Abnormal Data Detection Based on Adaptive Sliding Window and Weighted Multiscale Local Outlier Factor for Machinery Health Monitoring

Qinglin Xie, *Student Member, IEEE*, Gongquan Tao, Chenxi Xie, and Zefeng Wen

***Abstract*—Identifying abnormal data to improve data quality is of great importance for machinery health monitoring (MHM). Existing abnormal data detection methods generally depend on appropriate parameter settings and prior knowledge of data distribution, which result in relatively low adaptability to MHM data. To obtain more reliable MHM results, this article proposed a novel method to detect abnormal data. First, the concept of adaptive sliding window (ASW) is defined and the advantages of ASW in avoiding data leakage and data redundancy are derived. The ASW can optimally divide the overall data into several segments. Next, statistical factors in time- and frequency-domain are extracted from these segments to generate corresponding research objects. Then, an improved weighted multiscale local outlier factor (WMLOF) algorithm is proposed herein to evaluate the data anomaly degree of each object. The WMLOF has the ability to assess and fuse the local outlier factor characteristics at multiple scales, and eventually yields a more comprehensive WMLOF value to evaluate the anomaly degree of the MHM data. Finally, the effectiveness and superiority of the ASW and WMLOF are validated by a synthetic simulation of a faulty rolling bearing and two engineering projects pertaining to a railway vehicle gearbox, and bench test data. Comparative analysis shows that the proposed abnormal data detection method based on the ASW and WMLOF strategies outperforms five reported algorithms in outlier detection.**

***Index Terms*—Abnormal data detection, adaptive sliding window (ASW), data quality, structural health monitoring data, weighted multiscale local outlier factor (WMLOF).**

I. Introduction

IN THE past two decades, the amount of data obtained and stored in machinery systems has increased continually, and the study of machinery health monitoring (MHM) has entered the era of big data [1]. Meanwhile, owing to the complex conditions involved during data acquisition, transmission and storage, the modern MHM data exhibit the following five characteristics: high velocity; high variety; high volume; low value density; and low veracity [2]. Consequently, it is laborious to extract information representing the health status of machinery, particularly, when the MHM data contain anomaly. In this regard, conventional fault diagnosis methods based on signal processing techniques are unsuitable [3]. Intelligent diagnosis methods based on machine learning algorithms have gradually become a hot topic in recent years. Scholars have proposed various machine learning-based models and made considerable achievements in fault diagnosis of rolling bearings, wind turbines, gears, motors, etc., [4], [5].

Driven by big data, although machine learning-based methods offer some incomparable advantages in MHM, many problems remain to be solved, among which data quality is the key. Deep learning-based algorithms cannot evaluate data quality and exhibit the disadvantage of "garbage in, garbage out" [6], [7]. Results obtained based on substandard-quality MHM data may be incorrect or misleading. Therefore, data quality assurance methods must be investigated to achieve a robust MHM research. To examine data quality systematically and comprehensively, the characteristics of data must be analysed. Generally, the attributes of data quality can be summarised into data accuracy, timeliness, consistency, and completeness [8]. In engineering practice, the operating environment of machinery is typically harsh. As such, the equipment is vulnerable to random interference factors, which consequently results in abnormal MHM data and reduces the data accuracy [2]. Timeliness refers to the update status of data. Transmission failures, e.g., network congestion and interruption can reduce the timeliness of MHM data [9]. Consistency measures the uniformity of data format and structure. Data from multiple physical sources are stacked directly without any classification or preprocessing, which significantly reduces the data consistency [10]. Completeness indicates the continuity of data. Owing to the failure of data acquisition equipment,

Manuscript received 13 August 2022; revised 28 October 2022; accepted 15 December 2022. Date of publication 28 December 2022; date of current version 8 May 2023. This work was supported in part by the National Natural Science Foundation under Grant U21A20167, in part by the Natural Science Foundation of Sichuan Province under Grant 2022NSFSC1914, and in part by the China Postdoctoral Science Foundation under Grant 2021T140571. *(Corresponding author: Gongquan Tao.)*

The authors are with the State Key Laboratory of Traction Power, Southwest Jiaotong University, Chengdu 610031, China (e-mail: qlxie@my.swjtu.edu.cn; taogongquan@swjtu.edu.cn; chenxixie@my.swjtu.edu.cn; zfwen@swjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIE.2022.3231279.

Digital Object Identifier 10.1109/TIE.2022.3231279

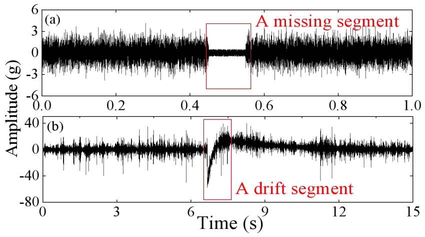

Fig. 1. Typical abnormal data. (a) Missing data. (b) Drift data.

e.g., sensor and CPU failures, missing data may be generated, thereby reducing data completeness [8]. Based on the definition and description of data quality, it is clear that missing data and drift (see Fig. 1) are typical issues of incomplete and inaccurate data. Undoubtedly, the anomaly of MHM data will directly affect the performance of fault diagnosis. Detecting and cleaning these anomalies can improve the data quality and yield more reliable results. Although studies regarding the algorithm and construction of the MHM model are abundant, studies regarding the quality assurance of MHM data, which serve as a foundation to the data-driven fault diagnosis studies, are scarce. Hence, this article focuses on issues pertaining to data quality assurance methods, particularly the two typical anomalies in MHM data, namely missing data and drift. Anomaly detection of MHM data belongs to the category of outlier detection. Conventional outlier detection is an important data mining activity, which is performed in numerous applications, including credit card fraud detection, video surveillance, weather prediction, and pharmaceutical research. Moreover, based on the definition of outliers by Hawkins [11], the following can be inferred: Anomalies in MHM data are an observation that deviates significantly from other observations. The anomaly might be generated by a different mechanism. Hence, anomalies in MHM data can be detected as outliers. Currently, studies regarding outlier detection are being pursued actively, and they can be classified into five categories based on the difference in the algorithm, namely statistic-, depth-, distance-, clustering- and density-based detection.

