IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 24, 2021, accepted November 12, 2021, date of publication November 16, 2021,
date of current version November 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3128749

An Enhanced Ensemble Learning-Based Fault Detection and Diagnosis for Grid-Connected PV Systems

KHALED DHIBI[1], MAJDI MANSOURI[2], (Senior Member, IEEE), KAIS BOUZRARA[1], HAZEM NOUNOU[2], (Senior Member, IEEE), AND MOHAMED NOUNOU[3], (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, Qatar
[3]Chemical Engineering Program, Texas A&M University at Qatar, Doha, Qatar

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

This work was supported by the Qatar National Library through the Qatar National Research Fund (QNRF) Research Grant.

ABSTRACT The main objective of this article is to develop an enhanced ensemble learning (EL) based intelligent fault detection and diagnosis (FDD) paradigms that aim to ensure the high-performance operation of Grid-Connected Photovoltaic (PV) systems. The developed EL based techniques consist in combining multiple learning models instead of using a single learning model. To do that, three EL-based FDD techniques are proposed. First, an EL technique that merges the benefits of Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Decision Tree (DT) is presented. The developed method contributes to the reduction of the overall diagnosis error and has the ability to combine various models. However, classical EL models ignore the time-dependence of PV measurements. In addition, the PV system data are frequently time-correlated. Therefore, kernel PCA (KPCA)-based EL and reduced KPCA (RKPCA)-based EL techniques are developed to take into consideration the dynamic and multivariate natures of the PV measurements. The two proposed KPCA -based EL and RKPCA-based EL techniques are addressed so that the features extraction and selection phases are performed using the KPCA and RKPCA models and the sensitive and significant characteristics are transmitted to the EL model for classification purposes. The presented results prove that the proposed EL based methods offer enhanced diagnosis performances when applied to PV systems.

INDEX TERMS Machine learning, ensemble learning, kernel principal component analysis (KPCA), fault detection, fault diagnosis, fault classification, grid-connected PV (GCPV).

I. LIST OF ABBREVIATIONS AND ACRONYMS

EL	Ensemle Learning.
ML	Machine Learning.
FDD	Fault Detection and Diagnosis.
PCA	Principal Component Analysis.
KPCA	Kernel PCA.
RKPCA	Reduced KPCA.
KPCs	Kernel Principal Components.
ℓ	Number of retained KPCs.
CPV	Cumulative Percentage of Variance.
KNN	K-Nearest Neighbors.
SVM	Support Vector Vachines.
DT	Decision Tree.
ED	Euclidean Distance.
PV	Photovoltaic.
GCPV	Grid-Connected PV.
CT	Computation Time.
FAR	False Alarm Rate.
MDR	Missed Detection Rate.
NN	Neural Network.
RNN	Recurrent NN.
FFNN	Feed-Foward NN.
MNN	Multiple Layers NN.
CFNN	Cascade Foward NN.
GRNN	Generalized Regression NN.
PNN	Probabilistic Neural Network NN.
CM	Confusion Matrix.

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

II. INTRODUCTION

Grid-Connected Photovoltaic (GCPV) systems have been receiving an increased interest during the last decade [1]–[3]. However, the operation of these systems is generally accompanied by different types of failures due to harsh environmental conditions or internal malfunctions. These failures (i.e. open-circuit/short-circuit faults, shading effects, inverter fault, grid-connection fault [4], [5]) might cause serious physical damage, present a risk of fire, and affect the efficiency of the solar modules and electrical power generation [6]. Therefore, the implementation of fault detection and diagnosis (FDD) techniques is becoming mandatory in ensuring safe and uninterruptible operation of Grid-Connected PV systems with low maintenance cost [7]–[9]. In recent years, several computational FDD techniques based on machine learning (ML) techniques have been successfully applied for PV systems [10], [11]. The ML techniques include Support Vector Machines (SVM) [12], Naive Bayes (NB) [13], K-Nearest Neighbors (KNN) [14], Decision Tree (DT) [15], Random Forest (RF) [16], discriminant analysis (DA) [17] and artificial neural networks (ANN) [18]–[21]. These techniques were reported to have performed better in fault detection and diagnosis of industrial systems than conventional techniques like linear regression [22]. SVM has been first introduced by Vapnik [12]. The main idea of SVM is to map the training data from the input space into a higher-dimensional feature space via a mapping function and then apply linear SVM in this space. SVM classifier seeks to find an optimal separating hyper-plane as the decision plane by maximizing the margin between two classes. SVM works relatively well when there is a clear margin of separation between classes and it is memory efficient [23]. K-Nearest-Neighbor (KNN) is a widely-used parametric classifier because of its simplicity and effectiveness. KNN algorithm does not need any training before making predictions, new data can be added seamlessly, which will not affect the accuracy of the algorithm [23]. A DT is constructed by the division of the dataset into smaller subsets until no further splitting can be implemented (unless a limit for splitting is set). DT method can handle both continuous and categorical variables and it can automatically handle missing values [24]. In the last two decades, several fault diagnosis techniques based on ensemble learning techniques are proposed. The ensemble learning (EL) approach has gained significant attention and it is becoming more and more popular [25]–[27]. The main idea behind EL algorithms is to improve machine learning results by correctly combining several models into one predictive model in order to become more accurate and robust [25], [28]. The widely known ensemble learning techniques include bagging, boosting, random subspace, and stacking. EL seeks to decrease variance (bagging), bias (boosting) and improve predictions (stacking) [29]. Recently, EL techniques have been successfully used in monitoring processes in several sectors, e.g., chemical industry [30], pharmacology [31], energy [32], finance [29], agriculture [33] and many others. The need for developed better ensemble learning models for fault diagnosis of PV systems has become more and more important in the research area. Therefore, several EL techniques have been explored in recent years. Generally, bagging and boosting are the most common methods used in the literature [34]. Despite the proven performances of numerous works that use ensemble learning techniques, most of these methods use only a specific type of classifier. Additionally, another main drawback of the existing FDD techniques based on ensemble learning methods is the direct use of the raw information from the process data. To overcome this challenge, several FDD techniques based on features extraction and selection step using a single classifier are proposed in the literature [35], [36].

Based on the above-mentioned discussion, This paper exclusively focuses on the FDD problem for Grid-Connected PV systems. The main contributions are threefold: In the first stage, ensemble learning includes SVM, DT and KNN will be proposed to distinguish between the different PV system operating modes using the extracted raw data. In the second stage, in order to overcome the limitations of the proposed ensemble learning technique due to the direct use of the raw data, an intelligent framework based on features extraction and selection step using the KPCA technique will be developed. In effect, combining KPCA and ensemble learning models could improve the performance of FDD and more specifically the decision-making accuracy. In the final stage, to improve the use of kernel PCA in terms of computation time and storage cost, a reduced extension will be proposed. To summarize, an enhanced machine learning technique for FDD will be developed. The novel technique optimally merges ensemble learning methods and multivariate statistical analysis (KPCA and reduced KPCA) to achieve an overall improved accuracy.

The remainder of this paper is outlined as follows: Section I presents the list of abbreviations and acronyms. Section II introduces the paper. Section III presents a review of related works. A brief overview of EL and some ML techniques is given in Section IV. Section V presents the proposed techniques are presented in section V. Section VI describes the validation of the proposed techniques. At last, some conclusions and future research directions are presented in section VII.

