IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 23, 2022, accepted April 5, 2022, date of publication April 13, 2022, date of current version May 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3167147

Interval-Valued Reduced Ensemble Learning Based Fault Detection and Diagnosis Techniques for Uncertain Grid-Connected PV Systems

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

[1]Electrical and Computer Engineering Program, Texas A&M University at Qatar, Doha, Qatar
[2]Department of Mathematical Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia
[3]Research Laboratory of Automation, Signal Processing and Image, National Engineering School of Monastir, Monastir 5019, Tunisia
[4]Chemical Engineering Program, Texas A&M University at Qatar, Doha, Qatar

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

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

ABSTRACT One of the most promising renewable energy technologies is photovoltaics (PV). Fault detection and diagnosis (FDD) becomes more and more important in order to guarantee high reliability in PV systems. FDD of PV systems using machine learning technique aims to develop effective models that can provide a better rate of accuracy. Recently, numerous machine learning based ensemble models have been applied in FDD using different combination techniques. Ensemble method is a tool that merges several base models in order to produce one optimal predictive model. In this study, we propose six effective Ensemble Leaning (EL)-based FDD paradigms for uncertain Grid-Connected PV systems. First, EL-based interval centers and ranges and interval upper and lower bounds techniques are proposed to deal with PV system uncertainties (current/voltage variability, noise, measurement errors, ...). Next, in order to more improve the diagnosis abilities, two interval kernel PCA (IKPCA)-based EL classifiers are developed. The IKPCA-EL techniques are addressed so that the features extraction and selection phases are performed using the IKPCA models and the sensitive and significant interval-valued characteristics are transmitted to the EL model for classification purposes. Finally, the number of observations in the training data set is reduced using Hierarchical K-means techniques in order to overcome the problem of computation time and storage cost. Therefore, two interval reduced KPCA-EL techniques are proposed. The study demonstrated the feasibility and efficiency of the proposed techniques for fault diagnosis of Grid-Connected PV systems.

INDEX TERMS Uncertain systems, ensemble learning, fault diagnosis, interval-valued data, kernel principal component analysis (KPCA), grid-connected PV (GCPV).

LIST OF ABBREVIATIONS AND ACRONYMS

FDD	Fault Detection and Diagnosis
FES	Feature extraction and selection
PCA	Principal Component Analysis
MPCA	Multiscale PCA
ANN	Artificial Neural Network
MNN	Multilayer Neural Network
CFNN	Cascade forward Neural Network
ℓ	Number of retained PCs
CPV	Cumulative Percentage of Variance
PV	Photovoltaic
GCPV	Grid-Connected PV
CT	Computation Time
NN	Neural Network
MNN	Multiple Layers 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 Dazhong Ma.