1) *Statistic-Based Detection:* Using a standard statistical distribution model, observations that deviate from the model are considered as outliers [12]. However, most classic statistics are sensitive to outliers and are only applicable to data with a single attribute [13]. Besides, the distribution of the investigated data should be determined in advance. In fact, these assumptions are often not satisfied. To this end, Lei et al. [14] introduced a difference operation, a windowing technique and a logarithm transform to reduce the feature tendency and volatility. They applied the three-sigma rule to the local means and volatilities for dirty data recognition.
2) *Depth-Based Detection:* This method relies on the computation of different layers of k–d convex hulls, where outliers are objects in the outer layer of these hulls [15]. Although it performs well on 2- or 3-D datasets, various issues arise when it extends to higher dimension, e.g., limited visualization methods, inadequacy of marginal methods and lack of a natural order [16]. Hussain [17] designed a nonparametric depth-based outlier detection method to address the above limitations. The availability of this method was verified by a hydrological data analysis.
3) *Distance-Based Detection:* Knorr et al. [18] systematically introduced the notion, algorithms and applications of distance-based outliers. The Mahalanobis distance (MD) and the Euclidean distance are commonly used to detect outliers. This type of method does not require prior knowledge regarding the distribution pattern of the dataset. The outlier is described as the object whose distance exceeds $d_{\min}$ from the target [19]. However, the performance of distance-based methods may be significantly reduced when several outliers exist within a multivariate dataset of interest [20].
4) *Clustering-Based Detection:* The outliers are by-products of clustering methods. Objects that do not belong to any class or belong to a small number of classes are regarded as outliers. In [20], a multistep clustering-based outlier detection scheme was presented. This method has been proved to be able to detect multiple outliers that may occur in multivariate normal data with high probability. However, it may fail to detect outliers because the main objective of clustering analysis is to reveal the distribution pattern of a dataset cluster [19].
5) *Density-Based Detection:* Breunig et al. [21] defined a local outlier factor (LOF) as a measure of outlier degree in terms of the density between an object and its neighbourhood objects. Emmott et al. [22] indicated that the LOF outperforms other common outlier detection methods when applied on real-world datasets. Thus, great progress has been made in the research and application of LOF in recent years, and many variants of LOF algorithms have been developed [23]. Xu et al. [2] proposed an improved kernel-based LOF method for Big Data cleaning of machinery condition monitoring. Note that in LOF methods, the dataset distribution is not necessary to be mastered in advance, but the number of the nearest neighbours k must be determined appropriately to ensure the detection accuracy [24].

From the above literature review, the LOF method is more suitable for our study. Nevertheless, the LOF algorithm is highly dependent on the number of the nearest neighbours k. Once k changes, the result of outlier detection will differ significantly. Therefore, we propose a weighted multiscale LOF (WMLOF) method to overcome this disadvantage. Further, we propose an adaptive sliding window (ASW) technique, which can optimally divide the overall MHM data into a series of segments, thereby facilitating specific analyses, reducing data dimensions, and increasing computational efficiency. Combining the improvements above, an abnormal data detection method for MHM data is established. The main contribution of this article is summarised as follows.

1) A novel ASW method applicable to MHM data is proposed. The ASW method can dexterously solve the problems of data leakages and data redundancy caused by a sliding window with an invariable length. The overall

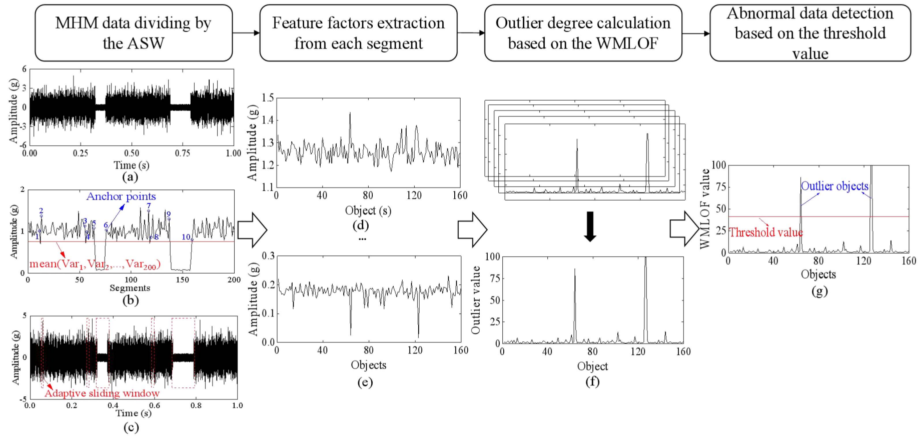
MHM data dividing by the ASW
Feature factors extraction from each segment
Outlier degree calculation based on the WMLOF
Abnormal data detection based on the threshold value
Amplitude (g)
Time (s)
(a)
Anchor points
mean(Var$_1$, Var$_2$,..., Var$_{200}$)
Segments
(b)
Adaptive sliding window
Time (s)
(c)
Object (s)
(d)
...
Objects
(e)
Outlier value
Object
(f)
Outlier objects
Threshold value
WMLOF value
Objects
(g)

Fig. 2. Flowchart of the proposed abnormal data detection method for MHM data based on ASW and WMLOF. (a) Raw MHM data. (b) Key points location. (c) Adaptive sliding windows. (d) Waveform factor. (e) Mean frequency. (f) Outlier value. (g) Abnormal data dection.

MHM data can be divided optimally into a series of segments by the ASW.

2) We propose a novel WMLOF method based on entropy weight theory to improve the shortages of the LOF method. Based on the intelligent weighting and integration of results comprising the multiscale parameter k, the WMLOF method can adaptively combine the LOF features at each scale.

3) An abnormal data detection method is established using the ASW and WMLOF strategies, and the application of outlier detection in MHM data is expanded. The quantitative abnormal degree of each object in the MHM data can be evaluated by comparing it with the corresponding threshold value.

The rest of this article is organized as follows: Section II describes the proposed abnormal data detection method. Section III presents the verification of the effectiveness of the proposed ASW and WMLOF strategies by using the detection of missing data from a synthetic faulty bearing dataset. In Section IV, the superiority and universality of the proposed method are demonstrated based on measurement data acquired from a railway vehicle gearbox and a bench test. Finally, Section V concludes this article.

II. Proposed Method

This section describes the abnormal data detection method based on the ASW and AMLOF strategies. The flowchart is shown in Fig. 2. First, the ASW technology is used to optimally divide the MHM data into a series of segments. Second, the time- and frequency-domain statistical factors in each segment are extracted to refine the data information. Third, based on these feature factors, the outlier degree of each object can be calculated using the WMLOF. Finally, anomalies in the MHM data can be detected by comparing the outlier degree with a threshold value.

A. MHM Data Dividing by the ASW

MHM data are typically large in volume, and valuable features of which are difficult to obtain by direct analysis. Hence, a sliding window technology is used in this article. First, the entire dataset is divided into a series of segments by a sliding window, which is helpful to target analysis and to improve the computational efficiency. Generally, the dataset is divided by a sliding window with an invariable length. However, the invariable length of sliding window technology often results in two non-negligible problems, i.e., data leakage and data redundancy, as depicted in Fig. 3. A sliding window with a length of L_1 can effectively identify region A, but it cannot completely include region B. A portion of data with the same attributes in region B is not included in sliding window L_1, which implies data leakage. Conversely, a sliding window with a length of L_2 can identify region B accurately. However, an over-inclusion occurs in region A. Some characteristics that do not belong to region A are summarised together by sliding window L_2, which implies data redundancy. Therefore, data leakage and redundancy are primary issues to be addressed. A sliding window with an adaptive length can be considered to address the effects of the invariable length of sliding window.

The ASW involves the generation of sliding windows of different lengths for different types of data. In this article, the attribute variation of data is utilised, and an ASW strategy is proposed to determine the sliding window size. In the MHM data, two important characteristics are observed when the data attributes change. First, the dataset containing a section of abnormal data must include two inflection regions, i.e., the intersection of normal and abnormal data, and the intersection of abnormal and normal data, which are known as the entry region (R_{in}) and exit region (R_{out}), respectively, as shown in Fig. 3. Second, the generation mechanism of abnormal data is different from normal data. Therefore, the deviation degree of the random variable in normal and abnormal data from its corresponding mathematical expectation, i.e., the variance, will likely be different, particularly in R_{in} and R_{out}. Based on the observation and analysis above, we propose an ASW strategy. The detailed procedures are as follows.

