

## RESEARCH ARTICLE

# Improvement of Kernel Principal Component Analysis-Based Approach for Nonlinear Process Monitoring by Data Set Size Reduction Using Class Interval

MOHAMMED TAHAR HABIB KAIB<sup>1</sup>, ABDELMALEK KOUADRI<sup>1</sup>, MOHAMED-FAOUZI HARKAT<sup>2</sup>,  
ABDERAZAK BENSMAIL<sup>1,3</sup>, AND MAJDI MANSOURI<sup>4</sup>, (Senior Member, IEEE)

<sup>1</sup>Signals and Systems Laboratory, Institute of Electrical and Electronics Engineering, University M'Hamed Bougara of Boumerdes, Boumerdes 35000, Algeria

<sup>2</sup>National Higher School of Technology and Engineering, Annaba 23000, Algeria

<sup>3</sup>Cement Production Plant SCAEK, Ain El Kebira, Setif 19008, Algeria

<sup>4</sup>Electrical and Computer Engineering Program, Texas A&M University at Qatar, Doha, Qatar

Corresponding author: Majdi Mansouri (majdi.mansouri@qatar.tamu.edu)

This work was supported in part by the Qatar National Library, and in part by the Qatar National Research Fund (QNRF) Research Grant.

**ABSTRACT** Fault detection and diagnosis (FDD) systems play a crucial role in maintaining the adequate execution of the monitored process. One of the widely used data-driven FDD methods is the Principal Component Analysis (PCA). Unfortunately, PCA's reliability drops when data has nonlinear characteristics as industrial processes. Kernel Principal Component Analysis (KPCA) is an alternative PCA technique that is used to deal with a similar data set. For a large-sized data set, KPCA's execution time and occupied storage space will increase drastically and the monitoring performance can also be affected in this case. So, the Reduced KPCA (RKPCA) was introduced with the aim of reducing the size of a given training data set to lower the execution time and occupied storage space while maintaining KPCA's monitoring performance for nonlinear systems. Generally, RKPCA reduces the number of samples in the training data set and then builds the KPCA model based on this data set. In this paper, the proposed algorithm selects relevant observations from the original data set by utilizing a class interval technique (i.e. histogram) to maintain a bunch of representative samples from each bin. The proposed algorithm has been tested on three tank system pilot plant and Ain El Kebira Cement rotary kiln process. The proposed algorithm has successfully maintained homogeneity to the original data set, reduced the execution time and occupied storage space, and led to decent monitoring performance.

**INDEX TERMS** Fault detection and diagnosis (FDD), data-driven techniques, time and storage space complexity, principal component analysis (PCA), kernel principal component analysis (KPCA), reduced KPCA (RKPCA), histogram, cement plant, three tanks system.

## I. INTRODUCTION

Industrial plants may suffer from faults at some point during their working time, affecting the production process. Because of this, monitoring systems are implemented to detect those faults and to avoid major casualties which raises the importance of ensuring their reliability and robustness.

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

PCA is a data-driven FDD technique, it lowers the number of variables in a given data set by mapping this data into a new lower-dimension space such that variables in this space possess most of the original data set variability, the basis of this space being a set of linearly independent vectors [1], [2]. PCA does this mapping by computing eigenvalues and eigenvectors of the input data set covariance matrix [3]. PCA-based FDD technique becomes an attractive technique for its flexibility and simplicity derived from the Single Values

Decomposition (SVD) problem [4]. Although PCA has a fine reputation and is widely used, it does not have the ability to keep up its performance when data under study has nonlinear characteristics as most industrial processes [5]. Many techniques have been proposed to overcome this kind of obstacle. Dong et al. [6] proposed principal curves PCA which merges both neural network and principal curves algorithm, Jia et al. [7] proposed a novel nonlinear PCA based on input training network and by defining non-parametric control limits for FDD, Geng et al. [8] proposed adaptive nonlinear PCA based on improved input training neural network [4], Scholkopf et al. [9] proposed KPCA that executes conventional PCA in higher dimensional feature space.

KPCA is a nonlinear extension of the PCA technique. It maps a given data set from input space to a higher dimensional feature space, Hilbert space, using kernel trick without involving a mapping function, and then it executes PCA in this new space [9]. KPCA is widely used for its simplicity compared to other nonlinear PCA techniques because it is based on solving eigenvalues problem just as conventional PCA, furthermore, no optimization is needed to achieve good results [9]. Because of its ability, KPCA consumes more computational time and occupies more storage space than the conventional PCA technique and may lose its accuracy if the data set is too large. Thus, KPCA may be disadvantageous when the training data set has a very large number of observations and eventually worthless when developing expensive computers [4], [9], [10]. Reduced KPCA (RKPCA) is introduced to overcome the drawbacks of the conventional KPCA technique. RKPCA uses a reduced number of observations while retaining as much useful information as the original training data set, then the KPCA model is built upon this reduced data set, helping to decrease the effect of the large-sized data set on the performance of the conventional KPCA [11].

To use conventional PCA, KPCA, and RKPCA algorithms for fault detection and diagnosis, it is required to use Statistical Process Monitoring (SPM) which is based on data-driven models by taking a number of measurements from different variables under healthy operating conditions of a system, in other words, it is based on monitoring the variations in a given data set. These variations can be categorized as common variations such as natural variations in a process and special variations such as faults [12]. Different statistics in SPM are used as fault detection indices, like the Hotelling ( $T^2$ -statistic), the Square Prediction Error (SPE or  $Q$ -statistic), and the combined index ( $\varphi$ -statistic). The  $T^2$ -statistic evaluates variations in principal components subspace, whereas the  $Q$ -statistic evaluates variations in residual subspace [13]. For each SPM data set, These fault detection indices are evaluated using different metrics which are: False Alarm Rate (FAR), Missed Detection Rate (MDR), and Detection Time Delay (DTD).

The proposed algorithm is based on the RKPCA technique where the retained observations are selected using class intervals similar to the histogram of the 1<sup>st</sup> PCA's principal component score. A histogram is a bar-plot type that groups data into commonly equally spaced intervals called bins on the horizontal axis, and the corresponding appearance frequencies on the vertical axis [14]. The idea of the proposed algorithm is to select a number of representative observations from each bin to form the reduced data set which is used to build the KPCA model for fault detection purposes. This study aims to form a reduced data matrix that maintains homogeneity to the original data set, reduces execution time and the occupied storage space, and manages to attain excellent FDD results. The reduction part of the proposed RKPCA algorithm is based on a class interval that does not require either optimization or predefined values. This work is proposed to create an RKPCA algorithm that can effectively reduce the number of observations in a large-sized training data set and enhance monitoring performance. In other words, the main contribution of this paper is the development of a reliable and effective method for selecting the most relevant observations using a class interval scheme while preserving homogeneity to the original data set. It is well-defined and duly analyzed theoretically. It improves RKPCA-based fault detection performance. Consequently, it reduces significantly both execution time and the needed storage space. Although the proposed algorithm is effective and easy to utilize, it is up to the user to select the appropriate reduced data matrix. As a numerical example, the developed algorithm was tested using a tank system data set and compared to both PCA and KPCA techniques, then it was applied and compared to other algorithms using a real industrial data set obtained from Ain El Kebira cement plant rotary kiln.

