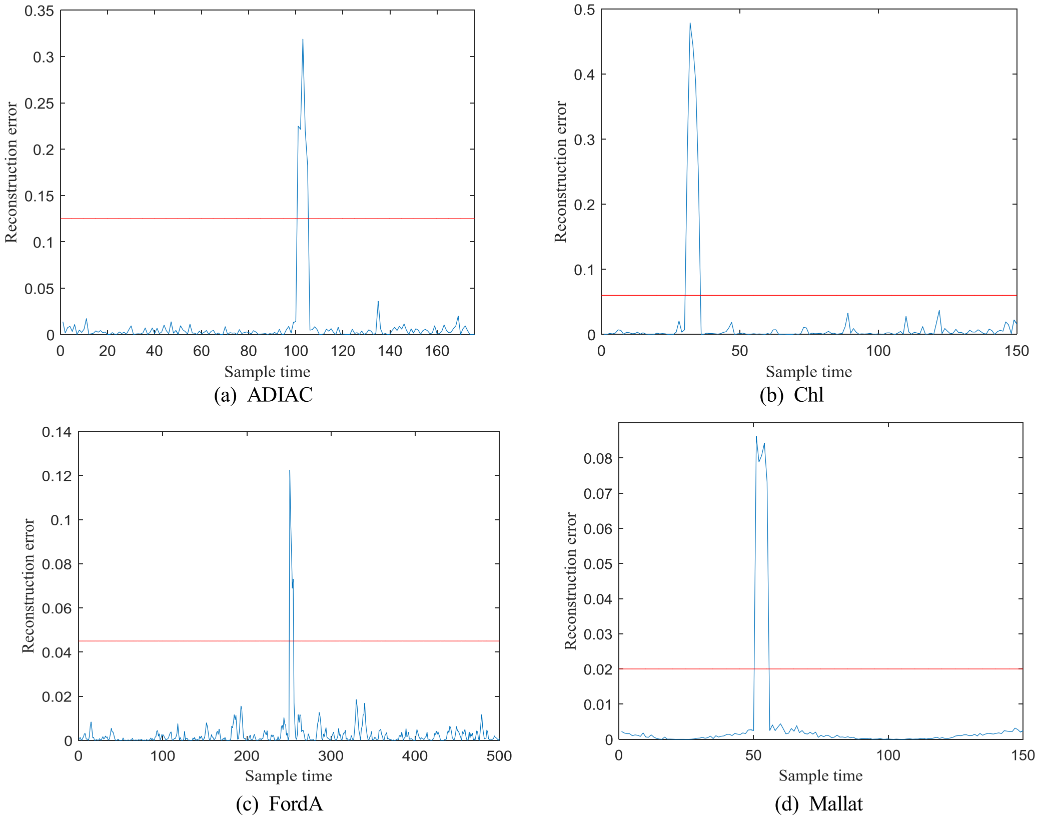
Methods	Datasets				Mean value
	ADIAC	Chl	FordA	Mallat	
SAE	98.21%	98.30%	93.67%	95.90%	96.52%
iForest	89.29%	92.75%	62.00%	94.40%	84.61%
GA	83.57%	86.15%	58.00%	91.20%	79.73%
K-means	97.14%	93.35%	59.67%	94.10%	86.06%
KNN	97.86%	99.45%	75.32%	100%	93.16%
PCA	96.43%	82.60%	96.50%	100%	93.88%

TABLE 5. The detection results of continuous outliers.

Results	Datasets			
	ADIAC	Chl	FordA	Mallat
TP	139	994	281	494
FP	1	6	19	6
FN	19	285	66	110
TN	121	715	234	390
Accuracy	92.86%	85.45%	85.83%	88.40%

As above, the detection results of continuous outliers are summarized in Table 5. In the paper, the result is recognized as correct when all the data points are correctly classified. Even if a data point is misclassified, the detection task will be considered to have failed. As can be seen from the table, the SAE model can still correctly classify normal data, which further demonstrates that the model has the ability of mapping feature well for normal data. For the recognition accuracies of abnormal data, they reduced compared to the detection results of the isolated outlier. But SAE still maintains a high recognition accuracy. It is easy to understand that the probability of misjudgment of SAE model will increase because the impacts of abnormal data on normal data and the interaction between abnormal data gradually increase throughout the feature mapping and data reconstruction.

