Annals of Nuclear Energy 181 (2023) 109503

Contents lists available at ScienceDirect

Annals of Nuclear Energy

journal homepage: www.elsevier.com/locate/anucene

Feature extraction for subtle anomaly detection using semi-supervised learning

Yeni Li [a,*], Hany S. Abdel-Khalik [a], Ahmad Al Rashdan [b], Jacob Farber [b]

[a] School of Nuclear Engineering, Purdue University, 205 Gates Road, FLEX Lab, West Lafayette 47906, USA
[b] Idaho National Laboratory, 1955 N. Fremont Ave., Idaho Falls 83415, USA

ARTICLE INFO

Keywords:
Subtle anomaly detection
High-order features
Feature extraction
Feature fusion

ABSTRACT

The demand for automated and effective monitoring techniques has soared with the increased digitization of industrial monitoring systems. State-of-the-art machine learning methods are effectively detecting abrupt changes in system states. However, these methods lack comparable maturity in detecting subtle changes that may be signs of incipient faults. This manuscript argues that the current anomaly detection methods can be enhanced by exploring weak patterns to enable subtle variation detection. Specifically, the concept of semi-supervised learning is employed, with labels representing knowledge about some anomalous conditions of a system. The basic idea is to extract a candidate set of weak patterns discarded by state-of-the-art baselining algorithms. With few labeled anomalous data, the algorithm selects the weak patterns and allows for their possible fusion using the highest sensitivity to the labeled anomalies. The method's applicability is demonstrated using a representative pressurized water reactor (PWR) model simulated by Dymola.

1. Introduction

Engineering systems are increasingly digitized in pursuit of benefits like improved efficiency, optimized economy, and higher system visibility. These digitization efforts must be equipped with powerful analytical tools to process the huge volumes of sensor data collected during operation. Machine learning (ML) techniques present a powerful set of analysis tools to harvest the sensor time-series data in search of patterns that can be used to automate responses to upset conditions. This approach is referred to as condition monitoring, wherein ML serves as a classifier to distinguish between normal behavior, including both steady-state and per-design transient conditions, and abnormal conditions (i.e., anomalous conditions resulting from undesirable scenarios like equipment failure or unanticipated accidents scenarios).

Generally, anomaly detection methods can be categorized into two types: model-based (Guo, 2020) and data-driven (Martin and Morris, 1996). In the context of condition monitoring, model-based approaches refer to techniques based on prior system knowledge, including causality or conditional dependence between measured variables to describe the system behavior. These model-based techniques combine measured data and physics models using methods like Bayesian networks (Vaddi et al., 2020; Wu et al., 2018) which require a precise description for the system behavior that anticipates the various mechanisms that may lead to degraded performance. In practice, constructing a whole system model is, at best, challenging and, practically speaking, infeasible. Thus, data-driven methods stand out as a pragmatic alternative as they may be applied in black-box mode and require no prior contextual knowledge about the system.

In unsupervised settings, data-driven methods accomplish the task of anomaly detection by first characterizing a baseline by relying on the identification of recurring patterns assumed to be associated with normal operation. These patterns, which are subtracted from the time series signal, leave a residual that is attributed to unimportant or non-modeled behavior and noises. Any non-conforming patterns that are above the residual level are treated as anomalies (Chandola et al., 2009). These state-of-the-art methods rely primarily on the dominant patterns for baseline behavior, and together serve as an excellent tool for detecting abrupt changes since they break the dominant patterns (Li and Huang, 2016; De Ketelaere et al., 2015; Peng et al., 2017) (e.g., a sudden spike or a new trend which overpowers the residual term). However, the weak patterns remain subsumed by the noise that dominates the residual term, making it difficult to distinguish between noise and subtle changes. Statistically, the algorithm is biased against subtle changes, as it is designed to primarily capture the dominant behavior by using an

* Corresponding author.
E-mail address: li2181@purdue.edu (Y. Li).

https://doi.org/10.1016/j.anucene.2022.109503
Received 2 June 2022; Received in revised form 2 October 2022; Accepted 4 October 2022
Available online 20 October 2022

algorithm that is insensitive to subtle changes.

This manuscript therefore proposes an alternative approach for detecting subtle changes, where the large space of weak components, referred to hereinafter as high-order features (HOFs), are carefully analyzed using semi-supervised learning to determine their sensitivity to subtle changes detection (Li et al., 2022; Li and Abdel-Khalik, 2021). In contrast, the low-order features (LOFs) denote dominant components and are used by the extant methods for baselining. The notions of low and high orders are borrowed from decomposition techniques such as singular value decomposition and principal component analysis. Both regard the low-indexed components as dominant, while high-indexed components are considered to be of less influence since they are more prone to noise contamination. For example, this manuscript employs a window-based decomposition where a running window is used to collect snapshots of the time series. A singular value decomposition is then employed to capture the LOFs and HOFs. If the transient behavior can be decomposed into thirty components, and the first three components are denoted as LOFs, leaving twenty-seven HOFs. Our goal is to identify which of these HOFs are most sensitive to the labeled subtle change anomalies. This methodology also allows fusing multiple HOFs and/or LOFs to maximize sensitivity to the anomalies.

Note that, in the unsupervised setting, all LOFs are aggregated together as the baseline for normal behaviors and all HOFs are aggregated together as the residual term, sometimes also called the null-space (Gawand et al., 2017; Zhang and Coble, 2020; Wang et al., 2019). Our rendering semi-supervised algorithm argues the need to explore the space of HOFs by analyzing each one. This is necessary because when a subtle change occurs, it modifies some components of the HOFs. Even if these changes are big, they can be overpowered by the randomness of all the other HOF components due to the law of large numbers, which is the sum of many random events reaching a normal distribution. Hence, even if a single or a few HOF components show a patterned behavior that can inform anomalies, the remaining HOF components are expected to overpower these few components, if one only analyzes their aggregated value, which is effectively equal to the quadratic summation of all the HOFs.

To pinpoint which HOF is anomaly-sensitive, one would rely on supervised learning in which all data are correctly labeled to gain a full picture of anomaly-sensitive HOFs. However, this is not feasible because a well-performing system is not expected to have huge incidents of anomalous behavior. What is more realistic is that the regular inspection activities would record some past anomalous incidents, while others might occur and get rectified without being logged. With this information, albeit limited, the proposed algorithm would identify the most sensitive HOFs for these recorded anomalies. Specifically, the algorithm calculates a set of candidate HOFs and monitors their values around the time the labels are available. This allows the algorithm to down-select the HOFs with the maximum sensitivity to the labeled anomalies. The candidate HOFs are calculated using window-based decomposition techniques called randomized window decomposition (RWD), which was developed in our earlier work (Li and Abdel-Khalik, 2021).

Furthermore, given the abundance of HOFs, we explore the possibility of designing anomaly-targeting HOFs (i.e., an HOF that is sensitive to a specific anomaly, including equipment and process anomalies). A simple fusion strategy is employed in this work to explore fusing multiple HOFs into a single HOF with higher overall sensitivity. This is depicted in Fig. 1, where the red stars denote the location of known anomalies, and the blue graphs display the anomaly scores for two candidate HOFs (shown in the left and middle graphs), each showing partial sensitivity to the anomalies. Two simple fusion operators are employed in this work, an addition operator, and a multiplication operator. Both operations are applied to the candidate HOFs to design new HOFs that capture the combined sensitivities of the individual HOFs, as shown in the right graph. It is expected that many fusion strategies of HOFs could be designed. Given the exploratory nature of the initial work, this manuscript will focus on proposing simple fusion strategies for the HOFs with future work proposed for more mechanistic approaches for their optimal design. In the field of anomaly detection, a key criterion to evaluate a feature is the distinction between normal and anomalous data, which requires the features to capture the essence of anomaly or normal behavior. If one needs to rely on a complicated model on the features, such as deep neural network, to achieve the goal of anomaly detection, then the feature does not capture the essence. Here to evaluate the applicability and demonstrate the usefulness of the fused features and the proposed feature extraction approach, we use simple straight forward machine learning techniques such as K-means clustering (for no labeling information) and KNN (K-nearest neighbors) classification on this feature extraction method and typical residual-based approach.