This article is organized as follows: Section II introduces some of the related work, section IV gives a literature review on PCA and KPCA for FD, section V introduces the proposed algorithm, section VI gives a brief description of the processes used in this paper, section VII shows results and gives discussion about them, and section VIII gives a general conclusion.

## II. RELATED WORK

In this section, some of the related works are presented. Most of the existing algorithms follow almost the same idea which is reducing a large-sized data set by omitting redundant and/or correlated samples. The difference between those algorithms is how to determine the samples to be omitted and the effectiveness of the reduction method. Felipe et al. [15] introduced an RKPCA algorithm based on the k-means clustering algorithm which needs an appropriate predefined number of clusters in order to have good results. Harkat et al. [16] proposed an RKPCA algorithm that uses PCA as data reduction on the transposed matrix of the input data set to select the uncorrelated observations, since the reduction is based on the PCA algorithm it may be

affected by the size of the covariance matrix which means it will drop its accuracy for very large data size, the reduced data matrix using PCA approach contains the de-mapping observations from PCA's principal components subspace to input space thus the new observations are slightly different from original ones. Lahdhiri et al. [17] presented an algorithm that keeps observations for which the kernel matrix has full rank, unfortunately this method retains a very large number of observations in the reduced matrix. Taouali et al. [18] introduced an algorithm that keeps observations that point to the largest variances with retained principal components of the kernel matrix of the total observations, the reduction part of this algorithm is highly related to the selected principal components so in order to have the best overall reduced data it can take a lot of time to obtain since for large data sets the number of principal components choices is large. Bencheikh et al. [4] proposed an algorithm that keeps observations with a given Euclidean distance between them so the homogeneity between the original and reduced data is not guaranteed. In [19], the authors use an RKPCA algorithm based on a variogram as a reduction technique. unluckily, the reduced data matrix is still quite large nearly 70% of the original data set because it deals only with spatial continuity. Thus, this approach highly depends on the smallest selected lag which in some processes is small and less than half of the number of samples, making the consequential data matrix the same as the original one and the user will not have any reduction in the data set. More recently, Kaib et al. in [20] introduced a reduction technique based on the fractal dimension which represents the intrinsic dimension of a large-sized training data set and is implicitly defined. Thus, the homogeneity of effectively reduced data sets is interpreted concisely and cautiously, as the power of statistical tests available for this purpose might be limited. Furthermore, this method is only applicable to chaotic systems.

The proposed algorithm tends to reduce the size of the training data set using the class interval, histogram, while maintaining homogeneity between the two data sets, this can help the proposed algorithm to achieve decent overall performances.

### III. TERMINOLOGIES AND NOTATIONS

#### A. ABBREVIATIONS

- CPV: Cumulative Percentage Variance.
- DTD: Detection Time Delay.
- FAR: False Alarm Rate.
- FD: Fault Detection.
- FDD: Fault Detection and Diagnosis.
- KPCA: Kernel Principal Component Analysis.
- MDR: Missed Detection Rate.
- PC: Principal Component.
- PCA: Principal Component Analysis.
- RKPCA: Reduced Kernel Principal Component Analysis.
- SPM: Statistical Process Monitoring.

#### B. MATHEMATICAL NOTATIONS

- $\hat{A}$ : Principal Components of A.
- $\tilde{A}$ : Residuals of A.
- $exp$ : Natural base exponent.
- $\sum$ : Covariance-Correlation matrix.
- $\mathcal{R}^d$ : Space of dimension  $d$ .
- $\|v\|$ : Euclidean norm of the vector  $v$ .
- $x$ : Row variable.
- $X$ : Matrix variable.
- $X^T$ : Transpose of matrix  $X$ .
- $X^{-1}$ : Inverse of matrix  $X$ .

#### C. PARAMETERS AND HYPERPARAMETERS

- $2\delta^2$ : Hyperparameter or ( $\gamma$ ) of the Radial Basis Function.
- $\sigma^2$ : variance of the input data set.
- $N_B$ : Number of Bins in the proposed algorithm.
- $\nu$ : Appearance frequency

### IV. LITERATURE REVIEW

#### A. PCA

PCA is known for its simplicity and low execution time value, it decomposes inter-correlated variables into a principal component set of variables [21].

Let  $X_o^{n \times m}$  be the data set matrix collected from the normal operating system of  $n$  observations from  $m$  sensors. Before proceeding with the conventional PCA, this data set matrix is normalized to zero mean and unit standard deviation to obtain  $\bar{X}^{n \times m}$ . Equation (1) demonstrates how the covariance matrix,  $\Sigma$ , is computed and decomposed using Singular Value Decomposition (SVD).

$$\Sigma_{m \times m} = \frac{1}{n-1} \bar{X}^T \bar{X} = P \Lambda P^T \quad (1)$$

The columns of  $P_{m \times m}$  are known as loading vectors and  $\Lambda$  is a diagonal matrix containing the eigenvalues of  $\Sigma$  in descending order,  $[\lambda_1 \dots \lambda_m]$ . The scores matrix,  $T$ , is then given by

$$T = XP \quad (2)$$

There exist different methods for the selection of the number of Principal Components (PCs) to be retained. The Cumulative Percentage Variance (CPV) is a widely used method, it selects the number of PCs,  $l$ , such that the sum of their variances,  $CPV(l)$ , is greater than or equal to a predefined value  $CPV_{limit}$  [16]. It is given as follows

$$CPV(l) = \frac{\sum_{i=1}^l \lambda_i}{\sum_{i=1}^m \lambda_i} \times 100(\%) \geq CPV_{limit}(\%) \quad (3)$$

After selecting PCs, the loading vectors can be decomposed into  $\hat{P}_{m \times l}$ , which represents the principal loading vectors, and  $\tilde{P}_{m \times (m-l)}$  that is the residual loading vectors.

$$P = \left[ \hat{P}_{m \times l} \quad \tilde{P}_{m \times (m-l)} \right] \quad (4)$$

In the same way, the scores and eigenvalues matrices are decomposed as in equations (5) and (6), respectively.

$$T = \begin{bmatrix} \hat{T}_{n \times l} & \tilde{T}_{n \times (m-l)} \end{bmatrix} \quad (5)$$

$$\Lambda = \begin{pmatrix} \hat{\Lambda}_{l \times l} & 0_{l \times (m-l)} \\ 0_{(m-l) \times l} & \tilde{\Lambda}_{(m-l) \times (m-l)} \end{pmatrix} \quad (6)$$

The  $T^2$ -statistic, the  $Q$ -statistic, and  $\varphi$ -statistic are given by equations (7), (8), and (9), respectively.

$$T^2 = x^T \hat{P} \hat{\Lambda}^{-1} \hat{P}^T x \quad (7)$$

$$Q = x^T \tilde{P} \tilde{P}^T x \quad (8)$$

$$\varphi = \frac{T^2}{T_\alpha} + \frac{Q}{Q_\alpha} \quad (9)$$

Their corresponding upper control limits  $T_\alpha$ ,  $Q_\alpha$ , and  $\varphi_\alpha$  are given as shown in equations (10), (11), and (12), respectively.

$$T_\alpha = \frac{(n^2 - 1)l}{n(n-l)} F_\alpha(l, n-l) \quad (10)$$

$F_\alpha(l, n-l)$  represents Fisher-Snedecor distribution with  $l$  and  $(n-l)$  degrees of freedom, and  $\alpha$  is the significance level.

$$Q_\alpha = \frac{\sigma_Q^2}{2\mu_Q} \chi^2 \left( \alpha, \frac{2\mu_Q^2}{\sigma_Q^2} \right) \quad (11)$$

where  $\sigma_Q^2$  is the variance of  $Q$ -statistic, and  $\mu_Q$  is its mean, and  $\chi^2$  is the chi-square distribution.

$$\varphi_\alpha = g \chi^2(\alpha, h) \quad (12)$$

where  $g$  and  $h$  can be found in [22].