As shown in Table 6, when there are continuous outliers in the data, the SAE method is still stable and achieves the superior average detection accuracy. It is worth noting that, for isolated outliers and continuous outliers, the detection



(a) ADIAC (b) Chl

(c) FordA (d) Mallat

FIGURE 4. The reconstruction errors of four datasets when continuous outliers in real dataset, where the red line represents the upper threshold and the blue line represents the reconstruction errors.

TABLE 6. The detection results of continuous outlier in four datasets.

Methods	Datasets				Mean value
	ADIAC	Chl	FordA	Mallat	
SAE	92.86%	85.45%	85.83%	88.40%	88.14%
iForest	85.00%	86.40%	56.50%	93.40%	80.30%
GA	81.79%	68.50%	52.83%	90.30%	73.36%
K-means	96.79%	88.90%	53.00%	91.00%	82.42%
KNN	94.27%	95.55%	58.67%	98.80%	86.82%
PCA	90.71%	56.50%	75.00%	97.90%	80.03%

accuracies of the other five algorithms on FordA dataset have experienced different degrees of decline. It is because that most of the outliers on the dataset belong to the local outlier whose value is within the normal range of the entire dataset. At this time, the other five algorithms are often difficult to identify such outliers, and the proposed algorithm is still valid for the situation. This illustrates that SAE has a unique advantage in processing local outliers in time series data.

Throughout the experiment, the SAE model can isolate abnormal data from normal data well and maintain an outstanding detection accuracy. Meanwhile, from the experimental results, SAE has better potential in dealing with local outliers. It also confirms that the proposed method has the ability to deal with real data and has better stability on different datasets.

B. THE EXPERIMENT ON SIMULATION DATASET

This group of experiments are to show the process of identifying the outliers more clearly and test the detection ability. In this set of experiments, the dataset was acquired by ANSYS program, and each piece of data corresponded to the strain of a certain wing structure under normal or certain fault conditions. The experiment collected strain data of 10 stations, each of which contained 60 strain values of a certain station under aerodynamic forces. The training dataset consisted of 972 sets of data, and each set of data included 600 values corresponding to strain values of 10 stations respectively. This data represented the station strain of the wing structure at different heights and speeds. Different SAE models will be trained separately for each station, and then the models will be put together to form a large detection model that can monitor data quality of 10 stations at the same time. In order to further verify the performance of the proposed algorithm, a contrast experiment was carried out between the proposed algorithm and the four other common algorithms of outlier detection. The four algorithms are principal component analysis (PCA), K-nearest neighbor algorithm (KNN) and K-means clustering method (K-means) and Genetic Algorithm (GA), respectively.

First, research was conducted on the situation of isolated outlier. When the data have an isolated outlier, its corresponding station data were shown in Fig. 5. After reconstruction

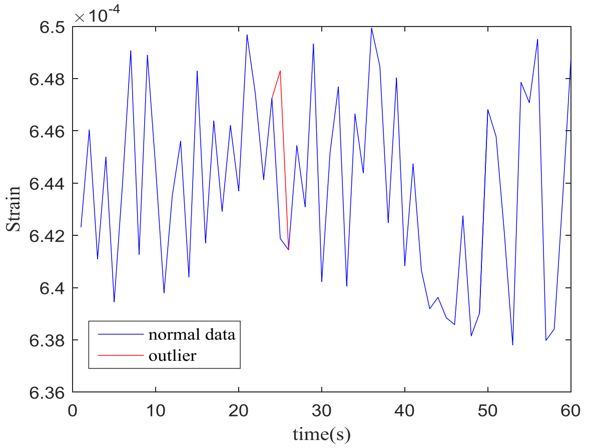


FIGURE 5. **The data characteristic with isolated outlier in simulation dataset.**

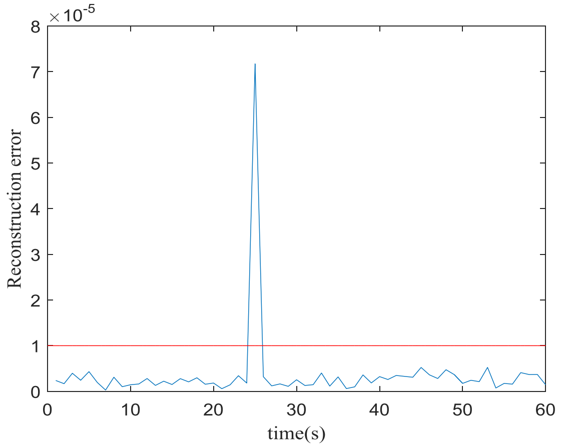


FIGURE 6. **The reconstruction error of data with isolated outlier in simulation dataset.**

by the SAE model, the reconstruction error was shown in Fig. 6. From the Fig. 6, the error corresponding to the outlier is significantly larger than the normal data's, and the detection effect is good.

