IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 30, 2021, accepted April 19, 2021, date of publication April 21, 2021, date of current version May 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074784

A Hybrid Fault Detection and Diagnosis of Grid-Tied PV Systems: Enhanced Random Forest Classifier Using Data Reduction and Interval-Valued Representation

KHALED DHIBI[1], RADHIA FEZAI[1], MAJDI MANSOURI[2], (Senior Member, IEEE), MOHAMED TRABELSI[3], (Senior Member, IEEE), KAIS BOUZRARA[1], HAZEM NOUNOU[2], (Senior Member, IEEE), AND MOHAMED NOUNOU[4], (Senior Member, IEEE)

[1]Research Laboratory of Automation, Signal Processing and Image, National Engineering School of Monastir, Monastir 5019, Tunisia
[2]Electrical and Computer Engineering Program, Texas A&M University at Qatar, Doha 23874, Qatar
[3]Electronic and Communications Engineering Department, Kuwait College of Science and Technology, Safat 13133, Kuwait
[4]Chemical Engineering Program, Texas A&M University at Qatar, Doha 23874, Qatar

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

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

ABSTRACT This paper proposes a novel fault detection and diagnosis (FDD) technique for grid-tied PV systems. The proposed approach deals with system uncertainties (current/voltage variability, noise, measurement errors,...) by using an interval-valued data representation, and with large-scale systems by using a dataset size-reduction framework. The failures encompassed in this study are the open-circuit/short-circuit, islanding, output current sensor, and partial shading faults. In the proposed FDD approach, named interval reduced kernel PCA (IRKPCA)-based Random Forest (IRKPCA-RF), the feature extraction and selection phase is performed using the IRKPCA models while the fault classification is ensured using the RF algorithm. The main contribution of the proposed approach is to provide a good trade-off between low computation time and high classification metrics. The performance of the proposed IRKPCA-RF approach is assessed using a set of emulated data of a grid-tied PV system operating under healthy and faulty conditions. The presented results show that the proposed IRKPCA-RF approach is characterized by enhanced diagnosis metrics, classification rate, and computation time compared to the classical techniques.

INDEX TERMS Random forest, interval-valued data, reduced kernel principal component analysis, fault diagnosis, feature extraction and selection, fault classification, PV systems.

I. INTRODUCTION

Photovoltaic (PV) has become the fastest growing renewable energy technology. Unfortunately, the operation of PV systems is generally accompanied by different types of failures due to the harsh environmental conditions or internal malfunctions [1], [2]. The most common PV systems' failures are the line-to-line or line-to-ground faults, short-circuits, junction box faults, shading effect, inverter fault, grid-connection fault, and open-circuit fault. In addition, hot spots are considered permanent faults. These faults are considered as the most challenging because they might cause serious physical damage and affect the efficiency of the solar modules and electrical power generation. Therefore, the operation of PV systems should be accompanied by the implementation of an accurate fault detection and diagnosis (FDD) algorithm in order to reduce power losses and avoid system collapse [2]–[4]. In recent years, many machine learning (ML) techniques were developed to deal with FDD in PV systems [3], [5]–[7]. Among these techniques, artificial neural network (ANN) [8], support vector machine (SVM) [9], [10] and random forest (RF) [11] are the most common

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

approaches, where RF has been showing better results in terms of fault diagnosis and classification accuracy [11]. A fault detection technique based-on ANN is proposed for detecting partial shading on a PV array [12], where the ANN input data are the solar irradiance, cell temperature, PV current and voltage [12]. In addition, short-circuit faults in a PV array were detected using three-layer feed-forward ANN [12]. In [13], a convolutional neural network (CNN) is adopted to address defective classification and to improve the performance of the flexibly and reliably of the aerial PV module images. The authors in [14] proposed a fault diagnosis scheme based on PCA technique and support vector machine (SVM) for a Cascaded Grid-Connected PV inverter. Based on the concept of bagging, the RF classifier is composed of a random subset of decision trees [15]. In [16], a random forest (RF) ensemble learning algorithm is proposed for FDD of PV arrays. It aims to detect and classify the faults of PV arrays by combining multiple learning algorithms to achieve a superior diagnostic performance. In order to guarantee good detection and diagnosis performances, the application of the RF classifier algorithm should be preceded by the preparation of data inputs, where the feature extraction and selection (FES) are the two most important steps [17], [18]. The goal of the feature extraction is to extract the parameters that correctly describe the system operating conditions, while the feature selection aims to select a small feature subset using a certain criterion. Many FES techniques have been proposed in the literature. Principal component analysis (PCA) [19], independent component analysis (ICA) [20] and partial least squares (PLS) [21] are the most commonly used feature extraction techniques. In [22], a fault classification method based on PCA technique and supervised machine learning was proposed for a grid-connected PV (GCPV) system. The PCA technique is also proposed for enhancing the diagnosis performance by extracting the most significant linear features from data. However, these techniques belong to the linear transformation family and consequently do not consider the nonlinear characteristics of the process. Thus, a nonlinear PCA version, named Kernel PCA (KPCA), has been developed to extract the nonlinear features [23], [24]. The KPCA aims to map the input space to a high-dimension feature space (using kernel function) where the linear PCA can be conducted [23]. KPCA method can effectively extract the nonlinear features contained in the mapping space in order to obtain better classification features [25]. However, the classical KPCA finds its limitations in industrial processes where the variables might be affected by errors/noise leading to an uncertain form of data. The uncertainty in the systems, which is represented by the interval-valued data, is the consideration of the minimum and maximum recorded values, while the single-valued data representation is obtained by a simplification of data during the mining procedure [26]. As alternative solutions, various nonlinear data-based interval-valued KPCA (IKPCA) methods have been developed [27]. The IKPCA technique transforms first the interval-valued data matrix on a numerical data matrix. Then, it projects the input numerical data onto the feature space through a nonlinear mapping function. Finally, the PCA is applied in the feature space. In addition, in [28], an enhanced FDD technique was proposed for wind energy conversion systems. In the developed approach, the interval-valued features were extracted based on interval Gaussian Process Regression (IGPR) model, and the selected features were fed to RF for classification purposes. Unfortunately, the larger is the size of the training data set, the lower is the effectiveness of the feature extraction and selection using the IKPCA as well as IGPR methods in terms of computation time. This drawback limits the implementation of these methods in real-world applications with massive data. To overcome this limitation, an improved technique based on a data size reduction framework is proposed in this paper. The proposed technique makes use of the Euclidean distance (ED) criteria to remove the irrelevant and redundant interval-valued samples. Then, an IKPCA model is employed to compute the nonlinear interval type features from reduced interval valued-data. Therefore, two versions of reduced IKPCA (IRKPCA) are proposed to extract features by transforming the single-valued data set into interval-valued latent variables with low computation time. The first IRKPCA$_{CR}$ approach concatenates the center and range matrices to compute the new numerical matrix and then fits an RKPCA model on the matrix. The second method, the IRKPCA$_{UL}$, fits two RKPCA models on the lower and upper bounds of the interval-valued variables. Next, it is important to select the most relevant and informative features before performing the classification task in order to improve the diagnosis effectiveness. Finally, for a high classification accuracy, the selected features are fed into a multi-class RF model for fault classification purposes.

In summary, the aims of this paper is to propose novel FDD approaches for uncertain PV systems. The main contribution is to provide a good trade-off between low computation time and high classification metrics. Therefore, two multi-class classifiers called IRKPCA$_{CR}$-RF and IRKPCA$_{UL}$-RF as well as a bank of one-class classifiers are proposed (there are as many classifiers as classes). The effectiveness of the proposed methods is investigated using a set of PV systems data where the faults are emulated at different stages (common coupling point, inverter, sensors, emulated PV arrays,. . .).