III. RELATED WORK

Fault diagnosis in GCPV systems become more and more important to ensure optimal energy harvesting, low maintenance cost and reliable power production. Several FDD using powerful machine learning (ML) and ensemble ML techniques in PV systems have been proposed in the literature to improve their reliability and performance. In [37], a technique based on SVM approach for the classification of islanding and grid fault events in LV distribution grid is proposed. For instance, an artificial neural network (ANN) was developed in [38], [39]. The main idea of this proposal is to detect partial shading faults. Besides, the authors used a three-layer feed-forward ANN to detect short-circuit faults in PV arrays [38]. In [40], the authors developed a fast FDD

based on the generalized local likelihood ratio test. In other studies, a diagnostic method based on two convolutional neural networks (CNN) has been proposed for fault classification in PV array [41]. In [42], the authors presented a fault diagnosis method for PV arrays using an extreme gradient boosting (XGboost) classifier. The developed technique is based on the string current, array voltage, temperature and irradiance measurements. Moreover, many works have attempted to use K-Nearest Neighbor (KNN) technique for fault classification purposes [43], [44]. The KNN classifier is a non-parametric technique that does not rely on the construction of a model during the training phase, and whose classification rule is based on a given similarity function between the training and the testing samples [44]. In [10], a new design of parity relation-based residual generator for fault detection method is proposed. The authors employed an iterative procedure that guarantees minimal regression error in the search for the optimal parameters to deal with linear and nonlinear systems. In [11], a fault diagnosis technique based on Semi-supervised Ladder Network With String Voltage and Current Measures is developed. A data-driven-based FDD approach was introduced in [45]. The proposed technique consists of monitoring the nonlinear processes based on the available sensing measurements only using a locally weighted projection regression (LWPR) for the partition of the input space and modified principal component analysis (MPCA) for fault detection. In [25], the authors propose ensemble machine learning (EML) algorithms to detect a series DC arc fault in a modern electrical system using local DC distribution. In [28], an ensemble learning technique that incorporates support vector machine (SVM), k-nearest neighbor (KNN), logistic regression (LR), decision tree (DT), and random forest (RF) is developed in order to diagnose faults in refrigeration systems. Bagging is one of the most well-known and successful ensemble learning techniques. Recently, many ensemble learning (ML) based Bagging techniques were developed [46], [47]. In [46], an enhanced bagging (eBagging) method is presented. The main idea behind this proposal is to use a new mechanism (error-based bootstrapping) instead of traditional random bootstrap technique when constructing training sets. A bagging based multi-objective differential evolution algorithm (MODE) with multiple sub-populations (BagMP-MODE) was proposed in [47]. This technique consists in incorporating the idea of bagging into the evolution process of MODE. For instance, data-driven approaches like principal component analysis (PCA) have been widely used for feature extraction and selection [36], [48]. In [36], an improved FDD technique was proposed by using the PCA technique for multivariate features extraction and selection, and single machine learning classifiers for faults classification. However, the PCA-based diagnosis technique has been only developed for linear systems while popular complex systems exhibit strong nonlinear correlations between their variables. Recently, Kernel PCA methods (KPCA) have been proposed to address the nonlinear relationships between process variables [49], [50]. The basic idea of the KPCA technique consists of (i) mapping the input data onto the feature space via a nonlinear kernel function, and (ii) perform PCA into a feature space. Although kernel PCA can extract nonlinear features in a high-dimensional space, it increases the space and time complexity compared to the PCA [51].

IV. PRELIMINARIES

In this section, we present the details of machine learning algorithms and ensemble learning techniques used in this work.

A. ENSEMBLE TECHNIQUES

The main idea behind multiple learning is to combine several models into a meta-algorithm in order to improve the classification results of any FDD techniques [28]. The ensemble learning methodology is based on three phases. The first one, member generation phase, consists of manipulating the training sets and building models with different learning algorithms. The second one, member selection phase, consists of selecting just models that are suitable for the prediction task. The third one, member combination phase, consists of combining the outputs of multiple classifiers into a final prediction [52], [53]. Besides, there are three steps to contribute to the task which require multiple classifiers. The existing steps are i) combining classifiers by deciding using different opinions ii) cooperating classifiers using one or more opinions iii) selecting classifiers by giving more importance to one or more classifiers according to various criteria like basic ensemble techniques. To combine the outputs of multiple classifiers into a final and more effective prediction, we use different basic ensemble techniques like average, weighted average, majority voting, and weighted majority voting [54]. There are three advanced EL techniques to combine machine learning classifiers which are bagging, boosting, and Random Subspace [55], [56]. Next, we present a brief discussion of the advanced combination techniques for ensemble learning.

1) BAGGING

The basic idea behind the bagging method is to combine bootstrapping and aggregation (decision trees) to get a generalized result. Bagging technique is mainly applied in classification and regression. It reduces variance to a large extent by increasing the accuracy of models through decision trees in order to increases accuracy which is a challenge to many predictive models [57], [58]. Bootstrapping is a sampling technique with replacement that gives the selection procedure the particularity of being random. Aggregation in bagging makes predictions accurate taking into account all possible outcomes. Thus, aggregation is performed to incorporate all possible outcomes of the prediction and randomize the outcome. The main advantages of bagging are the elimination of any variance and the reduction of model overfitting since it creates several classifiers with fixed bias and combines their outputs by averaging. This technique is powerful when the characteristics of the data have high variance and low bias. The main disadvantage of bagging is the expensive

calculation which can lead to more bias in the models when the proper bagging procedure is ignored. [25]. drawback of bagging is its random selection

2) BOOSTING

Boosting is a meta-algorithm that learns from precedent predictor mistakes to perform better predictions in the future [57], [58]. The main idea behind boosting technique is that each of the single models improves the performance of the ensemble. By boosting, every successive model depends on the preceding model where the errors of the previous model are corrected by each successive model in order to decrease the model's bias and to form one strong learner. Hence, the technique combines several weak learners to form one strong learner and so improves the predictability of models. Boosting takes many forms, including Adaptive Boosting (AdaBoost), gradient boosting, and Extreme Gradient Boosting (XGBoost). The boosting method is more reliable when the characteristics of the data have high bias and low variance [25].

3) RANDOM SUBSPACE

The random subspace method is an ensemble learning method that has a role to reduce the correlation between estimators in an ensemble by training them in feature space as a random sample instead of the entire feature set. Random Subspace can be presented in three steps: the first step is to select N subsets containing M features selected at random from F features, the second step is to train N weak learners using each random subset and the last step is to perform a prediction by majority vote.

B. CLASSIFICATION ALGORITHMS

1) SUPPORT VECTOR MACHINES (SVM)

Support vector machine (SVM) is one of the most powerful classification algorithms and it has been widely applied for fault diagnosis [59], [60].

The main idea of SVM is to map the training data from the input space into a higher-dimensional feature space via a mapping function and then apply linear SVM in this space. SVM classifier seeks to find an optimal separating hyper-plane as the decision plane by maximizing the margin between two classes. Consider a given training set of N samples $\{x_i, y_i\}_{i=1}^{N}$, with input data $x_i \in \mathbb{R}^m$ and output $y_i \in \{+1, -1\}$. The hyper-plane of the SVM is defined by

$$f(x) = w^T \phi(x) + b = 0 \tag{1}$$

where w is a weight vector and b denotes the bias vector.

The parameters w and b can be determined by solving a constrained optimization problem as,

$$\begin{cases} min \dfrac{1}{2}|w|^2 \\ y_i(w^T \phi(x_i + b)) \geq 1 \end{cases} \tag{2}$$

By introducing Lagrangian multipliers, can be rewritten as,

$$L(w, b, \alpha) = \frac{1}{2}|w|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i \left(w^T \phi(x_i) + b \right) \right) - 1 \Big] \tag{3}$$

where α_i denote the Lagrange coefficients.

