Weili Fang, Yixiao Shao, Peter E.D. Love, Timo Hartmann, Wenli Liu

Detecting anomalies and de-noising monitoring data from sensors: A smart data approach

Detecting Anomalies and De-Noising Monitoring Data from Sensors: A Smart Data Approach

Weili Fang[1], Yixiao Shao[2], Peter E.D. Love[3], Timo Hartmann[1], Wenli Liu[2]

[1]Department of Civil and Building Systems, Technische Universität Berlin, Gustav-Meyer-Allee 25, 13156 Berlin, Germany

[2]School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan, 430074, China

[3]School of Civil and Mechanical Engineering, Curtin University, GPO Box U1987, Perth, Western Australia 6845, Australia

Corresponding Author: Wenli Liu, Email: liu_wenli@hust.edu.cn

Detecting Anomalies and De-Noising Monitoring Data from Sensors: A Smart Data Approach

Abstract: When monitoring safety levels in deep pit foundations using sensors, anomalies (e.g., highly correlated variables) and noise (e.g., high dimensionality) exist in the extracted time series data, impacting the ability to assess geotechnical and structural safety risks. Our research aims to address the following question: *How can we detect anomalies and de-noise monitoring data from sensors in real time to improve its quality and use it to assess geotechnical safety risks?* In addressing this research question, we develop a hybrid smart data approach that integrates Extended Isolation Forest and Variational Mode Decomposition models to detect anomalies and de-noise data effectively. We use real-life data obtained from sensors to validate our smart data approach while constructing a deep pit foundation. Our smart data approach can detect anomalies with a root mean square error and signal-to-noise ratio of 0.0389 and 24.09, respectively. To this end, our smart data approach can effectively pre-process data enabling improved decision-making and the management of safety risks.

Keywords: Anomaly, deep pit foundations, de-noise, detection, smart data, safety risks

1.0 Introduction

Developments in sensing and data-processing technologies have enabled the effective monitoring of engineering data during the construction of deep pit foundations enabling geotechnical safety risks to be examined in greater detail (Zhou *et al*., 2019a;b; Asadzadeh *et al*., 2020). The data extracted and transmitted from sensors is imperfect, containing anomalies and noise, which severely jeopardizes its accuracy and completeness, which can also be exacerbated by random disturbances (Bao *et al*., 2019; Nessa *et al.,* 2020; Li *et al*., 2021; Liu *et al*., 2022; Seites-Rundlett *et al*., 2022).

Anomalies do not comply with the expected patterns (e.g., data missing) and possess various characteristics, which are common when using sensors. They generally occur due to faults, transmission errors, or structural damage (Kromanis and Kripakaran, 2013; Yi *et al*., 2013). Consequently, anomalous data may provide false information for decision-making and determination of safety risks (e. g., assessing geotechnical conditions). Thus, monitoring data needs to be automated and accurately detect anomalies to ensure it is robust and relevant for risk assessment (Nguyen and Goulet, 2019).

However, detecting anomalies is challenging due to the high dimensionality and correlations between the extracted engineering data (Thudumu *et al*., 2020). As sensors continuously monitor data, it is unfeasible to inspect and detect it in real time manually. Thus, several machine-learning approaches, such as neural network classifiers and decision trees, have been proposed to detect anomalies in real-time (Zuo and Xiong, 2019; Huang *et al*., 2020). While such approaches can detect anomalies quickly, they depend on several labeled (normal/abnormal) databases for their identification (Ahmed *et al*., 2016). Furthermore, labeled samples often contain noise generated by external disturbances. In the case of data monitoring in deep pit foundations, such disturbances are attributable to movements and vibrations generated by plants and equipment.

To detect anomalies, supervised and unsupervised approaches have been adopted widely. Commonly used supervised that can detect anomalies with high levels of performance (i.e., low false alarm rate) are the Support Vector Machine (Bhavsar and Waghmare, 2013) and Random Forest (Hasan *et al*., 2014). The datasets that employ supervised approaches in complex engineering environments require high-quality labeling. But datasets are often incomplete, requiring labeling to be undertaken manually, which is a time-consuming process. Contrastingly, the unsupervised approach does not require labeling. Their use has been advocated for detecting anomalies in sensor data (Chen *et al*., 2017). However, the detection rate is always low, and false-positive rates are high (Chen *et al*., 2017). Typical techniques used to process signals are the Wavelet Transform, Fourier Transform, and Empirical Mode Decomposition (EMD), though each has limitations (Abbate *et al*., 1997; Urciuolo and Marta, 2008; Hou and Guo, 2020). For example, the Fourier transform is unsuitable for handling non-stationary, non-linear signals with frequency over time (Urciuolo and Marta 2008), and the components of EMD are prone to modal aliasing (Hou and Guo 2020).

Against this contextual backdrop, our research addresses the following question: *How can we detect anomalies and de-noise monitoring data in real-time to improve its quality and use it to assess geotechnical safety risks?* In addressing this research question, a hybrid smart data approach that integrates the EIF and Variational Mode Decomposition (VMD) models is proposed to effectively detect anomalies and de-noise monitoring data to improve its quality to assess safety risks. The Extended Isolation Forest (EIF), an unsupervised anomaly detection algorithm, can detect anomalies and performs comparably to supervised algorithms (Carrera *et* *al*., 2022).

The basic idea of EIF is similar to the Isolation Forest (IF), which does not rely on building a profile for data to find non-conforming samples and remedies the shortcomings of the IF, which arise due to biases in the way the branching of the trees takes place (Hariri *et al*., 2021). The EMD is an algorithmic method to detect and decompose a signal into principal "modes" and is widely used in various time-frequency analysis applications. The EMD method is adaptive and applicable to non-linear and non-stationary processes (Huang *et al.*, 1998). The variational mode decomposition (VMD) (i.e., a non-recursive and noise robustness multi-resolution decomposition method) has better noise robustness performance than EMD in the application of vibration signal decomposition (Li *et al*., 2021). Furthermore, when compared with EMD, problems such as modal aliasing and end-point effects are better avoided (Cai *et al*., 2022), so the method is introduced in geotechnical engineering monitoring.

Our research commences by reviewing existing studies on anomaly data detection and denoising (Section 2). We then present a novel smart data approach that integrates EIF and VMD models to address anomalies and de-noise monitoring data to improve the ability to assess safety risks (Section 3). Our approach is smart as we focus on extracting *only* relevant engineering data for making decisions about geotechnical safety risks (Matthews *et al*., 2022). Next, the feasibility and effectiveness of our proposed approach are presented (Section 4). We subsequently discuss the implications of our approach and identify its limitations (Section 5) before submitting our conclusions (Section 6).

2.0 Monitoring Data from Sensors

The quality of engineering data obtained from sensors plays a pivotal role in monitoring the safety conditions in construction, especially in hazardous areas such as deep pit foundations. The detection of anomalous behavior from sensor data has received considerable attention in the literature. However, within the context of construction operations, research has been limited, though the problem of anomalies and noise reduction remains akin to other applications (e.g., Rabatel *et al*., 2011; Ahmed *et al*., 2016; Domingues *et al*., 2018; Hu *et al*., 2019).

2.1 Detecting Anomalies and Noise

Many approaches have been designed and developed to detect anomalies and are reported in the extant literature (Hill and Minsker, 2010; Cha and Wang, 2018). Existing sensor measurement approaches can be divided into three categories: (1) rule-based; (2) supervised learning-based; and (3) unsupervised learning-based (Huang *et al.,* 2017; Cha and Wang, 2018; Huang *et al*., 2020; Gao *et al*., 2022). For example, Mu and Yuen (2015) formulated an outlier-resistant extended Kalman filter to detect outliers caused by measurement errors. Similarly, Cha and Wang (2018) proposed an unsupervised anomaly-identification approach by modifying the original density-based fast clustering method. In this instance, Cha and Wang (2018) improved the ability to detect the location of structural damage by using a 'Gaussian kernel function of radius' to calculate the local density of data points. By the same token, under the assumption that measurement noise is Gaussian distributed, Huang *et al.* (2017) presented an anomaly-identification method in the noisy subspace of Principle Component Analysis. Examples of studies detecting anomalies from sensor measurement are shown in Table 1.

Notably, several challenges arise when using the above approaches to detect anomalies. For example, the rule-based approaches fail to recognize malicious events where no rules have been specified (Thottan and Ji 2003). Indeed, rule-based systems are restricted to only identifying events where rules exist. In the case of supervised learning, training data needs to be labeled, and algorithms cannot be used if this is not the case (Chandola *et al*., 2009; Ahmed *et al*., 2016). However, unsupervised learning approaches can train unlabeled data (Otoum *et al*., 2018). Despite unsupervised learning addressing this problem, training has challenges (Ahmed *et al*., 2016). Most unsupervised machine learning approaches to detect anomalies are evaluated on relatively small datasets in other domains (Inoue *et al*., 2017; Otoum *et al*., 2018). Moreover, with data being unlabeled, normal and abnormal signals can become mixed, rendering it difficult to demarcate the boundary between them. Therefore, normal data may contain anomalies in some specific scenarios.

Table 1. Examples of detecting anomalies in sensor measurement studies

Research approach	**Description**	**Author (Year)**
Pattern recognition neural network	Detection of multi-type data anomaly for structural health monitoring (SHM)	Gao *et al*. (2022)
Dynamic independent component analysis	Identification of two types of data anomalies in the SHM system of a cable-stayed bridge and then infer the structural damage	Huang *et al*. (2020)
Data visualization and deep learning network	Detect seven types of data anomalies in the SHM system of a long-span bridge	Bao *et al.* (2019)
Artificial neural network	A distributed similarity test and an artificial neural network were proposed to identify drift, spikes, and bias anomalies in wireless sensor networks	Fu *et al*. (2019)
Neural network	Estimate the state and detect the anomaly in a thermal power plant via a health monitoring system with multilayer perception	Banjanovic-Mehmedovic *et al*. (2017)
Autoregressive modelling and Kalman estimator	Detection of three types of data anomalies	Chang *et al*. (2017)

To this end, we aim to address the above challenges to develop an effective anomaly detection approach to improve the quality of engineering data extracted from sensors within deep pit foundations. Thus, to detect anomalies, we propose using the EIF described below.

2.1.1 Extended Isolation Forest

The EIF model was first proposed by Hariri *et al.* (2021). It extends the model-free anomaly detection algorithm, Isolation Forest (*i*Forest). The EIF extracts features from each monitoring dataset (e.g., the shaft force of steel shotcrete and building settlement) and builds a baseline model by creating an extended isolation forest tree collection. When new monitoring data is collected, it is mapped into each of these IFtrees, and an anomaly score is calculated. It will be defined as normal if its anomaly score is under a designated threshold value (Table 1). Otherwise, the monitoring data will be specified as abnormal.