The rest of the paper is structured as follows. Section II describes the IRKPCA-based FES methods. The RF-based fault classification is detailed in Section III. The performance of the proposed IRKPCA-RF methods is evaluated in Section IV, while interpretations and conclusions are drawn in Section V.

II. PRE-PROCESSING BY HYBRID DIMENSIONALITY REDUCTION

This section details the proposed hybrid dimensionality reduction for FES. The collected data are pre-processed by extracting the features using the IRKPCA technique and selecting the most active features during the training process.

A. FEATURE EXTRACTION BASED ON INTERVAL REDUCED KPCA TECHNIQUE

Two ED-based IRKPCA models are proposed to remove the irrelevant and redundant samples during the feature extraction process.

1) INTERVAL-VALUED DATA AND REDUCTION FRAMEWORK

The use of interval-valued data is motivated by the need of size reduction of massive datasets in some applications. An interval-valued variable $[x_{j,k}]$, can be determined using a lower and upper bound [27], such as $[x_j(k)] = [\underline{x}_{j,k}, \overline{x}_{j,k}]$, where $k \in \{1, \ldots, N\}$, $\underline{x}_{j,k} \leq \overline{x}_{j,k}$, and N is the number of samples.

Given an $N \times m$ classical training data matrix X, where m is the number of variables and N is the number of samples, the interval data matrix $[X]$ can be constructed as per:

$$[X] = \begin{pmatrix} [\underline{x}_{1,1}, \overline{x}_{1,1}] & . & . & [\underline{x}_{1,m}, \overline{x}_{1,m}] \\ . & . & & . \\ . & & . & . \\ [\underline{x}_{N1}, \overline{x}_{N,1}] & . & . & [\underline{x}_{N,m}, \overline{x}_{N,m}] \end{pmatrix} \tag{1}$$

where, the lower X^L and upper X^U bound matrices are respectively defined by:

$$X^L = \begin{pmatrix} \underline{x}_{1,1} & . & . & \underline{x}_{1,m} \\ . & . & & . \\ . & & . & . \\ \underline{x}_{N,1} & . & . & \underline{x}_{N,m} \end{pmatrix} \tag{2}$$

$$X^U = \begin{pmatrix} \overline{x}_{1,1} & . & . & \overline{x}_{1,m} \\ . & . & & . \\ . & & . & . \\ \overline{x}_{N,1} & . & . & \overline{x}_{N,m} \end{pmatrix} \tag{3}$$

The interval-valued variable $[x_{j,k}]$ can be also expressed by a couple $\{x^c_{j,k}, x^r_{j,k})\}$.

The center $x^c_{j,k}$ of the interval is given by:

$$x^c_{j,k} = \frac{1}{2}(\overline{x}_{j,k} + \underline{x}_{j,k}) \tag{4}$$

and the range $x^r_j(k)$ of the interval is defined by:

$$x^r_j(k) = \frac{1}{2}(\overline{x}_{j,k} - \underline{x}_{j,k}) \tag{5}$$

In this case, the center and range matrices are respectively defined by:

$$X^c = \frac{1}{2}\begin{pmatrix} \underline{x}_{11} + \overline{x}_{1,1} & . & . & \underline{x}_{1m} + \overline{x}_{1,m} \\ . & . & & . \\ . & & . & . \\ \underline{x}_{N,1} + \overline{x}_{N,1} & . & . & \underline{x}_{N,m} + \overline{x}_{N,m} \end{pmatrix} \tag{6}$$

$$X^r = \frac{1}{2}\begin{pmatrix} \underline{x}_{11} - \overline{x}_{1,1} & . & . & \underline{x}_{1,m} - \overline{x}_{1,m} \\ . & . & & . \\ . & & . & . \\ \underline{x}_{N,1} - \overline{x}_{N,1} & . & . & \underline{x}_{N,m} - \overline{x}_{N,m} \end{pmatrix} \tag{7}$$

By the concatenation of the center and range matrices, the new data matrix $\mathbf{X}_{CR}$ can expressed by:

$$\mathbf{X}_{cr} = \left[\mathbf{X}^c \ \mathbf{X}^r\right] \tag{8}$$

Before applying the KPCA technique, the size of the training dataset is first reduced by using the ED as a dissimilarity metric between observations. This indicator is adopted to select the relevant features. The ED between all observations are calculated and then the features with the greater ED values are selected. Let consider $x_{cr} = [x^c \ x^r] \in \mathcal{R}^{2m}$ a new data sample. The data including the dissimilarity between all pairs of observations can be represented using a dissimilarity matrix D as per:

$$D = \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1N} \\ d_{21} & d_{22} & \ldots & d_{2N} \\ . & . & . & . \\ d_{N1} & d_{N2} & \ldots & d_{NN} \end{bmatrix} \tag{9}$$

where d_{ij} represents the ED between the rows X_{CR_i} and X_{CR_j} of the data matrix X_{CR}. So, d_{ij} is given by:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \left(x'_{CR_{i,k}} - x'_{CR_{j,k}}\right)^2} \tag{10}$$

The new reduced data matrix X'_{cr} is expressed as:

$$X'_{CR} = \begin{bmatrix} x'_{CR_1} & x'_{CR_2} & . & . & . & x'_{CR_{N'}} \end{bmatrix}^T \in \mathbf{R}^{N' \times 2m} \tag{11}$$

where N' is the size of the reduced data matrix. For the interval-valued data based on upper and lower bounds, the lower bound ED (ED^L) and upper bound ED (ED^U) are respectively determined by:

$$ED^L_{i,j} = \sqrt{\sum_{k=1}^{m} \left(x^L_{i,k} - x^L_{j,k}\right)^2} \tag{12}$$

$$ED^U_{i,j} = \sqrt{\sum_{k=1}^{m} \left(x^U_{i,k} - x^U_{j,k}\right)^2} \tag{13}$$

Thus, the new lower $X^{L'}$ and upper $X^{U'}$ matrices are respectively constructed by:

$$X^{L'} = \begin{bmatrix} x'^L_1 & x'^L_2 & . & . & . & x'^{L'}_N \end{bmatrix}^T \in \mathbf{R}^{N' \times m} \tag{14}$$

$$X^{U'} = \begin{bmatrix} x'^L_1 & x'^L_2 & . & . & . & x'^{L'}_N \end{bmatrix}^T \in \mathbf{R}^{N' \times m} \tag{15}$$

2) INTERVAL REDUCED KPCA MODEL

Once the reduced interval training matrices are determined using interval centers/ranges and lower/upper bounds approaches, the RKPCA technique is applied to the new reduced data matrices. The IRKPCA-based interval centers and ranges IRKPCA$_{CR}$ applies the RKPCA technique to the new data matrix X'_{cr} which is formed by the concatenation of center and range reduced data matrices. The IRKPCA$_{CR}$

aims first to project the data matrix X'_{cr} to the feature space $\mathcal{X}'_{CR} = \left[\phi(x'_{CR_1})\ \phi(x'_{CR_2})\ \ldots\ \phi(x'_{CR_{N'}})\right]^T$.