As a result, the decision function can be obtained as follows:

$$f(x) = sign \left(\sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + b \right) \tag{4}$$

where x denotes the input vector to be classified.

2) K-NEAREST NEIGHBORS (KNN)

K-Nearest-Neighbor (KNN) is among the most models used for classification thanks to her performances and simplicity [61]. The main idea behind the KNN technique is to find the nearest neighbors for a given data based on some distance metric of interest [62], [63]. kNN is a nonparametric method used to identify in which class, already known, unknown data belong to it. To determine the KNN class, the Euclidean distance is used as follows,

Consider that the elements of known class are $x = [x_1\ x_2 \ldots x_k]$ and those of the data to be classified are $y = [y_1\ y_2 \ldots y_k]$. To define the distance between two samples, the Euclidean distance is used and it is defined as,

$$d(x, y) = \sqrt{(\sum_{i=1}^{k} (x_i - y_i)^2} \tag{5}$$

Then, a class is assigned at which the distance defined as in Eq.5 is minimal.

3) MODEL-TREE

Tree-based ML techniques are among the mostly used nonlinear models in many applications, where the Random Forest (RF) and Decision Tree (DT) are the most popular ones (they can be more accurate than neural networks) [64]. The goal of decision tree (DT) is to create a model that predicts the value of a target variable. DT model use two nodes, which are the decision node and leaf node [65]. Decision nodes have multiple branches and they are used to make any decision, while leaf nodes are the output of these decisions. The main idea behind the RF algorithm is to use a combination of randomized trees and make the prediction by a majority vote between all the produced decision trees [66]

V. RKPCA FOR FEATURES EXTRACTION AND SELECTION

KPCA method consists to map data into a feature space via a nonlinear mapping and then to calculate the kernel principal components (KPCs) [50]. Moreover, the advantage of using nonlinear kernel functions and integral operators allows KPCA to determine effectively the KPCs in the feature space. However, KPCA is not very effective when a large number of variables are recorded. Therefore, the computational times increases and the storage cost become important.

To overcome this challenging problem, we propose to use a model relationships between variables via a data-reduction framework.

Let us consider a data matrix $X = \left[x_1\ x_2\ \ldots\ x_N\right]^T \in \mathbb{R}^{N\times m}$, where m corresponds to the number of process variables and N represents the number of samples, The basic idea behind the proposed reduced KPCA (RKPCA) method is to extract a reduced number of observations (samples) between the m measurement variables such that the preserved observations have more relevant data information and by turn it is used as a new data matrix. To extract the most pertinent samples from data, Euclidean distance metric will be used. The Euclidean distance q_{ij} between the rows X_i and X_j of the data matrix X is given by

$$q_{ij} = \sqrt{\sum_{k=1}^{m}\left(x_{ik} - x_{jk}\right)^2} \tag{6}$$

Then, dissimilarity matrix Q which contains the measurement of dissimilarity between all pairs of the observations is presented as follows,

$$Q = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1N} \\ q_{21} & q_{22} & \cdots & q_{2N} \\ . & . & . & . \\ q_{N1} & q_{N2} & \cdots & q_{NN} \end{bmatrix} \tag{7}$$

Thus, the new reduced data matrix X' is defined as

$$X' = \left[x'(1) \quad x'(2) \quad . \quad . \quad . \quad x'(N')\right]^T \in \mathbf{R}^{N'\times m} \tag{8}$$

where N' is the size of the reduced training data matrix. the basic idea behind the RKPCA method is to apply the KPCA model in the reduced data matrix.

A. FEATURE EXTRACTION USING RKPCA

let consider a reduced data matrix $\left[X'\right]$. The mapped data in the feature space is arranged as $[\mathcal{X}] = \left[\phi([x_1])\ \phi([x_2])\ \ldots\ \phi([x_{N'}])\right]^T \in \mathbb{R}^{N'\times h}$, where $h \gg m$ is the dimension of the feature space. Using a kernel matrix whose elements are $k(\phi([x_i]), \phi(\left[x_j\right])), i, j = 1 \ldots N'$, the kernel principal components (KPC_s) can be computed using the following eigenvector expression:

$$\lambda\alpha = [K]\,\alpha \tag{9}$$

where α and λ are the eigenvector and eigenvalue of the kernel matrix K. The kernel matrix K of interval valued data is expressed as:

$$[K] = [\mathcal{X}]\left[\mathcal{X}^T\right] = \begin{bmatrix} k([x_1], ([x_1]) & \ldots & k([x_1], [x_{N'})] \\ . & & . \\ . & \ldots & . \\ . & & . \\ k([x_{N'}], [x_1]) & \ldots & k([x_{N'}], [x_{N'}]) \end{bmatrix} \tag{10}$$

where $k([x])$ is defined as:

$$k([x]) = (k([x_1], [x], \ldots, k([x_{N'}], [x]))^T \tag{11}$$

B. FEATURE SELECTION USING RKPCA

The eigenvector of the kernel matrix is given by is given by [67],

$$v = \lambda^{-1}\left[\mathcal{X}^T\right]\boldsymbol{\alpha} \tag{12}$$

The matrix of the ℓ principal eigenvectors of $[K]$ representing the largest eigenvalues $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_\ell\}$ is given by,

$$P = \left[{\lambda_1}^{-1}\left[\mathcal{X}^T\right]\boldsymbol{\alpha}_1, \ldots, {\lambda_\ell}^{-1}\left[\mathcal{X}^T\right]\boldsymbol{\alpha}_\ell\right] \tag{13}$$

where $\boldsymbol{\alpha}$ is the eigenvector of the matrix $[K]$ and λ is its corresponding eigenvalue. Then, the kernel principal components are defined as, [67],

$$t = \Lambda^{-1/2}P^T k([x]) \tag{14}$$

Additional to the ℓ first KPCs, squared prediction error (SPE) statistic, Hotelling's T^2 statistic and combined index φ, are used to choose the final effective features [68]. The statistical features are determined as follows:

$$T^2 = k([x])^T P\,\Lambda^{-1}P^T k([x]) \tag{15}$$

$$SPE = k([x], [x]) - k^T([x])Ck([x]) \tag{16}$$

$$\varphi = \frac{SPE}{\tau_\alpha^{SPE}} + \frac{T_{CR}^2}{\tau_\alpha^{T^2}} \tag{17}$$

$\tau_\alpha^{T^2}$ and τ_α^{SPE} represent thresholds of T^2 and SPE at the confidence level α, respectively.

$$\tau_\alpha^{T^2} = \frac{\ell(N_r-1)(N_r+1)}{N_r(N_r-\ell)}F_\alpha(\ell, N_r-\ell) \tag{18}$$

where $F_\alpha(\ell, N_r - \ell)$ an F-distribution with ℓ and $N_r - \ell$ degrees of freedom.

$$\tau_\alpha^{SPE} = g_{SPE}\chi^2_{h_{SPE},\alpha} \tag{19}$$

where $g_{SPE} = \frac{b}{2a}$ and $h_{SPE} = \frac{2a^2}{b}$, with a and b are the mean and variance of the SPE index, respectively.