For each statistic, a fault is reported when its value exceeds its corresponding upper control limit.

### B. KPCA

The main idea behind the KPCA algorithm is to perform PCA in feature space,  $\mathcal{F}$ , using kernel trick instead of the mapping function  $\phi$  [23].

Let  $X^{n \times m}$  be normalized data set to zero mean and unit standard deviation, then the mapping function  $\phi$  is defined as the following:

$$\phi : x_i \in \mathbb{R}^m \rightarrow \phi(x_i) \in \mathbb{R}^f \quad i = 1 \dots n \quad (13)$$

where  $x_i$  is a row vector from  $X^{n \times m}$  and  $f \gg m$ .

By making the assumption that the vectors in feature space are scaled to zero mean and unit variance, mapped data can be expressed as  $\Phi = [\phi(x_1) \phi(x_2) \dots \phi(x_n)]^T$  [9], the covariance matrix in feature space,  $\mathcal{F}$ , is then computed as the following:

$$\Sigma_\phi = \frac{1}{n} \Phi^T \Phi = \frac{1}{n} \sum_{i=1}^n \phi(x_i) \phi(x_i)^T \quad (14)$$

As property of this feature space is that the dot product  $(\phi(x_i) \cdot \phi(x_j))$  can be computed using kernel trick (kernel function) as the following [9]:

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j) \quad i, j = 1 \dots n \quad (15)$$

There are different types of kernel functions, but the most used one is the Radial Basis Function (Gaussian) [19], and it is given as:

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\delta^2}\right) \quad (16)$$

where the hyperparameter  $2\delta^2$  can be found using the formula  $2\delta^2 = 2r\sigma^2$ , such that  $r$  is empirically obtained, and  $m$  and  $\sigma^2$  are the number of variables and the variance of the data set matrix, respectively [23].

As conventional PCA, KPCA tends to solve the eigenvalues problem defined as in equation (17):

$$\Sigma_\phi V = \lambda V \quad (17)$$

Since the mapping itself is not known, the Gram matrix  $\Phi^T \Phi$  can be computed using kernel trick as in equation (15). with the use of this trick, the matrix  $K = \Phi^T \Phi$  is defined as the following equations:

$$K = \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{bmatrix} \quad (18)$$

The retained PCs are selected by the CPV method as shown in equation (3).

For FDD, the monitoring indices  $T^2$ ,  $Q$ , and  $\varphi$ -statistics are used. The way to compute them is shown in equations (19), (20), and (21) respectively. The Hotelling  $T^2$ -statistic is given as the following:

$$T^2 = k(x)^T \hat{P}_\phi \hat{\Lambda}_\phi^{-1} \hat{P}_\phi^T k(x) \quad (19)$$

where  $\hat{P}_\phi$  is the principal eigenvectors of  $K$ , and  $\hat{\Lambda}_\phi$  diagonal elements are the principal eigenvalues of  $K$ . The upper control limit of this statistic is given using equation (10).  $Q$ -statistic is obtained by:

$$Q = k(x, x) - k(x)^T \hat{P}_\phi \hat{\Lambda}_\phi^{-1} \hat{P}_\phi^T k(x) \quad (20)$$

Its upper control limit is given as in the conventional PCA by equation (11). For both equations (19) and (20),  $k(x)$  is given as  $k(x) = [k(x_{1,\cdot}, x) \ k(x_{2,\cdot}, x) \ \dots \ k(x_{n,\cdot}, x)]^T$  [22].  $\varphi$ -statistic is computed as:

$$\varphi = \frac{T^2}{T_\alpha} + \frac{Q}{Q_\alpha} \quad (21)$$

The upper control limit is given by equation (12).

For each of the aforementioned fault detection indices, three different monitoring performance metrics are evaluated. The False Alarm Rate (FAR) represents the rate of normal samples that exceed the control limit of a given fault detection index during the system normal operation conditions, it is described by equation (22). The Missed Detection rate (MDR) characterizes the faulty samples that don't exceed the control limit under defective operation conditions, it is described as in equation (23). The Detection Time Delay (DTD) is the

number of samples between the appearance of a fault and its detection, it is described by equation (24).

$$FAR = \frac{NF}{NOC} \tag{22}$$

$$MDR = \frac{FN}{FOC} \tag{23}$$

$$DTD = t_D - t_O \tag{24}$$

$NF$  is the number of non-faulty samples that exceed the control limit,  $NOC$  is the number of total non-faulty samples,  $FN$  is the number of faulty samples that didn't exceed the control limit,  $FOC$  is the number of total faulty samples,  $t_D$  is the sample of the detection of a fault while  $t_O$  is the sample of its occurrence.

### V. PROPOSED RKPCA ALGORITHM

Histogram, in statistics, is a type of bar graph that displays the value of appearance frequency of specific data within a given bin, so it helps to visualize data distribution and its skewness [24]. The width of the bins is controlled by the number of bins used as shown by the following equation:

$$B_w = \frac{M_V - m_V}{N_B} \tag{25}$$

where  $B_w$  is the bin width,  $M_V$  is the maximum value in data set,  $m_V$  is the minimum value in data set, and  $N_B$  is the number of bins. The y-axis can represent different types of values, such as appearance frequencies, probabilities, and percentages [24].

The proposed algorithm aims to reduce the size of large-sized data while preserving their most characteristics. It starts by computing the PCA model, using algorithm 1, for the original training data set and then selects  $1^{st}$  principal component score, because it contains the most information explained by its highest variance. After that, a histogram of this  $1^{st}$  principal component is plotted with a specified number of bins  $N_B$ . Lets define  $\varepsilon$  as the minimum appearance frequency,  $v_i$  as the appearance frequency of the  $i^{th}$  bin, and  $\gamma_i$  as shown in the following equation:

$$\gamma_i = \left\lceil \frac{v_i}{\varepsilon} \right\rceil \quad i = 1 \dots N_B \tag{26}$$

All the appearance frequencies will be scaled to the minimum appearance frequency  $\varepsilon$  such that the minimal frequency becomes 1 or 2 according to the frequency in this bin, and each other bin  $i$  has  $\gamma_i$  as frequency. The median is then taken as representative observations of a bin. The choice of having the same median ensures that the remaining samples are distributed through all the bins. After this, the proposed algorithm selects the corresponding rows in the scores matrix and performs the inverse of PCA mapping to obtain the reduced data matrix  $X_r$  with fewer observations than the original data set. Finally, use this data matrix to build the KPCA model, using algorithm 2, and test this model using faults described in section VI and loss functions (28, 29), these loss functions tend to minimize the number of abnormal

samples in FDD system. The proposed algorithm does not require optimization in the size reduction part because only a few values of  $N_B$  can lead to  $\varepsilon > 1$  and the size of  $X_r$  is directly related to  $\varepsilon$ .