*i*Forest samples *n* instances as a subset from the training dataset $\{\{X_1, X_2, \ldots, X_N\}$, where $\mathrm{X}_i = [\mathrm{X}_{i,1}, \mathrm{X}_{i,2}, \ldots, \mathrm{X}_{i,D}]^T$ denotes one *D*-dimensional data instance and then generates a binary tree from the root node. It randomly chooses one dimension from all *D* dimensions and randomly samples a split value from the uniform distribution $\mathrm{U}(\min_{i=1,\ldots,n} X_{i,d}, \max_{i=1,\ldots,n} X_{i,d})$.

Then the dataset is split into two parts: (1) $\{X_i | X_{i,d} < split\ value; i = 1, \ldots, n\}$ which is passed to the left branch of the node; and (2) $\{X_i | X_{i,d} \geq split\ value; i = 1, \ldots, n\}$ which is passed to the right branch of the node. The procedure is repeated iteratively to create each node of the tree until only one distinct instance remains in one node or reaches the height limit. The EIF used an axis-obliqued splitting method to solve the issue that the split process of the original IF will generate artifacts. Specifically, EIF creates a hyperplane with the form of a point-norm equation to split the data as shown in Eq. [1]:

$$(\mathrm{X} - \mathrm{P}) \cdot n^T = 0 \quad [1]$$

Where the point vector $\mathrm{P} = \{p_j | p_j \sim \mathrm{U}(\min_i X_{i,j}, \max_i X_{i,j}); i = 1, \ldots, n; j = 1, \ldots, D\}$, the norm vector $\mathrm{n} = \{n_j | n_j \sim N(0,1)\}$. Then the split rule becomes: $\{X_i | (\mathrm{X} - \mathrm{P}) \cdot n^T < 0; i = 1, \ldots n\} \rightarrow left\ branch;\ \{X_i | (\mathrm{X} - \mathrm{P}) \cdot n^T \geq 0; i = 1, \ldots n\} \rightarrow right\ branch.$

The *i*Forest algorithm assumes that the anomaly instances are rare. Such instances differ from those deemed normal in a given data set, making them more susceptible to isolation in several binary tree structures. In a random tree, instances are partitioned repeatedly until all instances are isolated. In contrast, nominal instances require many more splits to finally reach their leaf nodes (Li *et al*., 2020). For a given dataset, the algorithm takes *n* random samples of size *m*. A binary search tree is constructed for each random example, selecting a dimension and partition point for each comparison node in the tree. The anomaly score of a new data point is calculated by inserting it into each *n* random tree.

Isolation refers to the separation of an instance. Anomalous data has the nature of ‘few and special’, and it is easy to isolate outliers from normal data. The *i*Forest algorithm isolates data by recursively and randomly partitioning. Usually, normal data is typically dense and needs to be divided many times to be isolated. Conversely, abnormal data are outliers and only need to be randomly divided a few times to be isolated. In the whole process of isolation, a binary tree can represent the process of division. The earlier a point is divided, the more likely it is an abnormal point. An example of the partitioning process is presented in Figure 1, where the ‘red’ leaf node is most likely an outlier.

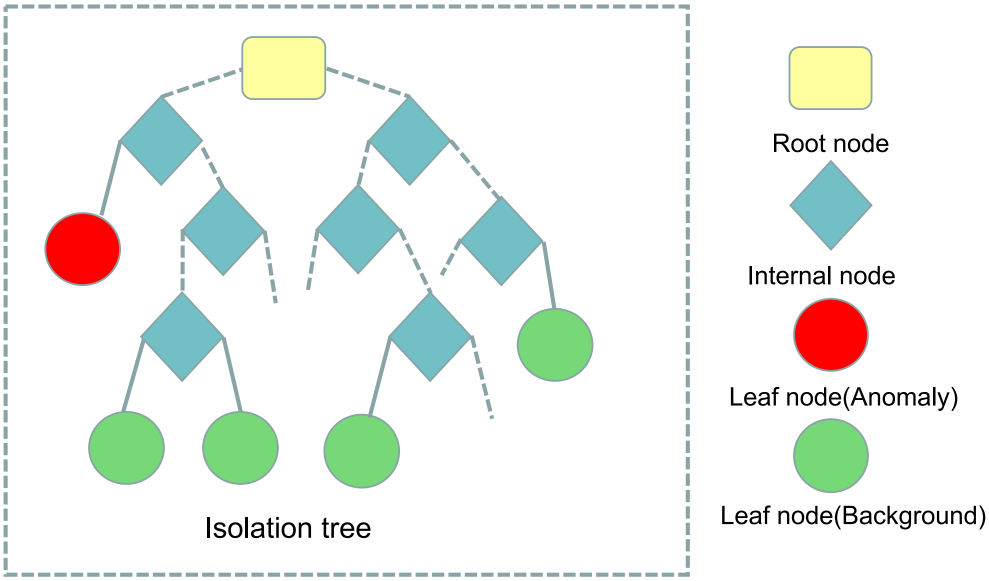


Figure 1. Example of the structure of an *i*Tree

The node (T) of the isolation tree is either external with no child or internal with one test and two daughter nodes (T_l, T_r), where the number of external nodes is n, the number of internal nodes is $n-1$, and the total number of nodes of an *i*Tree is $2n-1$ (Liu *et al*. 2012). A test consists of an attribute q and a split value p. Given a database $X = \{x_1..., x_n\}$ of n in instances from a d-variate distribution, to build an *i*Tree, we recursively divide X by randomly selecting an attribute q and a split value p, until either: (1) the tree reaches a height limit, (2) $|X| = 1$ or all data in X have the same values (Liu *et al*. 2012).

The Path length ($h(x)$) is determined by the number of edges x traverses an *i*Tree from the root node until the traversal is terminated at an external node. We borrow the analysis from Binary Search Tree (BST) to estimate the average path length ($E(h(x))$) of *i*Tree. The anomaly score (s) of an instance x is defined as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \qquad \text{Eq. [2]}$$

$$c(n) = 2H(n-1) - \left(\frac{2(n-1)}{n}\right) \qquad \text{Eq. [3].}$$

Specific details of the assessment process can be found in Liu *et al*. (2008).

2.2 De-noising Data`

Due to the spatial-temporal uncertainty and complexity of working conditions in deep pit foundations, raw monitoring data invariably contains noise. The noise will interfere with data analysis and decision-making accuracy in this instance. Methods such as low-pass filtering (De *et al*., 2010), Wiener filtering (Aschero *et al*., 2010), adaptive learning (Ortolan *et al*., 2003), and Kalman filtering (Singh *et al*., 2018) are traditionally used to de-nose signals. Despite their success, such approaches have limitations, as they filter out useful information or reduce valuable features (Andrate *et al*., 2006; Xiao *et al*., 2019).

To address the above limitations, a wavelet transforms a time-frequency domain method has been introduced to de-noise signals (Andrade *et al*., 2006; Maier *et al*., 2018). This method comprises three steps: (1) signal decomposition; (2) detail coefficient thresholding; and (3) signal reconstruction. When a wavelet transform is used, there is no requirement to incorporate artificial components into the original signal (Andrade *et al*., 2006; Maier *et al*., 2018). However, the limitation of such an approach is that the mother wavelet function must be pre-defined. Additionally, the selection of different mother wavelet functions can affect detection performance (Maier *et al*., 2018; Xiao *et al*., 2019).

Monitoring data is often mixed with multi-type noise. Thus, a novel signal analysis method is needed to decompose a multi-component signal into several band-limited intrinsic mode functions (BLIMFs). The VMD effectively determines the signal segmentation in the frequency domain and the components' separation. It has also been proven to simultaneously achieve accurate signal separation, better noise robustness, and higher computational efficiency. With this in mind, we will use the VMD in this research to de-noise the data.

Several studies have attempted to denoise the time series data using the traditional method for solving the signal denoising problem, which involves using linear time-invariant (LTI) filters (Selesnick *et al*., 2014; Prateek *et al*., 2021). An alternative approach uses wavelets; the main drawback of this approach is that it introduces pseudo-Gibbs artifacts at the singular points due to more local oscillations and smaller amplitude near signal discontinuities. And the sparsity-based methods, such as compressed sensing with dictionary elements from an oversampled discrete Fourier transform (DFT) matrix, cannot reconstruct the signal perfectly. The modal decomposition algorithm handles non-linear and non-smooth signals with good adaptive decomposition capability. It can decompose complex signals into intrinsic modal function forms sorted by frequency from high to low and extract the decomposed modal function to construct a filter. As a modal decomposition algorithm, VMD is selected in our research as it can separate tones of similar frequencies to represent time series characterization.

2.2.1 Variational Mode Decomposition

The VMD proposed by Dragomiretskiy and Zosso (2013) is a non-recursive decomposition method used for adaptive and quasi-orthogonal signal decomposition. It can simultaneously decompose a multi-component seismic trace into a finite number of band-limited intrinsic mode functions (IMFs). The VMD generalizes the classic Wiener filter into multiple adaptive bands. Wiener filtering is one of the most ubiquitous tools in signal processing, particularly for signal denoising and source separation. In the context of audio, it is typically applied in the time-frequency domain using the short-time Fourier transform (STFT) (Samuel and James, 2008). The VMD algorithm is more robust to noise than the EMD-based adaptive decomposition methods (Dragomiretskiy and Zosso, 2013). The concepts and theories related to VMD are as follows.

Definition 1: (Intrinsic Mode Function)

Intrinsic Mode Functions are amplitude-modulated-frequency-modulated (AM-FM) signals, which differs from the definition of EMD.

$$\mu_k(t) = A_k(t)\cos(\emptyset_k(t)) \qquad [4]$$

Where the phase $A_k(t)$ is an envelope of $\mu_k(t)$ and $\emptyset_k(t)$ is a non-decreasing function. The equation of phase $\emptyset_k(t)$ and instantaneous frequency $\omega_k(t)$ is as follow:

$$\omega_k(t) = \frac{d\emptyset_k(t)}{dt} \geq 0 \qquad [5]$$

Definition 2: (Total Practical IMF Bandwidth)

The total practical bandwidth of an IMF is estimated as Eq. [6]. Depending on the actual IMF, either of these terms may be dominant.

$$BW_{AM-FM} = 2(\Delta f + f_{FM} + f_{AM}) \qquad [6]$$

The workflow of the VMD, which we will follow, is presented in Figure 2.