I. INTRODUCTION

In recent decades, photovoltaic (PV) energy has gained great importance in the world with major developments in grid-connected applications, since it has desirable characteristics such as decreasing cost, environmental compatibility, short installation time, and low maintenance cost. Grid-Connected PV (GCPV) energy systems got the most interest and have been an increased attention. Faults in any components of GCPV systems such as grid-connection, converters, inverters, open-circuit/short-circuit, panels, modules and arrays, can earnestly affect the efficiency, security and reliability of the all GCPV plant. Therefore, fault detection and diagnosis (FDD) is very crucial for achieving the best functioning and ensure safe and continuous operation of GCPV systems. Addressing these issues, different techniques have been proposed to detect and diagnose faults in PV systems. The proposed methods vary in complexity, rapidity, and capability to identify a large number of faults. Most of traditional fault diagnosis techniques are based on machine learning algorithms [1]–[3]. In [4], a review on artificial intelligence techniques for GCPV Systems is presented. This study demonstrates that the ANN and its sub-architectures are the most widely used machine learning techniques to diagnose photovoltaic systems. However, most of the existing ANN techniques suffer from the problems of overfitting and complexity time. Therefore, improved versions of ANNs based on backpropagation algorithms like multilayer perceptron networks (MLPN) are proposed to overcome these challenges. ANN has been developed to predict the electrical outputs of a PV module and detect the faulty case in [5]. This proposed technique consists in determining the measured PV output values and the values predicted by the ANN method. The operating state is considered as a faulty state if the difference between the measured values and the predicted one exceeds a threshold value. The proposed method can accurately estimate PV production without complex mathematical calculations and can detect any decrease in output power. However, it is recommended to train the ANN periodically to preserve accuracy. In addition, other types of faults (open circuit, short circuit, ...) are not taken into account in the proposed method. In [6], a fault diagnosis method based on Modified neural networks has been proposed to detect faults in PV systems. In this proposal, the PV system was simulated using a solar Pro software package for gathering power generation data from PV modules under normal and faulty operating modes. In [7], a proposed method based on feature extraction using wavelet transform and classification attributes of radial basis function networks (RBFNs) is presented. In this proposal, the dynamic fusion of kernels is performed in order to improve the performance of the proposed method. In [8], a fault diagnosis model is proposed for fault detection and classification in PV systems. In this proposal, different faulty and normal datasets are normalized and preprocessed using several data-mining techniques and then fed into a probabilistic neural network (PNN) to predict and classify faults. However, the main drawback of this method is the significantly depending on the proper choice of the smoothing parameter to enhance the accuracy. In [9], the authors propose an online reduced kernel generalized likelihood ratio test technique for fault detection in PV systems with MPP operation data. Support vector machines (SVM) is one of the widely used technique in classification and nonlinear function estimation [10]. The main drawback of the classical SVM technique presented in the selection of features that may sometimes lead to wrong output especially when the data set has more noise. In [11], a fault diagnosis method based on experimental data, combined with the KNN technique is proposed. The main idea behind this proposal is to detect and classify different faults like open circuits, line-line, partial shading in real-time. KNN is one of the topmost used machine learning algorithms thanks to her simplicity and high capacity [12]. KNN does not require any assumption for underlying data distribution and any training data points for model generation. This in turn gives high performances when using real datasets [12], [13]. But KNN suffers from some limitations in the case of the large dataset because the instance calculation of distances between each samples would be very costly. The DT algorithm belongs to the family of supervised learning algorithms and it has been widely used in literature. The main idea of this technique is to predict the class or value of the target variable by learning simple decision rules allowed from training data. DT algorithm executes classification without requiring much computation and generates understandable rules. In [14], a decision tree method has been proposed to detect and classify open circuits, line-line short circuits, partial shading, and degradation. This proposal can accurately detect different conventional faults. The main drawback of this technique is the assumption that the PV array is operating at the (MPP), which is not ensured in real PV systems [15]. Besides, it generally undergoes problems of overfitting, especially in the case of a large data size.

During the last decades, ensemble learning (EL) models have gained significant attention from the scientific community [16]. EL is a technique that creates and combines multiple machine learning models in order to produce one optimal predictive model which gives improved results [16]. Bagging, boosting, stacking and random subspace are the main types of ensemble methods [17]. Bagging is used as a way to decrease the variance in order to improve the accuracy of models through decision trees. Boosting aims to learn from precedent predictor errors to make better predictions in the future (decrease bias). Stacking allows a learning algorithm to group together several other predictions of similar learning algorithms (improve predictions). Random subspace combines the predictions of multiple decision trees trained on different subsets of columns in the training dataset by simple majority voting in the final decision rule [18]. Bagging helps eliminate the overfitting of models in the procedure by decreasing variance. However, the resultant model using bagging ensemble methods can experience lots of bias when the proper procedure is ignored and it introduces a loss of