1) The original MHM data are first divided by a sliding window of minilength W. The value of W is associated with the length (sampling frequency × sampling time) of

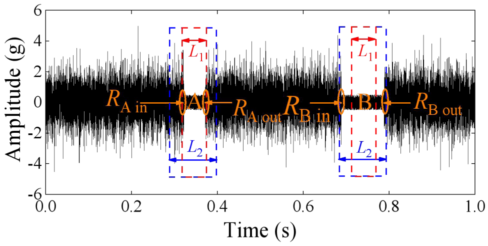

Fig. 3. Data leakage and data redundancy.

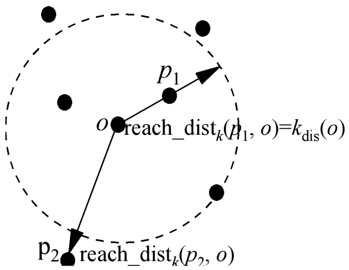
p_1
o
reach_dist$_k$(p_1, o)=k_{dis}(o)
p$_2$
reach_dist$_k$(p_2, o)

Fig. 4. Reach_dist$_k$(p_1, o) and reach_dist$_k$(p_2, o), for $k = 4$ [2].

the signal to be analysed. It is recommended that the ratio of the length of W to the original MHM data be 1/200 to 1/100 to ensure detection accuracy and efficiency.

2) The variance of the series of segments divided by a sliding window of mini-length W is calculated.
3) Based on the mean of the variance values, an ASW threshold V is obtained using the expression $V = \text{mean}$ $(\text{Var}_1, \text{Var}_2, \ldots, \text{Var}_n)$, where n is the number of segments.
4) The variance value of each segment is compared with the threshold V to obtain several intersection points. The first point on the left of odd intersections is considered as AP_{in} and the first point on the right of even intersections as AP_{out} to obtain a series of anchor points for the ASWs.
5) The obtained anchor points are used to generate the specific ASWs of different sizes. Data outside the AP_{in} and AP_{out} groups can be regarded as normal data with a same attribute and are unlikely to share anomalies. Thus, they can be divided by a conventional sliding window of mini-length W to improve the computational efficiency.

B. Feature Factors Extraction From Each Segment

It is beneficial to extract the time- and frequency-domain statistical factors in each segment to refine the data information and reduce the data dimensions. The abovementioned factors include the absolute mean, variance, standard deviation, kurtosis, skewness, root mean square, shape factor, peak factor, impulse factor, margin factor, kurtosis factor, clearance factor and mean frequency. These statistical features were selected because they are widely used in MHM and can effectively describe the characteristics of the MHM data. It is noteworthy that not the more feature factors are, the better. It is more important to select factors that can condense the information of the MHM data to the maximum. Otherwise, the calculation burden will increase, which may result in other adverse effects.

C. Outlier Degree Calculation Based on the WMLOF

We first introduce the basic theory of the LOF. The initial outlier degree of each object can be calculated based on the following steps 1 to 5 [21].

Step 1: We use o, p, and q to denote objects in a dataset. The notation $d(p, q)$ represents the distance between objects p and q. For a set of objects, we use C. Briefly, we use $d(p, C)$ to denote the minimum distance between objects p and q in C, as follows:

$$d\left(p, C\right) = \min\left\{d\left(p, q\right) | q \in C\right\}. \tag{1}$$

For any positive integer k, the k-distance of object p, denoted as $k_{\text{dis}}(p)$, is defined as the distance $d(p, o)$ between p and an object $o\in$ D such that:

1) for k objects o'$\in D \setminus \{p\}$ at the least, $d(p, o') \leq d(p, o)$ holds; and
2) for k–1 objects o'$\in$ D \ {p} at the most, $d(p, o') < d(p, o)$ holds.

Step 2: For the $k_{\text{dis}}(p)$ of object p, the $k_{\text{dis}}(p)$ neighbourhood of p contains each object whose distance from p does not exceed the k_{dis}, i.e.,

$$\boldsymbol{N}_k\left(p\right) = \left\{q \in D \backslash \left\{p\right\} | d\left(p, q\right) \leq k_{\text{dis}}\left(p\right)\right\} \tag{2}$$

where q is known as the k-neighbour of object p.

Step 3: Let k be a natural number. The reachability distance of object p with respect to object o is defined as

$$\text{reach\_dist}_k\left(p, o\right) = \max\left\{k_{\text{dis}}\left(o\right), d\left(p, o\right)\right\}. \tag{3}$$

Fig. 4 illustrates the idea of reachability distance with $k = 4$. Intuitively, if object p is distant from o (e.g., p_2 in the figure), then the reachability distance between them is in fact their actual distance. However, if they are "sufficiently" close (e.g., p_1 in the figure), then the actual distance is replaced by the k_{dis} of o. The statistical fluctuations of $d(p, o)$ for all p's that are near o can be reduced significantly, and the intensity of this smoothing effect depends considerably on parameter k.

Step 4: Combined with the above definitions, the local reachability density (**lrd**) of p can be expressed as the average reachability distance of k

$$\mathbf{lrd}_k\left(p\right) = \left(\frac{1}{k}\sum_{o\in N_{k(p)}}\text{reach\_dist}_k\left(p, o\right)\right)^{-1}. \tag{4}$$

Step 5: The LOF value of object p is calculated to evaluate its abnormal degree. The **LOF** is the average of the ratio between the **lrd** of p and the k nearest neighbours of p, expressed as

$$\mathbf{LOF}_k\left(p\right) = \frac{1}{k}\sum_{o\in N_{k(p)}}\frac{\mathbf{lrd}_k\left(o\right)}{\mathbf{lrd}_k\left(p\right)}. \tag{5}$$

The outlier degree of each object can be evaluated using the LOF. However, the LOF is extremely sensitive to the parameter k because the k cannot satisfy the detection requirements of all outliers. Hence, we constructed a more robust LOF, namely WMLOF. Based on the intelligent weighting and integration of results with different parameters k, the WMLOF can adaptively combine the LOF features at multiple scales. The core idea of the WMLOF method is to determine the appropriate weights. Compared with other weighting methods, the entropy weight method (EWM) is simple in terms of calculation and obviates the necessity of considering the subjective preference [25], [26]. It only requires objective data to calculate the weight [27], [28]. Furthermore, the EWM is a typical diversity-based weighting method that calculates attribute weights based on the diversity of attribute data among the alternatives. Therefore, the EWM is employed to determine the weights of the initial outlier degree at different k scale, the detailed steps are described as follows.

