Interval-Valued Features Based Machine Learning Technique for Fault Detection and Diagnosis of Uncertain HVAC Systems

SONDES GHARSELLAOUI1,2, MAJDI MANSOURI1, (Member, IEEE), MOHAMED ABDALLAH TRABELSI3, (Senior Member, IEEE), MOHAMED-FAOUZI HARKAT4, (Member, IEEE), SHADY S. REFAAT1, (Senior Member, IEEE), AND HASSANI MESSAOU5

1 Electrical and Computer Engineering Program, Texas A&M University at Qatar, Doha 23874, Qatar
2 Laboratory of Automatic Signal and Image Processing, Electrical Engineering Department, National Higher Engineering School of Tunis, Tunis 1008, Tunisia
3 Electronic and Communications Engineering Department, Kuwait College of Science and Technology, Safat 13133, Kuwait
4 LASMA, Badji Mokhtar - Annaba University, Annaba 23000, Algeria
5 Laboratory of Automatic Signal and Image Processing, National Engineering School of Monastir, Monastir 5035, Tunisia

Corresponding author: Sondes Gharsellaoui (sondes.gharsellaoui@qatar.tamu.edu)

Open Access funding provided by the Qatar National Library. The publication is the result of the Qatar National Research Fund (QNRF) research grant.

ABSTRACT The operation of heating, ventilation, and air conditioning (HVAC) systems is usually disturbed by many uncertainties such as measurement errors, noise, as well as temperature. Thus, this paper proposes a new multiscale interval principal component analysis (MSIPCA)-based machine learning (ML) technique for fault detection and diagnosis (FDD) of uncertain HVAC systems. The main goal of the developed MSIPCA-ML approach is to enhance the diagnosis performance, improve the indoor environment quality, and minimize the energy consumption in uncertain building systems. The model uncertainty is addressed by considering the interval-valued data representation. The performance of the proposed FDD is investigated using sets of synthetic and emulated data extracted under different operating conditions. The presented results confirm the high-efficiency of the developed technique in monitoring uncertain HVAC systems due to the high diagnosis capabilities of the interval feature-based support vector machines and k-nearest neighbors and their ability to distinguish between the different operating modes of the HVAC system.

INDEX TERMS HVAC systems, machine learning (ML), model uncertainties, feature extraction and selection, interval-valued principal component analysis (IPCA), fault detection and diagnosis (FDD).

I. INTRODUCTION

Generally, the energy demand of the residential and tertiary sector represents half of the total energy consumption where the HVAC systems represent the most energy consuming components (66% of the building’s energy consumption). However, the operational faults in HVAC systems could significantly decrease their efficiency. Research studies have proved that an efficiency increase of 5-15% is attainable by simply repairing faults and optimizing building control systems [1].

Thus, the deployment of FDD approaches is very essential to guarantee the safe operation of HVAC systems, improve the user comfort level, enhance the energy efficiency, and reduce the operating/maintenance costs [2]. Nevertheless, very few effective FDD approaches have been proposed for HVAC systems in the literature [3], [4].

FDD approaches can be classified into two main categories: data-driven [5], [6] and model-based approaches [7], [8]. Model-based FDD approaches consist in comparing systems measurements with system variables computed from the mathematical model, which is usually calculated using some fundamental understanding of the system under normal operating conditions [7], [9]. The difference between the
measurements and the predicted model prediction (so-called residuals), can be applied as a diagnosis metric for decision making [10], [11].

On the other hand, data-driven FDD methods only make use of the available diagnosis data [12]–[14]. The data are first applied to identify the model in the training phase, which is then used to diagnose faults in the testing phase.

Data-based FDD methods can be divided into two principal phases: feature extraction and selection, and faults classification [15], [16].

PCA is the mostly employed tool for feature extraction and selection [17]. The PCA is a data-based method that has been widely used for feature extraction and selection of diverse complex systems [18], [19]. Indeed, the features extraction and selection requires the building of the system PCA model under normal operating conditions (NOC). This model is then applied as a test reference for system monitoring and its identification is based on the estimation of the structure of the process by an eigen-decomposition of the covariance matrix of the training data [20].

Conventional PCA-based feature extraction and selection techniques have been only implemented for single-valued representations. During the data mining operations, these data are obtained after several simplification stages which leads to a high imprecision. Indeed, actual data is frequently affected by various uncertainties such as imprecision linked to the data estimation approach adopted, computation, and measurement errors. These uncertainties/errors have a bad effect on the PCA model, and therefore, on the FDD accuracy [21]. By considering a representation with interval values instead of a single value, this uncertainty could be treated more accurately while representing the actual data. Unlike single-valued, the interval-valued representation naturally provides extra information leading to a better decision making. The determination of the PCA model in this case implies the use of new suitable methods to the interval data.

Several conventional PCA versions have been extended to interval-valued representations over the past two decades [16]. The first variations are the centers PCA (CPCA) [22] and the vertices PCA (VPCA) [23]. The centers approach trusts on the interval centers, while the vertices approach is concentrated on the vertices of the hyper-rectangles made through the interval-valued data. Another method, named midpoints-radii PCA (MRPCA), deals with both interval ranges and interval centers [24]. It is an improvement of the centers PCA by incorporating the radius of data.

The authors in [25] presented an alternative technique by applying least squares for MRPCA, while an analytic method of PCA was proposed in [26] for interval-valued data established on an interval-valued covariance matrix. In [27], the authors applied the symbolic covariance to expand the traditional PCA to interval-valued data case. The complete-information principal component analysis (IPCA) presented in [28] is considered as a new PCA for interval-valued representation with an improved covariance matrix calculation. More precise monitoring can be obtained by representing the uncertainties in the form of intervals [29], where the PCA for interval-valued data is consequently applied for system feature extraction and selection. Nevertheless, this requires an expansion of the monitoring routine to the IPCA model. In this paper, the IPCA approach is applied to extract the more relevant and efficient interval-valued features from the HVAC system data. Then, the final selected features are fed to the ML techniques, namely support vector machines (SVM) [30], decision tree [31], K-Nearest Neighbors (KNN) [32], and Naive Bayes (NB) [33] for faults classification purposes.

Therefore, this paper proposes a higher safety and reliability multiscale IPCA-based ML technique for FDD of uncertain HVAC systems. The uncertainties are analyzed through interval-valued representation of data-sets and a further multiscale/wavelet decomposition is applied for a better diagnosis performance. The multiscale representation is considered as an effective technique to separate the important data features from the noise through filters. The random noise is characterized by its presence at different coefficients in the signal, while the deterministic data features are captured at large coefficients. The small wavelet coefficients usually correspond to noise, while the important data features are usually represented by large wavelet coefficients (in the detail signals). Thus, from the HVAC system measurements, the characteristics are extracted in an appropriate manner via the multiscale IPCA (MSIPCA) approach where an optimum number of characteristics is selected. Finally, different classifiers are used to classify the various occurring operating modes in HVAC systems.