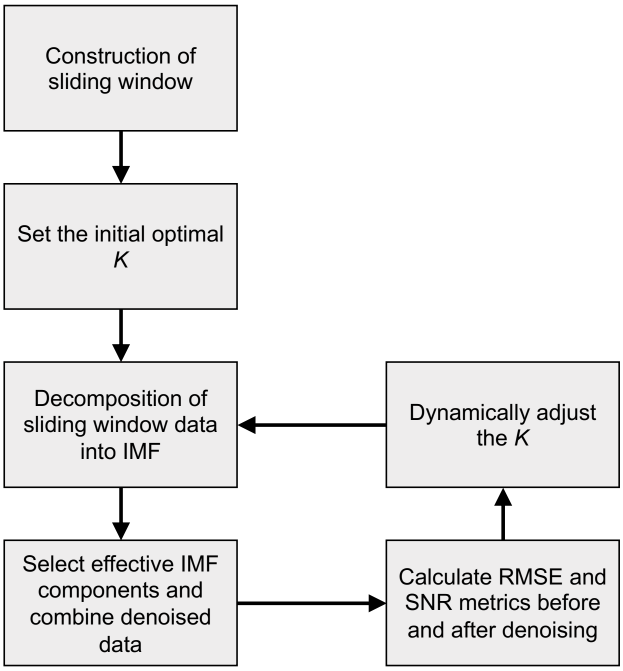


Figure 2. Workflow of implementation of VMD

The VMD comprises the following steps:

Step 1: modes update. The modes $(\hat{u}_k^{n+1}(w))$ are updated by Eq. [7], where α is the penalty factor, λ is the Lagrangian multiplier, μ_k is the IMF. Wiener filtering is embedded for an update as the mode directly in the Fourier domain with a filter tuned to the current center frequency w_k^n;

$$\hat{u}_k^{n+1}(w) = \frac{\hat{f}(w) - \sum_{i<k} \hat{u}_i^{n+1}(w) - \sum_{i>k} \hat{u}_i^{n}(w) + \left(\hat{\lambda}^n(\omega)/2\right)}{1+2\alpha\left(\omega-\omega_k^n\right)^2} \quad \text{Eq.[7]}$$

Step 2: Center frequencies update. The center frequencies are updated as the center of gravity of the corresponding mode's power spectrum, as shown in Eq. [8]

$$\omega_k^{n+1} = \frac{\int_0^{\infty} \omega \left|\hat{u}_k^{n+1}(w)\right|^2 d_\omega}{\int_0^{\infty} \left|\hat{u}_k^{n+1}(w)\right|^2 d_\omega} \qquad \text{Eq. [8]}$$

Step 3: Dual ascent update. For all $\omega \geq 0$, the Lagrangian multiplier $\hat{\lambda}^{n+1}$ is updated by Eq. [9] as a dual ascent to enforce exact signal reconstruction until $\sum_k \|\hat{u}_k^{n+1}(w) - \hat{u}_k^{n}(w)\|_2^2 / \|\hat{u}_k^{n}\|_2^2 < \varepsilon$.

$$\hat{\lambda}^{n+1} = \hat{\lambda}^{n} + \tau\left(\hat{f} - \textstyle\sum_k \hat{u}_k^{n+1}\right) \qquad \text{Eq. [9]}$$

Additional details about the VMD can be found in the works of Dragomiretskiy and Zosso (2013).

3.0 Research Approach

To recap, our research aims to develop a smart data approach to detect anomaly monitoring data and reduce noise to improve the quality of monitoring data extracted during the construction process of hazardous activities such as deep pit foundations. Our smart data approach extracts data relevant for decision-making to determine safety risks and consists of EIF and VMD. In the process of data collection, it is inevitable to produce some data that deviates from the rest of the observations in the sample to which it belongs. The reasons mainly include: (1) the failure of the equipment; (2) the abnormality of the collected data caused by the dynamic working environment.

To obtain a high-quality time-series monitoring dataset, it is necessary to perform abnormal processing on the data. While noise and outliers are similar in their statistical distribution and characteristics, they originate from fundamentally different causes. The workflow of our proposed method is presented in Figure 3. The research process we have adopted to develop our hybrid smart data approach consists of the following three steps (Figure 3):

- *Step 1 - Data segmentation*: Extracted monitoring data is divided into segments using a rectangle sliding window. Then, various numerical features, such as root mean square (RMS) and kurtosis of each data window, are determined.
- *Step 2(a) - Extended isolation forest construction*: The EIF is an outlier detector that builds an ensemble of *i*Trees for a given dataset. The EIF resolves the issues associated with assigning anomaly scores to given data points by using hyperplanes with random slopes (non-axis-parallel) to split data to create *i*Trees (Hariri *et al*., 2021)
- *Step 2(b) - Data anomaly detection*: In an EIF, data are subsampled and processed in a tree structure based on random cuts in the values of arbitrarily selected features in each dataset. Each tree is grown until each instance is isolated into a leaf node. The samples with shorter branches indicate anomalies.
- *Step 3 - Variational model decomposition construction:* As an adaptive signal processing method, the VMD removes harmonic noise and improves data quality. The VMD algorithm concurrently decomposes the input signal into several narrow-band modes. Each mode is band-limited around its center frequency, which leads to less spectral overlapping or instantaneous frequency fluctuation is observed in the VMD results. The VMD algorithm decomposes and reconstructs the monitoring data to achieve adaptive signal decomposition and noise reduction.

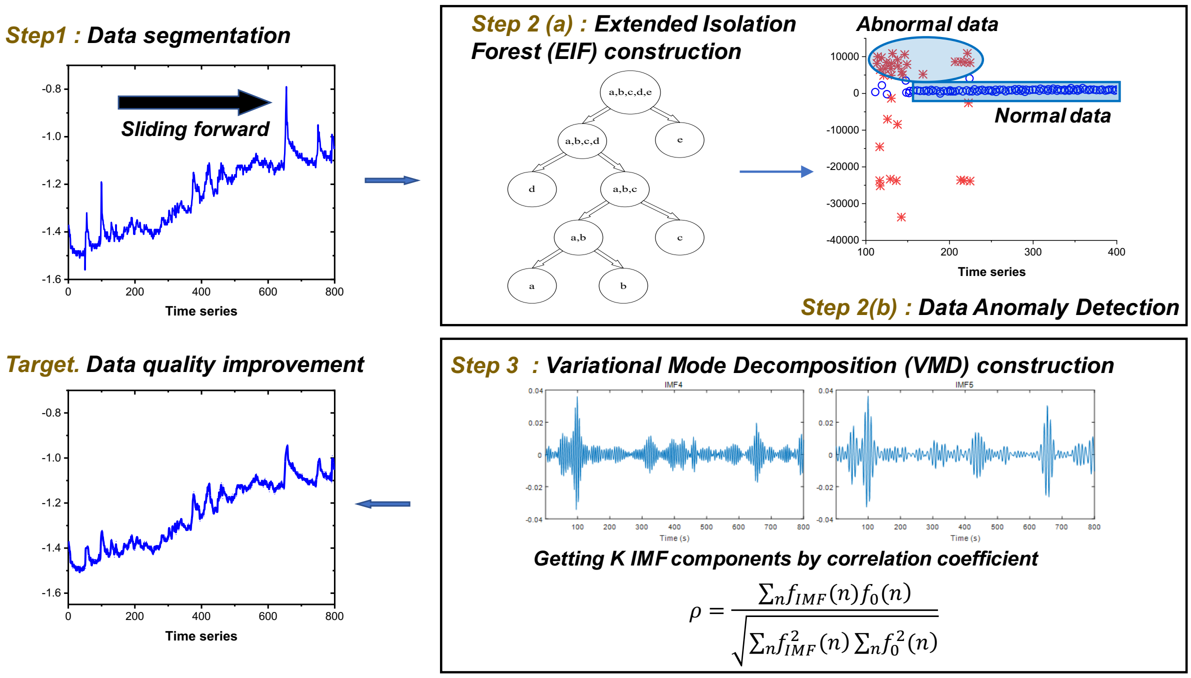


Figure 3. The workflow for the research

We now explain in greater detail our research approach for detecting anomalies and de-noising the engineering data extracted from sensors used to monitor deep pit foundations by focusing on EIF and VMD.

4.0 Case Study

We use an explanatory case study to demonstrate and validate our hybrid smart data approach (Dubé and Paré, 2003). A deep foundations pit of a subway project in Wuhan, China, while under construction, is selected. The project was chosen as sensors were used to monitor geotechnical safety risks, and the researchers worked closely with contractors on several other studies.

4.1 Case Description

The selected subway project is a T-shaped transfer between stations A and B. Subway station A is an underground three-story double-column 13m island platform station. The total length of the station is 239.2m, the full width of the standard section is 22.5m, the structure height is

22.63m-25.08m, and the roof is buried deep about 3m-4.1m, both ends of the station are shield tunnel receiving wells. Subway station B is a 14-meter island-style station with two underground floors and two columns. The total outsourcing of the station is 634.105m, and the full width of the standard section is 23.1m. The landform can be classified as a denudation accumulation ridge area (grade III terrace), and the ground elevation of the exploration area is between 26.0 and 30.7m.

Figure 4. Example of deep pit foundation

4.2 Experimental Set-up

We first install sensors while excavations are constructed to conduct this experiment, focusing explicitly on the supporting shaft's axial forces and building settlement. The layout of these sensors is presented as follows:

- *Support shaft axial force monitoring*: The axial force meter monitors the axial force of steel support. The meter is installed at the end of the steel support. In this case, four axial force monitoring points are established on the fourth and fifth layers of steel supports, with two monitoring points on each layer. The monitoring equipment used is the Vibrating

String Axial Force Meter produced by the Shenzhen JingSheng Tech Co., LTD, model MAS-AXF-40. The maximum range is 4000KN, a MAS of 0.1%, and a precision of ± 0.5.

- *Building settlement monitoring*: Based on the experience of experts, four-building settlement monitoring points were installed symmetrically on the four corners of the building, closest to the foundation pit. The monitoring equipment used is Photoelectric Static Level produced by TongWei Sensing, model ESJS-50. The measuring range is 50mm, and the precision is ± 0.1mm.

Examples of sensors installed in the case are presented in Figure 5. After the sensors are installed, the data are transmitted and stored in the web-based monitoring system, as shown in Figure 6.