2. Background

The premise of condition monitoring revolves around building a model that characterizes baseline behavior, with deviations representing anomalous behavior. This model can be constructed in either

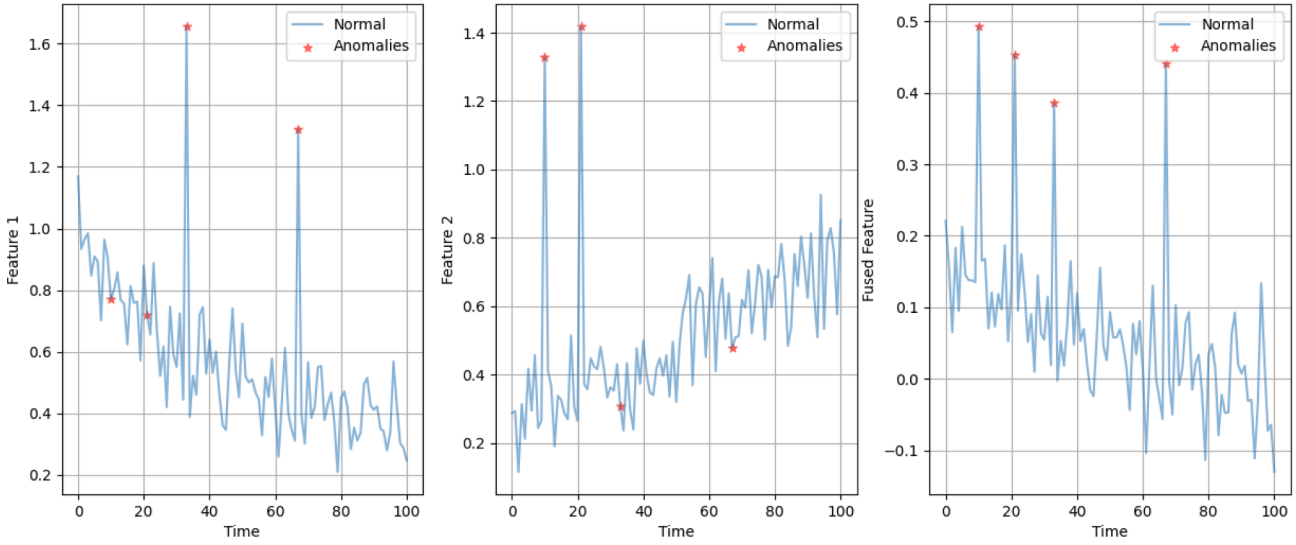

Fig. 1. Illustration for feature fusion.

supervised or unsupervised settings, based on whether an abundance of anomalous data are available. In either setting, working directly with the raw data is computationally infeasible, so the vast majority of condition monitoring techniques adopt a two-step strategy. In the first step, features are extracted from the data based on some criterion, such that the dimensionality of the features is far smaller than that of the original data, a process referred to in the literature as dimensionality reduction, data compression, encoding, feature engineering, or extraction. The goal of this first step is to reduce the data dimensionality for the second step. In the second step, a classifier is designed to distinguish between normal and anomalous behavior as a function of the extracted features. Clearly, if the extracted features are not representative of the process, the results of the classifier are expected to be poor.

In the unsupervised setting, the feature extraction step relies on capturing the recurring patterns in the data. The patterns are subtracted from the data, leaving a residual with desirable properties, such as a normal distribution and a standard deviation similar to the inherent noise. A residual-based threshold is used to set an anomaly score, with the score being a function of the residual deviation from the set threshold, possibly augmented by a statistical hypothesis test before sounding the alarm (Esmalifalak et al., 2017; Huang et al., 2012). In the supervised setting, data are split into two labeled batches, one denoting normal and the other anomalous behavior. The feature extraction step is still required. However, now the goal is not to detect dominant behavior in both the normal and anomalous data, but rather to find features that are separable into two batches. For example, in simple binary classification, one is interested in finding a feature that generates two numbers (e.g., 1 and –1), wherein the data from the normal batch consistently produce a value of 1 and the anomalous batch a value of –1. Clearly, the supervised setting is more effective than the unsupervised setting because the features can be fine-tuned to the anomalies.

In practice, the uncertainty about a system condition, human labor associated with labeling the data, as well as the normal scarcity of anomalous data for well-functioning systems, deems the supervised setting as not feasible. This has helped promote unsupervised methods as an attractive tool for anomaly detection. However, their success is premised on the ability to perform clustering using unlabeled data, which indirectly implies that one of the following two conditions must be satisfied: a) the anomalies must introduce large changes in the extracted features, or b) the anomalies must change the residual distribution (i.e., its mean value, standard deviation, or other higher order statistics). With no labels, the analyst is forced to rely on extracting dominant features to describe the recurring patterns under normal behavior. That choice, however, limits the analyst's ability to detect only abrupt changes. In response, this manuscript proposes the use of a semi-supervised learning setting, where the HOFs comprising the residual term are closely investigated as potential candidates for detecting subtle anomalies via a small set of labeled anomalies. This approach is expected to be practical because it combines the advantages of both unsupervised learning (i.e., not requiring an extensive library of labeled data) and supervised learning, allowing one to fine-tune the design of anomaly-targeting features.

Since feature extraction is key to all learning settings (i.e., supervised, unsupervised, and semi-supervised), they are briefly summarized here under two broad categories: parametric and nonparametric methods (Riedel et al., 1994; Basseville et al., 2000;Runger and Testik, 2003; Fantoni, 2005; Zhao, 2005; Shankar, 2004; Lu and Upadhyaya, 2005). The most commonly used parametric approach is regression, including basic polynomial regression, autoregressive (AR) (Bornn et al., 2009), and autoregressive–moving-average (ARMA) models (Caesarendr et al., 2010). This approach fits the historical time-series data against pre-determined regression models by employing a penalty function to calculate optimal values for the regression parameters. The penalty function attempts to minimize the discrepancy between the real and model-predicted values, with an additional penalty (called a regularization term) sometimes used to ensure robustness to noise. In this case, the patterns denote the basic model terms employed (e.g., linear and second-order term in a polynomial regression model) the degree of the AR or ARMA model, and the architecture of the neural network. The characteristics denote the optimal parameters extracted, such as the polynomial coefficients, the ARMA model parameters, or the neuron weights in a neural network. Parametric methods therefore effectively pre-determine the patterns, while the learning process finds how many of the assumed patterns exist in the data. In contrast, nonparametric methods allow the data to identify the patterns and extract the features. Examples include neural networks, the alternating conditional expectations (ACE) (Li et al., 2018), projection pursuit (Friedman, 1987), the Gaussian process (Jiang et al., 2020), as well as rank-revealing matrix decompositions such as singular value decomposition (SVD) (Li et al., 2022) and principal component analysis (PCA) (Wang et al., 2005).

A brief mathematical description is provided below to illustrate how the previous discussion on how features are extracted and used for anomaly detection. The core idea is to split the time series signal into two components: a regular recurring pattern and an unexplained residual. The regular pattern represents the structured (i.e., patterned, variations in the time series data) considered to represent normal behavior. The remaining part is assumed to be unexplained, meaning it has no structure and hence can be attributed to random noise. Mathematically, the regular pattern is identified by minimizing the norm of the unexplained residual with a distance-type metric like least-squares minimization. The premise of this approach is that, during normal operation, the regular pattern will manifest itself and the unexplained part will remain approximately bounded to its norm as determined during the training phase. Consider a parametric split of the form as an example:

$$x_t = a_i x_{t-i} + a_j x_{t-j} + a_k x_{t-k} + \varepsilon_t \tag{1}$$

This equation implies that the value of the variable x at time t is approximately determined during normal behavior by the weighted sum of certain lagged values of x, with the weights being constant in this simple example, where the subscript i, j, and k, are integers representing the lagged timesteps, and the residual ε_t represents the unexplained part of the signal. This is an example of a parametric method since it defines the relationship between the data a priori. In our context, the three-component vector containing these weights is referred to as a dominant pattern. The implication is that the predicted value of x at time t is simply the inner product (the projection) of the three lagged values with the feature vector, expressed in Eq. (2).

$$x_t = \begin{bmatrix} a_i & a_j & a_k \end{bmatrix} \begin{bmatrix} x_{t-i} \\ x_{t-j} \\ x_{t-k} \end{bmatrix} + \varepsilon_t \tag{2}$$

The idea of feature extraction is generic. It applies to simple linear models like the one above, and to general nonlinear functions such as those trained by deep neural networks, wherein the feature vectors are the weight vectors associated with the various neuron functions (i.e., $g_i(w_i^T x)$) described by a sigmoid function.

An effective anomaly classifier is one that can identify the features that are most sensitive to the sources of the anomalies rather than the dominant or "normal" behavior. When training neural networks in an unsupervised manner, the feature vectors are dominated by the data from normal behavior making it very difficult to identify features sensitive to subtle anomalous behavior. This follows because most ML algorithms rely on the back-propagation algorithm to update the weight vectors, which usually adopts a gradient descent algorithm to optimize the loss function and then employs Euclidean L_2 distance norm to determine the residual errors. The L_2 norm is known to be insensitive to small variations which tend to average out over the abundant volume of normal operating data, making it less effective for detecting subtle variations. On the other hand, the L_2 norm is very sensitive to large variations, making it attractive for the detection of sudden-change

anomalies.

Unlike the example above which represents a parametric approach, this work employs a nonparametric method for extraction of patterns, and subsequent selection of HOFs via few labels. Specifically, we employ a windowed SVD nonparametric approach as a basic ingredient in many feature extraction algorithms like principal component analysis, dynamic mode decomposition, and proper orthogonal decomposition. We employ a randomized version of this basic algorithm, introduced in earlier work (Li and Abdel-Khalik, 2021), denoted by the RWD. As applied to transient time series data, windows of fixed length are randomly placed over the time horizon to capture snapshots of the time series, forming a matrix **R**. Each column of **R** represents a snapshot, and the number of columns is the number of snapshots. A rank revealing decomposition is then applied to **R**, such as the SVD, which decomposes the matrix **R** into a product of three matrices, as implied by Eq. (3). The **U** matrix identifies the dominant patterns in column space of **R**, ordered by the pattern dominance as measured by the diagonal elements of **S**, denoted by the singular values

$$\mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{3}$$

This decomposition can be used to identify both the LOFs and HOFs. An LOF is the inner product between a dominant pattern, i.e., one of the low-indexed columns of **U**, and the original data in a given window. An HOF is the inner product between a weak pattern associated with a high-indexed low singular value. The HOFs and LOFs are all orthogonal in the L_2 sense. The features are then plotted versus time using a moving window approach. With access to known past anomalies encoded in the function $l(t)$, the second step of the algorithm picks the HOFs that show the highest sensitivities to the labeled anomalies.