The rest of the paper is presented as follows: Section II presents a brief background of interval-valued data representation. The feature extraction and selection based multiscale interval PCA is presented in Section III. The obtained results showing the performance of the developed FDD methodology are described in Section IV, while Section V concludes the paper.

II. INTERVAL-VALUED DATA DESCRIPTION AND NORMALIZATION

In practical, due to the eventual measurement errors, the actual value \( x_i^* \) could be different from the measured value \( x_i \). The measurement error is represented by \( \delta x_i \), \( x_i^* = x_i - \delta x_i \). Usually, a measurement error margin (an upper bound \( \delta_i \)) is provided by the sensor manufacturer. Thus, the real value \( x_i^* \) is in the interval \( x_i^* = [x_i^-x_i^+] \), where \( x_i^- = x_i - \delta_i \) and \( x_i^+ = x_i + \delta_i \).

Thus, for the sake of accuracy, it is better to present such measurements by an interval value instead of a single value. Since the closeness error is unbeknownst, it is assumed that its variation is restricted and can be defined by an interval \( [x^-, x^+] \) where \( x^- \) and \( x^+ \) designate the lower bound and upper bounds of \( x \), respectively.

A. INTERVAL VALUED-DATA DESCRIPTION

First, the properties of the interval valued variables are illustrated. An interval valued variable \( X_i \in \mathcal{R} \) is
defined by a sequence of sets of values delimited by ordered bounds couples called minimum and maximum: \( [x_j(1), x_j(2), \ldots, x_j(N)] \), where \( x_j(k) = [x_j^-(k), x_j^+(k)] \) \( \forall k \in [1, \ldots, N] \) and \( x_j^-(k) \leq x_j^+(k) \). The generic interval \([x_j(k)]\) can further be given by the couple \( \{x_j^-(), x_j^+()\} \) (biunivocal relationship) where:

\[
x_j^-(k) = \frac{1}{2}(x_j^+(k) + x_j^-(k))
\]

and

\[
x_j^+(k) = \frac{1}{2}(x_j^+(k) - x_j^-(k))
\]

For any interval valued variable

\[ [X_j] = \left( \begin{array}{cccc}
  x_j^-(1) & x_j^+(1) & \cdots & x_j^-(N) & x_j^+(N) \\
\end{array} \right)^T \]

The mean value is given by:

\[
E([X_j]) = \frac{1}{N} \sum_{k=1}^{N} E([x_j(k)])
\]

where \( E([x_j(k)]) = \frac{1}{2} \left( x_j^- + x_j^+ \right) \).

Accordingly, the centralized of \([x_j(k)]\) is \([x_j^-(k), x_j^+(k)]\) given by

\[
[x_j(k)] - E([X_j]) = \left[ x_j^-(k) - E([X_j]) , x_j^+(k) - E([X_j]) \right]
\]

Giving any interval-valued variables

\[ [X_q] = ([x_q^-(1), x_q^+(1)], [x_q^-(2), x_q^+(2)], \ldots, [x_q^-(N), x_q^+(N)])^T \]

and

\[ [X_q] = ([x_q^-(1), x_q^+(1)], [x_q^-(2), x_q^+(2)], \ldots, [x_q^-(N), x_q^+(N)])^T \]

The inner product is defined as:

\[
([X_j], [X_q]) = \sum_{k=1}^{N} ([x_j(k)], [x_q(k)]) = \frac{1}{4} \sum_{k=1}^{N} (x_j^-(k) + x_j^+(k))(x_q^-(k) + x_q^+(k))
\]

The squared norm for any interval-valued variable \([X_j]\) is defined by:

\[
([X_j], [X_j]) = \|[X_j]\|^2 = \sum_{k=1}^{N} \|[x_j(k)]\|^2 = \frac{1}{4} \sum_{k=1}^{N} (x_j^2(k) + x_j^-(k)x_j^+(k) + x_j^+(k))
\]

For all interval valued variables \([X_1], [X_2], \ldots, [X_m]\) of \( N \) observations and \( \forall a_j \in \mathbb{R}, j = 1, \ldots, m \) present an interval-valued variable \([Y(k)]\) as a linear function or combination of \([X_1(k)], \ldots, [X_m(k)], k = 1, \ldots, N, \) i.e.,

\[
[Y(k)] = \sum_{j=1}^{m} a_j [X_j(k)] = \left( [y^-(k) \cdots y^+(k)] \right)^T
\]

In order to overcome the problem of having the predicted lower bound values \( y^-(k) \) of response variable greater than the upper bound values \( y^+(k) \), the Moore’s linear combination rule used in interval arithmetic is adopted. Let \( a_j \in \mathbb{R} \) be a real scalar, then the interval-valued variable \([x]\) times \( a_j \) is given by [34]:

\[
a_j \left[ x^- + x^+ \right] = \begin{cases} 
  a_j x^- + a_j x^+ & \text{if } a_j > 0 \\
  a_j x^+ + a_j x^- & \text{if } a_j < 0
\end{cases}
\]

Then the lower bound values \( y^-(k) \) and the upper bound values \( y^+(k) \) can be expressed by:

\[
y^-(k) = \sum_{j=1}^{m} a_j \left( \tau x^- + (1-\tau) x^+ \right) \]

\[
y^+(k) = \sum_{j=1}^{m} a_j \left( (1- \tau) x^+ + \tau x^- \right)
\]

with

\[
\tau = \begin{cases} 
  0 & \text{if } a_j \leq 0 \\
  1 & \text{otherwise}
\end{cases}
\]

B. INTERVAL-VALUED DATA NORMALIZATION

Generally, certain standardization solutions must be carried out before the data processing in order to get scale-invariant results. Four alternative standardization methods for interval data were developed in [35], and are illustrated below.