The mean m, variance D^2, kurtosis K and skewness S of the first ℓ retained KPCs $t = [t_1, \ldots, t_N]^T$, where $t_k = [t_{k1}, \ldots, t_{k\ell}]$; $k = 1, \ldots, N$ are computed by [67],

$$m_j = \frac{1}{\ell}\sum_{i=1}^{\ell} t_{ji} \tag{20}$$

$$D_j^2 = \frac{1}{\ell}\sum_{i=1}^{\ell}(t_{ji} - m_j)^2 \tag{21}$$

$$K_j = \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\frac{t_{ji} - m_j}{D_j^2}\right)^4 \tag{22}$$

$$S_j = \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\frac{t_{ji} - m_j}{D_j^2}\right)^3 \tag{23}$$

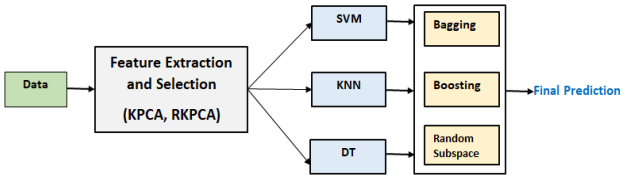


FIGURE 1. Schematic diagram of the KPCA-EL and RKPCA-EL algorithms.

1) PROPOSED METHODOLOGIES

The main objective of this paper is to develop a hybrid approach for fault diagnosis of grid-connected PV (GCPV) systems. The proposal methodology combines an ensemble learning technique and an improved data-driven method with dataset size reduction. Ensemble learning helps improve machine learning results by combining several models and it has already proven to be a powerful technique for creating classifiers. For this reason, we used three base learning classification techniques include Support Vector Machines (SVM), K-Nearest Neighbor (KNN), and Decision Tree (DT), with ensemble methods like Boosting, Bagging, and Random subspace for classification the dataset. However, the direct use of raw data by the proposed method limits their effectiveness. To improve the use of the proposed EL technique, we apply the KPCA method for feature extraction and selection in order to extract the most relevant and sensitive features from data. This, in turn, plays a pivotal role in improving the fault diagnosis results using the proposed ensemble learning (EL) technique. Although KPCA can extract nonlinear features in a high-dimensional space, it increases the space and time complexity. Therefore, to enhance the use of KPCA for feature extraction and selection in terms of computation time and storage cost a reduced KPCA (RKPCA) will be proposed. The main idea behind RKPCA is: i) select only the effective samples from raw data using Euclidean distance metric, ii) use reduced data to build KPCA model. Hence, the proposed technique for fault diagnosis achieves the best tread-off in terms of computation time and diagnosis metrics.

In the classification phase, once the global features are extracted and selected using KPCA or RKPCA techniques, it is applied as input data for the ensemble learning (EL) classifier. Thus, some arbitrary groups of the significant selected features are used to train the EL model. Finally, a comparison between the ensemble learning output results using the different selected arbitrary groups is made to make effective decisions. The main steps of the KPCA-EL and RKPCA-EL techniques are illustrated in Algorithm 1 and schematic diagram 1.

Time complexity analysis with Big-O notation is one of the most important concepts in learning in order to construct efficient code. In Big O analysis, we only consider the most dominant term, as the other terms and constants become insignificant asymptotically. Kernel PCA performs an eigen decomposition on the kernel expansion of the data, an $m \times N$ matrix (N is the total number of samples). To reduce the attendant $O(N^2)$ space time complexity, we propose reduced KPCA method with $O(N'^2)$ (N' is the reduced number of samples). The classical KPCA suffers from some limitations which restrict its practical applications when the number of samples is too large. In addition, in the training phase, KPCA requires to store and compute the eigenvectors of a $N * N$ kernel matrix, where N is the total number of samples. This computation needs a space complexity of $O(N^2)$ and a time complexity of $O(N^3)$, thus to reduce the attendant the space and time complexities, we propose a reduced KPCA method with $O(N'^2)$ of space complexity and $O(N'^3)$ of time complexity, where N' is the reduced number of samples. The standard SVM has $O(N^3)$ time and $O(N^2)$ space complexity where N is training set size using quadratic programming formulation. For the kNN algorithm, we have a time complexity of $O(N \times m)$, where N is the number of training examples and m is the number of dimensions in the training set. For simplicity, assuming $N \times m$, the complexity of the nearest neighbor search is $O(N)$. Note that the union of the subsets on each level of the tree is the entire training data of size m, and the time complexity for each level is thus $O(m \times N)$. Therefore, the standard decision-tree learning algorithm has a time complexity of $O(m \times N^2)$. The decision tree complexity of a function is the minimum depth of a decision tree that computes this function. When training a decision tree, a split has to be found while a maximum depth d has been reached. This split is finding by looking at each variable (there are N of them) to the different thresholds (there are up to n of them) and the information gain that is completed (evaluation in $O(n)$).

Algorithm 1 RKPCA-EL Algorithm

Input: $N \times m$ data matrix X.

Training data

1. Standardize the training data matrix,
2. Determine the new reduced data matrix X',
3. Compute the kernel matrix K' from X',
4. Extract the features using KPCA method,
5. Select the more effective features,
6. Introduce the selected features as input to the ensemble learning classifier,
7. Classify the faults using EL classifier,
8. Define the classification model,

Testing data

1. Obtain a new observation and standardize it using the mean and the variance calculated from the training data,
2. Calculate the kernel vector $k(x)$,
3. Extract the features using KPCA method,
4. Select the more effective feature,
5. Introduce the selected features as input to the EL classifier,
6. Classify the faults using EL classifier,
7. Compute the prediction model,
8. Determine the fault diagnosis results.

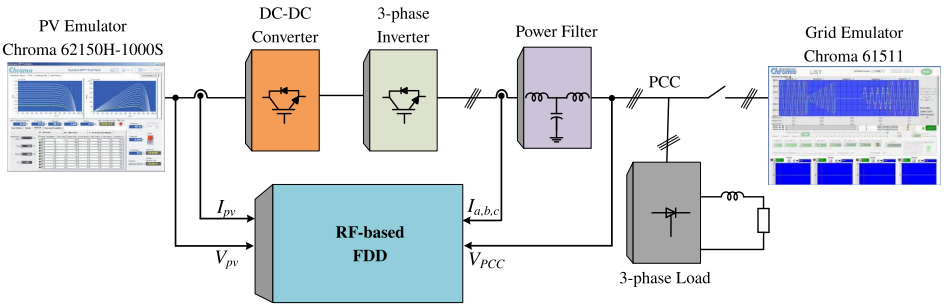


FIGURE 2. Synoptic of the grid-connected PV system under study.

VI. RESULTS AND DISCUSSION

A. PV IMPLEMENTATION AND DATA COLLECTION

Figure 2 shows the synoptic of the PV system under study, where PV and grid emulators are used to emulate the operation (under different operating modes) of PV panels and a 3-phase grid respectively. Table 1 shows the system variables considered in this study, where the measurements are recorded each 5-15s depending on the nature of the faults and their occurrence.

TABLE 1. Measured system variables.

Measures	Symbol	Variable	Description
Three-phase currents	I_a	x_1	The three-phase inverter's output currents
	I_b	x_2	
	I_c	x_3	
PV current	I_{pv}	x_4	The output current of the PV panel emulator
Three-phase voltages	V_a	x_5	The three-phase inverter's output voltages
	V_b	x_6	
	V_c	x_7	
PV voltage	V_{pv}	x_8	The output voltage of the PV panel emulator
Output voltage	V_{out}	x_9	The output voltage of the DC-DC converter

The faults were emulated at different system stages (common coupling point, inverter, sensors, emulated PV arrays, ...) to ensure a comprehensive analysis [36], [69]. A first fault F_1 was emulated by introducing an open-circuit fault on one of the inverter switches at the time (inverter fault). Another AC side fault F_3 was emulated by disconnecting the grid at the common coupling point (islanding referred as grid-connection fault). On the PV side, three types of faults were emulated. The fault F_2 was introduced at the sensor level (output current sensor fault) to emulate the sensor wiring/reading issues. Moreover, using the PV emulator features, a 10-20 % permanent partial shading was introduced to emulate the PV panel fault (F_4) while the connection faults (F_5) were emulated by introducing an open-circuit/short-circuit on PV cells connection.