Then, the covariance C of $\mathcal{X}'_{\mathcal{CR}}$ can be calculated in the following form:

$$\begin{aligned} C &= \frac{1}{N'-1}{\mathcal{X}'_{\mathcal{CR}}}^T \mathcal{X}'_{\mathcal{CR}} \\ &= \frac{1}{N'-1}\sum_{i=1}^{N'} \phi(x'_{CR_i})\phi^T(x'_{CR_i}) \end{aligned} \tag{16}$$

The eigenvalue λ and the corresponding eigenvector v of the covariance C can be computed to satisfy the following equation:

$$\lambda v = Cv = \frac{1}{N'-1}\sum_{i=1}^{N'} \phi(x'_{CR_i})\phi^T(x'_{CR_i})v \tag{17}$$

Note that the eigenvector equation 16 depends only on the dot products of mapped vectors in the feature space. In general, the mapping function $\phi(.)$ is not explicitly defined. Thus, the covariance matrix C may not be calculated implicitly. To avoid this problem, the kernel matrix is defined by $K = \mathcal{X}'_{CR}\mathcal{X}'^T_{CR}$, which is calculated using the kernel function $\phi(x'_{CR_i})^T\phi(x'_{CR_i}) = k(x'_{CR_i}, x'_{CR_i})$ and given by:

$$\begin{aligned} K &= \mathcal{X}'_{CR}\mathcal{X}'^T_{CR} \\ &= \begin{bmatrix} k(x'_{CR_1}, x'_{CR_1}) & \cdots & k(x'_{CR_1}, x'_{CR_{N'}}) \\ \vdots & \ddots & \vdots \\ k(x'_{CR_{N'}}, x'_{CR_1}) & \cdots & k(x'_{CR_{N'}}, x'_{CR_{N'}}) \end{bmatrix} \end{aligned} \tag{18}$$

Many kinds of the kernels $k(.,.)$ [29] exist, where the mostly applied Gaussian kernel function is expressed by:

$$k(x'_{CR_i}, x'_{CR_j}) = exp\left(-\frac{(x'_{CR_i} - x'_{CR_j})^T(x'_{CR_i} - x'_{CR_j})}{c}\right) \tag{19}$$

where c is the width of the Gaussian function.

To solve the eigenvector equation, it is assumed that $\alpha = \mathcal{X}'_{CR}v$. Multiply both sides of equation 17 with $\mathcal{X}'_{CR}$ leads to:

$$K\alpha = \lambda\alpha \tag{20}$$

where λ and α are the eigenvalue and the eigenvector of the kernel matrix K, respectively. Through the expression of $\alpha = \mathcal{X}'_{CR}v$, the eigenvector v can be expressed by:

$$v = \lambda^{-1}\mathcal{X}'^T_{CR}\alpha \tag{21}$$

Then, the matrix of the ℓ retained principal loading of the RKPCA is obtained in the feature space by $\hat{P} = [v_1, \ldots, v_\ell] \in \mathbb{R}^{N'\times\ell}$ and the $N' - \ell$ last principal loading is denoted by $\tilde{P} = [v_{\ell+1}, \ldots, v_{N'}] \in \mathbb{R}^{N\times(N'-\ell)}$.

$$\hat{P} = [\frac{1}{\lambda_1}\mathcal{X}'^T_{CR}\alpha_1, \ldots, \frac{1}{\lambda_\ell}\mathcal{X}'^T_{CR}\alpha_\ell] \tag{22}$$

Many studies have investigated the selection of the number of principal components (PCs). Therefore, in order to determine the number ℓ of significant PCs, the cumulative percent variance (CPV) criterion is used [30].

The principal and residual components, respectively $\hat{t} \in \mathbf{R}^\ell$ and $\tilde{t} \in \mathbf{R}^{N-\ell}$ of a test vector x_{CR} are then extracted by projecting $\phi(x_{CR})$ into the principal and residual spaces, as follows:

$$\begin{cases} \hat{t} = \hat{P}^T\phi(x_{CR}) \\ = \Lambda^{-\frac{1}{2}}P^T k(x_{CR}) \\ \tilde{t} = \tilde{P}^T\phi(x_{CR}) \end{cases} \tag{23}$$

where $k(x_{CR}) = \left[k(x_{CR_1}, x_{CR})\ldots k(x_{CR_{N'}}, x_{CR})\right]^T$, $P = \left[\alpha_1\ \alpha_2\ \ldots\ \alpha_\ell\right]$ and $\Lambda = diag(\lambda_1 \ldots \lambda_\ell)$.

For the proposed IRKPCA-based interval lower and upper bounds IKPCA_{LU}, two single-valued RKPCA models are applied to the lower and upper bounds of the reduced interval-valued data. To this end, the interval data matrices should be transformed into the feature space as follows:

$$\mathcal{X}^{L'} = \left[\phi(x_1^{L'})\quad \phi(x_2^{L'})\quad \ldots\quad \phi(x_{N'^{L'}})\right]^T \tag{24}$$

$$\mathcal{X}^{U'} = \left[\phi(x_1^{U'})\quad \phi(x_2^{U'})\quad \ldots\quad \phi(x_{N'}^{U'})\right]^T \tag{25}$$

$$k(\underline{x'}_i, \underline{x'}_j) = exp\left(\frac{-\|\underline{x'}_i - \underline{x'}_j\|^2}{2\sigma^2}\right) \tag{26}$$

$$k(\overline{x'}_i, \overline{x'}_j) = exp\left(\frac{-\|\overline{x'}_i - \overline{x'}_j\|^2}{\sigma^2}\right) \tag{27}$$

Thus, by using the kernel function for the lower bound data, the kernel matrix in the feature space is expressed by:

$$K^L = \mathcal{X}'^L\,\mathcal{X}'^{L^T} = \begin{bmatrix} k(\underline{x'}_1, \underline{x'}_1) & \ldots & k(\underline{x'}_1, \underline{x''}_N) \\ . & & . \\ . & \ldots & . \\ . & & . \\ k(\underline{x''}_N, \underline{x'}_1) & \ldots & k(\underline{x''}_N, \underline{x''}_N) \end{bmatrix} \tag{28}$$

$$K^U = \mathcal{X}'^U\,\mathcal{X}'^{U^T} = \begin{bmatrix} k(\overline{x'}_1, \overline{x'}_1) & \ldots & k(\overline{x'}_1, \overline{x'}_N) \\ . & & . \\ . & \ldots & . \\ . & & . \\ k(\overline{x'}_N, \overline{x'}_1) & \ldots & k(\overline{x'}_N, \overline{x'}_N) \end{bmatrix} \tag{29}$$

The eigen-decomposition of the kernel matrix K^L provides the necessary information to compute the projections of the lower $\phi(\underline{x})$ vector in the feature space which are given by:

$$\begin{cases} \hat{t}^L = (\Lambda^L)^{-\frac{1}{2}}(P^L)^T k(\underline{x}) \\ \tilde{t}^L = (\tilde{P}^L)^T\phi(\underline{x}) \end{cases} \tag{30}$$

where $k(\underline{x}) = \left[k(\underline{x'}_1, \underline{x})\ldots k(\underline{x'}_{N'}, \underline{x})\right]^T$,
$P^L = \left[\alpha_1^L\ \alpha_2^L \ldots \alpha_{\ell^L}\right]$ and $\Lambda^L = diag(\lambda_1^L \ldots \lambda_{\ell^L})$.

The eigenvalues and eigenvectors decomposition of the matrix K^U will be determined to obtain a RKPCA representation of the nonlinear uncertain data. Then, the new matrices of the eigenvectors P^U and eigenvalues Λ^U are used to compute

the scores of the upper $\phi(\overline{x})$ vector which are defined by:

$$\begin{cases} \hat{t}^u = (\Lambda^U)^{-\frac{1}{2}}(P^U)^T k(\overline{x}) \\ \tilde{t}^U = (\tilde{P}^U)^T \phi(\overline{x}) \end{cases} \tag{31}$$

where $k(\overline{x}) = \left[k(\overline{x'}_1, \overline{x}) \dots k(\overline{x'}_{N'}, \overline{x}) \right]^T$,
$P^U = \left[\alpha_1^u \ \alpha_2^U \dots \alpha_{\ell^U}^U \right]$ and $\Lambda^U = diag(\lambda_1^U \dots \lambda_{\ell^u}^U)$.

B. FEATURE SELECTION

The first ℓ IKPCs constitute the features extracted from the IRKPCA model. Several effective features can be computed from the IKPCs. In the current work, the Hotelling's T^2 statistic, squared prediction error statistic (SPE), combined index φ, sampled mean M, variance D^2, skewness S, and kurtosis K metrics are applied for feature selection [22], [31]. In the following, more details about the computation of these features are presented.