For the normalization of the kernel matrix  $K$  and testing kernel matrix  $K_t$ ,  $1_n$  is equal to  $\frac{1}{n}$  time a  $(n \times n)$  square matrix of ones, the same thing goes for  $1_t$  the only difference is the matrix of ones is now a  $(t \times n)$  matrix where  $t$  is the number of testing samples. The first value of  $N_B$  is chosen such that  $\varepsilon = 1$ , which means in this case that the reduced data matrix is the same as the original one and then it is decreased for larger  $\varepsilon$  values and repeat using smaller values of  $N_B$  to obtain matrices with fewer observations. After that, the performance of each resulting reduced data matrix is computed to pick the one with the best performance and having fewer observations than the original data set as shown in Tables 1 and 5. Algorithm 3 illustrates how the proposed algorithm works. The loss function presented in the following equation (27) aims to reduce the MDR, DTD, and FAR values for a given monitoring index which means that the lower the value of  $J_s$  is the better the monitoring metrics values are and vice versa. Change the number of bins,  $N_B$ , and repeat until a satisfying result is obtained. Algorithm 3 illustrates how the proposed algorithm works.

$$J_{stat} = a_1 FAR_{stat} + a_2 MDR_{stat} + a_3 \left(1 - e^{-0.1 DTD_{stat}}\right) \tag{27}$$

$$J = \frac{1}{3}(J_{T^2} + J_Q + J_\varphi) \tag{28}$$

In order to have the same effect of different indices over the loss function, the different weighting factors are selected as the following  $a_1 = a_2 = a_3 = 1$ . Furthermore, the algorithm performance is duly considered in assessing monitoring performance overall.

$$J_O = J + \frac{Et_r}{Et_n} + \frac{So_r}{So_n} \tag{29}$$

where  $Et_r$  and  $So_r$  represent the execution time for one testing sample from the testing part (online) and the occupied storage space of the training part of algorithm 2 using the reduced data matrix  $X_r$ , respectively. Similarly,  $Et_n$  and  $So_n$  are the same as  $Et_r$  and  $So_r$  only this time the original training data is used. The cost function (29) takes into account the overall monitoring performance, the execution time, and the needed storage space all at once. For a lower values of  $J_O$ , the algorithm is expected to have a decent monitoring performance and require less execution time and less storage space for the model that is the subject of this study.

The proposed algorithm can improve the performance of RKPCA because its purpose is to eliminate redundant observations and keep homogeneity between original and reduced data sets. As mentioned in [9], the size of the Gram matrix  $K$  can affect the accuracy of the KPCA's performance if it is a large-sized matrix.

**Algorithm 1** PCA model.

- Scale  $X_o^{n \times m}$  with zero mean and unit variance
- Compute covariance matrix,  $\Sigma$ , using equation (1)
- Compute eigenvalues and eigenvectors of the covariance matrix.
- Select the Principal component using the CPV method as in equation (3).
- Compute principal scores matrix  $T$ .

**Algorithm 2** KPCA for Fault Detection**Training Part:**

- Scale  $X_o^{n \times m}$  with zero mean and unit variance.
- Compute kernel matrix  $K$  using kernel function in equation (16).
- Scale the kernel matrix  $K$  using the following equation:  $K_n = K - 1_n K - K 1_n + 1_n K 1_n$
- Solve the eigenvalues problem defined in equation (17).
- Calculate different statistics used for fault detection and their limits.

**Testing Part:**

- Acquire testing samples.
- Scale these samples using the mean and standard deviation of the original data set.
- Compute testing kernel matrix using equation (16).
- Scale the obtained kernel matrix using the following equation:  $K_m = K_t - 1_t K - K_t 1_n + 1_t K 1_n$
- Compute different statistics, and report a fault if one of them exceeds its limit.

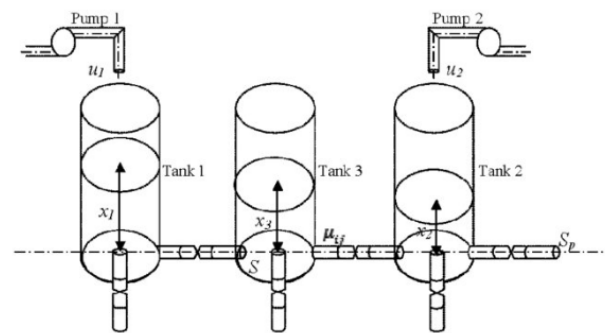
**Algorithm 3** Proposed RKPCA Algorithm

- Perform Algorithm 1.
- Select  $1^{st}$  Principal score vector
- Plot histogram of this vector (with specified number of bins  $N_B$ ).
- Pick the minimum appearance frequency  $\varepsilon.i \leq N_B$
- Compute median of the  $i^{th}$  bin.
- Compute  $\gamma_i$  as given in equation (26).
- Select the number of observations equal to  $\gamma_i$  with the same median.
- Select the corresponding rows in the scores matrix  $T$ .
- Perform the inverse mapping of PCA.
- Perform the inverse scaling to obtain  $X_r$ .
- Use  $X_r$  to build KPCA model using algorithm 2.
- Compute the value of the loss function  $J_O$  in equation (29) to evaluate the performance of the model.

**VI. PROCESS DESCRIPTION****A. THREE TANKS SYSTEM DTS-200**

This system corresponds to the plant made by Amira GmbH that includes three serially connected tanks (cylinders) and two pumps to fill two of the three tanks [25], [26]. Five sensors are used for this application, they are placed as

the following: (i) Level sensor at each of the three tanks.  $x_1$ ,  $x_2$ , and  $x_3$  are the measured variables from these sensors. (ii) Flow meters for each pump.  $u_1$  and  $u_2$  are the measured values. Figure 1 shows the diagram of this system. The data acquisition process is explained in [26]. Data sets of this system are organized as (i) Healthy training data set with a matrix size of  $9000 \times 5$ . (ii) Faulty data set with fault of 10% in the first tank level sensor with 2400 observations. (iii) Faulty data set with fault of 10% in the second tank level sensor with 2400 observations. For further explanation of the plant please refer to [26] and [27].

**FIGURE 1.** Three tanks system DTS-200.**B. AIN EL KEBIRA CEMENT PLANT**

Cement production is a complex process that starts by mining and then grinding raw materials including limestone and clay to a fine powder, called raw meal, which is then heated to a sintering temperature as high as  $1450^\circ C$  in a cement kiln to broke the chemical bonds of the raw materials and then they are recombined to form new compounds. The result is called clinker, which is grounded to a fine powder in a cement mill and mixed with gypsum to create cement.

The Ain El Kebira cement plant is located near Setif in eastern Algeria. it has a rotary kiln of  $5.4 m$  shell diameter and  $80 m$  length with  $3^\circ$  incline. The kiln is spun up to  $2.14 rpm$  using two  $560 kws$  asynchronous motors and the producing clinker of density varying from  $1300 kg.m^{-3}$  to  $1450 kg.m^{-3}$  under normal conditions. Two natural gas burners are used, the main one in the discharge end and the other one in the first level of the pre-heater tower without any tertiary air conduct.

Table 2 in paper [4] illustrates different variables of the process and a schematic diagram of this plant is presented in [28]. Data sets used for this work are as follows: (i) training data set with 768 observations. (ii) Testing data set with 11000 observations. (iii) Real process fault with 2048 observations. It includes the normal operating conditions part that last less than 420 samples, and the rest is the faulty part which constitutes multiple faults in various variables that increase gradually [30]. (iv) 10 simulated sensor faults data sets with each one lasting for 1000 observations these faults are described in [19].