Given the typically large number of HOFs compared to the dominant features, this approach allows tailoring the HOFs to the various types of anomalies to allow for causal identification of anomalous behavior. Finally, the selected HOFs are employed to train classifier or clustering based on the label availability. Also, with the selected HOFs, a threshold to determine the status of the data can be established based on the labeling information. The mathematical description of the steps is provided next.

3. Methodology

The proposed subtle anomaly detection algorithm includes two steps: 1) an unsupervised learning algorithm to identify a set of candidate HOFs using the RWD algorithm; 2) a semi-supervised approach consists of down-selecting and synthesizing new HOFs that are most sensitive to the available labeled anomalies.

Step 1: Identify HOFs using RWD algorithm

Starting with a matrix **M** of sequentially sliding k-sized windows, a matrix **R** is formed by randomly sampling columns (denoted by random indices a, b, c) from the matrix **M**. The RWD is then applied to the matrix **R** to find the dominant patterns as expressed in Eqs. (4)–(8). Based on a user-selected tolerance δ, a rank r is identified, which is selected to denote the number of dominant patterns, with the remaining (k-r) patterns representing the weak patterns. The time values of the HOFs ($\boldsymbol{\alpha}_H$) and LOFs ($\boldsymbol{\alpha}_L$) are calculated as the inner product between the columns of the matrix **M** and the weak and dominant patterns, respectively. One advantage of RWD is that the window-placing is random, such that the extracted dominant features are insensitive to the presence of rare anomalies. This is important since, in practical applications, most of the anomalies are unlabeled.

Note that each element of the α vectors is a time series representing the window-based projection of the original time series along one column of the **U** matrix, with length equal to (n-k + 1), where n is the number of times steps of the original time series, and k the size of the window. Therefore, if the window size is k, and r represents the number of dominant components, the RWD algorithm produces (k-r) possible HOFs, each represented by an (n-k + 1) time series, serving as potential candidates for the identification of anomalous behavior as done in the next step.

It is worth mentioning here that RWD may need to be applied multiple times on the resulting α time series vectors to render additional smoothing. Recall that the HOFs are defined as the higher order components of the SVD decomposition (i.e., those associated with low singular values) and, when plotted versus time, they display random behavior akin to noise. Earlier work demonstrated a multilevel approach for applying RWD to render additional smoothing of the HOFs prior to their use for classification. Details on that process may be found in earlier work (Li et al., 2022).

$$\boldsymbol{M} = \begin{bmatrix} x_1 & x_2 & \cdots & x_{n-k+1} \\ x_2 & x_3 & \cdots & x_{n-k+2} \\ \vdots & \vdots & \ddots & \vdots \\ x_k & x_{k+1} & \cdots & x_n \end{bmatrix} \tag{4}$$

$$R = \begin{bmatrix} x_a & x_b & \cdots & x_c \\ x_{a+1} & x_{b+1} & \cdots & x_{c+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{a+k} & x_{b+k} & \cdots & x_{c+k} \end{bmatrix} = USV^T \tag{5}$$

$$Find \max r \; to \; satisfy: \; \max \left\| R - U_r S_r V_r^T \right\| \leqslant \delta \tag{6}$$

$$\alpha_L = U_r^T \mathrm{M} \tag{7}$$

$$\alpha_H = U_{k-r}^T \mathrm{M} \tag{8}$$

It is noteworthy that even the RWD employs SVD for pattern discovery, it does not assume data linearity. Since for a single time series variable, there are two true dimensions: the time and the variable itself. The RWD aims to use many degrees of freedom (DOFs) that represent in the high-order components to capture the nonlinearity along these two dimensions. Consequently, all the principal components span the full spaces in a linear way, but the transient nonlinearity would be stored in the decomposed components. Furthermore, the number of DOFs is the same with the window length that is the only hyperparameter of RWD. If the window length, k, is too small, there will not have enough degrees of freedom to capture the transient variations; if the window length is too large, such as using the full length of the time series, this windowed approach will be rendered as traditional SVD approaches that used for identifying correlations between variables instead of the explore the temporal evolution of the variable.

Step 2: Down-selecting HOF using labeled anomalies

This step represents the core of the proposed semi-supervised algorithm, as it employs the known labeled anomalies to select the HOFs that are most sensitive to the labeled anomalies. A simple criterion is used here for selection, which measures the ratio of the peak value to the average noise level. Examples are shown in the figures in section V, representing different HOFs with different levels of sensitivity to the labeled anomalous data. The red dots represent the value of a given HOF in the anomalous range (where the location of a labeled anomaly is known). An anomaly-targeting HOF is expected to have an anomaly score that is significantly high at the location of known anomalies as compared to its average value.

Mathematically, for a certain feature vector α_i, this process is expressed in (9). If $\alpha_{i,max}^a \gg \alpha_{i,max}^n$, the feature vector, α_i, can be selected as a candidate, where the superscript a and n are denoted for “anomalous” and “normal,” respectively. Thus, a baseline line, η, can be established for the determination of normal or anomalous data, expressed in (10), where f is an engineering-oriented factor between 0 and 1. This process can be easily extended for cases with more labeled data, as expressed in (11), where p indicates the number of labeled normal regions and q means the number of labeled anomalous regions. As the number of labeled data increases, the gap between the normal and anomalous determination will get narrowed down to a “true” threshold that could

be achieved via supervised learning. On the other hand, when the number of labeled data decreases to 0, this approach steps back to unsupervised learning, which is performed as a variant of clustering.

$$\begin{gathered}\text{Assume: Labeled anomalous region: } g\leqslant t\leqslant h;\\ \text{Labeled normal region: } l\leqslant t\leqslant m.\\ \boldsymbol{\alpha}_i=\left[\alpha_{i,1},\alpha_{i,2},\cdots,\alpha_{i,g},\cdots,\alpha_{i,h},\cdots,\alpha_{i,l},\cdots,\alpha_{i,m},\cdots,\alpha_{i,n-k+1}\right],\ r<i\leqslant k\\ \alpha^a_{i,\max}=\max\left|\alpha_{i,g},\cdots,\alpha_{i,h}\right|,\ \alpha^n_{i,\max}=\max\left|\alpha_{i,l},\cdots,\alpha_{i,m}\right|\end{gathered}\tag{9}$$

$$\alpha^n_{i,\max}<\eta=f\alpha^a_{i,\max},0<f\leqslant 1\tag{10}$$

$$\max[\alpha^{n,p}_{i,\max}]<\eta=f\cdot\min[\alpha^{a,q}_{i,\max}]\tag{11}$$

Selection of the HOF exhibiting the highest sensitivity to the anomalous behavior represents a possible way forward. Another idea is to fuse multiple features via multiplication or linear superposition of their corresponding profiles, thereby pronouncing the impact of the anomalies. Given that the number of features is small, one can attempt simple fusion ideas like forming a set of binary-fused features (i.e., two at a time) or tertiary features, etc., and applying the proposed metric to identify the binary set with the highest sensitivity. In our study, this feature fusion is used after feature extraction and is prepared for further classification or clustering.

There is an important distinction here compared with existing semi-supervised learning techniques (Ruff et al., 2019). In some existing techniques, the time series data associated with the labeled anomalies are employed to train classifiers to learn the signature of the anomaly. In this algorithm, the anomalous data are not used to learn the anomalies. Instead, their locations are used to down-select the HOFs obtained from the unlabeled time series data. If the unlabeled time series data contain too many anomalies, the HOFs are expected to fail in capturing the anomalies, because they become tuned to the shapes of the anomalies. However, this is unlikely to happen in realistic applications.

For the readability of the methodology, a table of used mathematical symbol is listed below (Table 1):

Table 1
Mathematical symbol.

Symbol	Description
x_i (i = 1, 2, …)	i^{th} elements of a time series
n	The total number of elements in a time series
k	Window size
M	Hankel matrix constructed from a time series
R	Matrix constructed from randomly sampling windows
δ	User-defined tolerance
r	Rank of matrix **R** based on δ; the number of LOFs
k-r	The number of HOFs
U, **S**, $\mathbf{V}^T$	Left-singular vectors formed matrix, singular value matrix, right-singular vector formed matrix of matrix **R**
$\mathbf{U}_r$, $\mathbf{S}_r$, $\mathbf{V}_r^T$	Rank r approximated left-singular vectors formed matrix, singular value matrix, right-singular vector formed matrix
$\boldsymbol{\alpha}_L$	Matrix constructed from LOFs
$\boldsymbol{\alpha}_H$	Matrix constructed from HOFs
$\boldsymbol{\alpha}_i$	The i^{th} feature vector
$\alpha_{i,g}$ … $\alpha_{i,h}$	Labeled anomalous region of $\boldsymbol{\alpha}_i$, from the g^{th} element to the h^{th} element
$\alpha^a_{i\,\max}$	The maximum of the $\boldsymbol{\alpha}_i$ feature in the labeled anomalous region
$\alpha_{i,l}$ … $\alpha_{i,m}$	Labeled normal region of $\boldsymbol{\alpha}_i$, from the l^{th} element to the m^{th} element
$\alpha^n_{i\,\max}$	The maximum of the $\boldsymbol{\alpha}_i$ feature in the labeled normal region
p	The number of labelled normal regions
q	The number of labelled anomalous regions
$\max[\alpha^{n,p}_{i,max}]$	The largest feature value among all p labelled normal regions
$\min[\alpha^{a,q}_{i,max}]$	The minimum of the maximum among all q labeled anomalous regions
f	An engineering-oriented factor that tuned by user to determine threshold
η	Threshold of a features to determine the status of normal or anomalous data

4. Physics model

This study employs a Dymola-simulated PWR, shown in Fig. 2, to get the system transient behaviors. The solid blue lines represent the connection between different components. The sensor reading collection and control of these components are accomplished by "sensorBus" and "actuatorBus", respectively. The connection between sensors and the measurement center are represented by red dashed lines; the connection between actuators and actuation center are shown as green dashed lines. The layout of the control system for normal operation state is shown in Fig. 3. There are four physics quantities contained in the actuators: mass flow rate of the primary pump and feed water pump, reactivity, and the liquid heater in the pressurizer. Correspondingly, six variables are measured by the system, serving as input parameters of the control system for stabilizing the system state: total power, coolant temperature of core inlet and outlet, pressure in the pressurizer, steam generation amount, and the feedwater mass flow rate.