1) STANDARDIZATION USING THE DISPERSION OF INTERVAL CENTER AND RANGE

The interval-valued variables are standardized according to the procedure developed in [36]. The results are achieved with reference to certain arithmetic concepts of basic interval. Consider two basic notions: the mean interval and distance between intervals where the mean interval \([m_j]\) is represented by:

\[
[m_j] = \frac{1}{N} \sum_{k=1}^{N} [x_j(k)]
\]

while the distance between intervals is given by:

\[
d([x_j(k)], [y_j(k)]) = |x_j^-(k) - y_j^-(k)| + |x_j^+(k) - y_j^+(k)|
\]

where \(d([x_j(k)], [y_j(k)])\) satisfies the Euclidean distance properties. The definition given by equation (15) presents
the notation of scalar variance for interval-valued data. The variance is described as the sum of the squared distances from the mean interval, therefore the variance $\sigma^2$ for interval valued data is represented by: 

$$\sigma^2 = \frac{1}{n} \sum_{k=1}^{N} \left[ \left( x_j^-(k) - m_j \right) + \left( x_j^+(k) - m_j \right) \right]$$

The definition of variance can also be written by:

$$\sigma^2 = \frac{1}{N} \sum_{k=1}^{N} \left[ \left( x_j^-(k) - m_j \right)^2 + \left( x_j^+(k) - m_j \right)^2 \right] + 2 \sum_{k=1}^{N} \left[ x_j^-(k) - m_j \right] \left[ x_j^+(k) - m_j \right]$$

(16)

where $[m_j] = \left[ \frac{1}{N} \sum_{k=1}^{N} x_j^-(k), \frac{1}{N} \sum_{k=1}^{N} x_j^+(k) \right]$, $m_j^f = \frac{1}{n} \sum_{i=1}^{n} x_j^f_i$. With a little algebra we obtain:

$$\sigma^2 = \frac{1}{N} \sum_{k=1}^{N} \left( x_j^-(k) - m_j \right)^2 + \sum_{k=1}^{N} \left( x_j^+(k) - m_j \right)^2$$

$$+ 2 \sum_{k=1}^{N} \left[ x_j^-(k) - m_j \right] \left[ x_j^+(k) - m_j \right]$$

(17)

The expression in equation (17) asserts that the variance for interval valued data could be broken down into three components: variance between midpoints, variance between ranges and twice the connection among midpoints and ranges, defined by

$$\sum_{k=1}^{N} \left[ x_j^-(k) - m_j \right] \left( x_j^+(k) - m_j \right) \geq 0.$$

The remarked properties in (15) imply that the distance among intervals can be concluded to the Euclidean distance in the space $\mathbb{R}^m$. A standardized interval is

$$\left[ \frac{1}{\sigma} \left( x_j^-(k) - m_j - x_j^+(k) - m_j \right), \frac{1}{\sigma} \left( x_j^+(k) - m_j + x_j^+(k) - m_j \right) \right].$$

(18)

3) STANDARDIZATION USING THE DISPERSION OF THE INTERVAL BOUNDARIES

The third standardization approach transforms the $N$ intervals $[x_j(k)]$ for each variable $[X_j]$, such that the mean and the joint dispersion of the re-scaled interval boundaries are 0 and 1, respectively. The joint dispersion of a variable $[X_j]$ is given by:

$$\sigma^2_j = \frac{1}{n} \sum_{k=1}^{N} \left( x_j^-(k) - m_j \right)^2 + \left( x_j^+(k) - m_j \right)^2$$

(21)

Then, for $k = 1, \ldots, N$, the intervals $[x_j(k)] = [x_j^-(k), x_j^+(k)]$ are transformed into:

$$\left[ \frac{x_j^-(k) - m_j}{\sigma_j}, \frac{x_j^+(k) - m_j}{\sigma_j} \right]$$

(22)

4) STANDARDIZATION USING THE GLOBAL RANGE

The fourth standardization approach transforms the intervals $[x_j(k)] = [x_j^-(k), x_j^+(k)]$, $(k = 1, \ldots, N)$ for a given variable so that the range of the $n$ rescaled intervals is the unit interval $[0, 1]$. Either $Min_j = \min \{x_j^-(1), \ldots, x_j^-(N)\}$ and $Max_j = \max \{x_j^+(1), \ldots, x_j^+(N)\}$ be the lower and upper boundary values. The interval is transformed into standardized interval with boundaries with this notation:

$$\frac{x_j^-(k) - Min_j}{Max_j - Min_j} \text{ and } \frac{x_j^+(k) - Min_j}{Max_j - Min_j}$$

(23)

III. FEATURE EXTRACTION AND SELECTION USING MULTISCALE INTERVAL PCA

A. MULTISCALE REPRESENTATION

The interval data given by the matrix $[X] \in \mathbb{R}^{N \times m}$, where $N$ refers to the measurements and $m$ are the variables, are first multiscaled. Then, the initial signals are projected on a set of orthonormal scaling functions [37] as follows:

$$\phi_{ij}(t) = \sqrt{2^{-j}} \phi(2^{-j} t - k)$$

(24)

Another alternative is to make use of a low pass filter of length $r, h = [h_1, h_2, \ldots, h_r]$ by projecting the original signal into a set of wavelet based functions [37] given by:

$$\psi_{ij}(t) = \sqrt{2^{-j}} \psi(2^{-j} t - k)$$

(25)

A third solution is to derive a high-pass filter from the wavelet basis functions [37] and use it to fine scale the signal. Thus, the original signal can be reproduced by summing the detail signals at all scales and the scaled signal at the coarsest scale as follows [37]:

$$[X](t) = \sum_{K=1}^{n2^{-j}} a_{jk} \phi_{jk}(t) + \sum_{j=1}^{J} \sum_{k=1}^{n2^{-j}} d_{jk} \psi_{jk}(t),$$

(26)

where $n, k, j$, and $J$ are the original signal length, translation parameter, the dilation parameter and the number of scales respectively [38]. The wavelet transformation is considered
as an effective segregation between the deterministic characteristics and haphazard noise. Thus, it represents a strong transformation of the time-domain signals into the time-frequency domain [37].

B. FEATURE EXTRACTION

The interval PCA (IPCA) approach was proposed in [28] by using the traditional PCA to process the interval valued-day and extract more information in interval measurements. Given two interval-valued variables \([X_j]\) and \([X_f]\), according to IPCA [28], the inner product is represented as:

\[
\{[X_j], [X_f]\} = \sum_{k=1}^{N} \{[x_j(k)], [x_f(k)]\}
\]

(27)

where

\[
\{[x_j(k)], [x_f(k)]\} = \frac{1}{4} \left( x_j^-(k) + x_j^+(k) \right) \left( x_f^-(k) + x_f^+(k) \right)
\]

(28)

In the auto-correlation case given by \([X_j], [X_j]\), the inner product \(\|[X_j]\|^2\) for interval-valued data is represented as follows:

\[
\|[X_j]\|^2 = \sum_{k=1}^{N} \|x_j(k)\|^2
\]

(29)

where

\[
\|x_j(k)\|^2 = \frac{1}{3} \left( x_j^-(k) + x_j^-(k)x_j^+(k) + x_j^+2(k) \right)
\]

(30)

The covariance matrix \(\Sigma\) of \(X \in \mathbb{R}^{N \times m}\) is represented by equation (31) based on the above definitions of inner product and interval norm and with all data units pre-processed.

\[
\Sigma = \frac{1}{N} \left( \begin{array}{cccc}
\{[X_1], [X_1]\} & \{[X_1], [X_2]\} & \cdots & \{[X_1], [X_m]\} \\
\{[X_2], [X_1]\} & \{[X_2], [X_2]\} & \cdots & \{[X_2], [X_m]\} \\
\vdots & \vdots & \ddots & \vdots \\
\{[X_m], [X_1]\} & \{[X_m], [X_2]\} & \cdots & \{[X_m], [X_m]\}
\end{array} \right)
\]