The healthy operation was assigned to class C0 while the 5 faulty modes (referring to faults F_1-F_5) were assigned to classes C1 to C5 as per Table 2.

TABLE 2. Construction of database for fault diagnosis system.

Class	State	Training Data	Testing Data
C0	Healthy	1501	1501
C1	F_1	1501	1501
C2	F_2	1501	1501
C3	F_3	1501	1501
C4	F_4	1501	1501
C5	F_5	1501	1501

B. PERFORMANCE METRICS

For performance evaluation and comparison, the adopted criteria are: Accuracy (%), which represents the ratio of correctly predicted observation over the total number of observations. Recall (%) which represents the ratio of correctly predicted positive observations to the all observations in the pertinent class. Precision (%) which represents the number of correctly predicted positive observations divided by the number of total predicted positive observations. F_1 Score (%) which represents the weighted average of Precision and Recall, therefore, this score takes into account both false negatives and false positives. Computation time (CT (s)) which defines the time needed to execute the algorithm.

C. PARAMETER SETTING

For the kernel-based methods, the radial basis function (RBF) is applied and the kernel width is equal to the minimum distance between the training data.

The retained kernel component number ℓ is obtained by the 95% cumulative percent variance (CPV) criterion. In this study, 10-fold cross-validation was applied on the whole dataset. The minimum root mean-square error (RMSE) was considered as selection criterion for different ML classifiers. RF and DT were tested with 50 trees. For the NN, MNN, FFNN, CFNN, GRNN, PNN and RNN classifiers, the number of hidden layers chosen is ten and the number of hidden neurons in the hidden layer is equal to 50. The K and C parameters for SVM are selected with the lowest RMSE value and the K value for KNN is equal 3 for classic KNN and it is equal to 1, 3 and 5 for the ensemble learning. The number of bags in RF and Bagged multiple is chosen to be 50.

D. FAULT CLASSIFICATION RESULTS

To demonstrate the performance of the proposed ensemble learning method, the results are compared to bagging, Random Forest(RF), Decision Tree (DT), K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) techniques.

TABLE 3. Global Performances using different methods.

Methods	Global Performances			
	Accuracy	$F_1 score$	Recall	Precision
Ensemble learning (EL)	48.89	48.88	48.90	48.90
Bagging	47.33	47.29	47.32	47.31
RF	47.62	47.61	47.62	47.60
DT	45.04	45.09	45.07	45.05
KNN	47.43	47.36	47.39	47.42
SVM	47.22	47.16	47.17	47.19

Table 3 shows the multi-class classification results where it can be clearly noticed that the results obtained using the proposed method in terms of accuracy (48.89), $F_1 score$ (48.88%), recall (48.90%), and precision(48.90) are higher than the obtained results using other machine learning classifiers. It is easy to conclude that the proposed multiple learning methods enhance the fault classification performance.

Moreover, a new EL-based framework (KPCA) is proposed to further enhance the fault diagnosis performance of the proposed multiple learning technique, where the data set is scaled to zero mean and unit variance. Then, the models are constructed under normal operating conditions. The retained number of KPCs using the CPV criterion is equal to 28. To illustrate the FDD the efficacy of the developed methods, a 10-fold cross-validation approach was used to obtain the classification accuracy. The healthy operation was assigned to class C0 while the 5 faulty modes (F_1-F_5) were assigned to classes C1-C5 (Table 2). To get a good classification performance, it is important to select the best statistical characteristics from the extracted features. Accordingly, five arbitrary groups of features are performed and the best one is selected (Table 4).

TABLE 4. Selected features for fault classification.

Groups	Features Descriptions
Group 1	T^2
Group 2	SPE
Group 3	ϕ
Group 4	Sampled mean, kurtosis, variance and skewness of the ℓ retained KPCs
Group 5	The first ℓ KPCs

First, a fault database is collected and labeled using the emulation data. Then, the labeled data are applied as inputs for the proposed KPCA technique which can be splitted into a multi-class classifier stage (see Table 5). A first comparison is led between five arbitrary groups (see Table 5) using KPCA model in terms of accuracy. The comparison results from Table 5 show that group 5 of features provide a classification accuracy equal 96.96% which present the best one compared to other used groups of features (less than 60%).

TABLE 5. Accuracies using KPCA-EL method.

Method	Extracted Features					
	Accuracy	group 1	group 2	group 3	group 4	group 5
KPCA-EL		28.99	45.02	37.18	58.31	99.96

In order to more highlight the effectiveness of the proposed ML-based KPCA by decreasing the complexity and computational time, a proposed ML-based reduced KPCA method is done. The retained number of KPCs based RKPCA method using the CPV criterion is equal to 18. For multi-class classifiers, a comparison between the two proposed techniques in terms of accuracy and computation time is presented in Table 6. The results in Table 6, show that the proposed multiple model-based RKPCA achieves the best tread-off between accuracy (100% and computation time (110.36). Additionally, the computation time is reduced by more than 50% using EL-based RKPCA (110.36) compared to EL-based KPCA (221.85) technique. Thus, the proposed methods can not only reduce the computation cost but also guarantee the monitoring abilities.

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

	Methods	
Global Performance	Accuracy	CT(s)
KPCA-EL	**99.96**	**221.85**
RKPCA-EL	**100**	**110.36**

The confusion matrix (CM) is another performance measurement for machine learning classification. The CM provides more information not only about the performance of a predictive model, although about which classes correctly predicted, which incorrectly, and the type of errors made. Therefore, to more investigate the effectiveness of the proposed techniques, the confusion matrices of the EL-based KPCA and EL-based RKPCA techniques are presented in Tables 7 and 8 where the correct and miss-classified observations for different condition modes are presented. The rows present the predicted process statuses class while the columns show the true classes. Referring to the results given in Tables 7 and 8, it is clear for the healthy case (C0) that the enhanced classifiers KPCA-EL technique identifies 1500 measurement (true positive) from 1501 measurements and RKPCA-EL technique correctly identifies 1501 measurement for data sets. The results show that the proposed techniques are able to differentiate the six different modes and to get good classification results. In addition, the precision is 100% and the recall is 100% with 0.0% of misclassification using RKPCA-EL for all faulty cases.

Next, we consider a bank of one class classifiers. At this stage, the bank applies six classifiers. Each one is trained in order to classify a specific class labeled by 1 or -1 as shown in Table 9. Table 10 presents the global performance

TABLE 7. Confusion matrix of KPCA-EL using group 5.

Conf. Matrix		Predicted process statuses						Recall
True classes	C0	1500	1	0	0	0	0	99.93
	C1	0	1501	0	0	0	0	100
	C2	0	0	1500	1	0	0	99.93
	C3	0	0	0	1501	0	0	100
	C4	0	0	0	0	1500	1	99.93
	C5	0	0	0	0	0	1501	100
Precision		100	99.86	100	100	100	99.93	**99.96**

TABLE 8. Confusion matrix of RKPCA-EL using group 5.