To simulate the wide range of anomalies expected in a nuclear power plant, three types of anomalies are introduced here: a) anomalies gradually developing over longer time periods, and then gradually removed from the time series data, referred to as wide anomalies; b) anomalies developing over shorter periods of time and subsequently removed, referred to as narrow anomalies; and c) anomalies developing gradually, but not removed. The first two anomalies represent situations when the anomalies are discovered and removed by the regular maintenance work orders, whereas the third represent undiscovered anomalies.

5. Case studies

This methodology is practiced via three different types of anomalies: wide anomalies, narrow anomalies, and persistent anomalies. These study cases are demonstrated as follows.

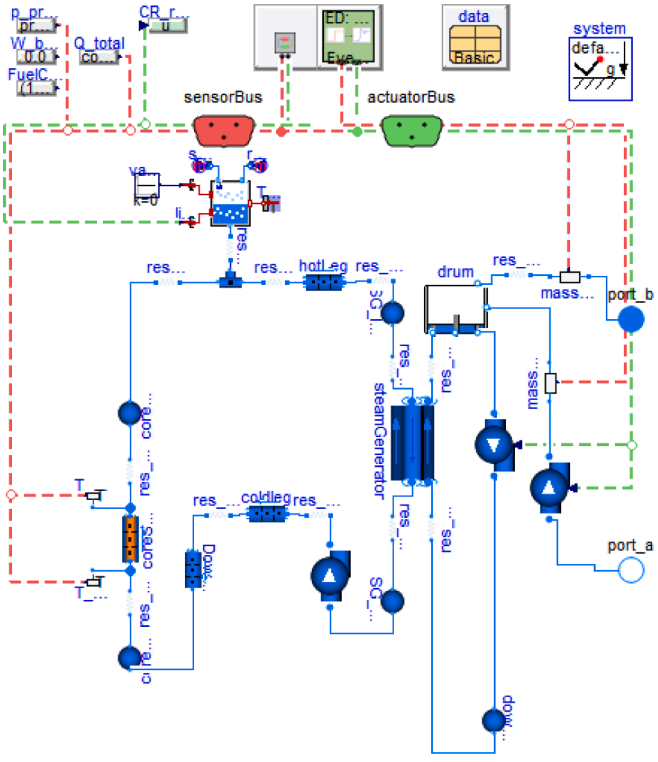

Fig. 2. Dymola simulated PWR layout (Greenwood et al., 2017).

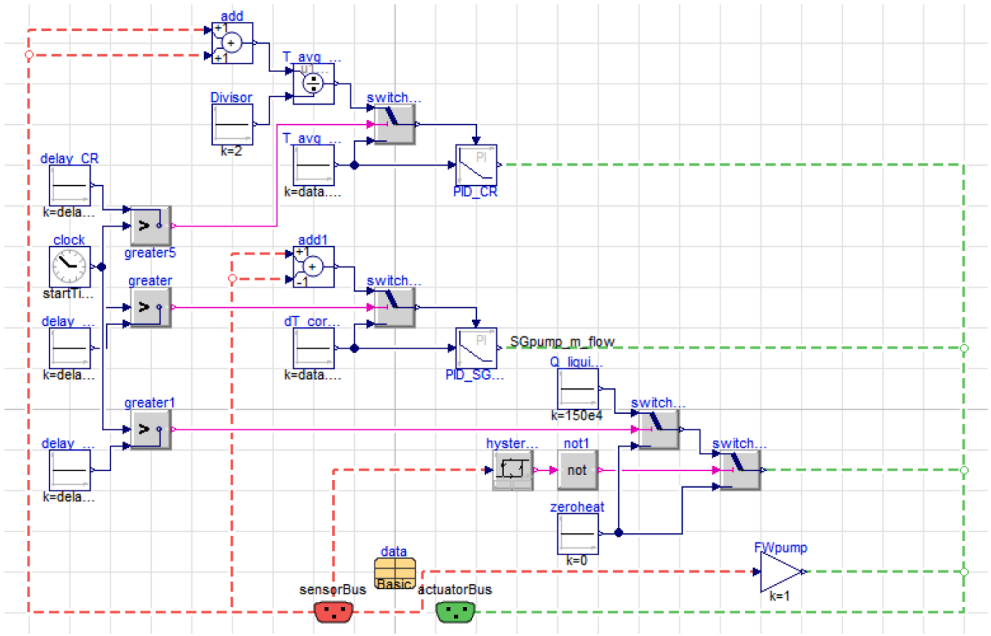

Fig. 3. Control system layout (Greenwood et al., 2017).

5.1. *Wide anomalies in the primary pump*

This methodology is first practiced by a scenario where the anomalies occur in the primary loop. The control system is modified to simulate this scenario. The control system reads the sensor data, calculates control commands based on different controllers (e.g., proportional–integral–derivative PID controller), then sends the commands to an actuator to carry out the commands. To simulate the on-site signals from the nuclear system, a series of Gaussian noises are added to the sensor readings to simulate the overall noise resulting during normal operations, which propagate to the whole system via the PID control actuators. as shown in Fig. 4.

The temporal evolutions of the total power under different conditions are shown in the first subplot of Fig. 5. The normal and the anomalous transient evolutions are shown as a blue curve and an orange curve, respectively. Both normal and anomalous power profiles are firstly denoised by Savitzky–Golay filter, and then standardized with zero mean and unit variance. The simulation time in this scenario lasts for 5000 s and the two anomalies are inserted as a 10 % gradual increase in the primary pump flow rate, followed by a gradual decrease, happening in 1000–2000 s and 3000–4000 s. The time for gradual increase and the decrease are the same, which is 500 s in each anomaly.

With the obtained temporal evolution, we must first smooth out the noise for both normal and anomalous profiles. Then, a sliding window is employed to take snapshots of the power evolution along the time axis for both normal and abnormal profiles and the snapshots of these

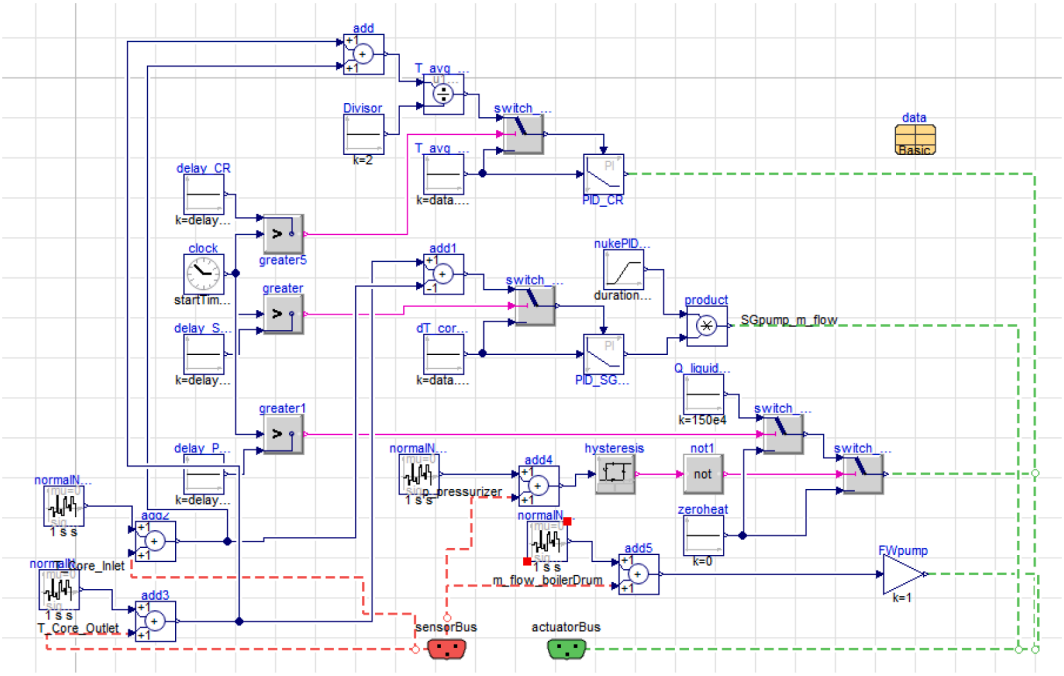

Fig. 4. Control system layout of the anomalous primary loop.