(31)

The interval-valued principal components \([T]\) in the IPA model are determined based on the linear combination algorithm for interval-valued variables [39]. Using equation (31), the IPA model first compute the covariance matrix \(\Sigma\) of the interval data matrix \([X]\) then carry out an eigen-decomposition, where \(\lambda_1, \ldots, \lambda_m\) and \(p_1, \ldots, p_m\) are the resulting eigenvalues and eigenvectors respectively. Based on Moore’s rule [39] the interval-valued principal components are presented by:

\[
\begin{align*}
\hat{x}_j^-(k) &= \sum_{i=1}^{m} p_{ij} \left( \tau x_i^-(k) + (1 - \tau) x_i^+(k) \right) \\
\hat{x}_j^+(k) &= \sum_{i=1}^{m} p_{ij} \left( (1 - \tau) x_i^-(k) + \tau x_i^+(k) \right)
\end{align*}
\]

(32)

with

\[
\tau = \begin{cases} 
0, & p_{ij} \leq 0, \\
1, & p_{ij} \geq 0 
\end{cases}
\]

From the IPCA model, the interval-valued estimations are given:

\[
\begin{align*}
\hat{x}_j^-(k) &= \sum_{q=1}^{m} C_{\ell qj} \left( \tau x_q^-(k) + (1 - \tau) x_q^+(k) \right) \\
\hat{x}_j^+(k) &= \sum_{q=1}^{m} C_{\ell qj} \left( (1 - \tau) x_q^-(k) + \tau x_q^+(k) \right)
\end{align*}
\]

(33)

with the same condition on \(\tau\), and given that \(C_{\ell} = P_{\ell}P_{\ell}^T\).

C. INTERVAL-VALUED PCA MODEL IDENTIFICATION

The selection of the appropriate number \(\ell\) of principal components [40] is the key step in identifying the PCA model. In this study, the reconstruction error variance is minimized based on the interval data as in [40] to determine \(\ell\) for the PCA model.

Generally, when the PCA based single-valued data is applied, the reconstruction method is applied to estimate a variable based mainly on the PCA model. The accuracy of the reconstruction depends on the capability of the PCA model to disclose iterative relations between all variables [40], [41]. In [18] the authors propose the IPCA approach using the variable reconstruction. The value of \(\ell\) that minimizes the variance of the interval-valued reconstructed error presents the number of principal components to be conserved in the IPCA model [18].

D. FEATURE SELECTION

The determination of the IPCA model is based on an eigen-decomposition of the covariance matrix \(\Sigma\) and the selection of the number \(\ell\) of components to be retained [18], [20]. The matrices of eigenvalues, eigenvectors and interval-valued principal components can be given by:

\[
\Lambda = \begin{bmatrix} \Lambda_{\ell} & 0 \\ 0 & \Lambda_{m-\ell} \end{bmatrix}
\]

(34)

\[
P = \begin{bmatrix} P_{\ell} & P_{m-\ell} \end{bmatrix}, \quad [T] = [[T_\ell] [T_{m-\ell}]]
\]

(35)

By taking into consideration the first \(\ell\) highest eigenvalues and their corresponding eigenvectors, the matrix \([X]\) is decayed as:

\[
[X] = [T_\ell] P_{\ell}^T + [E]
\]

(36)

where \([T_\ell] = [X] P_{\ell}\) and \([E]\) is the interval-valued residual matrix.

A sample vector \([x(k)] \in \mathbb{R}^m\) can be projected onto the principal and residual subspaces, respectively,

\[
\hat{x}(k) = P_{\ell} [t_\ell(k)] = C_{\ell} [x(k)]
\]

(37)

where \([\hat{x}(k)]\) is the estimation vector of \([x(k)]\), \(C_{\ell} = P_{\ell}P_{\ell}^T\) and,

\[
[t_\ell(k)] = P_{\ell}^T [x(k)] \in \mathbb{R}^\ell
\]

(38)
is the vector of the first $\ell$ interval-valued scores of latent variables. The vector of $m - \ell$ last interval-valued scores of latent variables, that represents the projection of interval-valued measurement data in the residual subspace, is defined by:

$$[t_{m-\ell}(k)] = P^T_{m-\ell} [x(k)] \in \mathbb{R}^{m-\ell}$$  \hspace{1cm} (39)

To achieve a good classification performance, it is essential to extract the statistical characteristics via the IPCA model by listing exhaustively a few possible values.

In this study, the selected and extracted features/characteristics from the IPCA model are the first retained interval-valued principal components, the interval squared weighted error (ISWE) statistic, the interval norm distance $D_N$, the City-Block distance $D_B$, the Hausdorff distance $D_H$, the Euclidean distance $D_E$ and the Wasserstein distance $D_W$. Next, the features are presented.

1) INTERVAL SQUARED WEIGHTED ERROR (ISWE) STATISTIC

The ISWE feature is the most essential measure in the remaining principal components subspace [18] and is given by:

$$ISWE(k) = \|t_{m-\ell}(k)\Lambda_{m-\ell}^{-1/2}\|^2$$ \hspace{1cm} (40)

2) INTERVAL NORM DISTANCE $D_N$

The interval distance is given as the interval norm of the difference between two interval-valued samples and is defined by:

$$D_N(k) = \sum_{j=1}^{m} a_N \left([x_j(k)], [y_j(k)]\right)$$ \hspace{1cm} (41)

where

$$a_N \left([x_j(k)], [y_j(k)]\right) = \frac{1}{3} \left( (e^-_j)^2(k) + (e^-_j)(e^+_j)(k) + (e^+_j)^2(k) \right)$$

and

$$\left[ e^-_j(k) \quad e^+_j(k) \right] = \left[ x_j(k) - y_j(k) \right]$$

$$= \left[ x_j(k) \quad x_j(k) \right] - \left[ y_j(k) \quad y_j(k) \right]$$

$$= \left[ \left( x_j(k) - y_j(k) \right) \quad \left( y_j(k) - x_j(k) \right) \right]$$

3) CITY-BLOCK DISTANCE $D_B$

$$D_B(k) = \sum_{j=1}^{m} a_B \left([x_j(k)], [y_j(k)]\right)$$ \hspace{1cm} (42)

where

$$a_B \left([x_j(k)], [y_j(k)]\right) = |x_j(k) - y_j(k)| + |\bar{x}_j(k) - \bar{y}_j(k)|$$

4) HAUSDORFF DISTANCE $D_H$

$$D_H(k) = \sum_{j=1}^{m} a_H \left([x_j(k)], [y_j(k)]\right)$$ \hspace{1cm} (43)

where

$$a_H \left([x_j(k)], [y_j(k)]\right) = \max \left\{ |x^-_j(k) - y^-_j(k)|, |x^+_j(k) - y^+_j(k)| \right\}$$