1) IRKPCA$_{CR}$-BASED FEATURE SELECTION

The above mentioned feature selection methods of the first ℓ retained KPCs $\hat{t}$, obtained from the IRKPCA$_{CR}$ model for interval valued data, are described as follows:

$$T^2 = k(x_{CR}) P \Lambda^{-1} P^T k(x_{CR}) \tag{32}$$
$$SPE = k(x_{CR}, x_{CR}) - k(x_{CR})^T C k(x_{CR}) \tag{33}$$

where $C = P\Lambda^{-1}P^T$.

$$\varphi = \frac{SPE}{\delta^2} + \frac{T^2}{\tau^2} \tag{34}$$

$$M_j = \frac{1}{\ell} \sum_{i=1}^{\ell} \hat{t}_{ji} \tag{35}$$

$$D_j^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} (\hat{t}_{ji} - M_j)^2 \tag{36}$$

$$S_j = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\frac{\hat{t}_{ji} - M_j}{D_j^2} \right)^3 \tag{37}$$

$$K_j = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\frac{\hat{t}_{ji} - M_j}{D_j^2} \right)^4 \tag{38}$$

2) IRKPCA$_{LU}$-BASED FEATURE SELECTION

In this case, the interval features are selected from the IRKPCA$_{LU}$ model. The selected features of the lower bound are described as follows:

$$(T^L)^2 = k(\underline{x}) P^L (\Lambda^L)^{-1} (P^L)^T k(\underline{x}) \tag{39}$$
$$SPE^L = k(\underline{x}, \underline{x}) - k(\underline{x})^T C^L k(\underline{x}) \tag{40}$$

where $C^L = P^L(\Lambda^L)^{-1}(P^L)^T$.

$$\varphi^L = \frac{SPE^L}{(\delta^L)^2} + \frac{(T^L)^2}{(\tau^L)^2} \tag{41}$$

where δ^L and τ^L are the threshold of the SPE^L and $(T^L)^2$ indices of the lower bound.

$$M_j^L = \frac{1}{\ell^L} \sum_{i=1}^{\ell^L} \hat{t}_{ji}^L \tag{42}$$

$$(D_j^L)^2 = \frac{1}{\ell^L} \sum_{i=1}^{\ell^L} (\hat{t}_{ji}^L - M_j^L)^2 \tag{43}$$

$$S_j^L = \frac{1}{\ell^L} \sum_{i=1}^{\ell^L} \left(\frac{\hat{t}_{ji}^L - M_j^L}{D_j^2} \right)^3 \tag{44}$$

$$K_j^L = \frac{1}{\ell^L} \sum_{i=1}^{\ell^L} \left(\frac{\hat{t}_{ji}^L - M_j^L}{(D_j^L)^2} \right)^4 \tag{45}$$

The selected features of the upper bound are defined as follows:

$$(T^U)^2 = k(\overline{x}) P^U (\Lambda^U)^{-1} (P^U)^T k(\overline{x}) \tag{46}$$
$$SPE^U = k(\overline{x}, \overline{x}) - k(\overline{x})^T C^U k(overlinex) \tag{47}$$

where $C^U = P^U(\Lambda^U)^{-1}(P^U)^T$.

$$\varphi^U = \frac{SPE^U}{(\delta^U)^2} + \frac{(T^U)^2}{(\tau^U)^2} \tag{48}$$

where δ^U and τ^U are the threshold of the SPE^U and $(T^U)^2$ indices of the upper bound.

$$M_j^U = \frac{1}{\ell^U} \sum_{i=1}^{\ell^U} \hat{t}_{ji}^U \tag{49}$$

$$(D_j^U)^2 = \frac{1}{\ell^U} \sum_{i=1}^{\ell^U} (\hat{t}_{ji}^U - M_j^U)^2 \tag{50}$$

$$S_j^U = \frac{1}{\ell^U} \sum_{i=1}^{\ell^U} \left(\frac{\hat{t}_{ji}^U - M_j^U}{D_j^2} \right)^3 \tag{51}$$

$$K_j^U = \frac{1}{\ell^U} \sum_{i=1}^{\ell^U} \left(\frac{\hat{t}_{ji}^U - M_j^U}{(D_j^U)^2} \right)^4 \tag{52}$$

In order to take into account the upper and lower bounds at the same time, new interval-valued statistical features are used. The unified representation of the interval features is defined by:

$$\Omega = \gamma \underline{\Omega} + (1 - \gamma) \overline{\Omega} \tag{53}$$

where Ω is the unified form of the statistical features that could be one of the available detection indices, the mean, variance, skewness, or kurtosis. $\gamma \in [0, 1]$ is the weight that defines the trade-off between the upper and lower bounds.

III. IRKPCA-BASED RANDOM FOREST FAULT CLASSIFICATION

In the proposed IRKPCA based Random Forest (IRKPCA-RF) for fault classification, the most relevant features including interval statistical features are firstly computed using the IRKPCA model, then they are introduced to

RF model for classification purposes. RF is a type of Artificial intelligence technique to identify the state of the PV system. It aims to create a forest using multiple decision trees [32]. RF can enhance the classification accuracy resulting to grow a set of trees and make them vote for the most promising class. In the classification stage, the direct use of the measured variables by the RF classifier lower its performance in case of data noises and redundancies. Thus, the extraction and selection the relevant features using the IRKPCA models are proposed in this paper. Then, the relevant features are introduced to RF classifier to perform fault classification. Thus, enhanced approaches based on the proposed IRKPCA methods the RF classifier are presented. The aim of the developed IRKPCA-RF techniques is to provide the best compromise between high classification metrics and low computational time. To discriminate between the healthy and faulty cases, the developed IRKPCA-RF approaches collect first the interval-valued PV data and then divide it (step 1 in Figure 1) into training and testing data sets. During the training phase, the size of the raw data is firstly reduced using the DR metrics while retaining the non-redundant information. Secondly, the IRKPCA models are applied to extract and select the relevant features (step 2 in Figure 1). In step 3, the RF uses the selected features for training. In step 4, the classification model is completed as per shown in Figure 1). After the IRKPCA models are constructed and well-trained, the testing data set is used to evaluate the model performance for fault classification. In step 5, the IRKPCA model is applied to extract and select the most relevant features. The interval statistical features are selected to provide good performance, reduce the number of features and increase accuracy. Finally, the final features are fed to the RF classifier to detect and classify faults. The main advantages of the proposed methods is the introduction of the features extraction and selection steps which can improve the fault classification accuracy by avoiding data noises and redundancies.

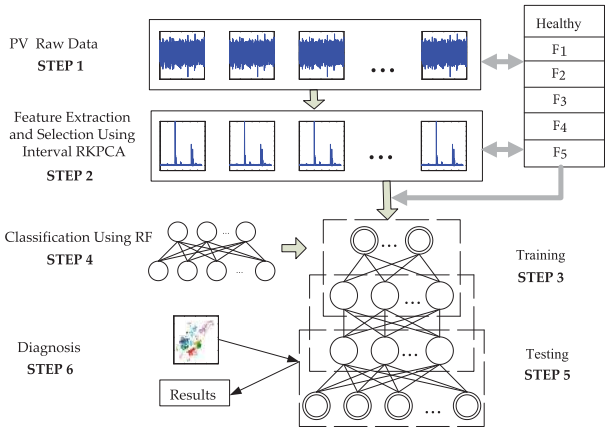


FIGURE 1. Flowchart of the proposed fault detection and diagnosis.

Algorithm 1 summarizes the overall procedure of the the IRKPCA-RF algorithms.