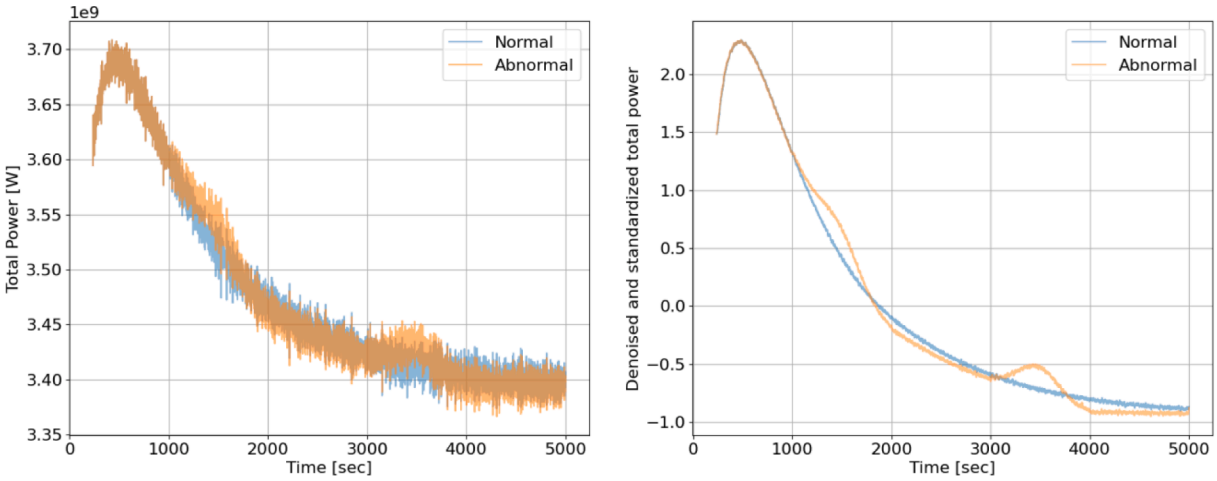

Fig. 5. Standardized total power with two wide anomalies.

profiles are aggregated separately to build two matrices, which works as a preparation step of RWD.

Examples of the HOFs extracted using the RWD algorithms are shown in Fig. 6 to depict different levels of sensitivity to the anomalies. The subplot (a) and (b) represent the HOFs with low and moderate sensitivity, respectively. The indices of the HOFs are shown as subscript of the y-axis label. The subplot (c) and (d) show the features with the high sensitivity. In these figures, the blue dots represent the HOF-captured normal transient behavior, and red dots indicate where the anomalous data contaminates the windowed signals. In the two anomalies in subplot (c) and (d), the two anomalies are shown as a wave, aligning with the anomalous pump flow rate, by which we can tell that nonlinearity of the time series can be captured by the HOFs.

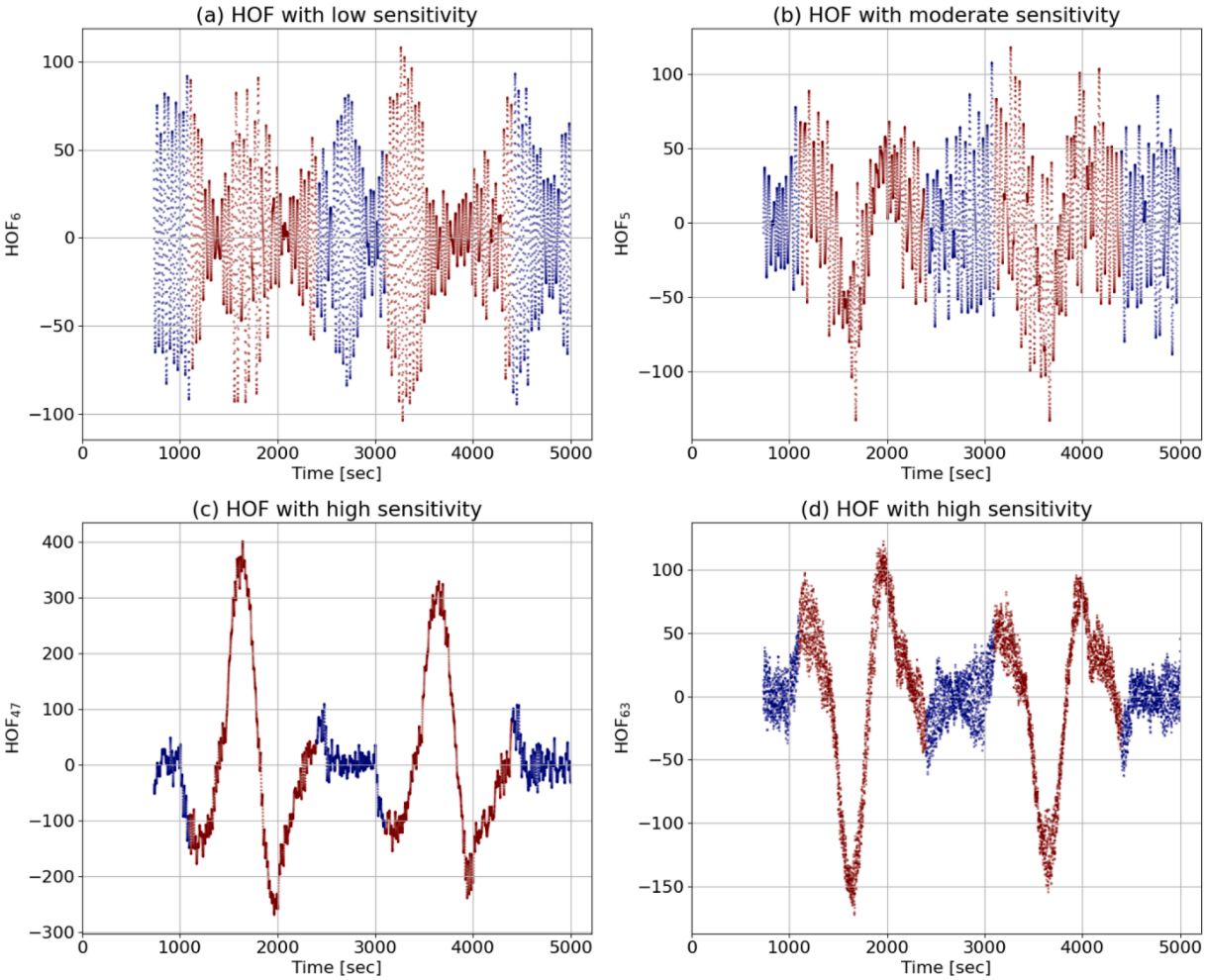

Fig. 6. Illustration of HOFs for two wide anomalies.

5.2. Narrow anomalies in the primary pump

Next, models exhibiting narrow anomalies are also developed. The simulation time in this scenario lasts for 5000 s and the two anomalies happen in 1900–2100 s and 3900–4100 s. The transient power is exemplified in Fig. 7. Similar to the previous scenario, the anomalies are inserted as a 10 % increase of the primary pump flow rate in the first 100 s, and a gradual decrease in the next 100 s. Representative HOFs are shown in Fig. 8. From the results, we can tell that the anomaly-sensitive HOFs usually captures the transient variation of the physics quantity, but the unsensitive ones usually represents high frequency behavior in both normal and anomalous regions, which indicate the occurrence of noises.

5.3. Persistent anomalies in the primary pump

Next are models showing various anomalies, including a persistent anomaly developed to depict the transient power, as exemplified in the first subplot of Fig. 9. Both normal and anomalous power profiles are firstly denoised by Savitzky–Golay filter, and then standardized with zero mean and unit variance. The preprocessed power profiles are shown in second subplot of Fig. 9. In this scenario, the simulation lasts for 10,000 s, and the four anomalies happen in 1900–2100 s, 4000–4100 s, 6000–6200 s, and 8000–8100 s. The first two anomalies are inserted by increasing the pump flow rate by 10 % within 50 s gradually and reducing the flow rate back to normal in the next 50 s gradually. The third anomaly is inserted by adding a sinusoidal oscillation of the flow rate in 200 s. Finally, the last anomaly represents as increasing the flow rate in 100 s gradually and let the system to reach its steady state without any further manipulation of the system. Fig. 10 shows four candidate HOFs for semi-supervised learning and the identification of unlabeled anomalies. From these four HOFs, we can tell that different types of anomalies trigger various anomalies at different levels. To amplify the subtle anomalies and reduce the triggered normal data, feature fusion is also adopted in this anomalous scenario, and the fused HOFs are shown in Fig. 11.

5.4. Comparative study between different feature extraction method

As stated in the introduction, there are two main advantages of RWD over the residual-based approach: (1) RWD decomposes the residual to a spectrum of HOFs, which can separate the anomaly-sensitive features from noises, enabling the differentiate between normal and subtle anomaly; (2) residual-based approach only provides one feature, the residual, to determine the state of monitored system. To provide the usefulness and applicability of the RWD approach for feature extraction, a comparative study between the RWD approach and the mainstream residual-based approach is conducted. The residual is generated from a rank r approximation.

To ensure a fair comparison between the residual-based approach and RWD-based approach, the indices of the HOFs representing RWD-based approach are all above identified rank r; in other words, the residual contains all HOFs, both anomaly-sensitive ones and the noise. The residual used for comparison is shown in Fig. 12. Then we provide different levels of label availability. The ratio of labeled data ranges from 0 % to 90 %. When the ratio of labeled data is 0 %, K-Means clustering method is used to distinguish the anomalies from normal data, with 60 % data for training and 40 % data for testing. The basic idea of K-Means algorithm is to separate the data into K specified clusters where each point belongs to one cluster, while minimizing the distance between the points and the cluster centroid (Sculley, 2010). Here to separate the data, the specified number of clusters is two, representing anomaly and normal behavior. When the ratio of labeled data is 10 %, a KNN classifier is used for classification with the 10 % labeled data, and the rest 90 % data are used for testing. Different from K-Means, KNN is a supervised learning method: the known labeled data are arranged in the feature space, and the label of the new point will be determined by comparing the classes of the K closest points, named as nearest neighbors (Goldberger et al., 2004). Here the number of neighbors is five, and Euclidian distance is used as the metric. There are two reasons for employing simple ML techniques to proceed anomaly detection. First, sophisticated ML techniques, like deep neural networks, usually come with large amount of hyperparameters. The classification/clustering performance can result from feature extraction or hyperparameter tuning. Also, the results from complicated ML techniques are not as robust and explainable as the ones from simple techniques like K-Means and KNN.