5) EUCLIDEAN DISTANCE $D_E$

$$D_E(k) = \sum_{j=1}^{m} a_E \left([x_j(k)], [y_j(k)]\right)$$ \hspace{1cm} (44)

where

$$a_E \left([x_j(k)], [y_j(k)]\right) = \left( x^-_j(k) - y^-_j(k) \right)^2 + \left( x^+_j(k) - y^+_j(k) \right)^2$$

6) WOSSEERSTEIN DISTANCE $D_W$

$$D_W(k) = \sum_{j=1}^{m} a_W \left([x_j(k)], [y_j(k)]\right)$$ \hspace{1cm} (45)

where

$$a_W \left([x_j(k)], [y_j(k)]\right) = \left( m_{x,j}(k) - m_{y,j}(k) \right)^2 + \frac{1}{3} \left( r_{x,j}(k) - r_{y,j}(k) \right)^2$$

and

$$m_{x,j}(k) = \frac{x^-_j(k) + x^+_j(k)}{2}, \quad m_{y,j}(k) = \frac{y^-_j(k) + y^+_j(k)}{2}$$

$$r_{x,j}(k) = \frac{x^-_j(k) - y^-_j(k)}{2}, \quad r_{y,j}(k) = \frac{x^+_j(k) - y^+_j(k)}{2}$$

IV. SIMULATION RESULTS

The various steps of the developed FDD technique are illustrated in Figure 1. The confusion matrix is used to compute the performance metrics of each classifier, where the classification accuracy is given the highest performance priority. Moreover, the Recall and Precision metrics are applied as per [42]:

$$Recall = \frac{TP}{TP + FN}$$ \hspace{1cm} (46)

$$Precision = \frac{TP}{TP + FP}$$ \hspace{1cm} (47)

In the above equations, the percentage metrics TP, FP, and FN refer to the number of accurately identified samples, the number of misidentified samples, and the erroneously identified samples respectively. The Recall metric is used to measure the distinct classification sensitivity (accuracy).
A. SIMULATED SYNTHETIC DATA

Two datasets were used to generate the database. The first data set is taken from a healthy operation scenario, while the second one contains the simulated data under three different faulty scenarios. The latter data are manually manipulated to emulate the behavior of each fault separately. Hence, the generated database is categorized via performing the appropriate multiscale pre-processing then exploited as a training data for the ML algorithm. In addition, the system is used to generate the faulty database via simulating the occurring faults scenarios. Then, the simulation results are labeled based on the applied type of fault. The labeled sets of data are used as inputs for the MSIPCA to distinguish between the different types of system operation. The example given in [37] is replicated using the below system and then the simulated synthetic data are generated. Two uncorrelated variables are generated using Gaussian measurements with zero mean and unit variance. The system contains combinations of adding/subtracting the first two variables with possible scaling as per: [37]:

\[
\begin{align*}
\tilde{x}_1(t) &= N(0, 1), \\
\tilde{x}_2(t) &= N(0, 1), \\
\tilde{x}_3(t) &= \tilde{x}_1(t) + \tilde{x}_2(t), \\
\tilde{x}_4(t) &= \tilde{x}_1(t) - \tilde{x}_2(t), \\
\tilde{x}_5(t) &= \tilde{x}_1(t) + 2\tilde{x}_2(t), \\
\tilde{x}_6(t) &= \tilde{x}_1(t) - 2\tilde{x}_2(t).
\end{align*}
\]

The six measured variables contained in the data matrix \( \tilde{X} \) are then disturbed by a white noise of zero mean and standard deviation of 0.2 as follows [37]:

\[
X(t) = \tilde{X}(t) + 0.2N(0, 1).
\]

Six variables are generated by using equation (48). These variables decompose on one healthy (designated to class \( C_0 \)) and three faulty scenarios of synthetic data (designated to \( C_i \), \( i = 1, ..., 3 \)) as described in Table 2. Taking into consideration the imprecision of 1\%, the interval data is generated. Different normalization techniques (Table 1) are used for data pre-processing before the IPCA model identification. In order to reduce the estimation error using the PCA model, more principal components should be kept in the model. However, the first principal components represent significant variation in the data and the last principal components represent the noise. Therefore, if more principal components are retained in the PCA model, the estimation error will be reduced in the training data but without generalization for testing data set (over-fitting). Thus, the number of principal components should be selected carefully. In this study, the value of \( \ell \) that minimizes the variance of the interval-valued reconstructed error presents the number of principal components to be conserved in the IPCA model. Via the decomposition of the eigenvalues, the variances are sorted in descending order. Then through the obtained model, the faulty data are transformed. Consequently, some characteristics are extracted and appropriately selected in order to represent simultaneously the different models in the two IPCA subspaces. Then the features are extracted and used for fault isolation using several classifiers. The selected features in this study are the first principal components combined with the ISWE and interval distances (including \( D_N, D_B, D_H, D_E \) and \( D_W \)). The training data set contains 2048 equally spaced observations (normal/faulty operation). The healthy or fault free data (samples 1-512) is assigned to class \( C_0 \). The generated fault
TABLE 3. Classification accuracy using IPCA with various normalization of interval-valued data.

| Classifiers | Global Performance Accuracy | IPCA/∥\text{T}_\ell\parallel | \text{T}_1 | \text{T}_2 | \text{T}_3 | \text{T}_4 |
|-------------|-----------------------------|-----------------------------|-------------|-------------|-------------|-------------|
| KNN Training | 68.67 | 64.64 | 46.11 | 73.73 |
| Testing     | 83.17 | 81.37 | 81.25 | 85.35 |
| NB Training | 61.08 | 65.38 | 65.4  | 64.5  |
| Testing     | 60.44 | 64.4  | 64.33 | 64.89 |
| DT Training | 74.36 | 72.72 | 72.02 | 80.56 |
| Testing     | 75.09 | 72.36 | 72.53 | 80.85 |
| SVM Training| 83.22 | 80.15 | 80.68 | 86.25 |
| Testing     | 79.24 | 77.66 | 77.34 | 82.66 |

(\text{F}_1, \text{ designated to class } C_1) \text{ represents a step-change in the mean of all four variables, } (\text{F}_2, \text{ assigned to class } C_2) \text{ consists of a variance change in the samples, and } (\text{F}_3, \text{ designated to class } C_3) \text{ represents an incipient fault. In the faulty scenario, the generated faults } (\text{F}_1), (\text{F}_2), \text{ and } (\text{F}_3) \text{ correspond to the observation samples } 513-1024, 1025-1537, \text{ and } 1538-2048 \text{ respectively. Where the testing dataset contains four operating modes from the training set and has } 2048 \text{ samples. A comparison between the performance of the MSIPCA-based ML approach and the IPCA-based ML is illustrated. In this study, several classifiers (KNN, NB, DT and SVM) are investigated. An accuracy performance comparison via the various selected features is illustrated in Tables 3 and 4.}