Figure 5. Examples of sensors installed in foundation pit

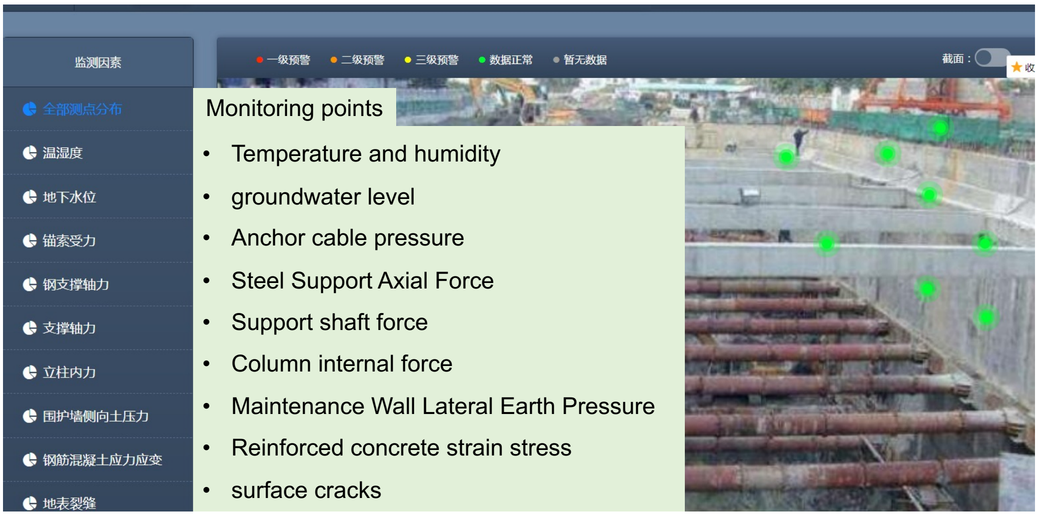


Figure 6. Web-based monitoring data system

4.3 Anomaly Monitoring Data Detection

The four installed sensors act as the monitoring points and our data source, one of which is used as the baseline, and the other three monitoring points (CJ1, CJ2, CJ3) are for analysis. As noted in Figure 7a, the settlement data has no apparent abnormality; The anomaly data detection under one (i.e., CJ1), two (i.e., CJ1 and CJ2, and three (CJI CJ2, and CJ3) dimensional analysis. An anomaly value is calculated for each point under different training set sizes in anomaly data detection. The distribution of anomaly value is used to analyze the effect of dimension selection and dataset size on anomaly detection. The results of anomaly monitoring data detection under different training sets and dimensional analysis are presented in Figure 7b.

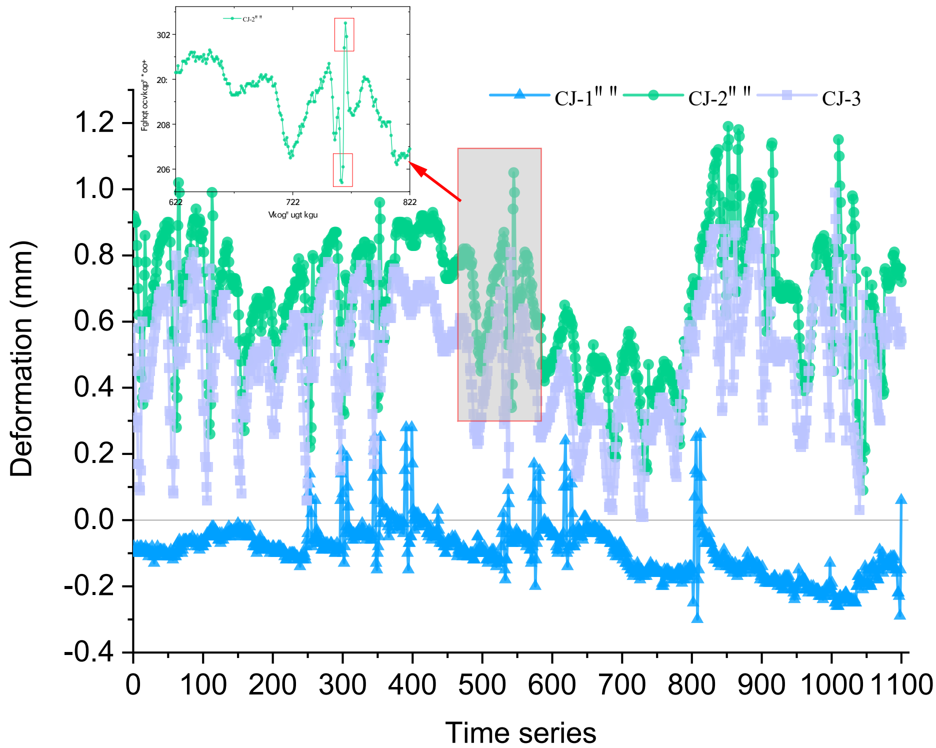
Here we define the data with anomaly scores higher than 0.6 as outliers and analyze the anomaly scores of the outliers. Figure 7b shows that one-dimensional data has a higher anomaly score than two-dimensional and three-dimensional data. As the sizes of the training set change, the anomaly scores of the one-dimensional data also change, but these values generally exceed

0.78. Conversely, the anomaly scores of two-dimensional and three-dimensional data are lower than 0.74. The maximum value of the anomaly score of the two-dimensional data fluctuates between 0.71 and 0.74. The maximum value of the anomaly score of the three-dimensional data ranges between 0.69 and 0.71. The anomaly scores for outliers in high-dimensional datasets are more concentrated with lower anomaly scores. Thus, we can conclude that the *i*Forest algorithm can process high-dimensional data (i.e., settlement monitoring data). The detection of outliers is smoother than the low-dimensional data, and the abnormal value is relatively lower.

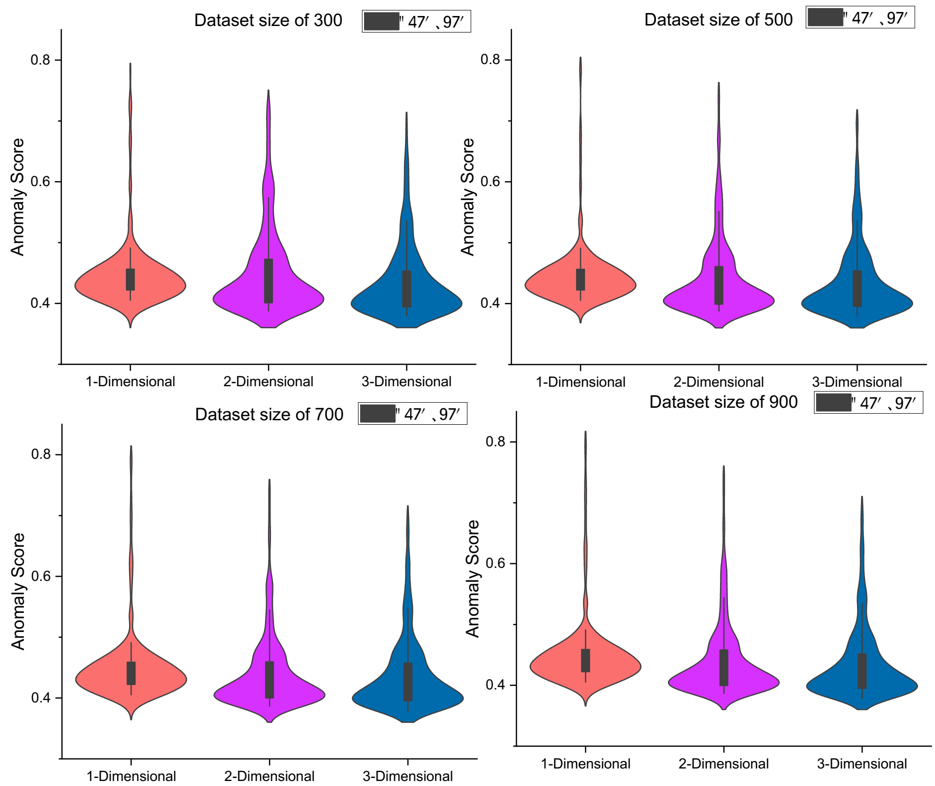
Again, four monitoring points are selected to analyze the steel support axial forces (ZCL-02-21, ZCL-02-22, ZCL-04-C6, ZCL-04-C7). Our results are presented in Figure 8a, and we can conclude that the monitoring data of ZCL-04-C6 is abnormal. As a result, we then analyzed the anomaly data detection under different dimensional conditions, with the results being presented in Figure 8b. As seen from Figures 8 and 9, in the detection data of the *i*Forest, the higher the dimension of monitoring data, the less sensitive it is to detecting anomalies. We can find the difference between single and multi-dimensional anomalies by analyzing high-dimensional monitoring data. The higher the dimension of monitoring data, the higher the anomaly value, and the easier it is to determine the cause of the abnormality.

4.4 Evaluation Performance

We compared the EIF with the *i*Forest algorithm to determine which method can better identify abnormal points. Figure 9 shows the two-dimensional abnormal point detection of the steel support axial force monitoring data using EIF and a standard *i*Forest algorithm. The left and right columns are the standard isolation forest and EIF algorithms.