Algorithm 1 IRKPCA-RF Algorithm

STEP 1. Inputs: $N \times m$ input interval data matrix $[X]$,
STEP 2. After data acquisition, compute the reduced interval data matrix to reduce the data dimensionality space while removing the relative samples which are redundant and even degrade the performance of the process,
STEP 3. Extracting the interval features using IRKPCA$_{CR}$ and IRKPCA$_{LU}$ algorithms,
STEP 4. Selecting the significant features from the features obtained from the IRKPCA$_{CR}$ and IRKPCA$_{LU}$ models,
STEP 5. TheRF classifier is trained using the selected features then evaluated by testing features.
STEP 6. Classifying the healthy and faulty operating conditions.

IV. RESULTS AND DISCUSSION

In this section, the performance of the proposed FDD methods is assessed using a set of emulated PV system data. The diagnosis assessment indicators include: 1) Normalized Classification Accuracy (NCA), which represents the ratio of the number of correct predictions to the total number of input samples, 2) Normalized Recall (NR), which is the percentage of fault measurements that are correctly classified over the total number of measurements in the pertinent fault class, 3) Normalized Precision (NP), which defines the number of samples properly classified divided by the number of classified samples, and 4) Computation Time (CT), which represents the time required to execute the FDD algorithm.

A. PV SYSTEM DATA COLLECTION

Figure 3 shows the synoptic of the grid-tied PV system under study, where Chroma PV and grid emulators are used. The system variables shown in Figure 3 were measured every 5-15s depending on the nature of the faults and their occurrence.

The faults were emulated at the common coupling point, inverter, sensors, and PV emulator [22], [31]. In the AC side, an open-circuit fault introduced on one inverter switch at the time is referred as an inverter fault F_1, while the islanding (grid-connection fault) is represented by an F_3 fault. In the PV side, the output current sensor wiring/reading errors are denoted by the F_2 fault. Additionally, a 10-20 % permanent partial shading fault (PV panel fault F_4) and open-circuit/short-circuit on PV cells connection faults (connection faults F_5) were emulated using the Chroma PV emulator functions.

1) Grid-side faults
   - F_1: Inverter fault (open-circuit fault on one switch at the time),
   - F_3: Grid-connection fault (switch to the standalone operation for protection reasons).
2) PV-side faults
   - F_2: Output current sensor fault (poor connection and/or erroneous reading),

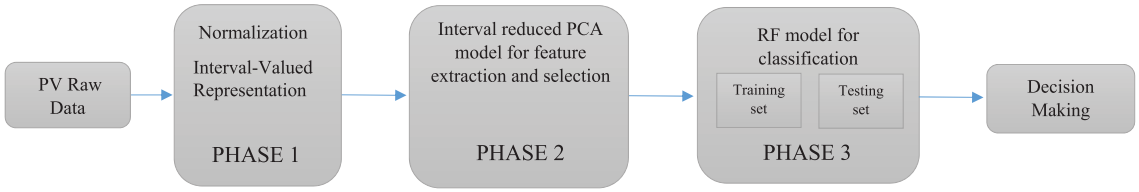


FIGURE 2. Schematic diagram of the IRKPCA-RF methodology.

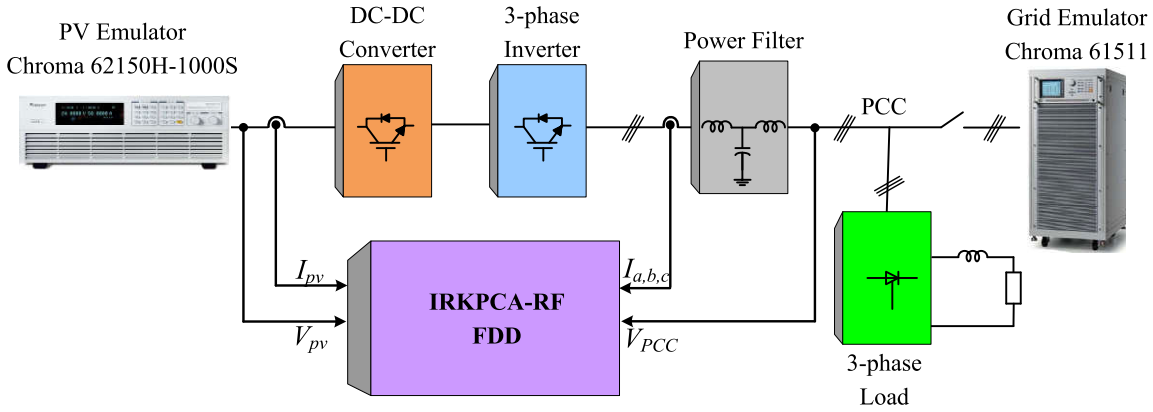


FIGURE 3. Synoptic of the studied grid-tied PV system.

- F_4: PV panel fault (permanent 10-20 % partial shading)
- F_5: PV panel connection fault (open-circuit, short-circuit, sudden disconnection)

The database is constructed by one healthy mode assigned to class C_0 and F_1-F_5 faulty modes assigned to classes C_1-C_5 respectively (Table 1).

TABLE 1. Constructed FDD database.

Class	State	Training Data	Testing Data
C_0	Healthy	1501	1501
C_1	F_1	1501	1501
C_2	F_2	1501	1501
C_3	F_3	1501	1501
C_4	F_4	1501	1501
C_5	F_5	1501	1501

B. FAULT DIAGNOSIS RESULTS

During the first step, the data set was standardized to zero mean and unit variance. Then, the interval training data set was used to create the $IKPCA_{CR}$ and $IKPCA_{LU}$ models, while the $IRKPCA_{CR}$ and $IRKPCA_{LU}$ models are created by means of the interval reduced data sets obtained using the ED. The number of IKPCs are determined using the cumulative percent variance (CPV) with 95% as an explained variance threshold. The retained number of IKPCs using both $IKPCA_{CR}$ and $IKPCA_{LU}$ models is equal to 31, while it is equal to 18 using $IRKPCA_{CR}$ and $IRKPCA_{LU}$ models. The faults are labeled within the built database and the best significant features are selected from the extracted characteristics to obtain good classification results. Therefore, five arbitrary groups of features are used and the best one is selected (Table 2).

TABLE 2. Selected features for fault classification.

Groups	Features Descriptions
Group 1	Sampled mean, IT^2
Group 2	Sampled mean, $ISPE$
Group 3	Sampled mean, $I\phi$
Group 4	Sampled mean, variance, kurtosis and skewness of the ℓ retained IKPCs
Group 5	The first ℓ IKPCs

The scatter plot of the IKPC1 and IKPC2 retained under different operating conditions of the GCPV system is illustrated in Figure 4. In addition, the various plan-projections of the features are given in Figures 5 to 7. It is clearly shown from these figures that all classes are clearly observed. Besides, we can show that the six classes are not totally distinguished. Therefore, the selected features will be introduced as inputs to RF classifier in order to enhance further the classification results.

The labeled data are then used as inputs for the proposed techniques which can be divided into two stages: a multi-class classifier stage (see Table 3) and a bank of one-class classifiers (see Table 8). The first step is to extract features from the data, and then select the most significant group from the extracted features which gives the best classification results. Therefore, an evaluation of the two proposed multi-class methods has been performed using groups 1-5 for training and testing phases (Table 3).

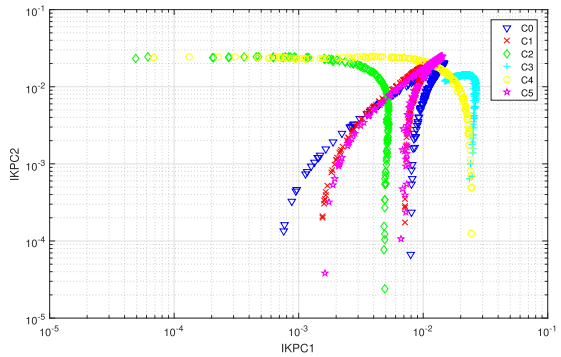


FIGURE 4. Scatter plot of IKPC1 and IKPC2.