Conf. Matrix		Predicted process statuses						Recall
True classes	C_0	1501	0	0	0	0	0	100
	C_1	0	1501	0	0	0	0	100
	C_2	0	0	1501	0	0	0	100
	C_3	0	0	0	1501	0	0	100
	C_4	0	0	0	0	1501	0	100
	C_5	0	0	0	0	0	1501	100
Precision		100	100	100	100	100	100	**100**

accuracy using the selected features of group 5 as inputs in the case of one class classifier scenario. The comparison results presented in Table 10 show that the two proposed techniques KPCA-EL and RKPCA-EL provide good results during the training and testing phases with a mean of accuracy equal to 99.97 and 99.99 using KPCA-EL and RKPCA-EL, respectively.

TABLE 9. Multiple one class classifier logic for fault diagnosis.

	Classes					
	C0	C1	C2	C3	C4	C5
Classifier for C0	**1**	-1	-1	-1	-1	-1
Classifier for C1	-1	**1**	-1	-1	-1	-1
Classifier for C2	-1	-1	**1**	-1	-1	-1
Classifier for C3	-1	-1	-1	**1**	-1	-1
Classifier for C4	-1	-1	-1	-1	**1**	-1
Classifier for C5	-1	-1	-1	-1	-1	**1**

TABLE 10. Accuracy using group 5 with different one class classifiers.

	Methods	
Class	KPCA-EL	RKPCA-EL
C0	99.94	99.96
C1	99.98	100
C2	99.99	100
C3	100	100
C4	99.92	100
C5	100	100
Average	**99.97**	**99.99**

In fault detection, the widely used metrics to assess the performance of diagnostic results are False Alarm Rate (FAR) and Missed Detection Rate (MDR). FAR is defined as the ratio between the number of misclassified measurements and the total number of measurements under healthy conditions (C_0). For each faulty scenario, the corresponding measurements classified in class C_0 are considered as missed detected. The MDR presents the ratio between the number

TABLE 11. Fault detection metrics.

Classifier	FAR %	MDR %				
	C_0	C_1	C_2	C_3	C_4	C_5
KPCA-EL	0.07	0	0.07	0	0.07	0
RKPCA-EL	0	0	0	0	0	0

of the missed detected measurements and the total number of measurements of its corresponding class. These metrics established for the two proposed methods KPCA-EL and RKPCA-EL are illustrated in Table 11. The selected features feeding into the multiple models provides good results in terms of MDR and FAR. Overall, the best performance was obtained using the proposed RKPCA-EL method with FAR and MDR equal to zero.

In order to more assess the obtained results and to support decision making process, we adopt the Friedman test methodology who is a non-parametric test at the significance level of $\alpha = 0.05$ [46], [70]. The obtained $p - values$ of the tests with the base classifiers Ensemble learning (EL), Bagging, RF, DT, kNN and SVM are showed in Table 12 by also representing their significance level. Regarding to Table 12, we can be conclude that the obtained results are considered as statistically significant.

TABLE 12. Adjusted p-values obtained from classification.

Methods	Friedman Statistical Test	
	p-value	Significance Level
Ensemble learning (EL)	0.00026	Very strong
Bagging	0.00029	Very strong
RF	0.00029	Very strong
DT	0.00386	Strong
KNN	0.00031	Very strong
SVM	0.00294	Very strong

For multi-class classifiers, a comparative study between the proposed methods and existing techniques such as Neural Network (NN), Multiple Layers Neural Network (MNN), Feed-Foward Neural Network (FFNN), Cascade Foward Neural Network (CFNN), Generalized Regression Neural Network (GRNN), Probabilistic Neural Network (PNN) and Recurrent Neural Network (RNN) has been conducted. The fault diagnosis performances of the different techniques are given in Table 13. Table 13 summarizes the performance according to the Accuracy, F_1 score, Recall, Precision, and computation time (CT). The comparative analysis showed that the proposed RKPCA-EL method totally outperformed the other models in terms of Accuracy (100 %), F_1 score (100 %), Recall (100 %) and Precision (100 %). Additionally, it is obvious that the performance of the fault classification, as well as the classification accuracy, were significantly enhanced using the proposed KPCA-EL and RKPCA-EL methods compared to deep learning models. From Table 13, it is shown that because the KPCA and RKPCA models can manipulate the nonlinearity of the PV system it improves the feature extraction accuracy and outperforms other deep

TABLE 13. Performance comparison of the different multi class classifiers.

Methods	Global Performance				
	Accuracy	F_1 score	Recall	Precision	CT (s)
RKPCA-EL	**100**	**100**	**100**	**100**	110.36
KPCA-EL	**99.96**	**99.95**	**99.96**	**99.97**	221.85
FFNN	77.14	77.16	77.20	77.19	53.28
MNN	79.09	79.15	79.12	79.13	21.46
NN	68.24	68.25	68.26	68.26	12.35
RNN	68.24	68.25	68.26	68.26	263.16
GRNN	63.27	63.31	63.25	63.23	30.34
PNN	62.18	62.16	62.18	62.21	26.00
CFNN	79.86	79.95	79.94	79.99	59.09

learning classifiers. Thus, KPCA-EL and RKPCA-EL classifiers are more useful for fault diagnosis. In addition, the application of the proposed techniques for fault classification makes the performance of fault diagnosis effective. The presented deep learning classifiers provide a classification accuracy less than 80% and a classification error greater than 20%. The poor classification results are due to the direct use of measured variables which indicates the success of the proposed KPCA-EL and RKPCA-EL methods which extract and select the more pertinent features before performing the classification. For NN, MNN, FFNN, CFNN, GRNN, PNN, and RNN classifiers, the highest classification rate was reached using CFNN and MNN with accuracy values of 79.86% and 79.09% and a misclassification rate of 20.14% and 20.91%, respectively. Thus, the use of these classifiers provides low classification accuracy which leads to poor fault diagnosis performances.

VII. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this work, three ensemble learning techniques are proposed to provide a reliable prediction for Grid-Connected Photovoltaic (PV) systems. Ensemble machine learning paradigms aims at developing effective and reliable models with higher accuracy than single machine learning. The main contributions are threefold: first, using the SVM, KNN, and tree models, we constructed an ensemble learners in order to obtain accurate performance than single learner to distinguish between the different PV system operating modes using the extracted raw data. Second, in order to further enhance the diagnosis results, intelligent FDD techniques were proposed, where the main steps are: feature extraction, features selection, and fault classification. For the features extraction and selection steps, KPCA and RKPCA methods are performed to extract and select the most significant features. Then, the most sensitive and significant characteristics are transmitted to the ensemble learning models for classification purposes. The developed approaches were developed to monitor a grid-connected PV system under healthy and faulty conditions. The experimental results demonstrated the feasibility and effectiveness of the proposed FDD techniques. The fault detection results obtained using the developed approaches provided some false alarm and missed detection rates and a few faults were not correctly detected. Hence, one future research perspective is to develop an online KPCA-based methods to update the model which may lead to a reduction in false alarm and missed detection rates. The second perspective is to extend the online KPCA-based methods to deal with uncertainties by using interval-valued data representation [71]. Besides, we propose to improve our contribution on detection and diagnosis purpose by using online models for more features extraction and selection in order to enhance the diagnosis metrics and classification rate of complex systems under different operating conditions. In the current work, the classical EL 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 [72]) or Mondrian forests (described in [73], [74]), may reduce the training and update time.