In order to further prove the superiority of the proposed algorithm to isolated outlier, the test dataset containing 1000 sets of data is constructed in the contrast experiment, including 500 sets of normal data and 500 sets of abnormal data with isolated outlier. Among the 500 normal data, there are 50 sets of data for each station, and the same is true for abnormal data. Since the dimension of the input data is 60 for each SAE model of station, the algorithm will analyze the reconstruction error by the Grubbs criterion.

In Fig. 7, the results of the five methods were summarized at different significance levels. It can be seen that the proposed algorithm is always superior to the other algorithms at each significance level from the figure.

The optimal significance level of the SAE algorithm is 0.01, and 0.005 is the optimal significance level of the other algorithms. In optimal significance level, the algorithm rankings are SAE, GA, KNN, K-means and PCA, and the corresponding detection accuracies are 95.00%, 87.50%, 70.30%, 67.90% and 51.90% respectively. Regarding to detection accuracy, the SAE algorithm performs very well on detection of isolated outlier and is very stable at various significance levels. Next is the GA algorithm, which also achieves a high detection accuracy and stable performance. The results of KNN and the K-means are very close and are lower than the GA in turn. The PCA algorithm is obviously not effective because the PCA algorithm is difficult to process data with high nonlinearity.

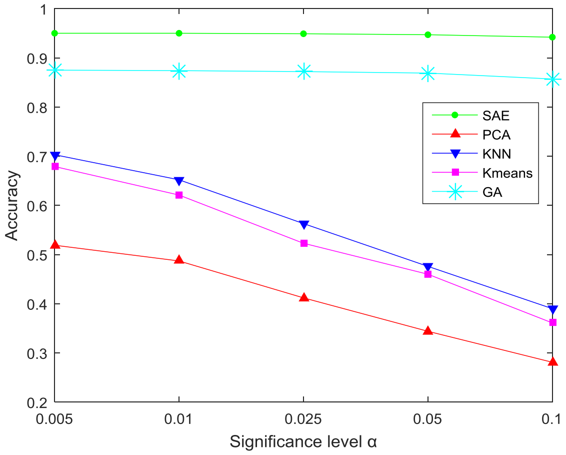


FIGURE 7. **The results of five algorithms to isolated outlier.**

When continuous outliers are added to the data, the data are as shown in Fig. 8, and the corresponding reconstruction error is presented in Fig. 9. As can be seen from Fig. 9, when there are continuous outliers in the data, the reconstruction error will produce a continuous peak, which ensures that successive outliers can be efficiently identified.

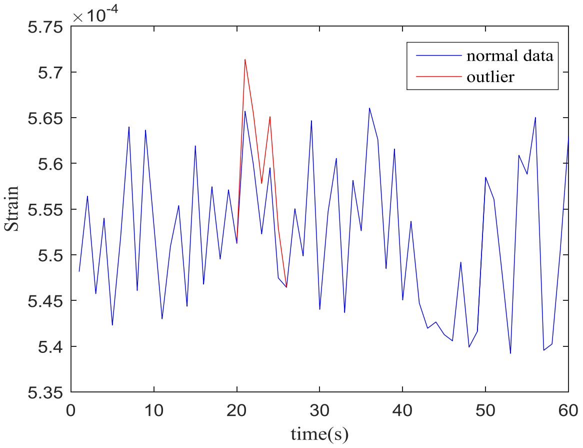


FIGURE 8. **The data characteristic with continuous outliers in simulation dataset.**

As above, the detection results of five algorithms are summarized in Fig. 10. In addition to the fact that the detection accuracy of the SAE algorithm is reduced at the significance

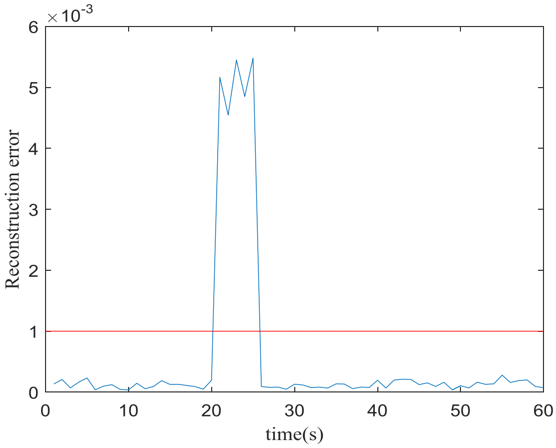


FIGURE 9. The reconstruction error of data with continuous outliers in simulation dataset.