In order to make more improvement in the classification accuracy, in addition to the first } \ell = 4 \text{ principal components, other features are added. Thus, the norm of the first retained interval-valued principal components } \|\text{T}_\ell\parallel \text{ are combined with the different interval distances given in equations 41, 42, 43, 44 and 45, respectively. Tables 5 and 6 show the results of the use of the new features as inputs to the KNN and SVM classifiers. It is clear that both classifiers give the best results when using } \|\text{T}_\ell\parallel \text{ combined with } D_N, D_H, D_E \text{ or } D_B. \text{ The accuracy of all these MSIPCA-based combinations are between } 97.16\% \text{ and } 100\%. \text{ From the above tables, it can be concluded that the MSIPCA-based ML presents an accurate classification better than those using the IPCA-based ML approaches. For instance, the MSIPCA-based KNN technique shows } 100\% \text{ of class accuracy for the classes } 0-3. 

B. EMULATED HEATING SYSTEM

1) SYSTEM DESCRIPTION

The TRNSys simulation software (transient simulation), TRNSys simulation studio, and TRNBuild interface are employed to imitate an actual building and to beget the heating system data. Thus, TRNBuild interfaces allow adding many proprieties like window and door properties, thermal conductivity, wall and layer material properties, and various gains, etc (non-geometrical properties). Based on the existing construction parameters, the TRNSYS model is run with a time step of 1h, using the meteorological data given by the US Department of Energy (DOE). The developed FDD method is validated in simulation by modeling in TRNSYS a building (located in France in the region of Amiens) during the cold season, with three zones, where the rooms are loaded with various profiles and schedules as a simulation. The data collected during one year of casual operation are used to guide and configure the FDD system by building a PCA model as per the above method used for synthetic data. So as to generate the faulty database, two fault cases were emulated in TRNSYS. The individual faults are executed statically by changing existing objects, i.e. schedules. The considered faults are i) Unplanned occupancy: this fault is tested by adding some unexpected persons or occupants in various hours, and ii) Opening the window when the HS is switched ON causing waste of energy. As the FDD issue can be considered as a classification trouble, three data classes are used: a healthy data class and two faulty data classes. The data time range is set from zero to 8000h with a time step of 1h. The description of the heating system variables are reported in Table 7.

2) FAULT CLASSIFICATION RESULTS

To validate the developed method of FDD, five different variables are simulated as given in Table 7. These variables
TABLE 7. Variables description.

| Variable | Description                                      |
|----------|--------------------------------------------------|
| z_1     | T_{amb}: Ambient temperature (°C)                |
| z_2     | RH: Relative humidity (%)                       |
| z_3     | W_{velocity}: Wind Velocity (m/s)               |
| z_4     | P: Atmospheric Pressure (Pa)                    |
| z_5     | Q_{heat}: Sensible heating demand of zone (kJ/hr)|

TABLE 8. Emulated database construction.

| Class | State  | Training Data | Testing Data |
|-------|--------|---------------|--------------|
| C_0   | Healthy| 2000          | 2000         |
| C_1   | SC_1   | 2000          | 2000         |
| C_2   | SC_2   | 2000          | 2000         |

TABLE 9. Classification accuracy using IPCA with various normalization of interval-valued data.

| Classifiers | Global Performance | IPCA/||T_i|| Accuracy |
|------------|--------------------|-----------------|
| KNN        | Training           | 39.63           |
|            | Testing            | 69.23           |
| NB         | Training           | 36.61           |
|            | Testing            | 36.5            |
| DT         | Training           | 41.36           |
|            | Testing            | 42.55           |
| SVM        | Training           | 67.2            |
|            | Testing            | 62.5            |

TABLE 10. Classification accuracy using MSIPCA with various normalization of interval-valued data.

| Classifiers | Global Performance | MSIPCA/||T_i|| Accuracy |
|------------|--------------------|-----------------|
| KNN        | Training           | 100             |
|            | Testing            | 71.33           |
| NB         | Training           | 35.65           |
|            | Testing            | 37.13           |
| DT         | Training           | 50.03           |
|            | Testing            | 46.86           |
| SVM        | Training           | 93.8            |
|            | Testing            | 67.53           |

represents one healthy (class C_0) and two separate modes of faulty operation (C_i, i = 1, ..., 2), as shown in Table 8.

Tables 9 and 10 show the selected features performance accuracy with different normalization of interval-valued data. One can conclude that the KNN and SVM classifiers present the best performances when using the standardization of interval-valued data-based MSIPCA approach using the global range of interval compared compared to the other standardization methods and to the interval-valued data-based PCA approach.

Tables 9 and 10 show the accuracy of performance. It can be noticed from Tables 11 and 12 that the MSIPCA-based ML gives a higher classification accuracy with comparison to the IPCA-based ML methods. From these results, it is clear that both classifiers give the best results when using \( ||T_i|| \) combined with \( D_N, D_H, D_E \) or \( D_B \). The accuracy of all these MSIPCA-based combinations are between 97.53% and 100%.

TABLE 11. Accuracy of the different IPCA-based extracted features with different classifiers.

| Classifiers | Global Performance | IPCA/\( ||T_i|| \) Accuracy |
|------------|--------------------|-----------------|
| KNN        | Training           | 55.83           |
|            | Testing            | 97.01           |
| SVM        | Training           | 81              |
|            | Testing            | 90.43           |

TABLE 12. Accuracy of the different MSIPCA-based extracted features with different classifiers.

| Classifiers | Global Performance | MSIPCA/\( ||T_i|| \) Accuracy |
|------------|--------------------|-----------------|
| KNN        | Training           | 100             |
|            | Testing            | 99.6            |
| SVM        | Training           | 97.86           |

TABLE 13. Confusion matrix using IPCA-based SVM through \( D_B \).

| Predicted class | True class | C_0 | C_1 | C_2 | Recall |
|-----------------|------------|-----|-----|-----|--------|
| C_0             | 1821       | 125 | 54  | 91.05|
| C_1             | 117        | 1817| 66  | 90.85|
| C_2             | 46         | 60  | 1894| 94.7 |
| Precision       | 91.78      | 90.75| 94.04| 92.2  |

TABLE 14. Confusion matrix using IPCA-based SVM through \( D_B \).

| Predicted class | True class | C_0 | C_1 | C_2 | Recall |
|-----------------|------------|-----|-----|-----|--------|
| C_0             | 1992       | 0   | 8   | 99.6 |
| C_1             | 4          | 1992| 4   | 99.6 |
| C_2             | 8          | 0   | 1992| 99.6 |
| Precision       | 99.4       | 100 | 99.4| 99.6  |

For fault diagnosis of the heating system, it can be noticed from Tables 11 and 12 that the MSIPCA-based ML gives a higher classification accuracy with comparison to the IPCA-based ML methods. From these results, it is clear that both classifiers give the best results when using \( ||T_i|| \) combined with \( D_N, D_H, D_E \) or \( D_B \). The accuracy of all these MSIPCA-based combinations are between 97.53% and 100%.