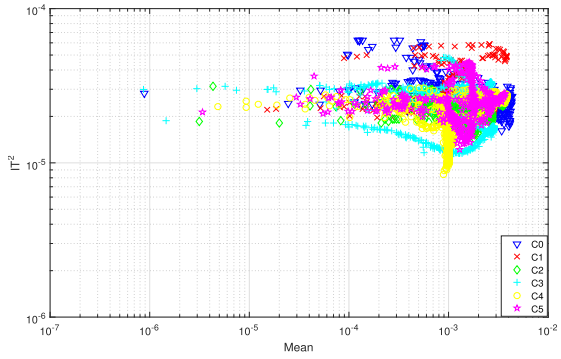


FIGURE 5. Scatter plot of mean and T^2 statistics.

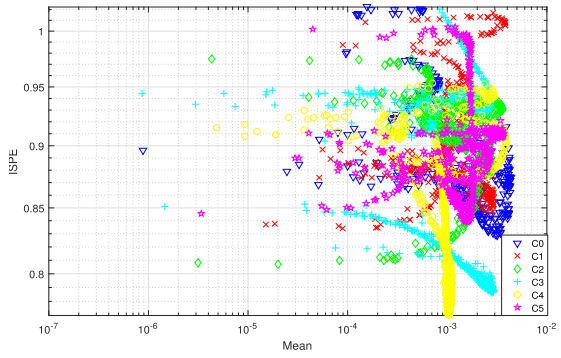


FIGURE 6. Scatter plot of mean and *ISPE* statistics.

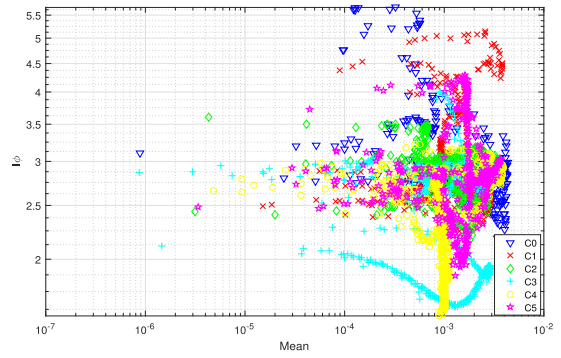


FIGURE 7. Scatter plot of mean and $I\phi$ statistics.

One can notice that using group 5, the developed techniques can achieve a perfect NCA. As shown in Table 3, both IRKPCA-RF and IKPCA-RF methods achieved perfect

TABLE 3. Classification accuracy of the proposed IRKPCA$_{CR}$-RF and IRKPCA$_{UL}$-RF techniques.

Method		Extracted Features				
	NCA	group 1	group 2	group 3	group 4	group 5
IRKPCA$_{CR}$-RF	Training	.6	.7	.59	.93	**1**
	Testing	.55	.67	.56	.84	**1**
IRKPCA$_{UL}$-RF	Training	.59	.75	.65	.91	**1**
	Testing	.55	.73	.64	.86	**1**
IKPCA$_{CR}$-RF	Training	0.56	0.79	0.68	0.88	**1**
	Testing	0.55	0.78	0.67	0.88	**1**
IKPCA$_{UL}$-RF	Training	0.54	0.81	0.67	0.89	**1**
	Testing	0.53	0.77	0.64	0.85	**1**

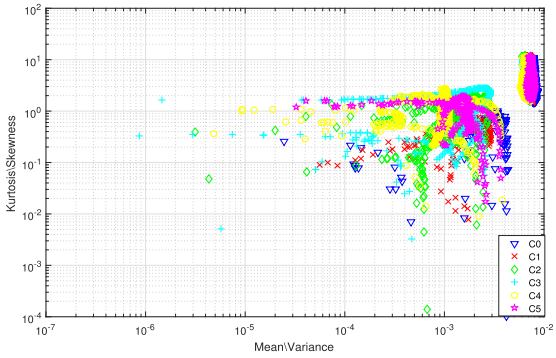


FIGURE 8. Scatter plot of mean, variance, kurtosis and skewness of the ℓ retained IKPCs statistics.

results in terms of NCA for the training (1) and testing (1) phases using the group of features 5.

The confusion matrix is another performance measurement for machine learning classification. The confusion matrices of the investigated techniques are presented in Tables 4 and 5, where the correct classified and mis-classified observations for the condition modes (C_0 to C_5) are presented. One can notice from these tables that the IKPCA-RF and IRKPCA-RF techniques identify 1501 observations among 1501 (true positive) in six different modes. Besides, the NP is 1 and its recall is 1 for all different modes. This confirms that the proposed

TABLE 4. Confusion matrix of IKPCA$_{CR}$-RF and IKPCA$_{UL}$-RF classifiers using group 5.

Conf. Matrix		Predicted process statuses						NR
True classes	C_0	1501	0	0	0	0	0	1
	C_1	0	1501	0	0	0	0	1
	C_2	0	0	1501	0	0	0	1
	C_3	0	0	0	1501	0	0	1
	C_4	0	0	0	0	1501	0	1
	C_5	0	0	0	0	0	1501	1
NP		1	1	1	1	1	1	**1**

TABLE 5. Confusion matrix of IRKPCA$_{CR}$-RF and IRKPCA$_{UL}$-RF classifiers using group 5.

Conf. Matrix		Predicted process statuses						NR
True classes	C_0	1501	0	0	0	0	0	1
	C_1	0	1501	0	0	0	0	1
	C_2	0	0	1501	0	0	0	1
	C_3	0	0	0	1501	0	0	1
	C_4	0	0	0	0	1501	0	1
	C_5	0	0	0	0	0	1501	1
NP		1	1	1	1	1	1	**1**

TABLE 6. Comparative classification accuracy and computation time results using group 5.

	Methods	
Global Performance	NCA (Training/Testing)	CT(s)
IRKPCA$_{CR}$-RF	**1/1**	**90.89**
IRKPCA$_{UL}$-RF	**1/1**	**108.14**
IKPCA$_{CR}$-RF	**1/1**	230.06
IKPCA$_{UL}$-RF	**1/1**	239.53
IPCA-RF	.92/.91	**89.82**
NN	.60/.68	**9.11**
RNN	.66/.68	259.08

techniques can distinguish between the six operating modes and provide a perfect classification accuracy.

To further highlight the classification performance of the proposed approaches, Table 6 shows a comparison in terms of classification accuracy and computation time between the developed techniques, two IKPCA-RF techniques, an IPCA-RF approach [33], Neural Network (NN) and Recurrent Neural Network (RNN). In fact, the NCA of both IKPCA-RF and IRKPCA-RF methods is 1 during the training and testing phases, while only 0.92 and 0.91 NCA is achieved by the IPCA-RF algorithm. One can notice from Table 6 that the proposed methods provide the best NCA compared to the NN and RNN methods. It is worth noting that the NN has a low computation time (9.11 s) while presenting a low classification accuracy (0.60/0.68). Moreover, the CT comparison shows low values for the IPCA and the proposed IRKPCA$_{UL}$-RF and IRKPCA$_{CR}$-RF methods compared to the IKPCA-based ones, where the best trade-off between NCA and CT is offered by the IRKPCA$_{CR}$-RF technique. Besides, the proposed methods based on data reduction framework provide a significant reduction in terms of computation time compared to the IKPCA-RF methods. As shown in Table 6, the computation time of the proposed IRKPCA-RF methods is reduced approximately to 60% compared to the ones recorded for the IKPCA-RF methods. Based on the above discussion, the best trade-off between the computation time and classification metrics is obtained using the proposed IRKPCA-RF approaches.