Step 6: Equation (6) shows the decision matrix (DM). Each row and column of the matrix is indicated to one object and one LOF scale, respectively. Therefore, the q_{pk} of the DM [$p = 1$, 2, …, n, $k = k_{\min}$, $k_{\min}+1$, …, $k_{\max}-1$, $k_{\max}$] contributes to the DM. n represents the number of objects. $K \in [k_{\min}, k_{\max}]$. Based on prior experience and studies [2], [3], $k_{\min}$ and $k_{\max}$ are set as 5 and 20 in this article, respectively,

$$\mathbf{DM} = \begin{bmatrix} q_{1k_{\min}} & q_{1k_{\min}+1} & \cdots & q_{1k_{\max}-1} & q_{1k_{\max}} \\ q_{2k_{\min}} & q_{2k_{\min}+1} & \cdots & q_{2k_{\max}-1} & q_{2k_{\max}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ q_{pk_{\min}} & q_{pk_{\min}+1} & \cdots & q_{pk_{\max}-1} & q_{pk_{\max}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ q_{nk_{\min}} & q_{nk_{\min}+1} & \cdots & q_{nk_{\max}-1} & q_{nk_{\max}} \end{bmatrix}. \quad (6)$$

Step 7: The linear normalization technique is utilised to render the dataset of DM dimensionless. This can effectively reduce errors caused by dimension or magnitude when analyzing different responses from different samples. The normalized DM (NDM) is shown as

$$\mathbf{NDM}_{pk} = \frac{q_{pk} - \mathrm{Min}\,(q_k)}{\mathrm{Max}\,(q_k) - \mathrm{Min}\,(q_k)}. \quad (7)$$

Step 8: The occurrence probability of the response ($\mathbf{Pr}_{pk}$) is obtained using (8). Equation (9) is utilised to acquire the entropy of the kth response ($\mathbf{En}_k$)

$$\mathbf{Pr}_{pk} = \frac{\mathbf{NDM}_{pk}}{\sum_{p=1}^{n} \mathbf{NDM}_{pk}} \quad (8)$$

$$\mathbf{En}_k = -\frac{1}{\log_e(n)} \sum_{p=1}^{n} \mathbf{Pr}_{pk} \log_e(\mathbf{Pr}_{pk}) \quad (9)$$

where the value of $\mathbf{En}_k$ is between zero and one.

Step 9: Equations (10) and (11) are employed to calculate the divergence degree ($\mathbf{Div}_k$) and the entropy weight of the kth response ($\mathbf{Ew}_k$), respectively,

$$\mathbf{Div}_k = |1 - \mathbf{En}_k| \quad (10)$$

$$\mathbf{Ew}_k = \frac{\mathbf{Div}_k}{\sum_{k=k_{\min}}^{k_{\max}} \mathbf{Div}_k}. \quad (11)$$

The EWM for weight computation debilitates the weak impact of some atypical attributes and yields progressively precise and sensible assessments [29]. Hence, the LOF results of each object under different k values are regarded as the responses of different samples. The **WMLOF** value of the object p can be obtained as follows:

$$\mathbf{WMLOF}_p = 100 \times \mathbf{Ew}(\mathbf{NDM}_p)^{\mathrm{T}} \quad (12)$$

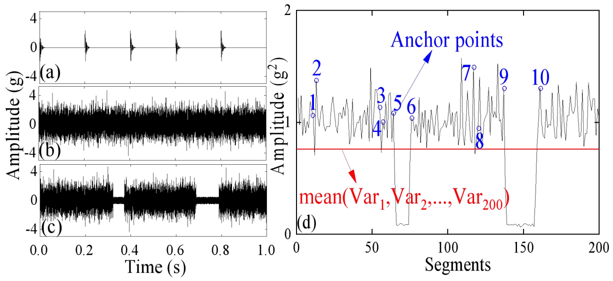

Fig. 5. (a) Partially simulated impulse signal. (b) Simulated faulty rolling bearing data. (c) Simulated faulty rolling bearing data with two missing segments. (d) Several anchor points based on ASW.

where the $\mathbf{WMLOF}_p$ is a scalar, $\mathbf{Ew}$ and $\mathbf{NDM}_p$ are row vectors. The purpose of multiplying by 100 is to enable a more intuitive understanding of the WMLOF characteristics.

D. Abnormal Data Detection Based on the Threshold Value

Based on the basic principles of the LOF, the density of anomaly is much lower than that of neighbouring normal data. Therefore, the outlier value of abnormal data is high [21]. Using a quantifiable threshold to detect the anomaly is more precise. Objects with WMLOF values greater than the threshold S are considered as anomaly objects. S is expressed as

$$S = \mu + \lambda\sigma \quad (13)$$

where μ and σ are the mean and standard deviation of the WMLOF values, respectively. Based on outlier detection principle, the parameter λ is set to 3 [30]. Hence, the anomaly whose WMLOF value differed significantly from the others can be detected.

III. Simulation Validation

To verify the effectiveness of the ASW strategy, a synthetic simulation of missing vibration data generated by a faulty rolling bearing was conducted. A remarkable feature of a faulty rolling bearing is periodic impulses, which are expressed as

$$\mathbf{y}(t) = y_0 e^{-\xi\omega_n t} \sin \omega_n \sqrt{1-\xi^2 t} \quad (14)$$

where y_0 is the amplitude of the fault impulse ($y_0 = 3$), ξ the damping coefficient ($\xi = 0.1$), ω_n the natural frequency of the rolling bearing, and f_{re} the corresponding resonance frequency ($f_{re} = 3000$ Hz). In addition, the frequency of fault characteristic f_o was set to 100 Hz, the sampling frequency f_s was 20 000 Hz, and the number of sampling points was 20 000. Fig. 5(a) depicts the simulated impulse signal partially. The data shown in Fig. 5(b) was obtained by adding Gaussian white noise, and the signal-to-noise ratio of the synthetic signal was zero. Fig. 5(c) exhibits two missing segments, which were created by replacing the original data from 0.322 to 0.372 s and 0.689 to 0.789 s with Gaussian white noise. In this regard, the proposed abnormal data detection method based on the ASW and WMLOF strategies is applied to detect the anomaly, i.e., data missing. The results are summarised in Figs. 5(d) and 6. Based on the principle of the

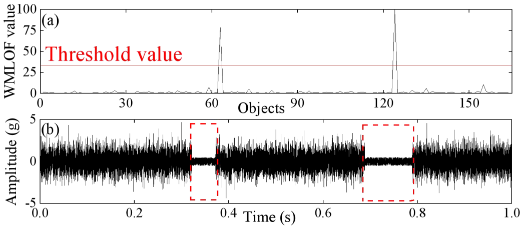

Fig. 6. Abnormal data detection based on the ASW and WMLOF. (a) WMLOF value. (b) Detected abnormal data segments.

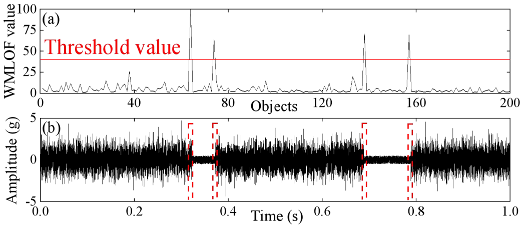

Fig. 7. Abnormal data detection based on short length sliding window and WMLOF. (a) WMLOF value. (b) Detected abnormal data segments.