**VII. RESULTS AND DISCUSSION**

For the two study cases, the computer used has Intel(R) Core(TM) i3-3217U CPU @ 1.80 GHz and 4 Go of RAM. Besides, the significant level  $\alpha$  is selected to be 1%.

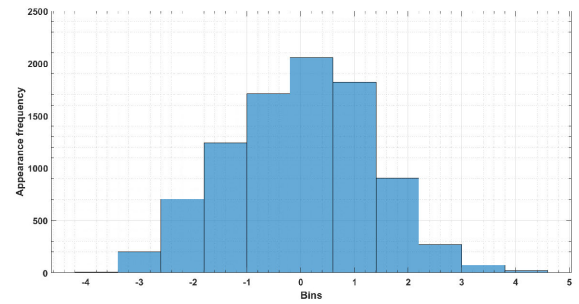
First, algorithm 3 is applied to the three tank system described in section VI. Table 1 exhibits some of the reduced matrices that are obtained using the proposed reduction method and their performance based on different FDD statistics and indices described in section IV. It can be noticed how the number of bins,  $N_B$ , affects the least appearance frequency,  $\epsilon$ , which determines the size of the reduced data matrix, the higher it gets the smaller the reduced data matrix is. The first reduced data matrix is the same as the original data set because  $\epsilon = 1$  and its performance is the same as the conventional KPCA. In this table, the comparison between different reduced data matrices is based mainly on the cost function described by the equation (29). Some of the reduced matrices have also better performance than conventional KPCA in terms of the different indices alone described by equation (27) as shown in columns (4, 5, and 6) but for this study the data matrix  $X_r$  with a number of observations of 1007 is chosen due to its overall performance based on cost function  $J_O$ . Figure 2 and Figure 3 show the histogram obtained from the first PC score of both the original data set and the selected reduced data matrix, respectively. As can be seen from both figures, both histograms have the same shape and skewness for the same bins. Furthermore, Table 2 is an illustration of how the figures 2 and 3 are alike by exhibiting the appearance frequencies and relative frequencies of both histograms. The relative frequencies from both histograms are close to each other which explains why both histograms have the same distribution. Hence, these results show that the proposed reduction method has successfully maintained the same distribution in the direction of the highest variance of the data.

Table 3 illustrates the comparison between conventional PCA, conventional KPCA, and the proposed RKPCA algorithm with the chosen reduced data matrix. This comparison is held upon the cost function described by equation (28). As anticipated, the PCA algorithm failed to perform as well as the KPCA and RKPCA algorithms due to nonlinearities. The proposed algorithm managed to perform as the conventional KPCA because it has only kept useful observations which are 11.19% of the total number of observations. Moreover, the proposed algorithm has kept homogeneity for all reduced matrices presented in Table 1 based on the test presented in [29]. These results manifest the potential of the proposed algorithm which leads to applying it to cement plant data and then comparing it to some of the existing RKPCA algorithms.

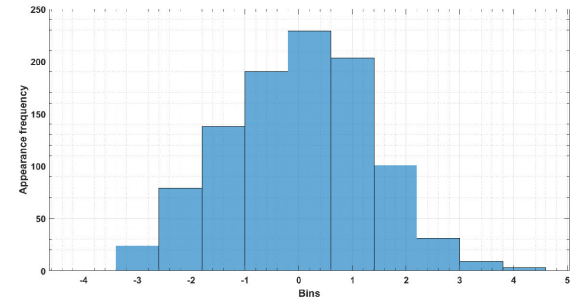
Table 4 shows the gain obtained using the proposed algorithm instead of the conventional KPCA in terms of the execution time and occupied storage space, taking into account that the KPCA algorithm and the RKPCA algorithms usually follow  $\mathcal{O}(n^3)$  time complexity and  $\mathcal{O}(n^2)$  for the storage space complexity and the difference in execution

**TABLE 1. Data set size reduction using the proposed algorithm for the three tank system.**

$N_B$	$\epsilon$	$X_r$ size	$J_{T2}$	$J_Q$	$J_\phi$	$J_O$
50	1	$9000 \times 5$	0.20	0.31	<b>0.21</b>	2.24
44	3	$3015 \times 5$	0.20	0.26	0.26	0.46
43	2	$4512 \times 5$	0.20	<b>0.24</b>	<b>0.21</b>	0.72
41	3	$3012 \times 5$	0.20	0.31	<b>0.21</b>	0.46
18	5	$1808 \times 5$	0.20	0.25	0.25	0.31
13	5	$1804 \times 5$	0.20	0.27	0.26	0.32
11	9	$1007 \times 5$	0.20	0.26	0.27	<b>0.27</b>
06	53	$172 \times 5$	<b>0.19</b>	0.30	0.31	<b>0.27</b>
05	71	$129 \times 5$	0.21	0.32	0.32	0.28



**FIGURE 2. Appearance frequency of the 1<sup>st</sup> PC score of the three tank system original data.**



**FIGURE 3. Appearance frequency of the 1<sup>st</sup> PC score of the three tank system reduced data.**

**TABLE 2. Relative frequency distribution of the 1<sup>st</sup> score data sets of the three tank system.**

Class Interval	Class Median Point	Frequency for Original Data	Frequency for Reduced Data	Relative Frequency for Original Data	Relative Frequency for Reduced Data
[-4.6, -3.4]	-3.8	9	0	0.001	0.000
[-3.4, -2.6]	-3.0	201	24	0.022	0.024
[-2.6, -1.8]	-2.2	704	79	0.078	0.079
[-1.8, -1.0]	-1.4	1240	138	0.138	0.014
[-1.0, -0.2]	-0.6	1780	190	0.190	0.189
[-0.2, +0.6]	0.2	2054	229	0.228	0.227
[+0.6, +1.4]	1.0	1819	203	0.202	0.202
[+1.4, +2.2]	1.8	902	101	0.100	0.100
[+2.2, +3.0]	2.6	269	31	0.030	0.031
[+3.0, +3.8]	3.4	73	9	0.008	0.009
[+3.8, +4.6]	4.2	21	3	0.002	0.003

time and required storage space between RKPCA and KPCA algorithms is a result of reducing the number of observations from the large-sized training data set. The time gained

TABLE 3. PCA, KPCA, and RKPCA performance comparison.

Method	Data matrix size	$J_{T^2}$	$J_Q$	$J_\varphi$	$J$
PCA	9000 × 5	0.91	0.48	0.52	0.64
KPCA	9000 × 5	<b>0.20</b>	0.31	<b>0.21</b>	<b>0.24</b>
Proposed algorithm	1007 × 5	<b>0.20</b>	<b>0.26</b>	0.27	<b>0.24</b>

TABLE 4. Gained execution time (%) for one sample monitoring via different fault indices and gained storage space (%) for the three tank system.

Method	$T^2$	$Q$	$\varphi$	Gained storage space
Proposed Algorithm	98.51	98.63	98.52	98.84

in this table is the required time to evaluate one sample in the online part of the KPCA algorithm. The gained time is then given as  $(1 - (E_{t_r}/E_{t_n})) \times 100(\%)$ . The same thing goes for the gained storage space, it is given as  $(1 - (S_{o_r}/S_{o_n})) \times 100(\%)$ . The gained storage space for the proposed algorithm is 98.84%.