Additionally, in order to further investigate the performance of the proposed FDD techniques, a bank of one-class classifiers containing six classifiers is considered. Each classifier is trained to classify a specific class with a label 1 or -1 according to the input features which are computed and compared (see Table 7). Table 8 presents the global performance accuracy using the selected features of group 5 as inputs in the case of one-class classifiers scenario. As shown in Table 8, all four methods give comparable average accuracy in the training and testing phases. Besides, it is clearly shown in this Table that more than 50% of CT reduction is offered by both IRKPCA-RF methods compared to the

TABLE 7. Multiple one-class classifiers logic for fault diagnosis.

	Classes					
	C_0	C_1	C_2	C_3	C_4	C_5
Classifier for C_0	**1**	-1	-1	-1	-1	-1
Classifier for C_1	-1	**1**	-1	-1	-1	-1
Classifier for C_2	-1	-1	**1**	-1	-1	-1
Classifier for C_3	-1	-1	-1	**1**	-1	-1
Classifier for C_4	-1	-1	-1	-1	**1**	-1
Classifier for C_5	-1	-1	-1	-1	-1	**1**

TABLE 8. NCA and average CT using group 5 with different one-class classifiers.

		Methods			
Class	Phase	IKPCA$_{CR}$-RF	IKPCA$_{UL}$-RF	IRKPCA$_{CR}$-RF	IRKPCA$_{UL}$-RF
C_0	Training	.99	.99	.99	1
	Testing	.99	.99	.99	.99
C_1	Training	.99	1	.99	1
	Testing	1	.99	.99	.99
C_2	Training	1	1	1	.99
	Testing	1	1	1	.99
C_3	Training	1	1	1	.99
	Testing	.99	.99	1	.99
C_4	Training	.99	1	.99	1
	Testing	.99	1	1	.99
C_5	Training	1	1	.99	1
	Testing	1	.99	.99	1
NCA	Training	**.99**	**.99**	**.99**	**.99**
	Testing	**.99**	**.99**	**.99**	**.99**
Average CT (s)		207.46	216.82	83.15	92.83

IKPCA-RF classifiers. From this table, one can notice that the proposed methods based on the data reduction framework achieve the best compromise between classification metrics and computation time.

V. CONCLUSION

In this paper, two interval reduced kernel PCA (IRKPCA)-based Random Forrest (RF) algorithms (IRKPCA-RF) were proposed for fault detection and diagnosis (FDD) of grid-tied PV systems. Firstly, two IRKPCA models with a data-reduction framework using the Euclidean distance (ED) metric were developed. The first proposed IRKPCA$_{CR}$ model used a kernel PCA model-based reduced interval centers and ranges of intervals. The second IRKPCA$_{UL}$ technique consisted of applying a kernel PCA on a reduced interval upper and lower bounds of intervals. The idea behind the developed IRKPCA models was to extract and select the most relevant features from data with the minimum computation time. Secondly, the final features were introduced as inputs into the RF algorithm for fault classification purposes. The feasibility and effectiveness of the proposed IRKPCA-RF techniques were evaluated under normal and faulty operating conditions. Based on the experimental results, the developed techniques were powerfully effective in terms of computation time and diagnosis metrics.

In future work, approaches to improve further the performance of the IRKPCA-RF method in fault diagnosis will be explored. In the current work, the classical RF algorithm was utilized to model the dynamic nature in both offline training and online update phase using the newly arrived

measurements. Instead, using online extensions of RF model in the first place, such as online incremental RF (presented in [34]) or Mondrian forests (described in [35], [36]), may reduce the training and update time. Moreover, there is a number of threats that may have an impact on the results of this study. The fault diagnosis approaches proposed in this study were built by using default parameters. Thus, it has not been investigated how these approaches are affected by the parameters variation. In consequence, other approaches might be better in diagnosing the faults. The parameters to be optimized include mainly the number of trees in the forest and the maximum depth of each tree. This task reduces the requirements for research experience during parameter tuning and avoids the need for tedious manual tuning. Moreover, the optimized RF model can achieve improved diagnosis performance. The optimization tools including particle swarm optimization (PSO), Genetic Algorithm (GA) and Multi-Objective Optimization (MOO) will be employed to optimize the RF parameters. The main challenge is to provide a good trade-off between low computation time and high classification metrics.