ASW strategy, W was set to 100 to obtain the anchor points. The acquired anchor points 1 & 2, 3 & 4, 5 & 6, 7 & 8, and 9 & 10 generated five different sliding windows, respectively. In fact, it is sufficient to determine the ASWs based on anchor points 5 & 6 and 9 & 10. The sliding windows generated by anchor points 1 & 2, 3 & 4, and 7 & 8 are by-products of the ASW. The data attributes of the above sliding windows are consistent with the normal data. Therefore, treating these sliding windows as individual objects does not affect the subsequent WMLOF calculation. The last step of the ASW strategy is to divide the data outside the anchor points by a conventional sliding window with the same length as W.

As shown in Fig. 6(a), two larger WMLOF values corresponding to the objects generated by anchor points 5 & 6 and 9 & 10 (in Fig. 5(d)) were indicated. By contrast, the WMLOF values of the normal data were small and the variation is mild. This indicated that the WMLOF values can be used to effectively distinguish between normal and abnormal objects. The corresponding detected missing data were denoted with a red dotted rectangle, as depicted in Fig. 6(b). It is clear that two segments with missing data of different scales, i.e., 0.320 to 0.375 s and 0.685 to 0.800 s, were successfully detected based on the proposed ASW and WMLOF strategies. Besides, the WMLOF values were calculated based on several invariable length of sliding windows to illustrate the disadvantages of the conventional sliding windows and the advantages of the ASW strategy. As shown in Fig. 7, the invariable window length was set to 100. The primary missing data were not detected except the data at both ends of the abnormal missing segments, i.e., 0.315 to 0.320 s, 0.365 to 0.370 s, 0.685 to 0.690 s and 0.780 to 0.785 s. It indicated that a small window length can result in data leakage, so that the anomaly cannot be detected completely. Consequently, the WMLOF values at both ends of the missing segments were higher than the threshold value, whereas the WMLOF values of most missing data were lower than the threshold value, which resulted in the improper detection of the missing data.

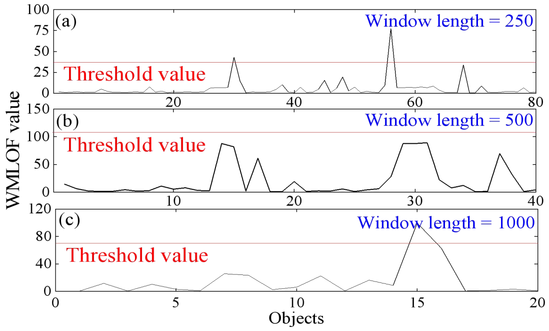

Fig. 8. Abnormal data detection based on large sliding windows and WMLOF: Window lengths of (a) 250; (b) 500; and (c) 1000.

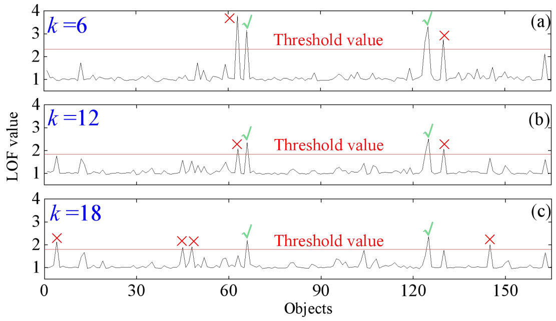

Fig. 9. Abnormal data detection based on the ASW and LOF with the nearest neighbours k of (a) 6, (b) 12, and (c) 18.

Larger length of sliding windows can also affect the detection results. Fig. 8 reveals the WMLOF values of each object under a sliding window with lengths of 250, 500, and 1000, respectively. None of the three lengths can correctly detect the missing data because the sliding window length is not specified based on the data attributes. The invariable length of sliding window cannot adapt to different scales of abnormal data. A small window length cannot completely contain the abnormal data, but it can only detect the ends of the abnormal fragment. Conversely, the larger length mixes the anomaly with the normal data, rendering the attributes of each segment more confusing. This is not conducive to correctly distinguishing the differences in attributes between the normal and abnormal MHM data. In addition, Fig. 9 gives the abnormal data detection results based on the ASW and traditional LOF. It can be seen that the incorrect outlier degree of abnormal data is obtained under several nearest neighbours k. The above shows the advantages of WMLOF in fusing LOF features at different scales. By comparing Figs. 6–9, it can be intuitively observed that the proposed method based on the ASW and WMLOF exhibits more adaptive and robust abilities than the conventional sliding windows and LOF algorithm. The necessity and ability of the developed ASW and WMLOF are verified by the ablation study. It can be seen that the proposed method combining the two strategies has advantages in detecting abnormal data at different scales, which is very important for the efficient processing of complex MHM data.

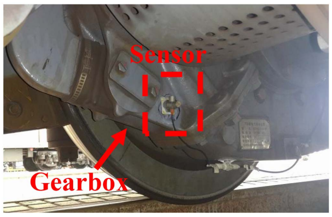

Fig. 10. Railway vehicle gearbox for vibration data acquisition.

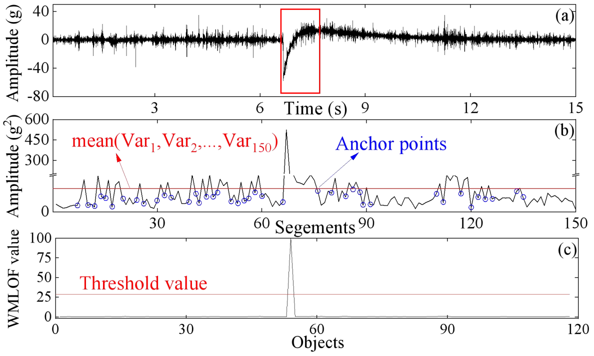

Fig. 11. Abnormal drift data detection based on the ASW and WMLOF. (a) Measured gear box Vibration data with drift data segment. (b) Several anchor points based on the ASW. (c) WMLOF value.

IV. Engineering Application

A. Anomaly Detection of Measurement Gearbox Data

Measurement data obtained from the gearbox of a metro vehicle were investigated to illustrate the practicability and universality of the proposed method. Fig. 10 shows a schematic diagram that used to acquire gearbox vibration data.

An accelerometer was installed on the gearbox, and the sampling frequency was 10 000 Hz. The railway vehicle is likely to be excited by wheel–rail irregularities when it operates on the line, which results in significant vibration responses of the vehicle components [31], [32]. Fig. 11(a) illustrates the case of data drift caused by sensor failure or connector damage. This anomaly often occurs in the measurement MHM data [33]. The abnormal data must be detected to improve the data quality so that the physical information in the MHM data can be extracted accurately. The proposed method was used to process the gearbox data. In addition, W was set to 1000. The anchor points of the gearbox data based on the ASW strategy are indicated in Fig. 11(b). Fig. 11(c) delineates the anomaly degree of each object based on the WMLOF method. The WMLOF value corresponding to one of the ASWs is significantly greater than the threshold. As shown in the red rectangular box in Fig. 11(a), the abnormal drift segment can be detected accurately and then cleaned for improving MHM data quality. This proves that the ASW strategy and WMLOF method proposed herein are not only suitable for the abnormal missing phenomenon in MHM data, but also can effectively detect data drift. Hence, they provide guidance on methods to detect anomalies from complex measurement MHM data.