REFERENCES

[1] B. K. Karmakar and A. K. Pradhan, "Detection and classification of faults in solar PV array using Thevenin equivalent resistance," *IEEE J. Photovolt.*, vol. 10, no. 2, pp. 644–654, Mar. 2020.

[2] X. Li, Y. Li, J. E. Seem, and P. Lei, "Detection of internal resistance change for photovoltaic arrays using extremum-seeking control MPPT signals," *IEEE Trans. Control Syst. Technol.*, vol. 24, no. 1, pp. 325–333, Jan. 2016.

[3] Y. S. Manjili, R. Vega, and M. M. Jamshidi, "Data-analytic-based adaptive solar energy forecasting framework," *IEEE Syst. J.*, vol. 12, no. 1, pp. 285–296, Mar. 2018.

[4] A. Triki-Lahiani, A. B. Abdelghani, and I. Slama-Belkhodja, "Fault detection and monitoring systems for photovoltaic installations: A review," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 2680–2692, Feb. 2018.

[5] M. Sabbaghpur Arani and M. A. Hejazi, "The comprehensive study of electrical faults in PV arrays," *J. Electr. Comput. Eng.*, vol. 2016, pp. 1–10, Dec. 2016.

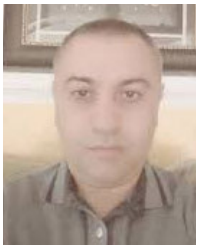
[6] R. Fazai, K. Abodayeh, M. Mansouri, M. Trabelsi, H. Nounou, M. Nounou, and G. E. Georghiou, "Machine learning-based statistical testing hypothesis for fault detection in photovoltaic systems," *Sol. Energy*, vol. 190, pp. 405–413, Sep. 2019.

[7] M. Mansouri, M. Trabelsi, H. Nounou, and M. Nounou, "Deep learning based fault diagnosis of photovoltaic systems: A comprehensive review and enhancement prospects," *IEEE Access*, vol. 9, pp. 64267–64277, 2021.
[8] K. Dhibi, R. Fezai, M. Mansouri, M. Trabelsi, K. Bouzrara, H. Nounou, and M. Nounou, "A hybrid fault detection and diagnosis of grid-tied PV systems: Enhanced random forest classifier using data reduction and interval-valued representation," *IEEE Access*, vol. 9, pp. 64267–64277, 2021.
[9] L. Oneto, F. Laureri, M. Robba, F. Delfino, and D. Anguita, "Data-driven photovoltaic power production nowcasting and forecasting for polygeneration microgrids," *IEEE Syst. J.*, vol. 12, no. 3, pp. 2842–2853, Sep. 2018.
[10] Y. Jiang, S. Yin, and O. Kaynak, "Optimized design of parity relation-based residual generator for fault detection: Data-driven approaches," *IEEE Trans. Ind. Informat.*, vol. 17, no. 2, pp. 1449–1458, Feb. 2021.
[11] S.-Q. Chen, G.-J. Yang, W. Gao, and M.-F. Guo, "Photovoltaic fault diagnosis via semisupervised ladder network with string voltage and current measures," *IEEE J. Photovolt.*, vol. 11, no. 1, pp. 219–231, Jan. 2021.
[12] 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.
[13] L. Jiang, L. Zhang, L. Yu, and D. Wang, "Class-specific attribute weighted naive Bayes," *Pattern Recognit.*, vol. 88, pp. 321–330, Apr. 2019.
[14] N. Suguna and K. Thanushkodi, "An improved k-nearest neighbor classification using genetic algorithm," *Int. J. Comput. Sci. Issues*, vol. 7, no. 2, pp. 18–21, 2010.
[15] J. R. Quinlan, "Simplifying decision trees," *Int. J. Man-Mach. Stud.*, vol. 27, no. 3, pp. 221–234, Sep. 1987.
[16] P. T. Noi and M. Kappas, "Comparison of random forest, K-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery," *Sensors*, vol. 18, no. 1, p. 18, 2018.
[17] W. Jia, Y. Deng, C. Xin, X. Liu, and W. Pedrycz, "A classification algorithm with linear discriminant analysis and axiomatic fuzzy sets," *Math. Found. Comput.*, vol. 2, no. 1, pp. 73–81, 2019.
[18] M. M. Hamed, M. G. Khalafallah, and E. A. Hassanien, "Prediction of wastewater treatment plant performance using artificial neural networks," *Environ. Model. Softw.*, vol. 19, no. 10, pp. 919–928, Oct. 2004.
[19] M. S. Nasr, M. A. E. Moustafa, H. A. E. Seif, and G. El Kobrosy, "Application of artificial neural network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT," *Alexandria Eng. J.*, vol. 51, no. 1, pp. 37–43, Mar. 2012.
[20] Y. Djebbar and R. M. Narbaitz, "Neural network prediction of air stripping KLa," *J. Environ. Eng.*, vol. 128, no. 5, pp. 451–460, May 2002.
[21] N. Moreno-Alfonso, "Intelligent waste-water treatment with neural-networks," *Water Policy*, vol. 3, no. 3, pp. 267–271, 2001.
[22] J. M. Hahne, F. Biessmann, N. Jiang, H. Rehbaum, D. Farina, F. Meinecke, K.-R. Müller, and L. Parra, "Linear and nonlinear regression techniques for simultaneous and proportional myoelectric control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 2, pp. 269–279, Mar. 2014.
[23] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *J. Intell. Learn. Syst. Appl.*, vol. 9, no. 1, pp. 1–16, 2017.
[24] H. Bhavsar and A. Ganatra, "A comparative study of training algorithms for supervised machine learning," *Int. J. Soft Comput. Eng.*, vol. 2, no. 4, pp. 2231–2307, 2012.
[25] V. Le, X. Yao, C. Miller, and B.-H. Tsao, "Series DC arc fault detection based on ensemble machine learning," *IEEE Trans. Power Electron.*, vol. 35, no. 8, pp. 7826–7839, Aug. 2020.
[26] O. Güngör, B. Akşanlı, and R. Aydoğan, "Algorithm selection and combining multiple learners for residential energy prediction," *Future Gener. Comput. Syst.*, vol. 99, pp. 391–400, Oct. 2019.
[27] F. Hasan, A. Kargarian, and A. Mohammadi, "A survey on applications of machine learning for optimal power flow," in *Proc. IEEE Texas Power Energy Conf. (TPEC)*, Feb. 2020, pp. 1–6.
[28] Z. Zhang, H. Han, X. Cui, and Y. Fan, "Novel application of multi-model ensemble learning for fault diagnosis in refrigeration systems," *Appl. Thermal Eng.*, vol. 164, Jan. 2020, Art. no. 114516.
[29] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," *J. Big Data*, vol. 7, no. 1, pp. 1–40, Dec. 2020.
[30] Z. Li, D. Wu, and T. Yu, "Prediction of material removal rate for chemical mechanical planarization using decision tree-based ensemble learning," *J. Manuf. Sci. Eng.*, vol. 141, no. 3, Mar. 2019, Art. no. 031003.
[31] S. Han, Y. Wang, W. Liao, X. Duan, J. Guo, Y. Yu, L. Ye, J. Li, X. Chen, and H. Chen, "The distinguishing intrinsic brain circuitry in treatment-naïve first-episode schizophrenia: Ensemble learning classification," *Neurocomputing*, vol. 365, pp. 44–53, Nov. 2019.
[32] D. B. Araya, K. Grolinger, H. F. Elyamany, M. Capretz, and G. Bitsuamlak, "An ensemble learning framework for anomaly detection in building energy consumption," *Energy Buildings*, vol. 144, pp. 191–206, Jun. 2017.
[33] S. Conţiu and A. Groza, "Improving remote sensing crop classification by argumentation-based conflict resolution in ensemble learning," *Expert Syst. Appl.*, vol. 64, pp. 269–286, Dec. 2016.
[34] M. A. Khairalla, X. Ning, N. T. Al-Jallad, and M. O. El-Faroug, "Short-term forecasting for energy consumption through stacking heterogeneous ensemble learning model," *Energies*, vol. 11, no. 6, p. 1605, 2018.
[35] S. Gharsellaoui, M. Mansouri, S. S. Refaat, H. Abu-Rub, and H. Messaoud, "Multivariate features extraction and effective decision making using machine learning approaches," *Energies*, vol. 13, no. 3, p. 609, Jan. 2020.
[36] 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*, vol. 59, pp. 313–321, May 2021.
[37] H. R. Baghaee, D. Mlakic, S. Nikolovski, and T. Dragicevic, "Support vector machine-based islanding and grid fault detection in active distribution networks," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 8, no. 3, pp. 2385–2403, Sep. 2020.
[38] 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.
[39] Y.-Y. Hong and M. T. A. M. Cabatac, "Fault detection, classification, and location by static switch in microgrids using wavelet transform and taguchi-based artificial neural network," *IEEE Syst. J.*, vol. 14, no. 2, pp. 2725–2735, Jun. 2020.
[40] L. Chen, S. Li, and X. Wang, "Quickest fault detection in photovoltaic systems," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 1835–1847, May 2018.
[41] 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.
[42] Y. Gan, Z. Chen, L. Wu, C. Long, S. Cheng, and P. Lin, "A novel fault diagnosis method for PV arrays using extreme gradient boosting classifier," Tech. Rep., 2019.
[43] V. Pashazadeh, F. R. Salmasi, and B. N. Araabi, "Data driven sensor and actuator fault detection and isolation in wind turbine using classifier fusion," *Renew. Energy*, vol. 116, pp. 99–106, Feb. 2018.
[44] A. Kusiak, Z. Zhang, and A. Verma, "Prediction, operations, and condition monitoring in wind energy," *Energy*, vol. 60, pp. 1–12, Oct. 2013.
[45] S. Yin, C. Yang, J. Zhang, and Y. Jiang, "A data-driven learning approach for nonlinear process monitoring based on available sensing measurements," *IEEE Trans. Ind. Electron.*, vol. 64, no. 1, pp. 643–653, Jan. 2017.
[46] G. Tuysuzoglu and D. Birant, "Enhanced bagging (eBagging): A novel approach for ensemble learning," *Int. Arab J. Inf. Technol.*, vol. 17, no. 4, pp. 515–528, Jul. 2020.
[47] K. Li and H. Tian, "A bagging based multiobjective differential evolution with multiple subpopulations," *IEEE Access*, vol. 9, pp. 105902–105913, 2021.
[48] 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.
[49] R. Rosipal and L. J. Trejo, "Kernel partial least squares regression in reproducing kernel Hilbert space," *J. Mach. Learn. Res.*, vol. 2, pp. 97–123, Mar. 2001.
[50] 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.
[51] K. Dhibi, R. Fezai, K. Bouzrara, M. Mansouri, A. Kouadri, and M.-F. Harkat, "Machine learning based multiscale reduced kernel PCA for nonlinear process monitoring," in *Proc. 17th Int. Multi-Conf. Syst., Signals Devices (SSD)*, Jul. 2020, pp. 302–307.
[52] M. Wozniak, M. Grana, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Inf. Fusion*, vol. 16, pp. 3–17, Mar. 2014.