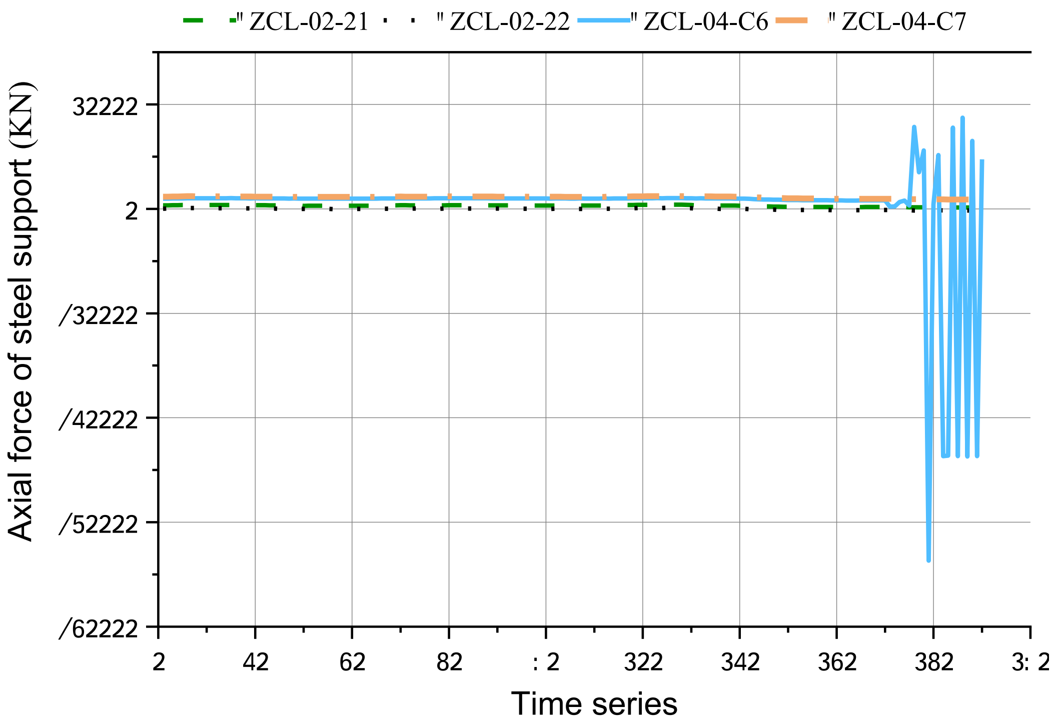


(a) Examples of monitoring data

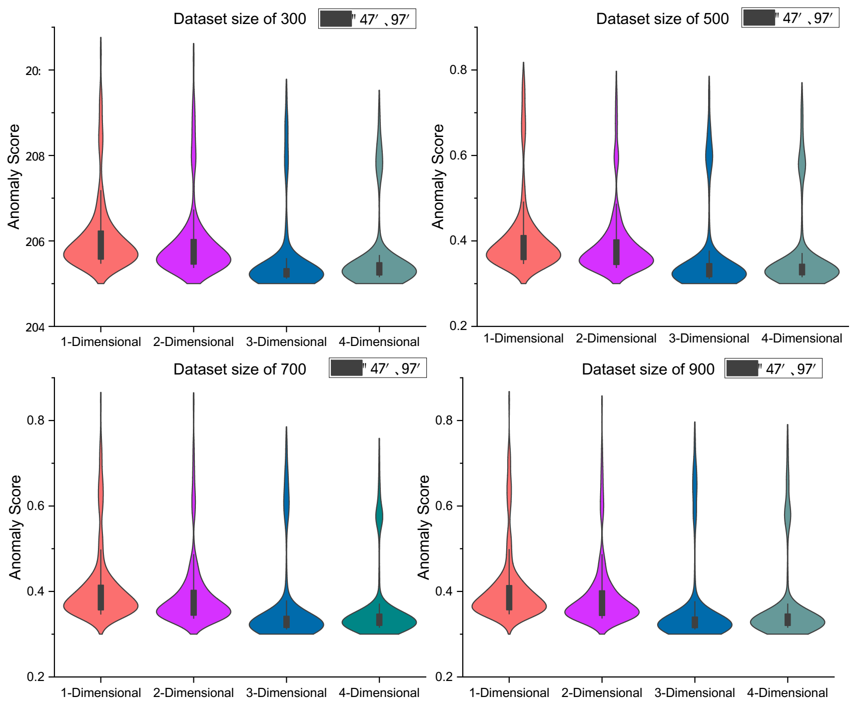


(b) Settlement data: different dimensional and training set sizes

Figure 7. Examples of monitoring and settlement data



(a) Examples of monitoring data



(b) Steel support axial force: different dimensional and training set sizes

Figure 8. Examples of monitoring steel support axial force data

Figure 9 shows that when the anomaly score is 0.75, all the anomalous data cannot be identified. We suggest that the reason is that the training set in this figure includes many abnormal data, and the data selection for model training cannot directly take the sliding window of a time point.

4.4.1 Quality of Training Database

The detection of anomalies using the EIF algorithm consists of two phases:

1. *Training*: An isolated tree is built based on subsamples of the training set;
2. *Testing*: An isolated tree calculates anomaly scores for each test sample. Hence, we design two group experiments to conduct this test: (i) a training database; and (ii) a training dataset without anomaly monitoring data. The size of the training database of these two group experiments is set to 800. The results are presented in Figure 10.

Figure 10 shows that the training set that excludes abnormal monitoring data achieves better performance on abnormal data detection. We process the data by setting a threshold for the anomaly score; that is, data with an anomaly score over 0.7 will be removed. During data analysis, we conclude that the different performance is from the abnormal points not excluded, leading to the other points being no longer 'isolated' as the other data set. Therefore, in detecting anomalies in *i*Forests, the quality of the training set should be maintained.

4.4.2 Size of the Training Database

In this experiment, the training set size ranges from 300 to 800. It is trained with a gradient of 100, with the abnormal detection results presented in Figure 11. Here we can see that with increased increments in the training set, the algorithm achieves better performance on anomaly detection until the 600 mark, where almost all anomalies are detected.