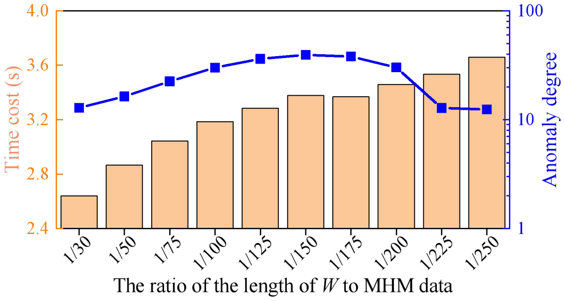

Fig. 12. Performance of the WMLOF under different lengths of W.

Besides, taking the data shown in Fig. 11(a) as the raw MHM data, several experiments were carried out to analyze the effect of different lengths of W on abnormal data detection. The corresponding results are shown in Fig. 12, where the abscissa represents the ratio of the length of W to the raw MHM data. The time cost indicates the time required to perform an outlier detection calculation. Moreover, it is known that the judgment of whether the object p is an anomaly lies in its WMLOF value and the threshold S based on the principle of the WMLOF. The greater the WMLOF value is than the threshold S, the higher the anomaly degree of the object p. AD is defined to illustrate the anomaly degree of the drift data segment under different values of W. AD is expressed as

$$AD = (\mathbf{WMLOF}_o - S)/S \tag{15}$$

where $\mathbf{WMLOF}_o$ indicates the WMLOF value of the drift data, and S denotes the same meaning as (13), i.e., the threshold value. It can be seen from Fig. 12 that as the length of W decreases, the time cost increases, and the AD first increases and then decreases. The reason is that the smaller the length of W, the more data segments are generated, and therefore the longer the computation time consumed. Although the larger W has higher efficiency, its detection resolution is insufficient and has certain disadvantages in accurate locating of abnormal data. Therefore, the length of W should not be too small or too large. Based on the above numerical experiments and analysis, we can obtain a recommended range for the value of W, i.e., the ratio of the length of W to the raw MHM data is 1/200 to 1/100 to ensure detection accuracy and efficiency.

B. Anomaly Detection of Bench Test Data

The proposed method was further validated using the displacement data from a bench test. The displacement data of the specimen was measured using a laser displacement sensor. The sampling frequency was 5000 Hz. Fig. 13 shows the displacement measurement system installed on a shaking table.

Fig. 14(a) depicts the original voltage waveform of the laser displacement. During the experiment, the disturbance was introduced to generate the abnormal data during 7 to 12 s by touching the sensor. It is difficult to distinguish the disturbances directly from the original data. The ASW and WMLOF were applied to detect the abnormal data. The minilength W in the ASW was set to 500. The obtained anchor points are pointed out in

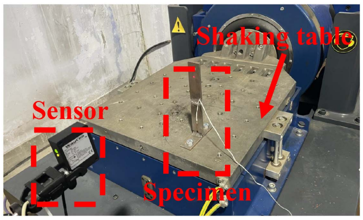

Fig. 13. Bench test for displacement data collection.

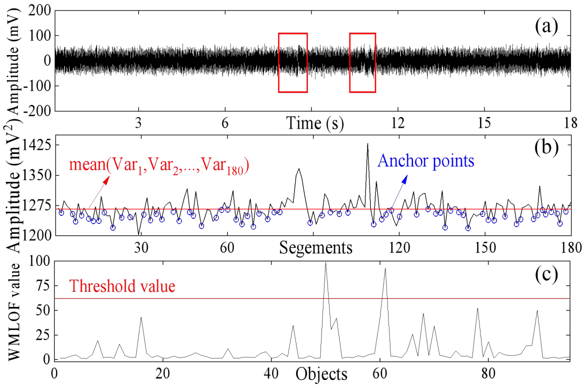

Fig. 14. Abnormal data detection based on the ASW and WMLOF. (a) Measured gear box vibration Data with abnormal data segment. (b) Several anchor points based on the ASW. (c) WMLOF value.

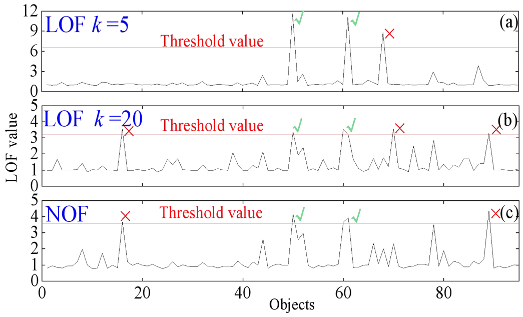

Fig. 15. Abnormal data detection. (a) LOF value when $k = 5$. (b) LOF value when $k = 20$. (c) NOF value.

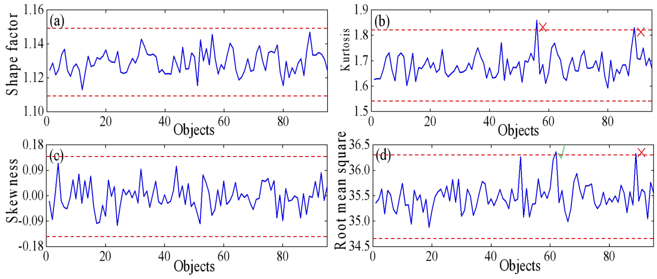

Fig. 16. Abnormal data detection based on three-sigma rule. (a) Shape factor. (b) Kurtosis. (c) Skewness. (d) Root mean square.

Fig. 14(b). Fig. 14(c) indicates the detected abnormal objects based on WMLOF values. The abnormal data detection results were indicated by a red rectangle, as shown in Fig. 14(a). It is confirmed that the laser displacement monitoring data during 7.8 to 8.8 s and 10.2 to 11.1 s were abnormal data, which is consistent with our expectations. The results show that the proposed method can achieve favourable performance in detecting abnormal MHM data, even if these anomalies are extremely slight or even invisible by the naked eye.

C. Comparative Analysis and Discussion

As mentioned in the introduction, the conventional LOF algorithms depend significantly on the parameter k. The value of k is typically determined based on researchers' experience. Therefore, researchers proposed the natural outlier factor (NOF) [19], [34] to adaptively obtain the appropriate value of k without requiring expert knowledge. The NOF is beneficial to avoid errors caused by human's subjective assessments. Both the conventional LOF and advanced NOF were used to detect the abnormal data given in Fig. 14(a). Fig. 15(a) and (b) shows the LOF values when $k = 5$ and 20, respectively. However, the abnormal data were not detected correctly in both cases. In addition to the two abnormal objects that are correctly detected, the LOF values of other objects are also greater than the threshold. The redundant amounts are 1 and 3 when $k = 5$ and 20, respectively. The above reported the false identification of the conventional LOF in outlier detection and hence the inability to effectively improve the data quality. The reason for this problem may be that the large value of k reduces the effect of noise on classification, although they render boundaries between classes less distinct. If the neighbourhood is extremely large with respect to the folds in the manifold on which the data points lie, then large values of k may cause short-circuit errors. Alternatively, small value of k reduces the neighbourhood correlation, or separates the data from the same class [34]. The NOF values can be found in Fig. 15(c). Similarly, it holds two false detection objects. Although the NOF can adaptively obtain the optimal value of k, the results are still a single scale and cannot fully reflect the outlier level of each object at multiple scales. Based on the comparative analysis and discussion, it is observed that the WMLOF method proposed herein can weigh and combine the LOF values, as well as aggregate the information of feature responses under multiple k scales. Summarily, the WMLOF method is not only superior to the conventional LOF method, but also offers significant advantages compared with the advanced NOF method.