Table 5 illustrates some of the results that are obtained by performing algorithm 3 on cement plant data set using different faults and real process fault mentioned in section VI, all the different statistics and metrics introduced in section IV are used for FD. As it is seen in Table 5, again the number of bins  $N_B$  affects the least appearance frequency  $\varepsilon$  which determines the reduced data matrix. The cost functions used in this table are defined by equations (27, 28, and 29) but the one used to select the appropriate reduced data matrix is described in equation (29) because it is based on the overall monitoring performance and takes into account the gained time and storage space complexities. For the  $T^2$  index the reduced data matrix with 199 samples has the best performance, and for the  $Q$  index the reduced data matrix with 263 samples outperforms other reduced data matrices, whereas, for the combined index  $\varphi$  two reduced data matrices with 261 and 157 samples have the best performance. For the overall monitoring performance only based on  $J$ , data matrices with the size 389, 261, and 157 samples have the best performance. From these data matrices, the selected one is the one that leads to the best performance based on  $J_O$  which is the data matrix with only 157 samples or one-fifth of the total number of observations. The reduction part has successfully maintained only 20.44% of the total number of observations.

After the selection of the data matrix  $X_r$ , Table 6 and figures 4 and 5 demonstrate the distribution of the original data and reduced data as well to compare and see how close they are. From figures 4 and 5, it can be noticed that there is a slight difference between the appearance frequency of the original data set and the one from the reduced data set, this is due to the redundant observations from the data set because this data was collected from real plant and one can notice that both figures have the same skewness and nearly the same distribution. on the other hand, Table 6 helps to

TABLE 5. Reduced data matrix selection of the cement plant.

$N_B$	$\varepsilon$	$X_r$ size	$J_{T^2}$	$J_Q$	$J_\varphi$	$J$	$J_O$
20	2	389 × 44	0.36	0.31	0.24	<b>0.30</b>	0.81
19	3	263 × 44	0.56	<b>0.23</b>	0.29	0.36	0.60
18	3	260 × 44	0.35	0.32	0.35	0.34	0.58
16	4	199 × 44	<b>0.34</b>	0.26	0.50	0.36	0.49
15	4	198 × 44	0.43	0.31	0.28	0.34	0.47
14	3	260 × 44	0.55	0.30	0.30	0.38	0.62
12	3	261 × 44	0.38	0.29	<b>0.23</b>	<b>0.30</b>	0.54
11	5	157 × 44	0.40	0.25	<b>0.23</b>	<b>0.30</b>	<b>0.41</b>
10	6	132 × 44	0.63	0.30	0.30	0.44	0.52
09	6	113 × 44	0.60	0.36	0.38	0.45	0.51
08	6	99 × 44	0.68	0.27	0.43	0.46	0.52

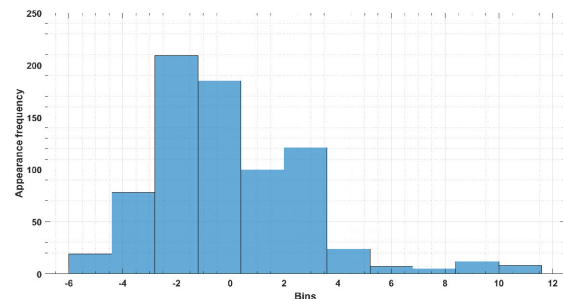


FIGURE 4. Appearance frequency of the 1<sup>st</sup> PC score of the cement plant original data.

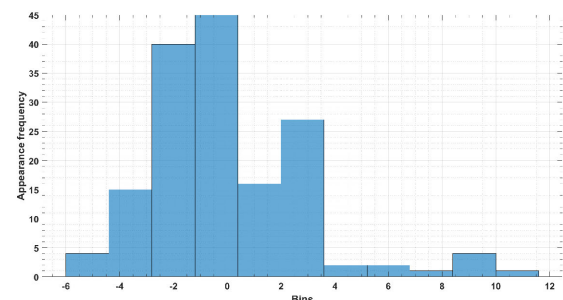


FIGURE 5. Appearance frequency of the 1<sup>st</sup> PC score of the cement plant reduced data.

TABLE 6. Frequency distribution of the 1<sup>st</sup> score of the cement plant system.

Class Interval	Class Median Point	Frequency of Original Data	Frequency of Reduced Data	Relative Frequency of Original Data	Relative Frequency of Reduced Data
[-6.00, -4.40]	-5.20	19	4	$2.47 \times 10^{-2}$	$2.55 \times 10^{-2}$
[-4.40, -2.80]	-3.60	78	15	$1.02 \times 10^{-1}$	$9.55 \times 10^{-2}$
[-2.80, -1.20]	-2.00	209	40	$2.72 \times 10^{-1}$	$2.55 \times 10^{-1}$
[-1.20, +0.40]	-0.40	185	45	$2.41 \times 10^{-1}$	$2.87 \times 10^{-1}$
[+0.40, +2.00]	1.20	100	16	$1.30 \times 10^{-1}$	$1.02 \times 10^{-1}$
[+2.00, +3.60]	2.80	121	27	$1.58 \times 10^{-1}$	$1.72 \times 10^{-1}$
[+3.60, +5.20]	4.40	24	2	$3.13 \times 10^{-2}$	$1.27 \times 10^{-2}$
[+5.20, +6.80]	6.00	7	2	$9.11 \times 10^{-3}$	$1.27 \times 10^{-2}$
[+6.80, +8.40]	7.60	5	1	$6.51 \times 10^{-3}$	$6.37 \times 10^{-3}$
[+8.40, +10.0]	9.02	12	4	$1.56 \times 10^{-2}$	$2.55 \times 10^{-2}$
[+10.0, +11.6]	10.8	8	1	$1.04 \times 10^{-2}$	$6.37 \times 10^{-3}$

understand the difference between appearance frequency and relative frequency of the two data sets and where they are alike and where they are not.



Then a comparison is done between the proposed algorithm and different algorithms. The comparison is based on loss function value (29) and homogeneity test introduced in [28], this test is chosen because it gives a non-parametric estimator of a given divergence for the case of a continuous distribution, the chosen function for divergence estimation is the asymmetric Kullback-Leibler. The same faults shown in section VI are used for the comparison between different algorithms and to evaluate the overall performance.