REFERENCES

[1] M. Sabbaghpur Arani and M. A. Hejazi, "The comprehensive study of electrical faults in PV arrays," *J. Electr. Comput. Eng.*, vol. 2016, Dec. 2016, Art. no. 8712960.
[2] R. Fezai, M. Mansouri, M. Trabelsi, M. Hajji, H. Nounou, and M. Nounou, "Online reduced kernel GLRT technique for improved fault detection in photovoltaic systems," *Energy*, vol. 179, pp. 1133–1154, Jul. 2019.
[3] H. Momeni, N. Sadoogi, M. Farrokhifar, and H. F. Gharibeh, "Fault diagnosis in photovoltaic arrays using GBSSL method and proposing a fault correction system," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5300–5308, Aug. 2020.
[4] D. S. Pillai, F. Blaabjerg, and N. Rajasekar, "A comparative evaluation of advanced fault detection approaches for PV systems," *IEEE J. Photovolt.*, vol. 9, no. 2, pp. 513–527, Mar. 2019.
[5] Z. Yi and A. H. Etemadi, "Fault detection for photovoltaic systems based on multi-resolution signal decomposition and fuzzy inference systems," *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1274–1283, May 2017.
[6] F. Aziz, A. Ul Haq, S. Ahmad, Y. Mahmoud, M. Jalal, and U. Ali, "A novel convolutional neural network-based approach for fault classification in photovoltaic arrays," *IEEE Access*, vol. 8, pp. 41889–41904, 2020.
[7] S. Leva, M. Mussetta, and E. Ogliari, "PV module fault diagnosis based on microconverters and day-ahead forecast," *IEEE Trans. Ind. Electron.*, vol. 66, no. 5, pp. 3928–3937, May 2019.
[8] S. Motahar and H. Bagheri-Esfeh, "Artificial neural network based assessment of grid-connected photovoltaic thermal systems in heating dominated regions of Iran," *Sustain. Energy Technol. Assessments*, vol. 39, Jun. 2020, Art. no. 100694.
[9] M. Van, D. T. Hoang, and H. J. Kang, "Bearing fault diagnosis using a particle swarm optimization-least squares wavelet support vector machine classifier," *Sensors*, vol. 20, no. 12, p. 3422, Jun. 2020.
[10] S. Munikoti, L. Das, B. Natarajan, and B. Srinivasan, "Data-driven approaches for diagnosis of incipient faults in DC motors," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5299–5308, Sep. 2019.
[11] H. Cui, Y. Wang, G. Li, Y. Huang, and Y. Hu, "Exploration of cervical myelopathy location from somatosensory evoked potentials using random forests classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 11, pp. 2254–2262, Nov. 2019.
[12] H. Mekki, A. Mellit, and H. Salhi, "Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules," *Simul. Model. Pract. Theory*, vol. 67, pp. 1–13, Sep. 2016.
[13] X. Li, Q. Yang, J. Wang, Z. Chen, and W. Yan, "Intelligent fault pattern recognition of aerial photovoltaic module images based on deep learning technique," in *Proc. 9th Int. Multi-Conf. Complex., Inform. Cybern.*, 2018, pp. 1–6.
[14] W. Yuan, T. Wang, D. Diallo, and C. Delpha, "A fault diagnosis strategy based on multilevel classification for a cascaded photovoltaic grid-connected inverter," *Electronics*, vol. 9, no. 3, p. 429, Mar. 2020.
[15] R. K. Patel and V. K. Giri, "Feature selection and classification of mechanical fault of an induction motor using random forest classifier," *Perspect. Sci.*, vol. 8, pp. 334–337, Sep. 2016.
[16] Z. Chen, F. Han, L. Wu, J. Yu, S. Cheng, P. Lin, and H. Chen, "Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents," *Energy Convers. Manage.*, vol. 178, pp. 250–264, Dec. 2018.
[17] S. Ye, J. Jiang, Z. Zhou, C. Liu, and Y. Liu, "A fast and intelligent open-circuit fault diagnosis method for a five-level NNPP converter based on an improved feature extraction and selection model," *IEEE Access*, vol. 8, pp. 52852–52862, 2020.
[18] B. Luo, H. Wang, H. Liu, B. Li, and F. Peng, "Early fault detection of machine tools based on deep learning and dynamic identification," *IEEE Trans. Ind. Electron.*, vol. 66, no. 1, pp. 509–518, Jan. 2019.
[19] M. Z. Sheriff, C. Botre, M. Mansouri, H. Nounou, M. Nounou, and M. N. Karim, "Process monitoring using data-based fault detection techniques: Comparative studies," in *Fault Diagnosis and Detection*. Rijeka, Croatia: InTech, 2017, pp. 237–261.
[20] S. Z. Li, X. Lu, X. Hou, X. Peng, and Q. Cheng, "Learning multiview face subspaces and facial pose estimation using independent component analysis," *IEEE Trans. Image Process.*, vol. 14, no. 6, pp. 705–712, Jun. 2005.
[21] A. Kembhavi, D. Harwood, and L. S. Davis, "Vehicle detection using partial least squares," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1250–1265, Jun. 2011.
[22] M. Hajji, M. F. Harkat, A. Kouadri, K. Abodayeh, M. Mansouri, H. Nounou, and M. Nounou, "Multivariate feature extraction based supervised machine learning for fault detection and diagnosis in photovoltaic systems," *Eur. J. Control*, Apr. 2020.
[23] C. Kim and D. Klabjan, "A simple and fast algorithm for L1-norm kernel PCA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1842–1855, Aug. 2020.
[24] X. Deng, X. Tian, S. Chen, and C. J. Harris, "Deep principal component analysis based on layerwise feature extraction and its application to nonlinear process monitoring," *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 6, pp. 2526–2540, Nov. 2019.
[25] 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.
[26] T. Ait-Izem, M.-F. Harkat, M. Djeghaba, and F. Kratz, "Sensor fault detection based on principal component analysis for interval-valued data," *Qual. Eng.*, vol. 30, no. 4, pp. 635–647, Oct. 2018.
[27] M.-F. Harkat, M. Mansouri, M. Nounou, and H. Nounou, "Fault detection of uncertain nonlinear process using interval-valued data-driven approach," *Chem. Eng. Sci.*, vol. 205, pp. 36–45, Sep. 2019.
[28] M. Mansouri, R. Fezai, M. Trabelsi, M. Hajji, M.-F. Harkat, H. Nounou, M. N. Nounou, and K. Bouzrara, "A novel fault diagnosis of uncertain systems based on interval Gaussian process regression: Application to wind energy conversion systems," *IEEE Access*, vol. 8, pp. 219672–219679, 2020.
[29] J.-M. Lee, C. Yoo, and I.-B. Lee, "Fault detection of batch processes using multiway kernel principal component analysis," *Comput. Chem. Eng.*, vol. 28, no. 9, pp. 1837–1847, Aug. 2004.
[30] A. Maulud, D. Wang, and J. A. Romagnoli, "A multi-scale orthogonal nonlinear strategy for multi-variate statistical process monitoring," *J. Process Control*, vol. 16, no. 7, pp. 671–683, Aug. 2006.
[31] K. Dhibi, R. Fezai, M. Mansouri, M. Trabelsi, A. Kouadri, K. Bouzara, H. Nounou, and M. Nounou, "Reduced kernel random forest technique for fault detection and classification in grid-tied PV systems," *IEEE J. Photovolt.*, vol. 10, no. 6, pp. 1864–1871, Nov. 2020.
[32] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[33] S. Gharsellaoui, M. Mansouri, M. Trabelsi, M.-F. Harkat, S. S. Refaat, and H. Messaoud, "Interval-valued features based machine learning technique for fault detection and diagnosis of uncertain HVAC systems," *IEEE Access*, vol. 8, pp. 171892–171902, 2020.
[34] M.-A. Kaufhold, M. Bayer, and C. Reuter, "Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102132.

[35] B. Lakshminarayanan, D. M. Roy, and Y. W. Teh, "Mondrian forests: Efficient online random forests," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3140–3148.
[36] Y. Zhong, H. Yang, Y. Zhang, and P. Li, "Online random forests regression with memories," *Knowl.-Based Syst.*, vols. 201–202, Aug. 2020, Art. no. 106058.

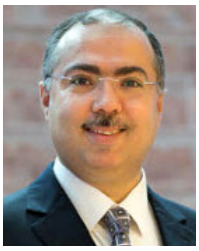
KHALED DHIBI is currently pursuing the Ph.D. degree with the Faculty of Sciences of Monastir (FSM), Monastir, Tunisia. His work focuses on the implementation of data-driven techniques for fault detection and diagnosis of industrial processes.

RADHIA FEZAI is currently an Assistant Research Scientist with the Electrical Engineering Program, Texas A&M University at Qatar. Her work focuses on the use of applied mathematics and statistics concepts to develop statistical data and model-driven techniques and algorithms for modeling, fault detection, and diagnosis with the aim of improving the operation of industrial systems.

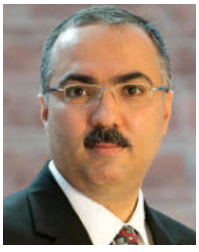
MAJDI MANSOURI (Senior Member, IEEE) received the degree in electrical engineering from SUPCOM, Tunis, Tunisia, in 2006, the M.Sc. degree in electrical engineering from ENSEIRB, Bordeaux, France, in 2008, the Ph.D. degree in electrical engineering from UTT Troyes, France, in 2011, and the H.D.R. (Accreditation 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 150 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.

MOHAMED TRABELSI (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from INSAT, 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 has been holding different Research positions at Qatar University and Texas A&M University at Qatar. Since September 2018, he has been an Associate Professor with the Kuwait College of Science and Technology. He has published more than 100 journal articles and conference papers. He is the author of two books and two book chapters. His research interests include systems control with applications arising in the contexts of power electronics, energy conversion, renewable energies integration, and smart grids.

KAIS BOUZRARA is currently a Professor of electrical engineering with the Laboratory of Automatic Signal and Image Processing, National Engineering School of Monastir, Monastir, Tunisia. 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 emphasis on process modeling, monitoring, and estimation.



HAZEM NOUNOU (Senior Member, IEEE) is currently a Professor of electrical and computer engineering with Texas A&M University at Qatar. He has more than 19 years of academic and industrial experience. He has significant experience in research on control systems, database control, system identification and estimation, fault detection, and system biology. He has been awarded several NPRP research projects in these areas. He has successfully served as the lead PI and a PI on five QNRF projects, some of which were in collaboration with other PIs in this proposal. He has published more than 200 refereed journal articles and conference papers and book chapters. He has served as an Associate Editor and on the technical committees of several international journals and conferences.



MOHAMED NOUNOU (Senior Member, IEEE) is currently a Professor of chemical engineering with Texas A&M University at Qatar (TAMU). He has more than 19 years of combined academic and industrial experience. He has successfully worked as the lead PI and a PI on several QNRF projects (six NPRP projects and three UREP projects). He has published more than 200 refereed journals and conference publications and book chapters. His research interests include systems engineering and control, with emphasis on process modeling, monitoring, and estimation. He is a Senior Member of the American Institute of Chemical Engineers (AIChE).