Via the standardization methods, using the dispersion of interval center, the KNN and SVM classifiers’ accuracy are improved compared to the precision achieved applying the standardization methods 1, 3, and 4. To further improve the accuracy of the classification, city-block distances are added as a new feature. KNN and SVM accuracy’s using this feature present good results comparing to the others. This combination considers the variation of the data in the two IPCA subspaces. Tables 13 and 14 present the SVM confusion matrices applying \( D_B \) as a feature based IPCA and MSIPCA in testing respectively. For the testing healthy data, assigned to class \( C_0 \), the SVM based IPCA (respectively, based MSIPCA) classifier (see Tables 13 and 14) identifies 2000 samples from 6000 (true positive). Moreover, the accuracy of detection is 91.78% (respectively, 99.4%) and its recall is 91.05% (respectively, 99.6%) which also represents the
classification accuracy. Thus, for this class, only 8.22% (respectively 0.6%) of misclassification is observed (false alarms). For the first fault (F1) designated to class C1, the precision is 100% and the recall is 99.6% with 0% of misclassification for the training data set in the case of MSIPCA. The achieved results ratify the effectiveness of the proposed method for FDD of the heating system.

V. CONCLUSION
In this paper, a novel fault detection and diagnosis (FDD) technique was developed for uncertain HVAC systems. The developed method, called multiscale interval principal component analysis (MSIPCA)-based machine learning (ML), was applied for feature extraction and selection and the ML method was used for fault classification. The proposed MSIPCA-ML technique was developed for diagnosing uncertain HVAC systems under various operating conditions

Various cases were considered to prove the robustness and efficiency of the proposed FDD method. The effectiveness of the FDD method was investigated using synthetic and emulated heating system interval valued data. The developed FDD method presented a good diagnosis efficiency and better classification accuracy under different modes.

As future works, improved interval nonlinear feature extraction and selection approaches will be developed to deal with uncertainties and non-linearity natures of HVAC systems. Therefore, interval kernel PCA and kernel PCA-based machine learning classifiers will be developed for fault detection and diagnosis of uncertain and nonlinear HVAC systems.

REFERENCES
[1] A. Spoor, G. Zucker, and R. Hofmann, “Automated HVAC control creation based on building information modeling (BIM): Ventilation system,” IEEE Access, vol. 7, pp. 74747–74758, 2019.

[2] J. E. Braun, “Automated fault detection and diagnostics for vapor compression cooling equipment,” J. Sol. Energy Eng., vol. 125, no. 3, pp. 266–274, Aug. 2003.

[3] Y. Yan, P. B. Luth, and K. R. Pattipati, “Fault diagnosis of components and sensors in HVAC air handling systems with new types of faults,” IEEE Access, vol. 6, pp. 21682–21696, 2018.

[4] C. B. Jones and C. Carter, “Trusted interconnections between a centralized controller and commercial building HVAC systems for reliable demand response,” IEEE Access, vol. 5, pp. 11063–11073, 2017.

[5] S. Pan, Z. Ye, and J. Zhou, “Fault detection filtering for a class of nonhomogeneous Markov jump systems with random sensor saturations,” Int. J. Control, Autom. Syst., vol. 18, no. 2, pp. 439–449, Feb. 2020.

[6] Y. Shen and K. Khorasani, “Hybrid multi-mode machine learning-based fault diagnosis strategies with application to aircraft gas turbine engines,” Neuro. Netw., vol. 130, pp. 126–142, Oct. 2020.

[7] M. Schmid, E. Gebauer, C. Hanzl, and C. Endisch, “Active model-based fault diagnosis in reconfigurable battery systems,” IEEE Trans. Power Electron., early access, Jul. 30, 2020, doi: 10.1109/TPEL.2020.3012964.

[8] M. Mansouri, M.-F. Harkat, H. N. Nounou, and M. Nounou, “Reduced kernel random forest technique for fault detection and classification in grid-tied PV systems,” IEEE J. Photovolta., early access, Aug. 4, 2020, doi: 10.1109/JPHOTOV.2020.3011068.

[9] L. Ren, Z. Y. Xu, and X. Q. Yan, “Single-sensor incipient fault detection,” IEEE Sensors J., vol. 11, no. 9, pp. 2102–2107, Sep. 2011.

[10] M. F. Harkat, M. Mansouri, K. Abdoayeh, M. Nounou, and H. Nounou, “New sensor fault detection and isolation strategy–based interval-valued data,” J. Chemometrics, vol. 34, no. 5, p. e3222, May 2020.

[11] M. Mansouri, M. Haji, M. Trabelsi, M. F. Harkat, A. Al-khazraji, A. Livera, H. Nounou, and M. Nounou, “An effective statistical fault detection technique for grid connected photovoltaic systems based on an improved generalized likelihood ratio test,” Energy, vol. 159, pp. 842–856, Sep. 2018.

[12] S. Joe Qin, “Statistical process monitoring: Basics and beyond,” J. Chemometrics, vol. 17, nos. 8–9, pp. 480–502, 2003.

[13] A. Emami-Naeini, M. M. Akhter, and S. M. Rock, “Effect of model uncertainty on failure detection: The threshold selector,” IEEE Trans. Autom. Control, vol. 35, no. 12, pp. 1106–1115, Dec. 1988.

[14] P. Cazes, A. Chouaikria, E. Diday, and Y. Schektman, “Extension de l’analyse en composantes principales à des données de type intervalle,” Revue de Statistique appliquée, vol. 45, no. 3, pp. 5–24, 1997.

[15] A. Douzal-Chouaikria, “Extension des méthodes d’analyse factorielles à des données de type intervalle,” Ph.D. dissertation, Univ. Paris IX Dauphine, Paris, France, 1998.

[16] F. Palumbo and C. N. Lauro, “A PCA for interval-valued data based on midpoints and radii,” New Develop. Psychometrics, pp. 641–648, 2003.

[17] K. Dhibi, R. Fezai, M. Mansouri, M. Trabelsi, A. Kouadri, K. Bouzara, H. Nounou, and M. Nounou, “An effective statistical fault detection approach for interval-valued photovoltaic systems based on an improved generalized likelihood ratio test,” Energy, vol. 159, pp. 842–856, Sep. 2018.

[18] S. Joe Qin, “Statistical process monitoring: Basics and beyond.,” J. Chemometrics, vol. 17, nos. 8–9, pp. 480–502, 2003.