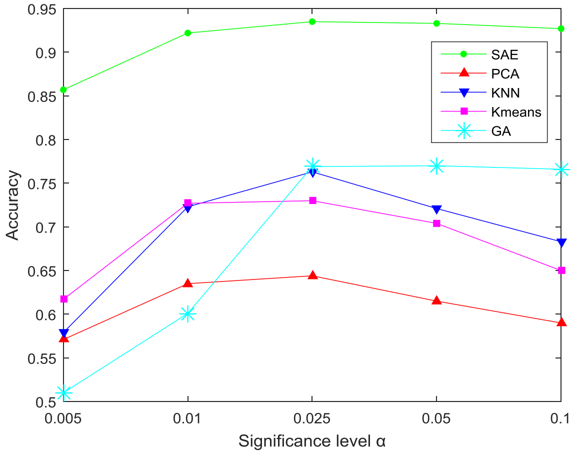


FIGURE 10. The results of five algorithms to continuous outliers.

TABLE 7. Detection effect of SAE algorithm under optimal significance levels.

Results	The case of outlier	
	Isolated	Continuous
TP	499	499
FP	1	1
FN	50	64
TN	450	436
Accuracy	94.90%	93.50%

level of 0.005, the detection accuracy is still very good in other cases. The reason for this is that more abnormal data were misjudged as normal data in the significance level. On the whole, when there are continuous outliers in the data, the optimal significance levels of all algorithms are 0.025 except GA which is 0.05. The rankings of the algorithms are also SAE, GA, KNN, K-means and PCA, and the corresponding detection accuracies are 93.50%, 77.00%, 76.30%, 73.00% and 64.40%, respectively.

Comprehensively considering the detection effects of the proposed algorithm on the isolated outlier and continuous outliers, the final significance level of the simulation dataset is set to 0.05. At the significance level, the algorithm's confusion matrix and detection accuracy can be obtained from the Table 7.

It can be seen that the SAE method has a good ability to recognize the abnormal data in the simulation dataset, which further confirms the usability of proposed method. Throughout the simulation experiment, the proposed algorithm is superior to the four methods for isolated outlier and continuous outliers, which proves the superiority of the proposed algorithm.

IV. CONCLUSION

In this paper, we proposed a method based on stacked autoencoder to detect abnormal data appearing in the monitoring data. First, the SAE model is trained by normal data to obtain the encoder parameters and decoder parameters of the model. After the input vector is mapped by the encoder and decoder, a reconstructed vector is generated. According to the size of the reconstruction error, normal data and abnormal data can be distinguished. It can be seen from the experiments that the proposed method has a good detection effect on isolated outlier and continuous outliers. The recognition effect of the proposed algorithm for continuous outliers may decrease with the increase of the proportion of outliers compared with the isolated outlier, but it still maintains a high detection accuracy.

In order to train a good SAE model, the key is to set the parameters of the model, such as model depth, learning rate, iterations, batches, and so on. These parameters are super-parameters. There is currently no a great way to explain how to set them up, so the optimal parameters are obtained in constant trials. A good model of deep learning usually requires a lot of data to train, and the method mentioned in this article is no exception. However, SAE models achieved a considerable detection result without too much training data in the experiments, which shows that the proposed method is not too dependent on the size of the training set.

The proposed method has an advantage in dealing with large-scale, high-dimensional nonlinear data compared with the traditional method, and has the potential ability to detect outliers online and offline. When the abnormal data detection is performed offline, the aim usually is to remove the abnormal data from the data set to ensure the reliability of subsequent data modeling or analysis. The need for algorithm runtime in this way is usually not as strong as online detection methods. When the abnormal data detection is performed online, it need isolate the abnormal data in time and feed back to the corresponding processing mechanism to avoid false-alarm and false negative due to abnormal data. This sets a higher request on the real-time of the algorithm.