Based on results shown in Table 7, the method presented in [20] has a relatively very small number of retained observations compared to the original data set, meaning it could be challenging to detect true differences in homogeneity if they exist. So, the homogeneity statistical test is ignored. The proposed algorithm has significantly reduced the number of retained observations compared to some algorithms presented in the same table. It has reduced the number of samples to approximately one-fifth of the total observations. For Hotelling  $T^2$ -index, the proposed algorithm has decent monitoring performances compared to the other algorithms. Furthermore, the ones that do better than the proposed algorithm are the conventional KPCA and the variogram-based RKPCA which have more than the triple of samples in the training data set. For the Squared Prediction Error  $Q$ -index, the proposed algorithm has the second-best monitoring performance after the k-means-based RKPCA which requires a pre-defined number of clusters to have such performance. Whereas, the proposed algorithm has the best monitoring performance through the combined index  $\varphi$  along with the variogram-based RKPCA which again has a larger number of samples. According to the overall monitoring performance given by  $J$ , the proposed algorithm has the second-best monitoring performance after the variogram-based RKPCA. Furthermore, it can be clearly seen that the proposed algorithm has the best value  $J_0$  which means that the proposed algorithm has balanced between the monitoring performance and retained samples. To visualize how the proposed algorithm performs in the case of a real process fault explained in Section VI, Figure 6 is introduced. It can be noticed that overall, the proposed algorithm has successfully detected the fault through the different monitoring indices despite a slightly high FAR value brought by the index  $Q$  compared to other indices. In the case of homogeneity between original and reduced data sets, the proposed algorithm has kept more homogeneous variables with the original data set than other algorithms such that only one variable is not homogeneous whereas other algorithms have at least two non-homogeneous variables which give the credibility of this reduced data to replace the original data set and to represent the same system without losing its relevant features. This homogeneity leads the proposed algorithm to perform much well in comparison with other RKPCA algorithms.

From Table 7 and 8, it is noticed that the proposed algorithm has successfully balanced the overall performance, the gained computation time with respect to other algorithms,

TABLE 7. Performance metrics comparison of different algorithms.

Method	Retained Samples	$J_{T^2}$	$J_Q$	$J_\varphi$	$J$	$J_0$	homogeneity
KPCA	768	0.34	0.28	0.26	0.30	2.30	0
Reduced Rank [17]	613	0.56	0.30	0.27	0.38	1.64	$x_{15}, x_{44}$
K-means [15]	131	0.63	0.24	0.24	0.37	0.44	$x_1, x_{10}, x_{15}, x_{44}$
Euclidean Distance [4]	131	0.66	0.29	0.28	0.41	0.48	$x_{15}, x_{44}$
Variogram [19]	548	0.34	0.29	0.23	0.29	1.30	$x_{15}, x_{44}$
Correlation Dimension [20]	11	1.67	0.60	0.60	0.96	0.97	//
Proposed Algorithm	157	0.40	0.25	0.23	0.30	0.41	$x_{44}$

TABLE 8. Comparison of gained execution time (%) for one sample monitoring using different fault indices and gained storage space (%) via cement plant data of different algorithms.

Method	$T^2$	$Q$	$\varphi$	Gained storage space
Reduced Rank [17]	36.59	33.33	36.17	35.90
Variogram [19]	46.34	46.67	51.06	48.65
K-means [15]	96.98	97.62	97.06	96.61
Euclidean Distance [4]	96.98	97.62	97.06	96.61
Correlation Dimension [20]	99.78	99.17	99.18	99.94
Proposed Algorithm	96.93	97.42	96.74	95.40

the gained storage space of 95.40%, and homogeneity to the original data set.

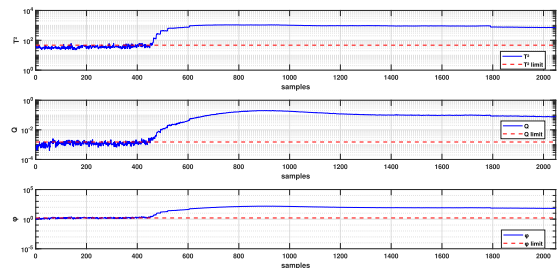


FIGURE 6. Different fault detection indices before and after the occurrence of a real process fault.

### VIII. CONCLUSION

This paper introduces a new RKPCA algorithm for fault detection. As the conventional KPCA, the proposed algorithm has conserved the ability to work with non-linear data sets while reducing the computation time and storage space needed to execute this algorithm. Not only it does enhance the execution time and occupied storage space, but it maintains the original data set's characteristics by preserving the homogeneity of this data set. The basic idea of the proposed algorithm is to select a number of representative observations in each class interval of the first principal score histogram from the original data set; in other words, it selects those observations based on the distribution of the original data set. By using this, a reliable KPCA model is attained. This model is then compared to the conventional PCA and KPCA algorithms through the three tank system data set. After that, the proposed algorithm is applied to the data set collected from the Ain El Kebira cement plant rotary kiln and compared to other existing algorithms. The overall performance of the proposed algorithm in terms of monitoring, computation time, and storage space was outstanding compared to the other techniques. Furthermore,

it has successfully maintained homogeneity with the original data set.

Future research can be extended to many pertinent similarity techniques that could be developed and adapted to the size reduction purposes whose main objective is to obtain better performance in terms of monitoring metrics and computational costs. Therefore, this topic remains more attractive to the big data research community