[19] K. Haji, M. Trabelsi, M. F. Harkat, A. Al-khazraji, A. Livera, H. Nounou, and M. Nounou, “An effective statistical fault detection method for grid connected photovoltaic systems based on an improved generalized likelihood ratio test,” Energy, vol. 159, pp. 842–856, Sep. 2018.

[20] F. Palumbo and C. N. Lauro, “A PCA for interval-valued data based on midpoints and radii,” New Develop. Psychometrics, pp. 641–648, 2003.

[21] P. D’Urso and P. Giordani, “A least squares approach to principal component analysis for interval valued data,” Chemometrics Int. Lab. Syst., vol. 70, no. 2, pp. 179–192, Feb. 2004.

[22] F. Gioia and C. N. Lauro, “Principal component analysis on interval data,” Comput. Statist., vol. 21, no. 2, pp. 343–363, Jun. 2006.

[23] J. Le-Rademacher and L. Billard, “Symbolic covariance principal component analysis and visualization for interval-valued data,” J. Comput. Graph. Statist., vol. 21, no. 2, pp. 413–432, Apr. 2012.

[24] H. Wang, R. Guan, and J. Wu, “CIPCA: Complete-information-based principal component analysis for interval-valued data,” Neurocomputing, vol. 86, pp. 158–169, Jun. 2012.

[25] T. A. Izem, W. Bougeloun, M. F. Harkat, and M. Djezhaba, “Fault detection and isolation using interval principal component analysis methods,” IFAC-PapersOnLine, vol. 48, no. 21, pp. 1402–1407, 2015.

[26] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[27] J. R. Quinlan, “Simplifying decision trees,” Int. J. Man-Mach. Stud., vol. 27, no. 3, pp. 221–234, Sep. 1987.

[28] N. Suguna and K. Thanushkodi, “An improved k-nearest neighbor classification using genetic algorithm,” Int. J. Comput. Sci. Issues, vol. 7, no. 2, pp. 18–20, 2010.

[29] L. Jiang, L. Zhang, L. Yu, and D. Wang, “Class-specific attribute weighted naive Bayes,” Pattern Recognit., vol. 88, pp. 321–330, Apr. 2019.

[30] T. M. Cover, A. R. Kauffman, and M. J. Cheadle, “Introduction to interval analysis,” Philadelphia, PA, USA: SIAM, 2009.

[31] F. D. A. T. de Carvalho, P. Brito, and H.-H. Bock, “Dynamic clustering of interval data based on L2 distance,” Comput. Statist., vol. 21, no. 2, pp. 231–250, Jun. 2006.
A. Kouadri, M. Hajji, M.-F. Harkat, K. Abodayeh, M. Mansouri, S. Valle, W. Li, and S. J. Qin, “Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods,” Ind. Eng. Chem. Res., vol. 38, no. 11, pp. 4389–4401, Nov. 1999.

S. Mallat, “A theory of multiresolution signal decomposition: The wavelet representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 11, no. 7, pp. 674–693, Jul. 1989.

R. Moore, Interval Analysis. Englewood Cliffs, NJ: USA: Prentice-Hall, 1966.

S. J. Qin and R. Dunia, “Determining the number of principal components for best reconstruction,” J. Process Control, vol. 10, nos. 2–3, pp. 245–250, Apr. 2000.

S. Valle, W. Li, and S. J. Qin, “Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods,” Ind. Eng. Chem. Res., vol. 38, no. 11, pp. 4389–4401, Nov. 1999.

A. Kouadri, M. Hajji, M.-F. Harkat, K. Abodayeh, M. Mansouri, H. Nounou, and M. Nounou, “Hidden Markov model based principal component analysis for intelligent fault diagnosis of wind energy converter systems,” Renew. Energy, vol. 150, pp. 598–606, May 2020.

SONDES GHARSELLOAUI received the degree in electrical engineering from the National Engineering School of Monastir (ENIM), University of Monastir, Tunisia, in 2015, where she is currently pursuing the Ph.D. degree with the Laboratory of Automatic Signal and Image Processing, Electrical Engineering Department, National Higher Engineering School of Tunis, Monfleury, Tunisia. She joined the Department of Electrical and Computer Engineering, Texas A&M University at Qatar, as a Research Assistant, in 2019. Her research interests include process modeling, machine learning, fault diagnosis, process modeling and monitoring, multivariate statistical approaches, control systems, control systems theory, big data, and energy management systems.

MAJDI MANSOURI (Member, IEEE) received the degree in electrical engineering from SUPCOM, Tunis, Tunisia, in 2006, the M.Sc. degree in electrical engineering from ENSEIRB, Bordeaux, France, in 2008, the Ph.D. degree in electrical engineering from UTN 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 150 publications. He is also the coauthor 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.

MOHAMED ABDALLAH TRABELSI (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from INSA, Tunisia, in 2006, and the M.Sc. degree in automated systems and the Ph.D. degree in energy systems from INSA Lyon, France, in 2006 and 2009, respectively. From October 2009 to August 2018, he held different Research positions with Qatar University and Texas A&M University at Qatar. Since September 2018, he has been with the Kuwait College of Science and Technology, as an Associate Professor. He has published more than 90 journal and conference papers. He is the author of two books and two book chapters. His research interests include systems control with applications in power electronics, energy conversion, renewable energies integration, and smart grids.

MOHAMED-FAOUZI HARKAT (Member, IEEE) received the degree in automatic control engineering from Annaba University, Annaba, Algeria, in 1996, and the Ph.D. degree from the 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. He has over 20 years of research and practical experience in systems engineering and process monitoring. In 2004, he joined the Electronics Department, Badji Mokhtar – Annaba University, where he is currently a Professor. He is the author of more than 100 refereed journal and conference publications and book chapters. He served on technical committees and an associate editor of several international journals and conferences.

SHADY S. REFAAT (Senior Member, IEEE) was an Electrical Design Engineer with Industry for a period of 12 years. He is currently an Assistant Research Scientist with the Department of Electrical and Computer Engineering, Texas A&M University at Qatar. He is also a member with the Institution of Engineering and Technology (IET) and the Smart Grid Center-Extension in Qatar (SGC-Q). He has published over 50 journal and conference papers. He has successfully realized many potential research projects. His research interests include electrical machines, power systems, smart grid, energy management systems, reliability of power grid and electric machinery, fault detection, and condition monitoring in conjunction with fault management and development of fault tolerant systems.

HASSANI MESSAOUD prepared the Ph.D. thesis with the University of Nice-Siophia Antipolis, France, in 1993, and the Habilitation thesis with the School of Engineers, Tunis, Tunisia, in June 2001. He is currently a Professor and the Head of the Research Laboratory, LARATSI, National Engineering School of Monastir, Monastir, Tunisia. His main research interests include process identification and control and signal and image processing.