Although it takes a long time to train SAE model, the trained model does not occupy a large amount of resources when it is used. Hence it does not affect the stability of the system when it is added to an existing system as a pre-processing module, which means that the method has a good ability to adapt to various scenarios. Of course, it does not mean that the SAE model is always better than other algorithms. Perhaps the traditional algorithms are more suitable for some relatively simple monitoring data.

In the experiments, we also found it is because the abnormal data are reconstructed to normal data to some extend that a large reconstruction error will be generated at the outlier. It shows that the proposed method can not only identify the abnormal data, but also restore it. It is the focus of our work to research the ability that the SAE algorithm restores abnormal data for some time to come.

REFERENCES

[1] D. A. Tibaduiza-Burgos and M. A. Torres-Arredondo, "Investigation of an expert health monitoring system for aeronautical structures based on pattern recognition and acousto-ultrasonics," *Smart Mater. Struct.*, vol. 24, no. 8, Jul. 2015, Art. no. 085020.

[2] C. Zhang, Y. Liu, F. Wan, B. Chen, and J. Liu, "Isolation and identification of compound faults in rotating machinery via adaptive deep filtering technique," *IEEE Access*, vol. 7, pp. 139118–139130, 2019.

[3] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 2, pp. 159–170, 2nd Quart., 2010.

[4] T. Yu, X. Wang, and A. Shami, "Recursive principal component analysis-based data outlier detection and sensor data aggregation in IoT systems," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 2207–2216, Dec. 2017.

[5] W. Wang, K. Velswamy, K. Hao, L. Chen, and W. Pedrycz, "A hierarchical memory network-based approach to uncertain streaming data," *Knowl. Based Syst.*, vol. 165, pp. 1–12, Feb. 2019.

[6] M. Morshedizadeh, M. Kordestani, R. Carriveau, D. S.-K. Ting, and M. Saif, "Application of imputation techniques and adaptive neuro-fuzzy inference system to predict wind turbine power production," *Energy*, vol. 138, pp. 394–404, Nov. 2017.

[7] J. Dai, H. Song, G. Sheng, and X. Jiang, "Cleaning method for status monitoring data of power equipment based on stacked denoising autoencoders," *IEEE Access*, vol. 5, pp. 22863–22870, 2017.

[8] K. Mahmoodi and H. Ghassemi, "Outlier detection in ocean wave measurements by using unsupervised data mining methods," *Polish Maritime Res.*, vol. 25, no. 1, pp. 44–50, Mar. 2018.

[9] Z. Huan, C. Wei, and G.-H. Li, "Outlier detection in wireless sensor networks using model selection-based support vector data descriptions," *Sensors*, vol. 18, no. 12, p. 4328, Dec. 2018.

[10] J. K. Dutta and B. Banerjee, "Improved outlier detection using sparse coding-based methods," *Pattern Recognit. Lett.*, vol. 122, pp. 99–105, May 2019.

[11] L. Sun, K. Zhou, X. Zhang, and S. Yang, "Outlier data treatment methods toward smart grid applications," *IEEE Access*, vol. 6, pp. 39849–39859, 2018.

[12] Z. Sun and H. Sun, "Stacked denoising autoencoder with density-grid based clustering method for detecting outlier of wind turbine components," *IEEE Access*, vol. 7, pp. 13078–13091, 2019.

[13] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 73–79, 2011.

[14] M. Ahsan, M. Mashuri, H. Kuswanto, D. D. Prastyo, and H. Khusna, "Outlier detection using PCA mix based T^2 control chart for continuous and categorical data," in *Proc. Commun. Statist. Simulation Comput.*, Apr. 2019, pp. 1–28.

[15] M. Hu, Z. Ji, K. Yan, Y. Guo, X. Feng, J. Gong, X. Zhao, and L. Dong, "Detecting anomalies in time series data via a meta-feature based approach," *IEEE Access*, vol. 6, pp. 27760–27776, 2018.

[16] J. Ahn, M. H. Lee, and J. A. Lee, "Distance-based outlier detection for high dimension, low sample size data," *J. Appl. Statist.*, vol. 46, no. 1, pp. 13–29, Mar. 2018.

[17] H. Liu, X. Li, J. Li, and S. Zhang, "Efficient outlier detection for high-dimensional data," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 12, pp. 2451–2461, Dec. 2018.