To evaluate the results, we use accuracy and F_1 score for the comparison between feature extraction methods. The accuracy is to evaluate the alarm time for the anomaly and the silent time for the normal region. Since the normal and anomalous data are imbalanced, F_1 score, the harmonic mean of precision and recall, is used taking the data imbalance into consideration. The label prediction accuracy is shown in Fig. 13. Here the residual-based approach is shown as green curve with triangle markers; the HOF-based approach is shown as blue curve with circle markers; the accuracy with fused HOFs is represented as orange curve with square markers. From this result we can tell that the two HOF-based approaches have higher accuracy than the residual-based approach, and

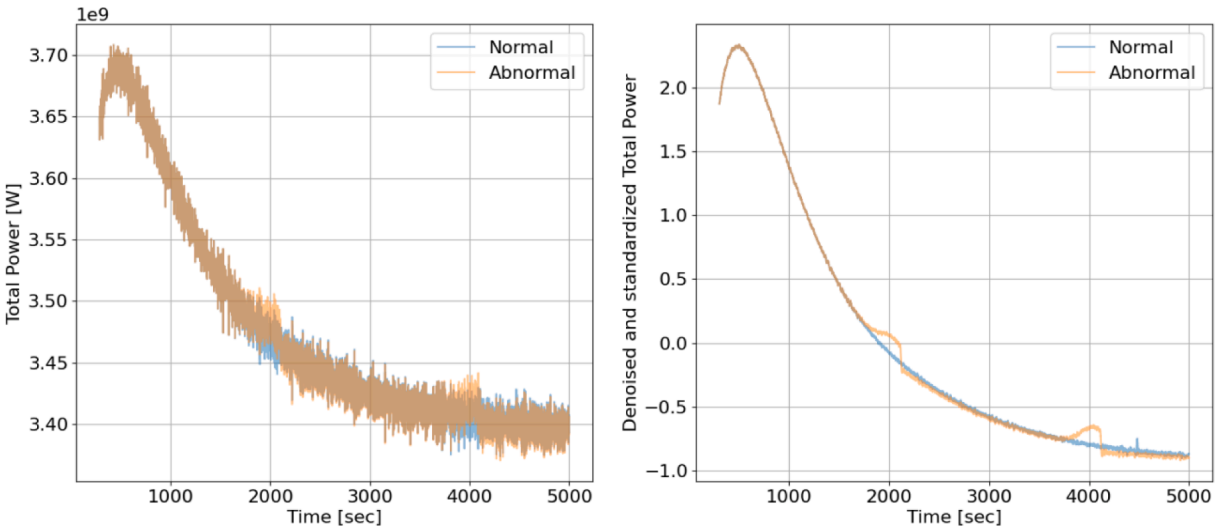

Fig. 7. Abnormal scenario with two narrow anomalies.

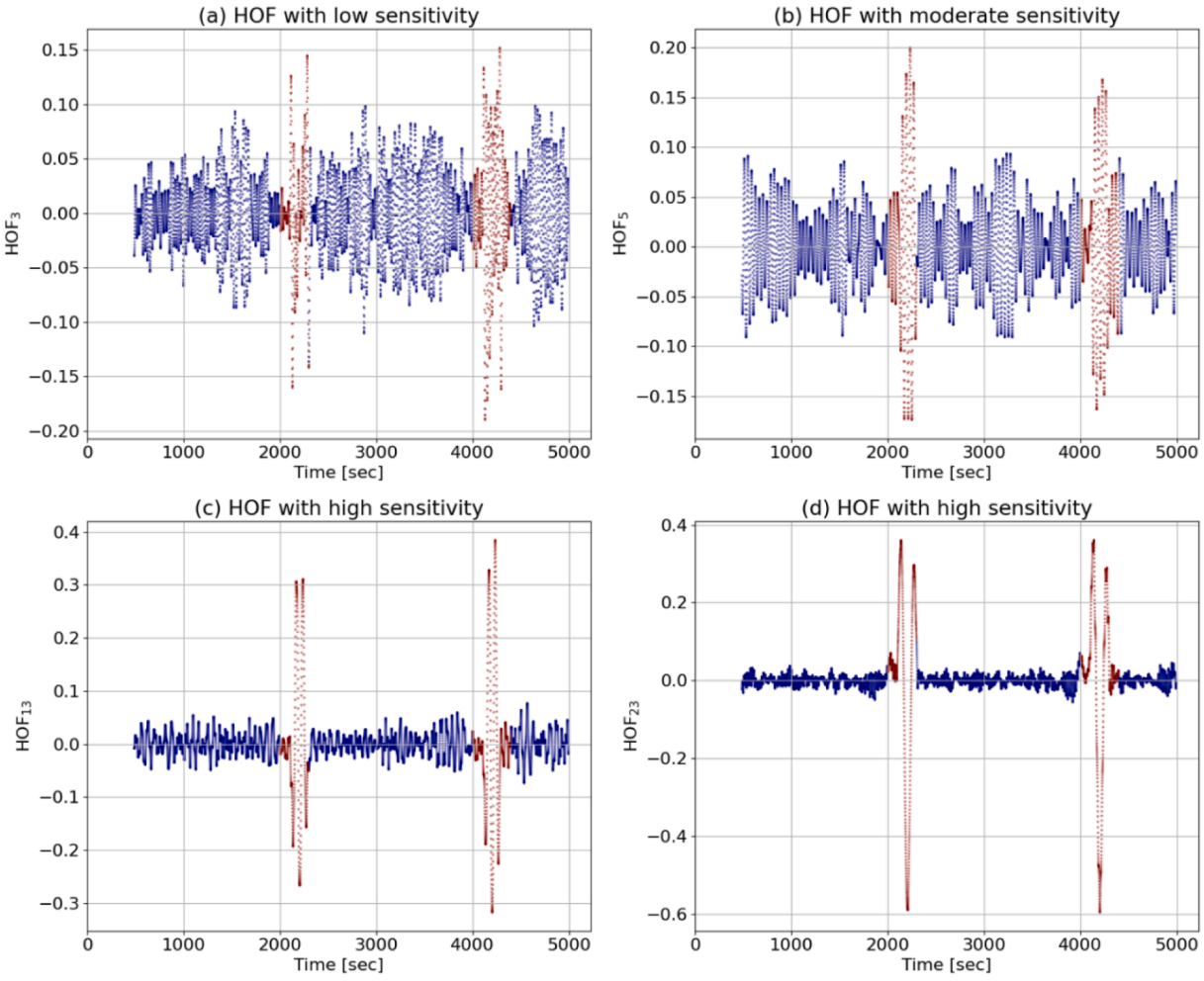

Fig. 8. Illustration of HOFs.

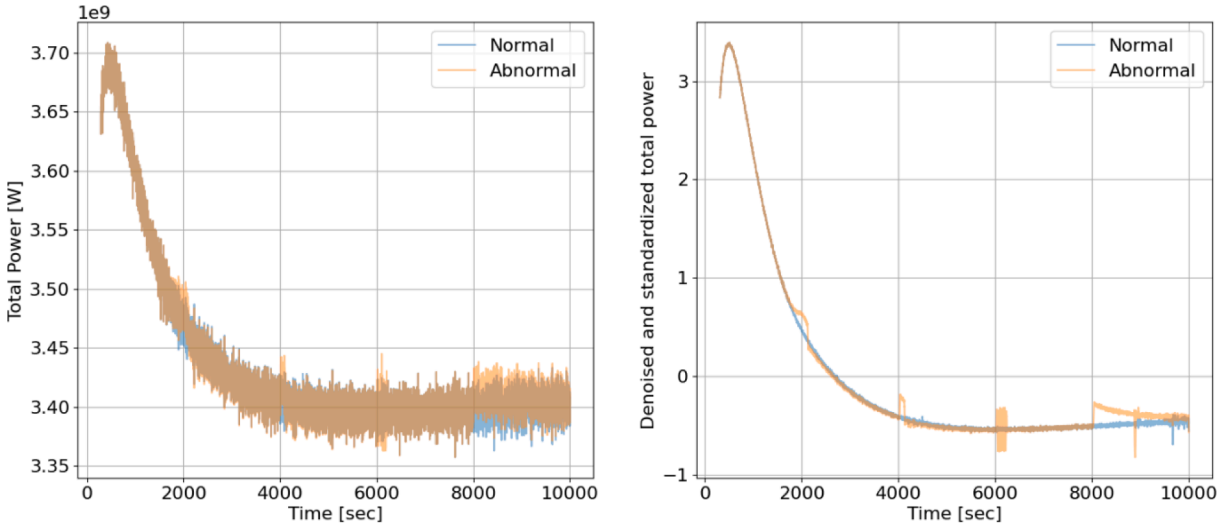

Fig. 9. Abnormal scenario with four narrow anomalies.

the feature fusion method can further enhance the stability of the anomaly detection performance. For the nuclear system in this scenario, the anomalous data are much less than the normal data, and correct prediction of anomalies, i.e., the true positives, are more important than true positives. Thus, F_1 score in Fig. 14 provides the feature performance with respect to data imbalance. The F_1 score results also indicate that the HOF-based approaches have more predicted anomalous data. Both figures show that as the labeling information increases, the classification have a better performance, which aligns with the common understanding of the ML techniques.