## REFERENCES

- [1] L. Iannucci, "Chemometrics for data interpretation: Application of principal components analysis (PCA) to multivariate spectroscopic measurements," *IEEE Instrum. Meas. Mag.*, vol. 24, no. 4, pp. 42–48, Jun. 2021.
- [2] X. Zhang and X. Ren, "Two dimensional principal component analysis based independent component analysis for face recognition," in *Proc. Int. Conf. Multimedia Technol.*, Jul. 2011, pp. 934–936.
- [3] I. Dagher and R. Nachar, "Face recognition using IPCA-ICA algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 996–1000, Jun. 2006.
- [4] F. Bencheikh, M. F. Harkat, A. Kouadri, and A. Bensmail, "New reduced kernel PCA for fault detection and diagnosis in cement rotary kiln," *Chemometric Intell. Lab. Syst.*, vol. 204, Sep. 2020, Art. no. 104091.
- [5] X. Deng, N. Zhong, and L. Wang, "Nonlinear multimode industrial process fault detection using modified kernel principal component analysis," *IEEE Access*, vol. 5, pp. 23121–23132, 2017.
- [6] D. Dong and T. J. Mcavoy, "Nonlinear principal component analysis based on principal curves and neural networks," *Comput. Chem. Eng.*, vol. 20, no. 1, pp. 65–78, Jan. 1996.
- [7] F. Jia, E. B. Martin, and A. J. Morris, "Non-linear principal components analysis for process fault detection," *Comput. Chem. Eng.*, vol. 22, pp. S851–S854, Mar. 1998.
- [8] Z. Geng and Q. Zhu, "Multiscale nonlinear principal component analysis (NLPCA) and its application for chemical process monitoring," *Ind. Eng. Chem. Res.*, vol. 44, no. 10, pp. 3585–3593, May 2005.
- [9] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.
- [10] T.-J. Chin and D. Suter, "Incremental kernel principal component analysis," *IEEE Trans. Image Process.*, vol. 16, no. 6, pp. 1662–1674, Jun. 2007.
- [11] K. Dhibi, R. Fezai, K. Bouzrara, M. Mansouri, A. Kouadri, and M.-F. Harkat, "Machine learning based multiscale reduced kernel PCA for nonlinear process monitoring," in *Proc. 17th Int. Multi-Conf. Syst. Signals Devices (SSD)*, Jul. 2020, pp. 302–307.
- [12] J. C. Benneyan, "Statistical process control as a tool for research and healthcare improvement," *Qual. Saf. Health Care*, vol. 12, no. 6, pp. 458–464, Dec. 2003.
- [13] J. Wang, C. Zhang, X. Ma, Z. Wang, Y. Xu, and R. Cattley, "A multivariate statistics-based approach for detecting diesel engine faults with weak signatures," *Energies*, vol. 13, no. 4, p. 873, Feb. 2020.
- [14] M. H. Trauth, *MATLAB@Recipes for Earth Sciences*. Berlin, Germany: Springer, 2006.
- [15] A. Perez and C. Felipe, "Fault diagnosis with reconstruction-based contributions for statistical process monitoring," *Diss. Abstr. Int.*, vol. 73, no. 1, 2011.
- [16] M.-F. Harkat, A. Kouadri, R. Fezai, M. Mansouri, H. Nounou, and M. Nounou, "Machine learning-based reduced kernel PCA model for nonlinear chemical process monitoring," *J. Control, Autom. Electr. Syst.*, vol. 31, no. 5, pp. 1196–1209, Oct. 2020.
- [17] H. Lahdhiri, I. Elaissi, O. Taouali, M. F. Harkat, and H. Messaoud, "Nonlinear process monitoring based on new reduced rank-KPCA method," *Stochastic Environ. Res. Risk Assessment*, vol. 32, no. 6, pp. 1833–1848, Jun. 2018.
- [18] O. Taouali, I. Jaffel, H. Lahdhiri, M. F. Harkat, and H. Messaoud, "New fault detection method based on reduced kernel principal component analysis (RKPCA)," *Int. J. Adv. Manuf. Technol.*, vol. 85, nos. 5–8, pp. 1547–1552, Jul. 2016.
- [19] M. T. H. Kaib, A. Kouadri, M. F. Harkat, and A. Bensmail, "RKPCA-based approach for fault detection in large scale systems using variogram method," *Chemometric Intell. Lab. Syst.*, vol. 225, Jun. 2022, Art. no. 104558. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0169743922000697>
- [20] M. Tahar, H. Kaib, A. Kouadri, M. F. Harkat, A. Bensmail, and M. Mansouri, "Improving kernel PCA-based algorithm for fault detection in nonlinear industrial process through fractal dimension," *Process. Saf. Environ. Protection*, vol. 179, pp. 525–536, Nov. 2023.
- [21] S. A. Alsenan, I. M. Al-Turaiki, and A. M. Hafez, "Feature extraction methods in quantitative structure–activity relationship modeling: A comparative study," *IEEE Access*, vol. 8, pp. 78737–78752, 2020.
- [22] C. F. Alcalá and S. J. Qin, "Reconstruction-based contribution for process monitoring with kernel principal component analysis," *Ind. Eng. Chem. Res.*, vol. 49, no. 17, pp. 7849–7857, Sep. 2010.
- [23] K. E. Pilario, M. Shafiee, Y. Cao, L. Lao, and S.-H. Yang, "A review of kernel methods for feature extraction in nonlinear process monitoring," *Processes*, vol. 8, no. 1, p. 24, Dec. 2019.
- [24] J. Chen. (2021). *Histogram Investopedia*. [Online]. Available: <https://www.investopedia.com/terms/h/histogram.asp>
- [25] A. Kouadri, K. Baiche, and M. Zelmat, "Blind source separation filters-based-fault detection and isolation in a three tank system," *J. Appl. Statist.*, vol. 41, no. 8, pp. 1799–1813, Aug. 2014.
- [26] A. Kouadri, M. A. Aitouche, and M. Zelmat, "Variogram-based fault diagnosis in an interconnected tank system," *ISA Trans.*, vol. 51, no. 3, pp. 471–476, May 2012.
- [27] A. Kouadri, A. Namoun, M. Zelmat, and M. A. Aitouche, "A statistical-based approach for fault detection in a three tank system," *Int. J. Syst. Sci.*, vol. 44, no. 10, pp. 1783–1792, Oct. 2013.
- [28] R. Kaced, A. Kouadri, K. Baiche, and A. Bensmail, "Multivariate nuisance alarm management in chemical processes," *J. Loss Prevention Process. Industries*, vol. 72, Sep. 2021, Art. no. 104548.
- [29] M. Wornowizki and R. Fried, "Two-sample homogeneity tests based on divergence measures," *Comput. Statist.*, vol. 31, no. 1, pp. 291–313, Mar. 2016.
- [30] A. Kouadri, A. Bensmail, A. Kheldoun, and L. Refoufi, "An adaptive threshold estimation scheme for abrupt changes detection algorithm in a cement rotary kiln," *J. Comput. Appl. Math.*, vol. 259, pp. 835–842, Mar. 2014.



**MOHAMMED TAHAR HABIB KAIB** received the B.S. degree in electrical and electronic engineering and the M.S. degree in control engineering from the University of Boumerdes, Boumerdes, Algeria, in 2019, where he is currently pursuing the Ph.D. degree. He is the author of one article. His research interests include multivariate statistical process monitoring for fault detection and diagnosis and big data reduction.



**ABDELMALEK KOUADRI** is currently a Professor in electrical engineering with the Institute of Electrical and Electronics Engineering, University M'Hamed Bougara of Boumerdes, Algeria. He has more than 15 years of combined academic and industrial experience. He has published more than 80 refereed journals and conference publications and book chapters. His research interests include systems engineering and control, with an emphasis on process modeling, monitoring, and estimation.

He served as a technical committee member for several international journals and conferences.



**MOHAMED-FAOUZI HARKAT** received the Eng. degree in automatic control from Annaba University, Algeria, in 1996, the Ph.D. degree from Institut National Polytechnique de Lorraine (INPL), France, in 2003. From 2002 to 2004, he was an Assistant Professor with the School of Engineering Sciences and Technologies of Nancy (ESSTIN), France. In 2004, he joined the Electronics Department, Badji Mokhtar—Annaba University, Annaba, Algeria, where he is currently a Professor. He has over 20 years of research and practical experience in systems engineering and process monitoring. He is the author of more than 100 refereed journals and conference publications and book chapters. He served as an associate editor and in technical committees for several international journals and conferences.



**ABDERAZAK BENSMAIL** received the Graduate Diploma degree in automation engineering and the Magister degree in infotonics from the National Institute of Hydrocarbon and Chemistry at Boumerdes, Algeria, in 1996 and 2004, respectively. He is currently pursuing the Ph.D. degree with the National Institute of Electrical and Electronics Engineering, University of Boumerdes, Algeria. Since, he has been working for many companies in plants construction, commissioning, and start up. He is also a Research Member of the Signals and Systems Laboratory, while he is also a Control Systems Engineer in industry. His main research interests include development and adaptation of applicable statistical techniques to fault detection, monitoring, and control of industrial manufacturing processes.



**MAJDI MANSOURI** (Senior Member, IEEE) received the degree in electrical engineering from the Higher School of Communications of Tunis (SUPCOM), Tunis, Tunisia, in 2006, the M.Sc. degree in electrical engineering from the National School of Electronics, Computer Science, Telecommunications, Mathematics and Mechanics of Bordeaux (ENSEIRB), Bordeaux, France, in 2008, the Ph.D. degree in electrical engineering from the University of Technology of Troyes (UTT), Troyes, France, in 2011, and the H.D.R. (accreditation to supervise research) degree in electrical engineering from the University of Orleans, France, in 2019. He joined the Electrical Engineering Program, Texas A&M University at Qatar, in 2011, where he is currently an Associate Research Scientist. He is the author of more than 200 publications. He is also the author of the book *Data-Driven and Model-Based Methods for Fault Detection and Diagnosis* (Elsevier, 2020). His research interests include development of model-based, data-driven, and machine learning techniques for fault detection and diagnosis.

• • •