[18] C. Wang, Z. Liu, H. Gao, and Y. Fu, "Applying anomaly pattern score for outlier detection," *IEEE Access*, vol. 7, pp. 16008–16020, 2019.

[19] Q. Zhu, X. Fan, and J. Feng, "Outlier detection based on k-neighborhood MST," presented at the 15th Int. Conf. Inf. Reuse Integr., Redwood City, CA, USA, Aug. 2014.

[20] E. Schubert and M. Gertz, "Intrinsic t-stochastic neighbor embedding for visualization and outlier detection," presented at the 10th Int. Conf., SISAP, Munich, Germany, Oct. 2017.

[21] G. Bhattacharya, K. Ghosh, and A. S. Chowdhury, "Outlier detection using neighborhood rank difference," *Pattern Recognit. Lett.*, vol. 60, pp. 24–31, Aug. 2015.

[22] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, p. 3, 2012.

[23] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," presented at the 8th IEEE Int. Conf. Data Mining, Pisa, Italy, Dec. 2008.

[24] Z. Liu, X. Liu, J. Ma, and H. Gao, "An optimized computational framework for isolation forest," *Math. Problems Eng.*, Feb. 2018, Apr. 2018, Art. no. 2318763.

[25] N. Pattanavijit, P. Vateekul, and K. Sarinnapakorn, "A linear-clustering algorithm for controlling quality of large scale water-level data in Thailand," presented at the 12th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE), Songkhla, Thailand, Jul. 2015.

[26] Y. F. Wang, Y. Jiong, G. P. Su, and Y. R. Qian, "A new outlier detection method based on OPTICS," *Sustain. Cities Soc.*, vol. 45, pp. 197–212, Feb. 2019.

[27] G. Gan and M. K.-P. Ng, "K-means clustering with outlier removal," *Pattern Recognit. Lett.*, vol. 90, pp. 8–14, Apr. 2017.

[28] H. N. Akouemo and R. J. Povinelli, "Data improving in time series using ARX and ANN models," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3352–3359, Sep. 2017.

[29] K. Mahapatra, N. R. Chaudhuri, and R. G. Kavasseri, "Online bad data outlier detection in PMU measurements using PCA feature-driven ANN classifier," presented at the IEEE Power Energy Soc. Gen. Meeting, Chicago, IL, USA, Feb. 2018.

[30] X. Deng, P. Jiang, X. Peng, and C. Mi, "An intelligent outlier detection method with one class support tucker machine and genetic algorithm toward big sensor data in Internet of Things," *IEEE Trans. Ind. Electron.*, vol. 66, no. 6, pp. 4672–4683, Jun. 2019.

[31] L. Chomatek and A. Duraj, "Multiobjective genetic algorithm for outliers detection," presented at the IEEE Int. Conf. Innov. Intell. Syst. Appl. (INISTA), Gdynia, Poland, Jul. 2017.

[32] H. Lee and E. Kim, "Genetic outlier detection for a robust support vector machine," *Int. J. Fuzzy Logic Intell. Syst.*, vol. 15, no. 2, pp. 96–101, Jun. 2015.

[33] D. Cucina, A. di Salvatore, and M. K. Protopapas, "Outliers detection in multivariate time series using genetic algorithms," *Chemometrics Intell. Lab. Syst.*, vol. 132, pp. 103–110, Mar. 2014.

[34] B. Xiao, Z. Wang, Q. Liu, and X. Liu, "SMK-means: An improved mini batch k-means algorithm based on mapreduce with big data," *Comput. Mater. Continua*, vol. 56, no. 3, pp. 365–379, Nov. 2018.

[35] W. U. Xiaoya, "Mining high-dimensional outliers based on the combination of genetic algorithm and simulated Annealing algorithm," *MicroComput. Inf.*, vol. 21, pp. 139–141, Jul. 2010.

[36] C. Mei and L. Bo, "Abnormal data detection method based on ant colony algorithm," *Comput. Eng.*, vol. 42, no. 8, pp. 166–169, Aug. 2016.

[37] H. Zhu and B. Liu, "Outlier detection based on artificial bee colony intelligent technology," *J. Frontiers Comput. Sci. Technol.*, vol. 12, pp. 1984–1992, Nov. 2017.

[38] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Inf. Fusion*, vol. 42, pp. 146–157, Jul. 2018.