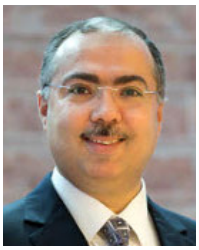
[53] P. Arya, A. Bhagat, and R. Nair, "Improved performance of machine learning algorithms via ensemble learning methods of sentiment analysis," *Int. J. Emerg. Technol.*, to be published.
[54] V. Gunes, M. Ménard, P. Loonis, and S. Petit-Renaud, "Combination, cooperation and selection of classifiers: A state of the art," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 17, no. 8, pp. 1303–1324, Dec. 2003.
[55] Z. H. Zhou, "Ensemble learning," in *Encyclopedia of Biometrics*, vol. 1. Berlin, Germany: Springer, 2009, pp. 270–273.
[56] J. Shin, "Random subspace ensemble learning for functional near-infrared spectroscopy brain-computer interfaces," *Frontiers Human Neurosci.*, vol. 14, p. 236, Jul. 2020.
[57] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. London, U.K.: Chapman & Hall, 2019.
[58] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ, USA: Wiley, 2014.
[59] J. Li, Y. Guo, J. Wall, and S. West, "Support vector machine based fault detection and diagnosis for hvac systems," *Int. J. Intell. Syst. Technol. Appl.*, vol. 18, nos. 1–2, pp. 204–222, 2019.
[60] J. Saari, D. Strömbergsson, J. Lundberg, and A. Thomson, "Detection and identification of windmill bearing faults using a one-class support vector machine (SVM)," *Measurement*, vol. 137, pp. 287–301, Apr. 2019.
[61] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.
[62] Y. Wang, Z. Pan, and Y. Pan, "A training data set cleaning method by classification ability ranking for the k-nearest neighbor classifier," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1544–1556, May 2020.
[63] P. Müller, K. Salminen, V. Nieminen, A. Kontunen, M. Karjalainen, P. Isokoski, J. Rantala, M. Savia, J. Väliaho, P. Kallio, J. Lekkala, and V. Surakka, "Scent classification by k nearest neighbors using ion-mobility spectrometry measurements," *Expert Syst. Appl.*, vol. 115, pp. 593–606, Jan. 2019.
[64] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–97, 2020.
[65] W.-Y. Loh, "Classification and regression trees," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 14–23, 2011.
[66] S. Lakshmanaprabu, K. Shankar, M. Ilayaraja, A. W. Nasir, V. Vijayakumar, and N. Chilamkurti, "Random forest for big data classification in the Internet of Things using optimal features," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 10, pp. 2609–2618, 2019.
[67] K. Dhibi, R. Fezai, M. Mansouri, A. Kouadri, M.-F. Harkat, K. Bouzara, H. Nounou, and M. Nounou, "A hybrid approach for process monitoring: Improving data-driven methodologies with dataset size reduction and interval-valued representation," *IEEE Sensors J.*, vol. 20, no. 17, pp. 10228–10239, Sep. 2020.
[68] H. H. Yue and S. J. Qin, "Reconstruction-based fault identification using a combined index," *Ind. Eng. Chem. Res.*, vol. 40, no. 20, pp. 4403–4414, 2001.
[69] 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.
[70] N. Settouti, M. E. A. Bechar, and M. A. Chikh, "Statistical comparisons of the top 10 algorithms in data mining for classification task," *Int. J. Interact. Multimedia Artif. Intell., Special Issue Artif. Intell. Underpinning*, vol. 4, pp. 46–51, Sep. 2016.
[71] 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.
[72] 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.
[73] 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.
[74] 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.

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 and Computer 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.



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 journal 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 at 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, data based 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 at TAMU-Texas A&M University at Qatar. He has more than 19 years of combined academic and industrial experience. He has published more than 200 refereed journal and conference publications and book chapters. He has successfully served as the lead PI and a PI on several QNRF projects (six NPRP projects and three UREP projects). 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).