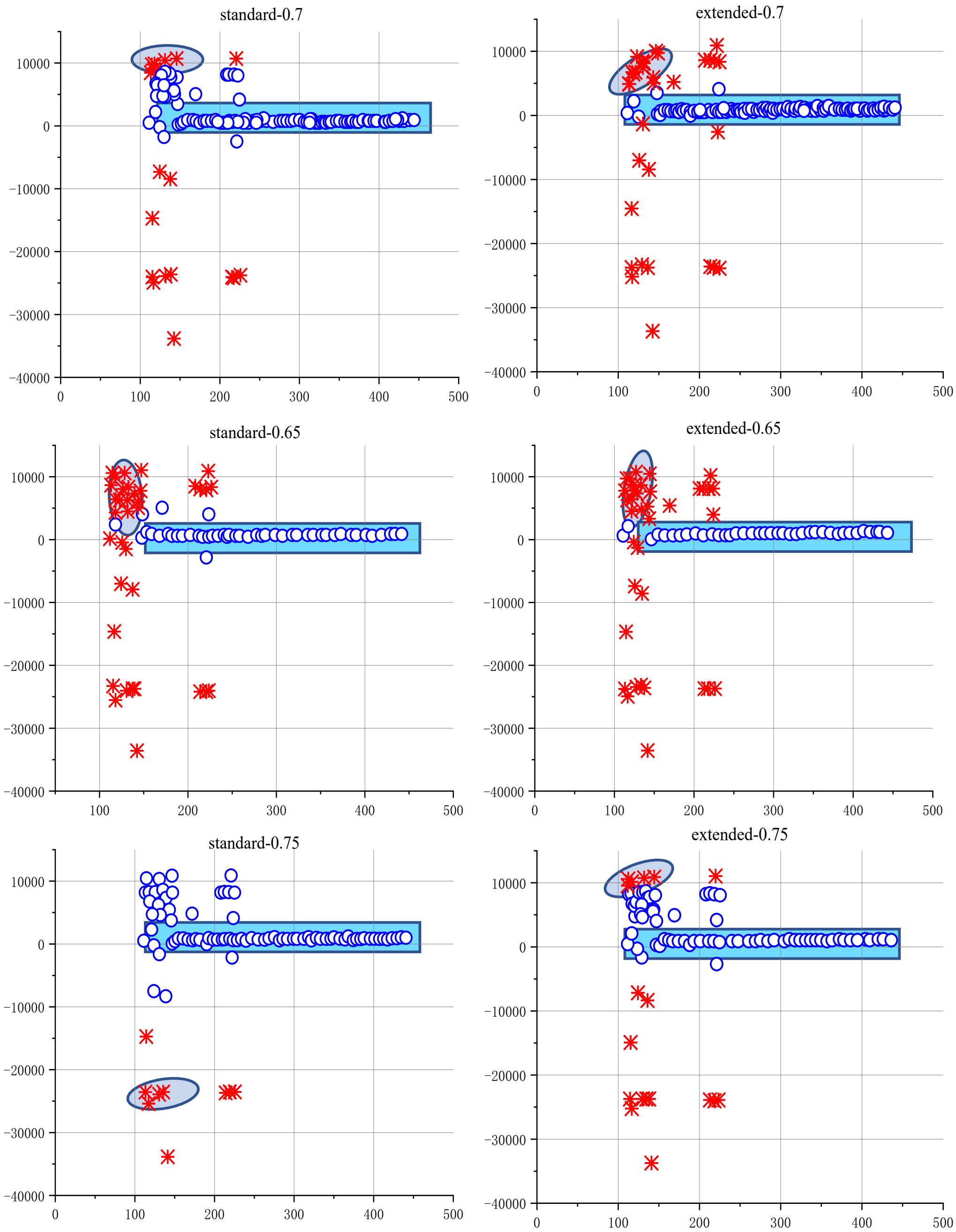


Figure 9. The comparison results of EIF and *i*Forest algorithms

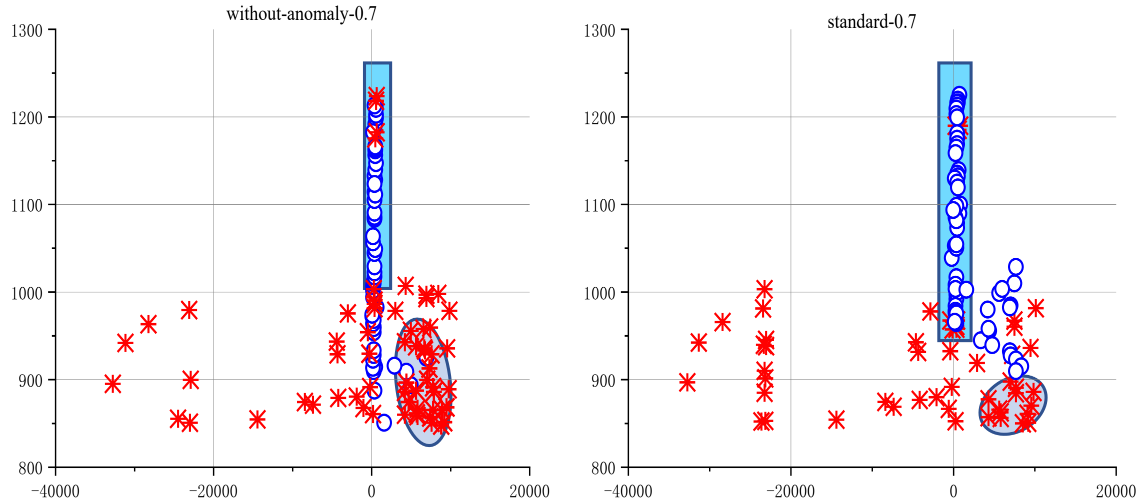


Figure 10. The comparison result of using two training databases

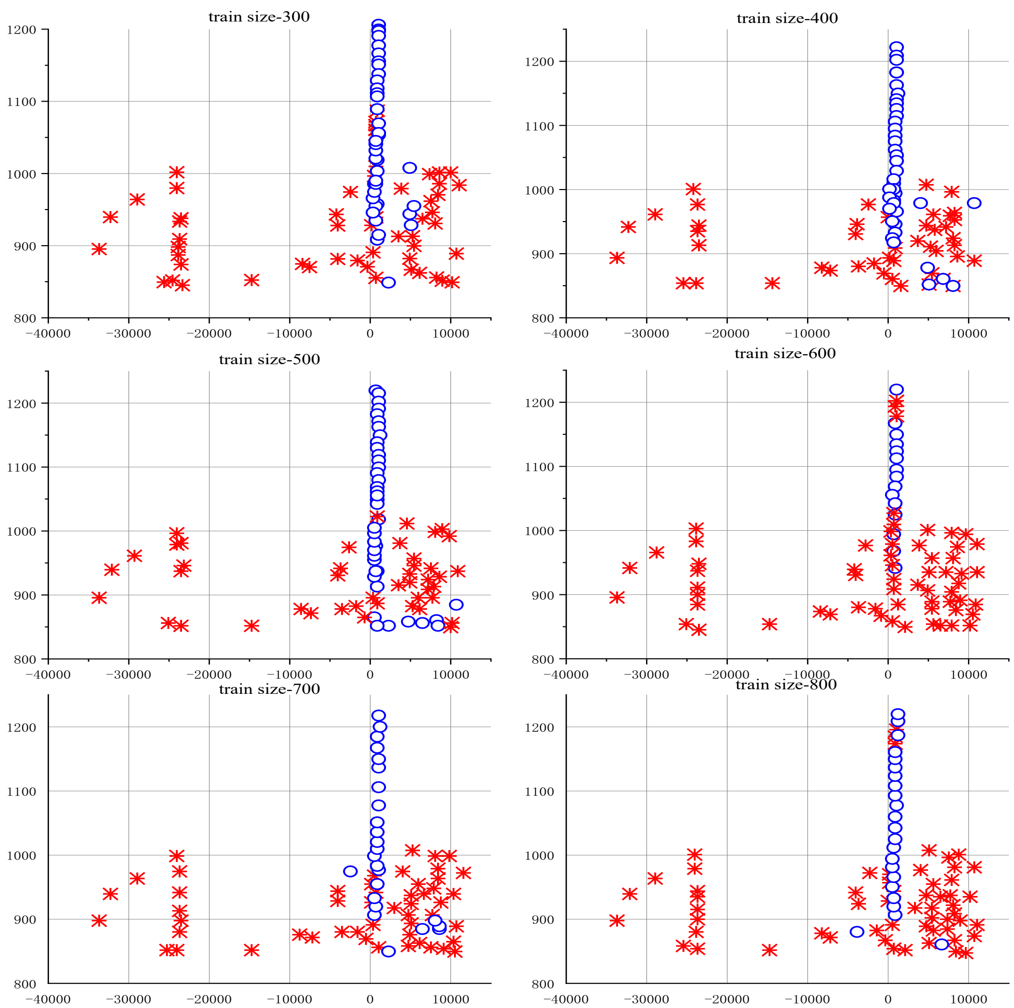


Figure 11. A comparison of the results for different training databases

4.4.3 Effectiveness of EI*F*

We use the steel support axial force dataset to verify the effectiveness of the EIF algorithm compared with the KNN algorithm and the ABOD (angle-based outlier detection) algorithm. We calculate the results using a two-dimensional dataset and a four-dimensional dataset, respectively, for analysis. The evaluation measures are area under the curve (AUC) and Accuracy. The AUC is an index normally used to evaluate the efficiency of classifiers, defined as the area under the receiver operating characteristic (ROC) curve. Accuracy is the proportion of instances in the monitoring data detected correctly. The results are shown in Table 2.

Table 2. AUC and Accuracy of EIF, KNN, and ABOD

Algorithm	**AUC**	**Accuracy**
ABOD(2-Dimensional)	0.8629	89.5
EIF(2-Dimensional)	0.8893	94.4
KNN(2-Dimensional)	0.8827	93.7
ABOD(4-Dimensional)	0.8964	94.9
EIF(4-Dimensional)	0.8997	96.3
KNN(4-Dimensional)	0.8634	89.6

As can be seen from Table 2, the AUC and Accuracy of the four algorithms are basically consistent, and there is no noticeable difference. The accuracy of EIF and AMOD improved with the increase of the dataset dimension, while the accuracy of KNN decreased. The experimental results show that the EIF algorithm can effectively improve the execution efficiency of anomaly detection with a high-dimensional dataset. Therefore, EIF is suitable for anomaly detection on large-scale monitoring data.

4.5 Monitoring Data De-noising

We select a dataset containing 1000 monitoring points based on the period and frequency of data collection for denoising sample data. Some studies found that the decomposition number K significantly influences the decomposition results (Xia *et al*. 2021; Wang *et al*. 2019). When the number of decomposition modes (*K)* is too low, under-decomposition will occur, and some 'modes' cannot be recognized effectively (Xia *et al*. 2021; Wang *et al*. 2019). When *K* is too large, a particular 'mode' in the signal may be 'pulled' into multiple IMF components, resulting in excessive decomposition (Li *et al*., 2019). In this study, the number of decompositions includes 1 to 9 in advance to choose the best *K*. The waveforms of the nine IMFs decomposed by the VMD are presented in Figure 12. K is the number of modes. In the original EMD description, a mode is defined as a signal whose number of local extrema and zero-crossings differ at most by one. In later related works, the definition is slightly changed into so-called Intrinsic Mode Functions (IMF); through the decomposition of different *K* ensemble members, the correlation coefficient between each IMF component and the sample data is calculated, with the results being presented in Table 3.

From Table 3, we can conclude that when *K* is 2 and 3, the decomposed IMF components are valid according to the threshold value. It indicates that the dataset is underpinning, and the high-frequency noise is not isolated. So, these two modes are not analyzed later. We set the threshold as 0.1 of the value, which is the largest of all correlation coefficients. When the value of IMF is less than the threshold, we define the value as a failure. (Yu 2008) When *K* is 4~9, the effective IMFs are all 3. The first three IMF components are reconstructed. The reconstructed signal and the original data signal are calculated by the root mean square error (RMSE) and the signal-to-noise ratio (SNR). The RMSE and SNR are defined by Eq. [6] and Eq. [7]:

$$RMSE = \sqrt{\frac{1}{n}\sum_n (f_0(n) - f_1(n))^2} \qquad \text{Eq. [6]}$$

$$SNR = 10 \times \log_{10}\left(\frac{\frac{1}{n}\sum_n f_0^2(n)}{\frac{1}{n}\sum_n (f_0(n) - f_1(n))^2}\right) \qquad \text{Eq. [7]}$$

Where, f_0 is the original signal data, f_1 is the reconstructed signal data.

We can conclude that when K is 4, the RMSE is the smallest, and the signal-to-noise ratio SNR is the largest, so the denoising effect is the best, and the optimal K value should be selected as 4.

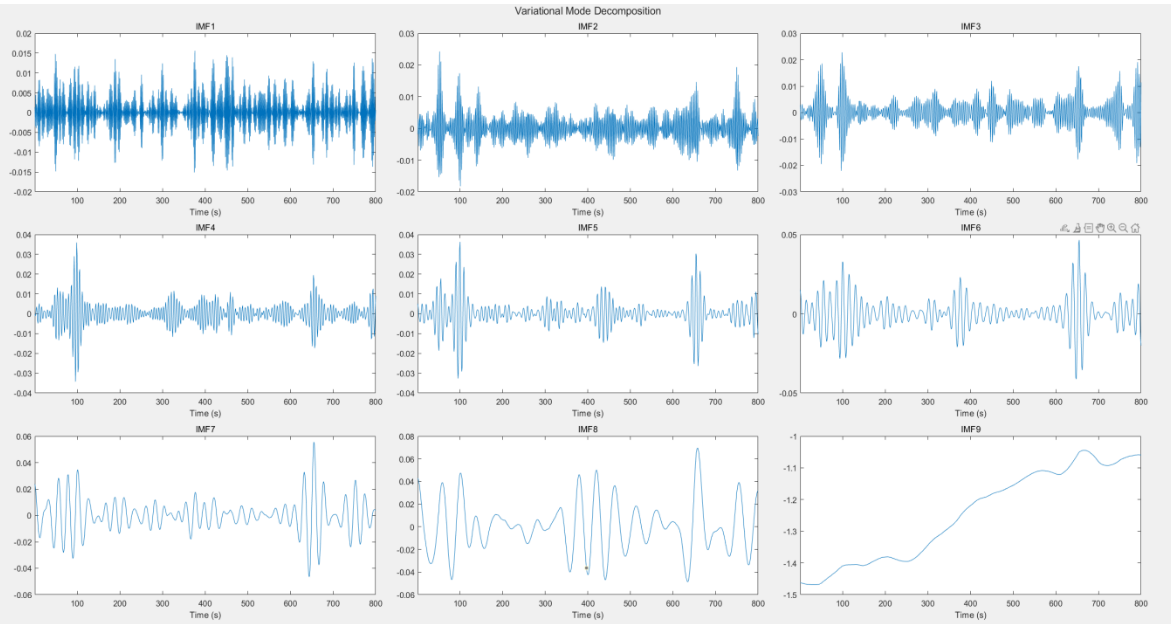


Figure 12. Waveforms of the nine IMFs decomposed by the VMD

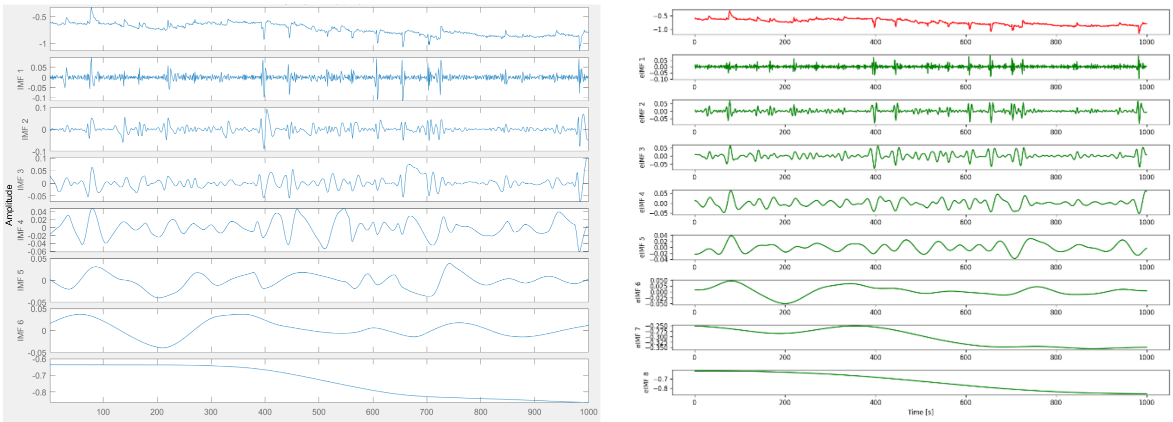
Table 3. Correlation coefficients between IMF and original signal data under different K ensemble members

K	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8	IMF9
2	0.987	0.156							
3	0.982	0.179	0.119						
4	0.981	0.181	0.115	0.090					
5	0.980	0.181	0.111	0.085	0.075				
6	0.980	0.183	0.108	0.083	0.076	0.065			
7	0.980	0.184	0.107	0.081	0.074	0.065	0.047		
8	0.980	0.185	0.106	0.077	0.067	0.067	0.058	0.044	
9	0.980	0.185	0.106	0.077	0.065	0.063	0.059	0.046	0.038

Table 4. RMSE and signal-to-noise ratio under different K values

K	4	5	6	7	8	9
RMSE	0.0389	0.0413	0.0438	0.0442	0.0453	0.0455
SNR	24.09	23.58	23.07	22.98	22.77	22.74

We use the same dataset to verify the effectiveness of the VMD algorithm compared with the EMD algorithm and the EEMD (ensemble EMD) algorithm. The evaluation measures are RMSE and SNR. The decomposition result of the EMD and EEMD for the signal is presented in Figure 13, and the results are shown in Table 5. As can be seen from the results, EEMD denoising is superior to EMD, and VMD denoising is better than EEMD, which has high SNR and low RMSE.