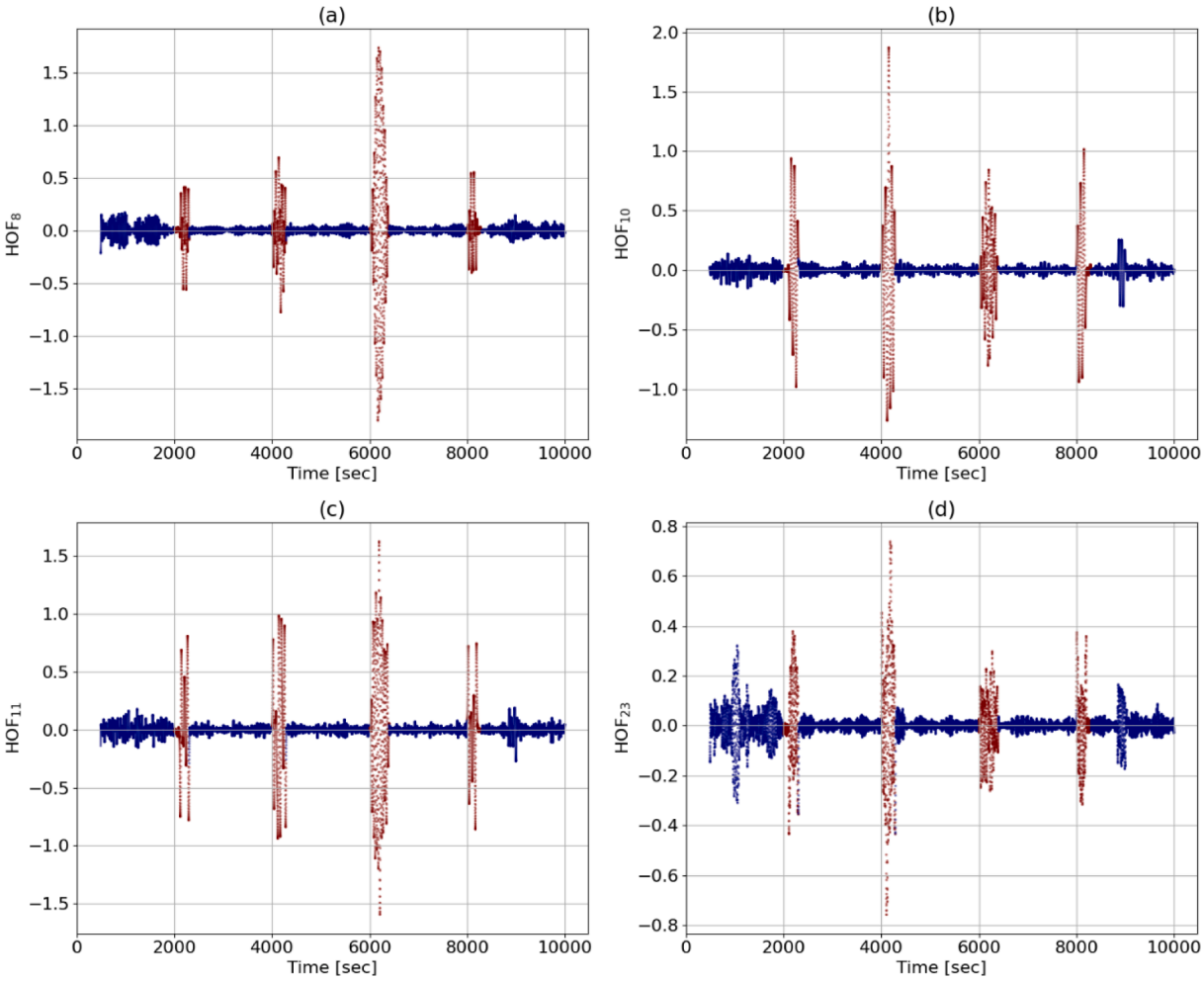

Fig. 10. Candidate HOFs for anomaly detection and feature fusion.

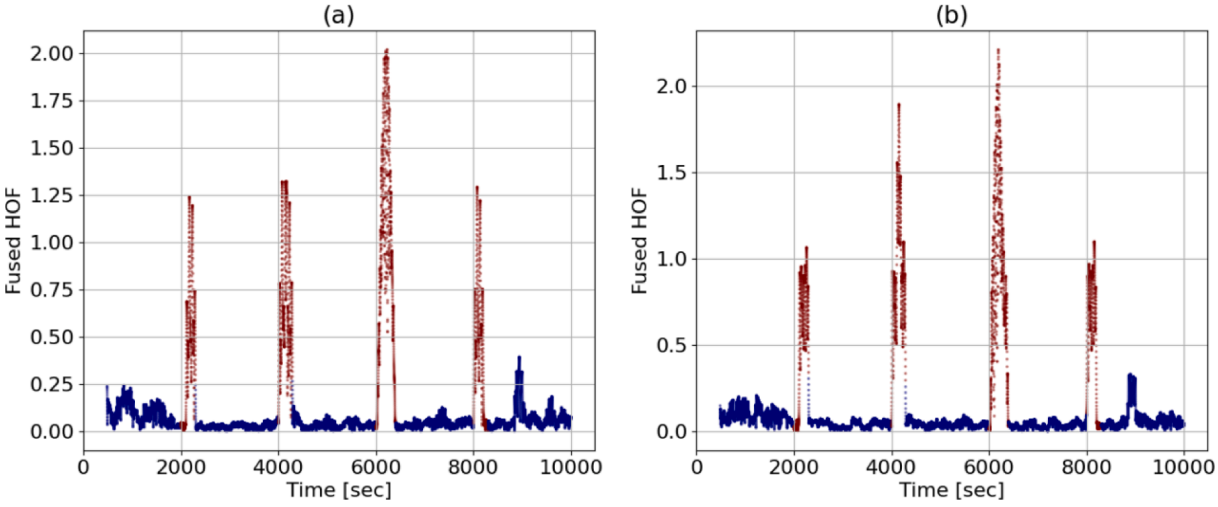

Fig. 11. Fused candidate HOFs.

6. Conclusion

The authors recognize the increased value of anomaly detection techniques in supporting operation of engineering systems that increasingly rely on digitization and automation. A gap area is also recognized in the anomaly detection literature, especially in detecting subtle anomalies in the presence of few labels. The state-of-the-art methods have proven to be effective in detecting both subtle and abrupt changes when a highly visible anomaly exists. Abrupt changes can be detected using purely data-driven models. However, the area of subtle anomaly detection using data-driven methods is less mature. To address this gap, a new algorithm is proposed that explores the weak

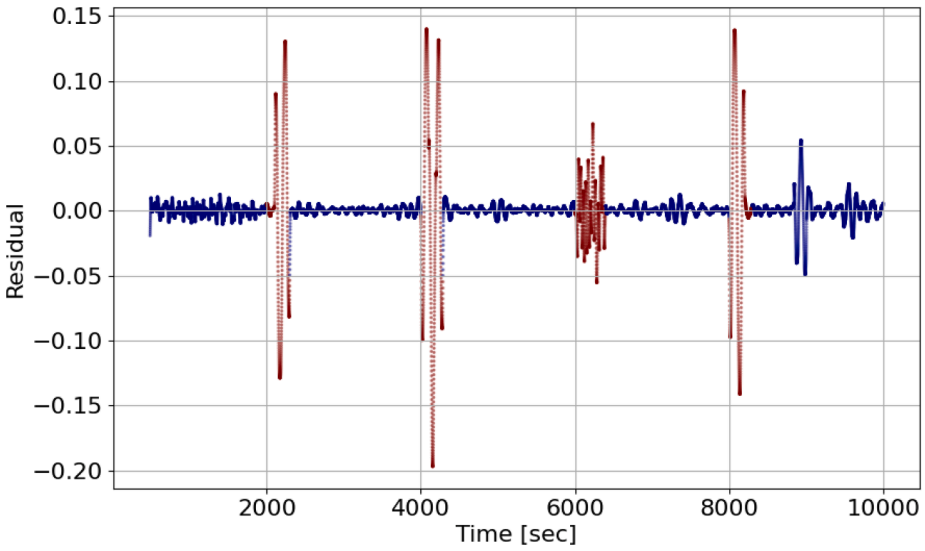

Fig. 12. Residual from rank approximation.

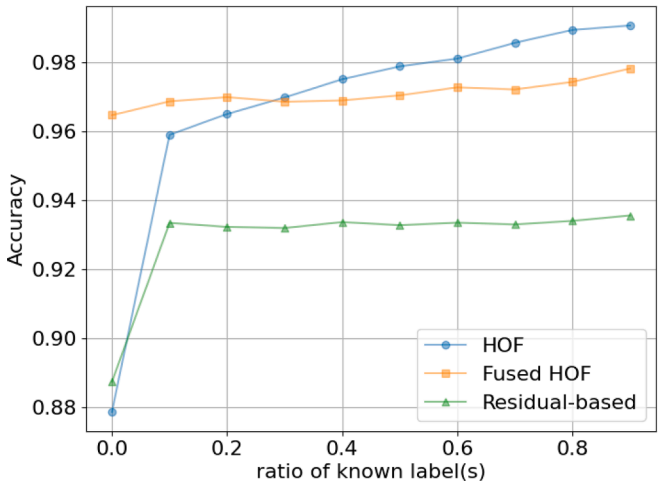

Fig. 13. Label prediction accuracy with different label availability.