For a more comprehensive comparison, three methods recommended in [2] and were used to detect the anomaly in MHM data, including the three-sigma rule, MD, and k-means. Besides, as a commonly used outlier detection method, density-based spatial clustering of applications with (DBSCAN) is also introduced. The original MHM data were divided into different segments using the ASW strategy. Time- and frequency-domain statistical features were extracted from each segment to generate different objects. First, the three-sigma rule using a single time-domain feature was employed. The results in Fig. 16 show that the abnormal data cannot be detected using features including shape factor and skewness, whereas the normal data were wrongly detected as abnormal data when using kurtosis. Only parts of the abnormal data were detected and false detection existed when using root mean square. Moreover, as a classical distance-based

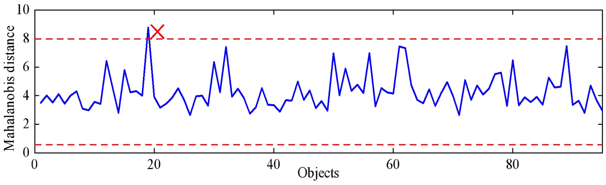

Fig. 17. Abnormal data detection based on MD.

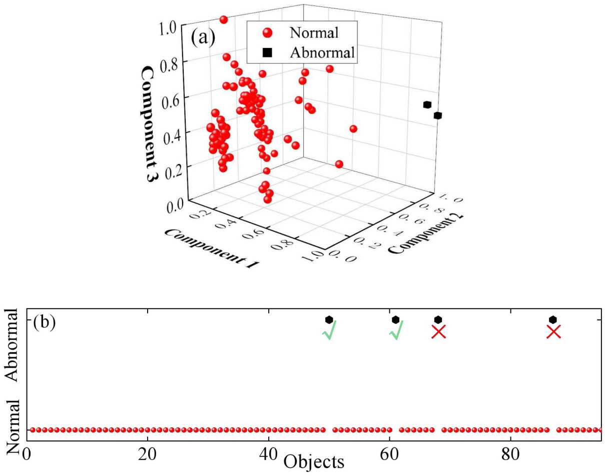

Fig. 18. Abnormal data detection based on *k*-means clustering. (a) Clustering of normal and abnormal data. (b) Abnormal data detection.

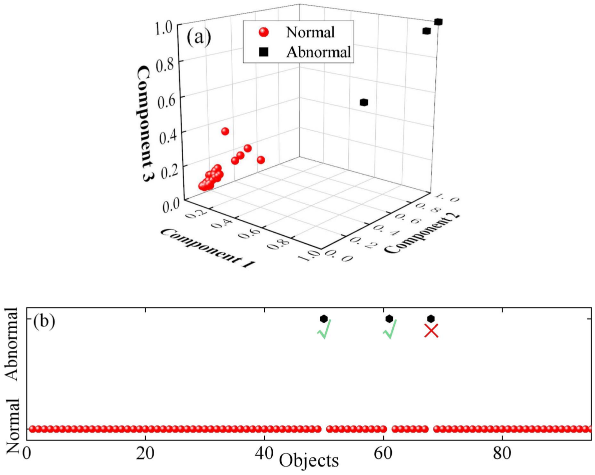

Fig. 19. Abnormal data detection based on DBSCAN. (a) Clustering of normal and abnormal data. (b) abnormal data detection.

method, the MD of the series of objects is calculated as

$$\mathbf{MD}\left(\mathbf{g}_p\right) = \sqrt{\left(\mathbf{g}_p - \mathbf{g}_{\text{mean}}\right) \sum{}^{-1} \left(\mathbf{g}_p - \mathbf{g}_{\text{mean}}\right)^{\mathrm{T}}} \tag{16}$$

where $\boldsymbol{g}_{\text{mean}}$ and Σ are the sample mean and sample covariance, respectively.

The MD of the objects is shown in Fig. 17. The three-sigma rule was adopted to detect the abnormal data whose MD is greater than the threshold. The objects including the normal data are detected as anomaly, while the abnormal data are failed to be detected. Accordingly, the MD method is unable to identify the abnormal data because the MD-based bound can only describe the contour of the Gaussian cluster.

Further, as a common clustering-based method, the *k*-means is employed to detect the anomaly in MHM data. First, the principal component analysis (PCA) was adopted to extract the first three principal from the statistical features [2]. The results of PCA are shown in Fig. 18(a), in which the cluster with red sphere is the normal data. We set the anomaly as black cube for a more intuitive observation. Then, the parameter *k* of *k*-means method was set to 2 because the whole objects are expected to fall into two clusters, including normal and abnormal clusters. The detection results in Fig. 18(b) show that some normal data are wrongly classified as anomaly. It can be seen that the *k*-means method cannot detect the anomaly accurately for the abnormal data with tiny disturbances as shown in Fig. 14(a).

Finally, the DBSCAN is introduced to further verify the availability of the WMLOF. The DBSCAN algorithm involves the selection of a minimum number of neighborhood points (MinPts) and radius of neighbourhood (Eps). We determine MinPts as 17 and Eps as 8 based on the principles proposed in [35]. The detection results are shown in Fig. 19. A normal sample is incorrectly detected as an anomaly. Although the DBSCAN achieves better results than the three-sigma rule, MD and k-mean clustering, it also has the limitation that the performance is sensitive to the parameter arrangement. As indicated in [35], the clustering quality of the DBSCAN is poor when the density of the dataset is unbalanced, which means that the DBSCAN is unable to perform adaptive abnormal data detection based on the dataset characteristics.

V. Conclusion

To improve data qualities, this article develops a novel abnormal MHM data detection method including ASW and WMLOF. An ASW was first proposed to avoid data leakage and redundancy caused by a sliding window with an invariable length, and several optimally divided segments are obtained. Then, a WMLOF was used to extract the outlier features of ASW-based objects and evaluate the anomaly degree thanks to the effectiveness of the WMLOF in assessing and fusing the LOF characteristics at multiple scales. The simulation data of a faulty rolling bearing and measurement data collected from a railway vehicle gearbox and bench test were employed to evaluate the availability of the proposed method. The results demonstrate that the proposed abnormal data detection method based on the ASW and WMLOF strategies can achieve good performance in detecting the typical anomalies of data missing and drift even the degree of anomaly was very weak. Compared with the LOF, NOF, three-sigma rule, MD, and clustering based algorithms, the proposed method gives better outlier detection ability. Although the proposed method was shown to be beneficial to data-driven MHM research, determining how to reconstruct the detected abnormal data into usable data was a more challenging study. Besides, we were considering enhancing the computational efficiency of the WMLOF algorithm to meet the timeliness demands for practical engineering problems.

References

[1] Y. Lei, F. Jia, J. Lin, S. Xing, and S. X. Ding, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical Big Data," *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3137–3147, May 2016.

[2] X. Xu, Y. Lei, and Z. Li, "An incorrect data detection method for Big Data cleaning of machinery condition monitoring," *IEEE Trans. Ind. Electron.*, vol. 67, no. 3, pp. 2326–2336, Mar. 2020.