(a) (b)

Figure 13. Decomposition result by EMD(a) and EEMD(b)

Table 5. RMSE and signal-to-noise ratio of different algorithm

	VMD	EMD	EEMD
SNR	24.09	14.26	15.91
RMSE	0.0389	0.1573	0.1479

5.0 Discussion

It has been suggested that big data analytics provides the basis to identify patterns and derive insights about safety issues in construction (Guo *et al*., 2016; Fang *et al*., 2020; Fang *et al*., 2021; Liu *et al*., 2023). However, despite the espoused benefits of big data and there has been an increasing drive for construction organizations to embrace and apply its dimensions in their respective projects (Ngo *et al*., 2020), its adoption should be treated with a degree of skepticism "as big data is not always better data" (Ghasemaghaei and Calic, 108: p.147). Many construction organizations remain unprepared to effectively utilize big data derived from sensors for assessing geotechnical safety risks (Matthews *et al*., 2022).

We suggest that the outcomes of our research can support decision-making in identifying unsafe conditions based on big data. Our employed EIF, an anomaly detection method, is used to effectively identify anomalous data and retain the normal fluctuation characteristics within its time series. It can be helpful for subsequent data processing and provide high-quality data sources for subsequent data analysis. In addition, the denoise processing of monitoring data significantly reduces data errors and improves the accuracy of identifying unsafe conditions. Hence, the motivation to develop our hybrid smart data approach is to use monitoring data extracted from sensors to help construction organizations assess geotechnical safety risks.

One of the challenges is that a significant amount of data collected from sensors used to detect geotechnical conditions contains noise, rendering it challenging to determine the correct information needed to train algorithms and undertake risk analysis. In addressing this void, we have developed a hybrid smart data approach that can detect noise and de-noise data extracted from sensors monitoring a building's geotechnical conditions, impacting its structural safety. The contributions of our research are twofold.

Firstly, we have developed an EIF approach to detect noise in monitoring data. Existing anomaly detection algorithms detect anomalies by understanding the distribution of their properties and isolating them from a normal data sample. Our employed EIF uses a model-free algorithm that does not rely on building a profile for data to find non-conforming samples. Instead, it utilizes anomalous data with various characteristics compared with normal data samples. In this instance, our employed EIF has computationally efficient and high accuracy without a profile of normal instances demonstrated in this case.

Secondly, a VMD approach and dynamic threshold processing are used to de-noise the monitoring data to improve its validity. The value of the *K* has an important influence on decomposing the data, as shown in Tables 2 and 3, which can prevent under-decomposition and over-decomposition problems of VMD (Dragomiretskiy and Zosso, 2013). Reconstructing the IMF components can effectively decompose the original data. By calculating the RMSE and SNR of the original data and the reconstructed signal, we get the optimal mode number of VMD. This means that when we reuse VMD for new applications, we need to pay attention to the setting of the *K* value to improve the effectiveness of data noise reduction.

5.1 Limitations

Despite the novelty of our research, it needs to be acknowledged that several limitations exist. The study was limited to a single project in the Wuhan subway and two types of monitoring geotechnical data (i.e., building settlement and steel support axial forces). Future research, therefore, is required to examine the generalisability of our approach in different projects and a broader range of activities that use sensors to monitor the geotechnical conditions that influence structural components. In addition, the experimental results demonstrate that the proposed method performs satisfactorily (i.e., RMSE and SNR are 0.0389 and 24.09, respectively). However, we did not conduct comparative experiments to evaluate our hybrid smart data approach's performance (i.e., accuracy and computational efficiency) with other state-of-the-art measurement methods (e.g., deep learning-based). We suggest this limitation can be addressed by conducting additional experiments in our future work.

6.0 Conclusion

Anomaly identification and denoising are necessary tasks to improve the quality of monitoring data extracted from sensors in construction. Our research aims to develop a novel smart data approach that can effectively detect anomalies and de-noise monitoring data to improve its quality to assess geotechnical safety risks. Our smart data approach consists of an:

- Extended Isolation Forest algorithm, which extracts features from each monitoring dataset and is used to identify abnormal points; and
- Variational Mode Decomposition to remove harmonic noise, thus improving data quality.

A case of the Wuhan subway project is used to validate the effectiveness and feasibility of our proposed approach. The results demonstrated that by applying EIF and VMD, a high degree of accuracy could be achieved in detecting anomalies and denoising data. Our results show that our new method can detect anomalies with an RMSE and SNR are 0.0389 and 24.09, respectively. It was revealed that the EIF and VMD could accurately detect anomalies and de-noise monitoring data.

Even though our approach could not recognize all anomalies, our hybrid smart data approach can provide site management to improve their ability to assess geotechnical safety risks. Furthermore, we suggest that our approach can improve the quality of data extracted from sensors in deep foundation pits with minimal error. Thus, our proposed novel smart data approach effectively reduces noise from monitoring data extracted from sensors.

Acknowledgment

The authors would like to acknowledge the financial support of the Alexander von Humboldt Foundation and the National Natural Science Foundation of China (U21A20151, 51978302).

References

Abbate, A. Koay, J, Frankel, J., Schroeder, S.C., and Das, P. (1997) Signal detection and noise suppression using a wavelet transform signal processor: Application to ultrasonic flaw detection. *IEEE Transactions on Ultrasonics Ferroelectrics and Frequency Control*. **44**(1), pp.14-26.

Ahmed, M., Mahmood, A. N., and Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, **60**, pp.19-31.

Andrade, A. O., Nasuto, S., Kyberd, P., Sweeney-Reed, C. M., and Van Kanijn, F. R. (2006). EMG signal filtering is based on empirical mode decomposition. *Biomedical Signal Processing and Control*, **1**(1), pp.44-55.

Aschero, G., and Gizdulich, P. (2010). Denoising of surface EMG with a modified Wiener filtering approach. *Journal of Electromyography and Kinesiology*, **20**(2), pp.366-373.

Asadzadeh, A., Arashpour, M., Li, H., Ngo, T., Bab-Hadiashar, A., and Rashidi, A. (2020). Sensor-based safety management. *Automation in Construction*, **113**, 103128.

Bao, Y., Tang, Z., Li, H., and Zhang, Y. (2019). Computer vision and deep learning-based data anomaly detection method for structural health monitoring. *Structural Health Monitoring*, **18**(2), pp.401-421.

Banjanovic-Mehmedovic, L., Hajdarevic, A., Kantardzic, M., Mehmedovic, F., and Dzananovic, I. (2017). Neural network-based data-driven modelling of anomaly detection in thermal power plant. *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, **58**(1), pp.69-79.

Bhavsar, Y. B., and Waghmare, K. C. (2013). Intrusion detection system using data mining technique: Support vector machine. *International Journal of Emerging Technology and Advanced Engineering*, **3**(3), pp.581-586

Braun, S. (2011). The synchronous (time-domain) average revisited. *Mechanical Systems and Signal Processing*, **25**(4), pp.1087-1102.

Cai, L., Hu, D., Zhang, C. Yu S, Xie J. (2022). Tool vibration feature extraction method based on SSA-VMD and SVM. *Arabian Journal for Science and Engineering*, pp.1-11.

Carrera, F., Dentamaro, V., Galantucci, S., Iannacone, A., Impedovo, D. and Pirlo, G. (2022). Combining unsupervised approaches for near real-time network traffic anomaly Detection. *Applied. Science*. **12**(3), pp. 1759.

Cha, Y. J., and Wang, Z. (2018). Unsupervised novelty detection–based structural damage localization using a density peaks-based fast clustering algorithm. *Structural Health Monitoring*, **17**(2), pp.313-324.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys* (CSUR), **41**(3), pp.1-58.

Chang, C. M., Chou, J. Y., Tan, P., and Wang, L. (2017). A sensor fault detection strategy for structural health monitoring systems. *Smart Structures and Systems*, **20**(1), pp.43-52.

Chen, W., Kong, F., Mei, F., Yuan, G., and Li, B. (2017). *A novel unsupervised anomaly detection approach for intrusion detection system*. In 2017 IEEE 3rd International Conference on Big Data Security on the Cloud, IEEE International Conference on High Performance and Smart Computing, and IEEE International Conference on Intelligent Data and Security, 26th-28th May, Beijing, China, pp. 69-73.

De Luca, C. J., Gilmore, L. D., Kuznetsov, M., and Roy, S. H. (2010). Filtering the surface EMG signal: Movement artifact and baseline noise contamination. *Journal of Biomechanics*, **43**(8), pp.1573-1579.

Domingues, R., Filippone, M., Michiardi, P., and Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition,* **74**, pp.406-421.

Dragomiretskiy, K., and Zosso, D. (2013). Variational mode decomposition. *IEEE Transactions on Signal Processing*, **62**(3), pp.531-544.

Dubé, L., and Paré, G. (2003). Rigor in information systems positivist case research: Current practices, trends, and recommendations. *MIS Quarterly*, **27**(4), pp.597–636.

Fang, W., Love, P. E., Luo, H., & Ding, L. (2020). Computer vision for behaviour-based safety in construction: A review and future directions. *Advanced Engineering Informatics*, 43, 100980.

Fang, W., Love, P. E., Ding, L., Xu, S., Kong, T., & Li, H. (2021). Computer Vision and Deep Learning to Manage Safety in Construction: Matching Images of Unsafe Behavior and Semantic Rules. *IEEE Transactions on Engineering Management* (In Press).

Fu, Y., Peng, C., Gomez, F., Narazaki, Y., and Spencer Jr, B. F. (2019). Sensor fault management techniques for wireless smart sensor networks in structural health monitoring. *Structural Control and Health Monitoring*, **26**(7), e2362.

G. Han, J. Tu, L. Liu, M. Martínez-García and C. Choi. (2022). An Intelligent Signal Processing Data Denoising Method for Control Systems Protection in the Industrial Internet of Things.

IEEE Transactions on Industrial Informatics, 18(4), pp. 2684-2692,

Gao, K., Chen, Z. D., Weng, S., Zhu, H. P., and Wu, L. Y. (2022). Detection of multi-type data anomaly for structural health monitoring using pattern recognition neural network. *Smart Structures and Systems*, **29**(1), pp.129-140.