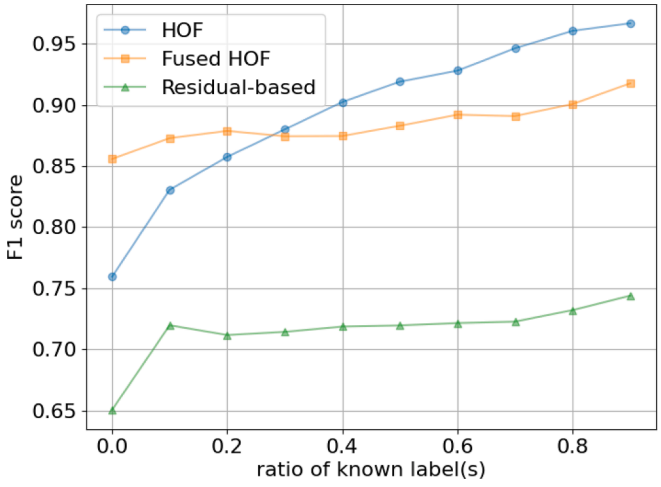

Fig. 14. Label prediction F_1 score with different label availability.

patterns, discarded as noise by unsupervised learning techniques, for their possible association with labeled anomalies from past operation. This situation is typical in most engineering systems, such as nuclear reactors, where maintenance activities keep records of past anomalies. The mathematical and implementation details are provided in the manuscript, and the applicability is demonstrated using a Dymola-simulated PWR with different types of anomalies. The results in subsections V. A and V. B indicate that the anomalous threshold can be established based on labeled data of one type of anomaly, allowing for the subsequent detection of unlabeled anomalies. In subsection V. C, the results indicate that the establishment of the threshold will vary with the type of anomalies. In subsection V. D, a comparative study between residual-based approach and the RWD-based approach is conducted. The results indicate the RWD-based anomaly detection method has a higher accuracy and F_1 score than the residual-based approach. This work demonstrates the basic idea of employing HOFs to detect subtle anomalies. Future work will focus on the development of an algorithm for the automatic selection and fusion of HOF candidates using ML-based algorithms.

7. Disclaimer

CRediT authorship contribution statement

Yeni Li: Conceptualization, Methodology, Software, Validation,

Investigation, Visualization, Formal analysis, Writing – original draft, Writing – review & editing. **Hany S. Abdel-Khalik:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Ahmad Al Rashdan:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Jacob Farber:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Basseville, M., Abdelghani, M., Benveniste, A., 2000. Subspace-based fault detection algorithms for vibration monitoring. Automatica 36 (1), 101–109. https://doi.org/10.1016/S0005-1098(99)00093-X.

Bornn, L., Farrar, C.R., Park, G., Farinholt, K., 2009. Structural health monitoring with autoregressive support vector machines. J. Vibr. Acoust. Trans. ASME 131 (2), 0210041–0210049. https://doi.org/10.1115/1.3025827.

Caesarendr W., Widodo A., Thoma P. H., Yang B.-S., Machine degradation prognostic based on RVM and ARMA/GARCH model for bearing fault simulated data, 2010.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM Comput. Surv. 41 (3), 1–58.

De Ketelaere, B., Hubert, M., Schmitt, E., 2015. Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data. J. Qual. Technol. 47 (4), 318–335. https://doi.org/10.1080/00224065.2015.11918137.

Esmalifalak M., Member S., Liu L., Member S., 2014. Detecting stealthy false data injection using machine learning in smart grid, 11(3) 1–9.

Fantoni, P.F., 2005. Experiences and applications of PEANO for online monitoring in power plants. Prog. Nucl. Energy 46 (3–4), 206–225. https://doi.org/10.1016/j.pnucene.2005.03.005.

Friedman, J.H., 1987. Exploratory projection pursuit. J. Am. Stat. Assoc. 82 (397), 249–266. https://doi.org/10.1080/01621459.1987.10478427.

Gawand, H.L., Bhattacharjee, A.K., Roy, K., 2017. Securing a cyber physical system in nuclear power plants using least square approximation and computational geometric approach. Nucl. Eng. Technol. 49 (3), 484–494. https://doi.org/10.1016/j.net.2016.10.009.

Goldberger, J., Hinton, G.E., Roweis, S., Salakhutdinov, R.R., 2004. Neighbourhood components analysis. Adv. Neural Inf. Process. Syst. 17.

Greenwood M. S., Cetiner M. S., Fugate D. L., Hale R. E., Harrison T. J., Qualls A. L., 2017. TRANSFORM - TRANsient Simulation Framework of Reconfigurable Models.

Guo J., "Model-Based Cyber-Security Framework for Nuclear Power Plant, 2020, [Online]. Available: https://deepblue.lib.umich.edu/handle/2027.42/162955%0Ahttps://deepblue.lib.umich.edu/bitstream/handle/2027.42/162955/gjunjie_1.pdf?sequence=1.

Huang, Y., Lai, L., Li, H., Chen, W., Han, Z., 2012. Online quickest multiarmed bandit algorithm for distributive renewable energy resources. In: 2012 IEEE 3rd International Conference on Smart Grid Communications, SmartGridComm 2012, pp. 558–563. https://doi.org/10.1109/SmartGridComm.2012.6486044.

Jiang, B.T., Zhou, J., Huang, X.B., Wang, P.F., 2020. Prediction of critical heat flux using Gaussian process regression and ant colony optimization. Ann. Nucl. Energy 149, 107765. https://doi.org/10.1016/j.anucene.2020.107765.

Li, Y., Abdel-Khalik, H.S., 2021. Data trustworthiness signatures for nuclear reactor dynamics simulation. Prog. Nucl. Energy 133, 103612.

Li, Y., Bertino, E., Abdel-Khalik, H., 2018. Effectiveness of model-based defenses for digitally controlled industrial systems: nuclear reactor case study. Nucl. Technol. 206 (1), 82–93.

Li, J., Huang, X., 2016. Cyber attack detection of I&C systems in NPPS based on physical process data. Int. Conf. Nucl. Eng., Proc., ICONE 2, 1–4. https://doi.org/10.1115/ICONE24-60773.

Li, Y., Sundaram, A., Abdel-Khalik, H.S., Talbot, P.W., 2022. Real-time monitoring for detection of adversarial subtle process variations. Nucl. Sci. Eng. 196 (5), 544–567.

Lu, B., Upadhyaya, B.R., 2005. Monitoring and fault diagnosis of the steam generator system of a nuclear power plant using data-driven modeling and residual space analysis. Ann. Nucl. Energy 32 (9), 897–912. https://doi.org/10.1016/j.anucene.2005.02.003.

Martin, E.B., Morris, A.J., 1996. An overview of multivariate statistical process control in continuous and batch process performance monitoring. Trans. Inst. Meas. Control 18 (1), 51–60. https://doi.org/10.1177/014233129601800107.

Peng, M.-J., Wang, H., Yang, X.u., Liu, Y.-K., Guo, L.-Z., Li, W., Jiang, N., 2017. Real-time simulations to enhance distributed on-line monitoring and fault detection in pressurized water reactors. Ann. Nucl. Energy 109, 557–573.

Riedel, K.S., Basseville, M., Nikiforov, I.V., Basseville, M., 1994. Detection of abrupt changes: theory and application. Technometrics 36 (3), 326. https://doi.org/10.2307/1269388.

Ruff L. et al., Deep Semi-Supervised Anomaly Detection, 2019, [Online]. Available: http://arxiv.org/abs/1906.02694.

Runger, G.C., Testik, M.C., 2003. Control charts for monitoring fault signatures: Cuscore versus GLR. Qual. Reliab. Eng. Int. 19 (4), 387–396. https://doi.org/10.1002/qre.591.

Sculley, D., 2010. Web-scale k-means clustering. In: ProceedIngs of the 19th International Conference on World Wide Web, pp. 1177–1178. https://doi.org/10.1145/1772690.1772862.

Shankar R. 2004. On line monitoring of instrument channel performance - Volume 3: Applications to nuclear power plant technical specification instrumentation, *Epri*, 3 (3).

Vaddi P. K. et al. 2020. Dynamic bayesian networks based abnormal event classifier for nuclear power plants in case of cyber security threats, *Progr. Nucl. Energy*, 128 (August) 103479, doi: 10.1016/j.pnucene.2020.103479.

Wang W., di Maio F., Zio E. 2019. *A non-parametric cumulative sum approach for online diagnostics of cyber attacks to nuclear power plants*. doi: 10.1007/978-3-319-95597-1_9.

Wang, X., Kruger, U., Irwin, G.W., 2005. Process monitoring approach using fast moving window PCA. Ind. Eng. Chem. Res. 44 (15), 5691–5702. https://doi.org/10.1021/ie048873f.

Wu, G., Tong, J., Zhang, L., Zhao, Y., Duan, Z., 2018. Framework for fault diagnosis with multi-source sensor nodes in nuclear power plants based on a Bayesian network. Ann. Nucl. Energy 122, 297–308. https://doi.org/10.1016/j.anucene.2018.08.050.

Zhang, F., Coble, J.B., 2020. Robust localized cyber-attack detection for key equipment in nuclear power plants. Prog. Nucl. Energy 128 (July), 103446. https://doi.org/10.1016/j.pnucene.2020.103446.

Zhao K. An Integrated Approach to Performance Monitoring and Fault Diagnosis of Nuclear Power Systems, 2005.