[39] X. Wang and H. Liu, "Soft sensor based on stacked auto-encoder deep neural network for air preheater rotor deformation prediction," *Adv. Eng. Inform.*, vol. 36, pp. 112–119, Apr. 2018.

[40] F. Zhao, Y. Liu, K. Huo, S. Zhang, and Z. Zhang, "Radar HRRP target recognition based on stacked autoencoder and extreme learning machine," *Sensors*, vol. 18, no. 1, p. 173, 2018.

[41] X. Huang, T. Hu, C. Ye, G. Xu, X. Wang, and L. Chen, "Electric load data compression and classification based on deep stacked auto-encoders," *Energies*, vol. 12, no. 4, p. 653, Feb. 2019.

[42] Z. Xiang, X. Zhang, W. Zhang, and X. Xia, "Fault diagnosis of rolling bearing under fluctuating speed and variable load based on TCO Spectrum and Stacking Auto-encoder," *Measurement*, vol. 138, pp. 162–174, May 2019.

[43] Y. Zhou, R. Qin, H. Xu, S. W. Sadiq, and Y. Yu, "A data quality control method for seafloor observatories: The application of observed time series data in the East China Sea," *Sensors*, vol. 18, no. 8, p. 2628, Aug. 2018.

[44] *UCR Time Series Classification Archive*. Accessed: 2018. [Online]. Available: https://www.cs.ucr.edu/~eamonn/time_series_data_2018/

FANGYI WAN received the B.S. degree in water resources and hydroelectric engineering and the M.S. degree in printing and packaging engineering from the Xi'an University of Technology, Xi'an, China, in 1994 and 1996, respectively, and the Ph.D. degree in mechanics engineering from Xi'an Jiaotong University, Xi'an, in 2003.

He was a Visiting Scholar with the Virginia Polytechnic Institute and State University, Virginia, USA, from 2005 to 2006, and with the University of Liège, Liège, Belgium, from 2013 to 2014. He joined the School of Aeronautics, Northwestern Polytechnical University, Xi'an, as a Faculty Member, where he is currently an Associate Professor. His main research interests include vibration analysis and control, design, modeling, test, and health management of aircraft structures.

GAODENG GUO received the B.S. degree in quality and reliability of aircraft from Shenyang Aerospace University, Shenyang, China, in 2017. He is currently pursuing the M.S. degree with the School of Aeronautics, Northwestern Polytechnical University, Xi'an, China.

His research interests include health management, fault diagnosis, pattern recognition, and intelligence algorithm.

CHUNLIN ZHANG received the B.S. and Ph.D. degrees in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2011 and 2017, respectively.

From 2014 to 2016, he was a Visiting Scholar with the University of Michigan, Ann Arbor, MI, USA. He joined the School of Aeronautics, Northwestern Polytechnical University, Xi'an, where he is currently an Assistant Professor. His main research interests include structural vibrations, dynamics and fault diagnosis, and nonlinear vibration energy harvesting.

QING GUO received the B.S. degree in computer science engineering, the M.S. degree in aircraft design, and the Ph.D. degree in aeronautics engineering from Northwestern Polytechnical University, Xi'an, China, in 2004, 2007, and 2012, respectively.

He is a faculty with the School of Aeronautics, Northwestern Polytechnical University, where he is currently an Associate Professor. His research direction is aircraft design and flight experimental research. One of the important is aircraft scale reduction verification study. And he is good at aviation models and flight operations of drones. The team named NPU Innovation he led received several World Series awards, such as UMSIC (um) in Egypt, ACC (Air Cargo Challenge) in Europe, NFC (New Flying Competition) in Germany, and many Chinese competitions.

JIE LIU received the B.Eng. degree in electronics and precision engineering from Tianjin University, Tianjin, China, in 1998, the M.Sc. degree in control engineering from Lakehead University, Thunder Bay, ON, Canada, in 2005, and the Ph.D. degree in mechanical engineering from the University of Waterloo, Waterloo, ON, in 2008.

He is currently an Associate Professor with the Department of Mechanical and Aerospace Engineering, Carleton University, Ottawa, Canada. He is also a part-time Researcher with the School of Aeronautics, Northwestern Polytechnical University, Xi'an, China. His research interests include dynamics, signal processing, vibration analysis and control, linear/nonlinear system control, machine condition monitoring, instrumentation and measurement, mechatronic systems, and artificial intelligence.