Ghasemaghaei, M., and Celic, G. (2020). Assessing the impact of big data on firm innovation performance: Big data is not always better data. *Journal of Business Research*, **108**, pp.147–162

Guo, S.Y., Ding, L.Y., Luo, H., and Jiang, X.Y. (2016). A Big-Data-based platform of workers' behavior: Observations from the field. *Accident Analysis and Prevention*, **93**, pp.299-309.

Han, L., Zhang, R., Wang, X., Bao, A., and Jing, H. (2019). Multi-step wind power forecast based on VMD-LSTM. *IET Renewable Power Generation*, **13**(10), 1690-1700.

Hariri, S., Kind, M. C., and Brunner, R. J. (2021). Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*. **33**(4), pp.1479-1489.

Hasan, M. A. M., Nasser, M., Pal, B., and Ahmad, S. (2014). Support vector machine and random forest modeling for intrusion detection system (IDS). *Journal of Intelligent Learning Systems and Applications,* **6**(1), Article ID:42869

He, Q., Wang, X., and Zhou, Q. (2014). Vibration sensor data denoising using a time-frequency manifold for machinery fault diagnosis. *Sensors*, **14**(1), pp.382-402.

Hill, D. J., and Minsker, B. S. (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling and Software,* **25**(9), pp.1014-1022.

Hou, S., and Guo, W. (2020). Optimal denoising and feature extraction methods using modified CEEMD combined with duffing system and their applications in fault line selection of non-solid-earthed network. *Symmetry*. **12**(4), pp.536.

Hu, H., Zhang, L., Yan, H., Bai, Y., and Wang, P. (2019). De-noising and baseline drift removal method of MEMS hydrophone signal based on VMD and wavelet threshold processing. *IEEE Access*, **7**, pp.59913-59922.

Huang, H. B., Yi, T. H., and Li, H. N. (2020). Anomaly identification of structural health monitoring data using dynamic independent component analysis. *ASCE Journal of*

Computing in Civil Engineering, **34**(5), 04020025.

Huang, H. B., Yi, T. H., and Li, H. N. (2017). Sensor fault diagnosis for structural health monitoring based on statistical hypothesis test and missing variable approach. *Journal of Aerospace Engineering*, **30**(2), B4015003.

Huang N E , Shen Z , Long S R , Wu MC, SHIH HH, Zheng Q, Yen NC, Tung CC, Liu HH. (1998). The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis. *Proceedings Mathematical Physical & Engineering Sciences*, 454(1971), 903-995.

Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M., and Sun, J. (2017). Anomaly detection for a water treatment system using unsupervised machine learning. 2017 *IEEE International Conference on Data Mining Workshops* (ICDMW), 18^{th}-21^{st} November, New Orleans, USA, pp. 1058-1065

Liu, F. T., Ting, K. M., and Zhou, Z. H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data* (TKDD), **6**(1), pp.1-39.

Li H, Xu Y, An D, Zhang L, Li S, Shi H. (2020). Application of a flat variational modal decomposition algorithm in fault diagnosis of rolling bearings. *Journal of Low Frequency Noise, Vibration and Active Control*, 39(2):335-351.

Li, S., Zhang, K., Duan, P., and Kang, X. (2019). Hyperspectral anomaly detection with kernel isolation forest. *IEEE Transactions on Geoscience and Remote Sensing*, **58**(1), pp.319-329.

Li, T., Zhao, Z., Sun, C., Yan, R., and Chen, X. (2021). Hierarchical attention graph convolutional network for fuse multi-sensor signals for remaining useful life prediction. *Reliability Engineering and System Safety*, **215**, 107878

Liu, F. T., Ting, K. M., and Zhou, Z. H. (2008). Isolation forest. Proceedings of the 8^{th} IEEE International Conference on Data Mining, 15^{th}-19^{th} December, Pisa, Italy pp. 413-422.

Liu, D., Wang, S., and Cui, X. (2022). An artificial neural network supported Wiener process-based reliability estimation method considering individual difference and measurement error. *Reliability Engineering and System Safety*, **218**, 108162

Liu, W., Li, A., Fang, W., Love, P. E., Hartmann, T., & Luo, H. (2023). A hybrid data-driven model for geotechnical reliability analysis. *Reliability Engineering & System Safety*, **231**, 108985.

Kromanis, R., and Kripakaran, P. (2013). Support vector regression for anomaly detection from measurement histories. *Advanced Engineering Informatics*, **27**(4), pp.486-495.

Maier, J., Naber, A., and Ortiz-Catalan, M. (2017). Improved prosthetic control based on myoelectric pattern recognition via wavelet-based denoising. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **26**(2), pp.506-514.

Matthews, J., Love, P. E.D, Porter, S. R., and Fang, W. (2022). Smart data and business analytics: A theoretical framework for managing rework risks in mega-projects. *International Journal of Information Management*, **65**, 102495.

Mu, H. Q., and Yuen, K. V. (2015). Novel outlier-resistant extended Kalman filter for robust online structural identification. *ASCE Journal of Engineering Mechanics*, **141**(1), 04014100.

Nessa, A., Adhikari, B., Hussain, F., and Fernando, X. N. (2020). A survey of machine learning for indoor positioning. *IEEE Access,* **8**, pp.214945-214965.

Ngo, J., Hwang, B.-G., and Zhang, C. (2020). Factor-based big data and predictive analytics capability assessment tool for the construction industry. *Automation in Construction*, **110**, 103042

Nguyen, L. H., and Goulet, J. A. (2019). Real-time anomaly detection with Bayesian dynamic linear models. *Structural Control and Health Monitoring*, **26**(9), e2404.

Otoum, S., Kantarci, B., and Mouftah, H. (2018). Adaptively supervised and intrusion-aware data aggregation for wireless sensor clusters in critical infrastructures. *IEEE International Conference on Communications* (ICC), 20^{th}-24^{th} May Kansas City, USA, pp.1-6

Ortolan, R. L., Mori, R. N., Pereira, R. R., Cabral, C. M., Pereira, J. C., and Cliquet, A. (2003). Evaluation of adaptive/nonadaptive filtering and wavelet transform techniques for noise reduction in EMG mobile acquisition equipment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **11**(1), pp.60-69.

Prateek G V., Ju Y E., and Nehorai A. (2021). Sparsity-Assisted Signal Denoising and Pattern Recognition in Time-Series Data. *Circuits Systems and Signal Processing*, **2021**(9), pp.1-50.

Rabatel, J., Bringay, S., and Poncelet, P. (2011). Anomaly detection in monitoring sensor data for preventative maintenance. *Expert Systems with Applications*, **38**(6), pp.7003-7015

Rai, H. M., and Chatterjee, K. (2019). Hybrid adaptive algorithm based on wavelet transform and independent component analysis for denoising of MRI images. *Measurement*, **144**, pp.72-82.

Samuel T. Thurman and James R. Fienup. (2008). Wiener filtering of aliased imagery. *International Society for Optics and Photonics*, **7076**, pp.70760J.

Selesnick I W., Graber H L., Pfeil D S., and Barbour R. L. (2014). Simultaneous Low-Pass Filtering and Total Variation Denoising. *IEEE Transactions on Signal Processin*g, **62**(5), pp. 1109-1124.

Seites-Rundlett, W., Bashar, M., Torres-Machi, C., and Corotis, R.B. (2022). Combined evidence model to enhance pavement condition prediction from highly uncertain sensor data. *Reliability Engineering and System Safety*, **217**, 108031

Singh, P., Shahnawazuddin, S., and Pradhan, G. (2018). An efficient ECG denoising technique based on non-local means estimation and modified empirical mode decomposition. *Circuits, Systems, and Signal Processing*, **37**(10), pp.4527-4547.

Thottan, M., and Ji, C. (2003). Anomaly detection in IP networks. *IEEE Transactions on Signal Processin*g, **51**(8), pp. 2191-2204.

Thudumu, S., Branch, P., Jin, J., and Singh, J. J. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, **7**(1), pp.1-30.

Urciuolo, M. (2008). Restriction of the Fourier transform. *Revista de la Unión Matemática Argentina*, **49**(2), pp. 39-44.

Wang, Z., He, G., Du, W., Zhou, J., and Kou, Y. (2019). Application of parameter optimized variational mode decomposition method in fault diagnosis of gearbox. *IEEE Access*, **99**(7), pp. 44871-44882.

Xia, S.; Yang, J.; Cai, W.; Zhang, C.; Hua, L., and Zhou, Z. (2021). Adaptive Complex Variational Mode Decomposition for Micro-Motion Signal Processing Applications. *Sensors,* **21**(5), pp. 1637.

Xiao, F., Yang, D., Guo, X., and Wang, Y. (2019). VMD-based denoising methods for surface electromyography signals. *Journal of Neural Engineering*, **16**(5), 056017.

Xu, Z., Li, C., and Yang, Y. (2020). Fault diagnosis of rolling bearing of wind turbines based on the variational mode decomposition and deep convolutional neural networks. *Applied Soft Computing*, **95**, 106515.

Yi, T. H., Li, H. N., and Gu, M. (2013). Recent research and applications of GPS‐based monitoring technology for high‐rise structures. *Structural Control and Health Monitoring*, **20**(5), pp.649-670.

Yu, L, Lin L. (2008). Improvement on empirical mode decomposition based on correlation coefficient. *Computer and Digital Engineering*, **36**(12), pp.28-29.

Yue, G. D., Cui, X. S., Zou, Y. Y., Bai, X. T., Wu, Y. H., and Shi, H. T. (2019). A Bayesian wavelet packet denoising criterion for mechanical signal with non-Gaussian characteristic. *Measurement*, **138**, pp.702-712.

Zhou, C., Kong, T., Jiang, S., Chen, S., Zhou, Y., and Ding, L. (2020). Quantifying the evolution of settlement risk for surrounding environments in underground construction via complex network analysis. *Tunnelling and Underground Space Technology*, **103**, 103490.

Zhou, Y., Li, S., Zhou, C., and Luo, H. (2019a). Intelligent approach based on random forest for safety risk prediction of deep foundation pit in subway stations. *ASCE Journal of Computing in Civil Engineering*, **33**(1), 05018004.

Zhou, Y., Li, C., Ding, L., Sekula, P., Love, P. E.D, and Zhou, C. (2019b). Combining association rules mining with complex networks to monitor coupled risks. *Reliability Engineering and System Safety*, **186**, 194-208.

Zuo, R., and Xiong, Y. (2018). Big data analytics of identifying geochemical anomalies supported by machine learning methods. *Natural Resources Research*, **27**(1), pp.5-13.