[3] O. Avci et al., "A review of vibration-based damage detection in civil structures: From traditional methods to machine learning and deep learning applications," *Mech. Syst. Signal Process.*, vol. 147, 2021, Art. no. 107077.

[4] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mech. Syst. Signal Process.*, vol. 138, 2020, Art. no. 106587.

[5] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, "Real-time motor fault detection by 1-D convolutional neural networks," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7067–7075, Nov. 2016.

[6] M. M. Najafabadi, F. Villanustre, T. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in Big Data analytics," *J. Big. Data*, vol. 2, no. 1, pp. 1–21, Feb. 2015.

[7] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for Big Data processing," *J. Adv. Signal Process.*, vol. 67, pp. 1–16, Dec. 2016.

[8] I. Taleb, R. Dssouli, and M. A. Serhani, "Big data pre-processing: A quality framework," in *Proc. IEEE Int. Congr. Big Data*, 2015, pp. 191–198.

[9] W. Fan, F. Geerts, and J. Wijsen, "Determining the currency of data," *ACM Trans. Database Syst.*, vol. 37, no. 4, pp. 1–46, Dec. 2012.

[10] A. Wahyudi, G. Kuk, and M. Janssen, "A process pattern model for tackling and improving Big Data quality," *Inf. Syst. Front.*, vol. 20, pp. 457–469, Jan. 2018.

[11] D. Hawkins, *Identification of Outliers*. London, U.K., Chapman and Hall, 1980.

[12] V. Barnett and T. Lewis, *Outliers in Statistical Data*. New York, NY, USA: Wiley, 1996.

[13] R. Domingues, M. Filipponea, P. Michiardi, and J. Zouaoui, "A comparative evaluation of outlier detection algorithms: Experiments and analyses," *Pattern Recognit.*, vol. 74, pp. 406–421, Sep. 2017.

[14] Y. Lei, X. Zhou, X. Xu, and F. Jia, "A dirty data recognition method for machinery condition monitoring in Big Data era," in *Proc. IEEE 43rd Annu. Conf. Ind. Electron. Soc.*, 2017, pp. 7061–7066.

[15] T. Johnson, I. Kwok, and R. T. Ng, "Fast computation of 2-dimensional depth contours," in *Proc. 4th Int. Conf. Knowl. Discov. Data Mining*, 1998, pp. 224–228.

[16] X. Dang and R. Serfling, "Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties," *J. Statist. Plan. Inference*, vol. 140, no. 1, pp. 198–213, Jan. 2010.

[17] I. Hussain, "Outlier detection using nonparametric depth-based techniques in hydrology," *IJACSA*, vol. 11, no. 9, pp. 456–462, 2020.

[18] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," *VLDB J.*, vol. 8, pp. 237–253, Feb. 2000.

[19] J. Huang, Q. Zhu, L. Yang, and J. Feng, "A non-parameter outlier detection algorithm based on natural neighbor," *Knowl. Based Syst.*, vol. 92, pp. 71–77, Jan. 2016.

[20] J. M. Jobe and M. Pokojovy, "A cluster-based outlier detection scheme for multivariate data," *J. Amer. Statist. Assoc.*, vol. 110, no. 512, pp. 1543–1551, Jan. 2016.

[21] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, Jun. 2000.

[22] A. Emmott, S. Das, T. Dietterich, A. Fern, and W. K. Wong, "A meta-analysis of the anomaly detection problem," Aug. 2016, *arXiv:1503.01158v2*.

[23] J. Zhu et al., "Review and Big Data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data," *Annu. Rev. Control*, vol. 46, pp. 107–133, Oct. 2018.

[24] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2250–2267, Sep. 2014.

[25] G. H. Tzeng and J. J. Huang, *Multiple Attribute Decision Making: Methods and Applications*. New York, NY, USA: Springer-Verlag, Aug. 2011.

[26] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[27] P. Chen, "Effects of the entropy weight on TOPSIS," *Expert Syst. Appl.*, vol. 169, Apr. 2021, Art. no. 114186.

[28] P. Chen, "A novel coordinated TOPSIS based on coefficient of variation," *Mathematics*, vol. 7, no. 7, pp. 1–17, Jul. 2019.

[29] R. Kumar et al., "Revealing the benefits of entropy weights method for multi-objective optimization in machining operations: A critical review," *J. Mater. Res. Technol.*, vol. 10, pp. 1471–1492, Jan./Feb. 2021.

[30] H. P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Outlier detection in arbitrarily oriented subspaces," in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 379–388.

[31] G. Tao, Z. Wen, X. Jin, and Y. Yang, "Polygonisation of railway wheels: A critical review," *Railway Eng. Sci.*, vol. 28, pp. 317–345, Sep. 2020.

[32] Y. Ye, B. Zhu, P. Huang, and B. Pengt, "OORNet: A deep learning model for on-board condition monitoring and fault diagnosis of out-of-round wheels of high-speed trains," *Measurement*, vol. 199, 2022, Art. no. 11268.

[33] Q. Xie, G. Tao, B. He, and Z. Wen, "Rail corrugation detection using one-dimensional convolution neural network and data-driven method," *Measurement*, vol. 200, 2022, Art. no. 111624.

[34] Q. Zhu, J. Feng, and J. Huang, "Natural neighbor: A self-adaptive neighborhood method without parameter K," *Pattern Recognit. Lett.*, vol. 80, pp. 30–36, Sep. 2016.

[35] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.

Qinglin Xie (Student Member, IEEE) received the B.S. degree in engineering mechanics and the M.S. degree in vehicle engineering in 2018 and 2021, respectively, from Southwest Jiaotong University, Chengdu, China, where he is currently working toward the Ph.D. degree in vehicle operation engineering with the State Key Laboratory of Traction Power, Southwest Jiaotong University.

His research interests include condition monitoring and intelligent fault diagnostics of wheel/rail shortwave irregularity.

Gongquan Tao received the B.S. degree in engineering structure analysis and the M.S. and Ph.D. degrees in vehicle operation engineering from Southwest Jiaotong University, Chengdu, China, in 2011, 2013 and 2018, respectively.

Since 2019, he has been with the State Key Laboratory of Traction Power, Southwest Jiaotong University, where he is currently an Associate Professor. His research interests include wheel-rail interaction, wheel and rail health management, and fault diagnosis.

Chenxi Xie received the B.S. degree in engineering mechanics and the M.S. degree in vehicle engineering in 2016 and 2019, respectively, from Southwest Jiaotong University, Chengdu, China, where he is currently working toward the Ph.D. degree in vehicle operation engineering with the State Key Laboratory of Traction Power, Southwest Jiaotong University.

His research interests include vibration fatigue of metro vehicle bogie components under non-Gaussian excitation.

Zefeng Wen received the B.S. degree in mechanical engineering and the M.S. and Ph.D. degrees in vehicle operation engineering from Southwest Jiaotong University, Chengdu, China, in 1998, 2000 and 2006, respectively.

Since 2006, he has been with the State Key Laboratory of Traction Power, Southwest Jiaotong University, where he is currently a Full Professor and the Deputy Director. His research interests include wheel-rail interaction and vibration and noise.