interpretability of a model [19]. Boosting algorithm seeks to reduce the model's bias and it is used when high bias and low variance are presented. In addition, Boosting generates an unified model with fewer errors as it focuses on maximizing benefits and reducing shortcomings in a single model [19]. The main advantage of the random subspace technique is the random selection of subsets of features, resulting in weakly correlated multiple weak learners [20]. In conclusion, when the challenge in a single model is overfitting, the bagging method performs better than the boosting technique. Boosting faces the challenge of handling over-fitting since it comes with over-fitting in itself. When the challenge is to obtain low-correlated multiple weak learners, the random subspace technique method is better than boosting and bagging. A Comparison study between single and ensemble learning algorithms is presented in [17], [21]. It is showed that ensemble learning techniques can outperform classical single machine learning methods in many cases [21]. The first one is when the training algorithm fails to find the best solution (computational problems). The second one is when the available training data are too small compared to the search space (statistical problems). The last one is when the learning algorithms miss affecting fitness functions (representation problems) [21]. Another study demonstrates that boosting ensemble techniques outperformed bagged ensemble techniques to predict the stock market [22]. In the literature, ensemble learning algorithms are widely used to affect maximum performance and they have been applied in a variety of real-world applications [17], [23]. In [24], the authors propose a technique to improve the predictive performance of existing conventional machine learning (ML) algorithms as an arc fault detection method. This proposal is based on the superposition of conventional ML algorithm to create an enhanced classifier that decreases the bias and decision variance. Another fault detection method based on ensemble machine learning is introduced in [25]. In [17], an enhanced ensemble learning method was proposed to provide a better and higher rate of prediction accuracy of stock-market prediction. In this proposal, boosting, bagging, stacking, blending, and simple maximum voting combination techniques are used to construct twenty-five different ensemble classifiers using DT, SVM, and multilayer perceptron (MLP) neural networks. Despite numerous studies revealing the dominance of ensemble learning methods over single learning methods, most of these works only ensemble a specific type of classifier. In addition, the previously investigated ensemble learning-based fault classification approaches use only single-valued data, and the uncertainties of the system are not taken into account. The uncertainty in the systems, which is presented 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. Thus, the interval-valued data representation offers a better overview of the measured phenomenon compared to the representation of the average value. However, inaccuracy, uncertainty, or parameters variability might characterize the important information describing the real systems [26]. Thus, classical data is not able to present these dissimilarities and for this reason, it is important to represent the data as interval-valued data. In [27], a KNN approach to deal with uncertainties by using data in the form of intervals. In other studies, a new approach for constructing regression and classification models for interval-valued data using support vector machine method is proposed [28]. An uncertainty analysis technique based on a non-parametric statistical modelling method for photovoltaic array output is proposed in [29]. This proposal aims to resolve the problem of differences between the parameter estimation (PE) results and the real output distributions by using nonparametric kernel density estimation (NKDE) methods. Besides, another main drawback of the classical ensemble learning classifiers is the direct use of the raw information from the process data. In the literature, different FDD techniques based on feature extraction and selection steps using a single classifier have been proposed [30], [31]. The main idea behind the extraction and selection steps is to extract and to select the most pertinent and informative data features, which will consequently enhance the use of the ML algorithm in the classification step for diagnosis purposes [32]. Literature has shown that the applications of some techniques for feature extraction and selection have significantly enhanced the accuracy of classification. In [33], a fault classification method based on multiscale interval PCA (MSIPCA) and ML method was proposed for uncertain HVAC systems. The MSIPCA technique is also proposed for enhancing the diagnosis performance by extracting the most significant linear features from data. However, popular complex systems show strong nonlinear correlations between their variables. Various nonlinear Kernel PCA (IKPCA) methods have been presented [34], [35]. The main objective of the IKPCA method is to i) transform the interval-valued data matrix on a numerical data matrix, (ii) map the input numerical data onto the feature space using a nonlinear mapping function, and (iii) use PCA into a feature space [35].

In this work, we propose innovative ensemble learning paradigms to deal with the problem of fault detection and diagnosis of uncertain PV systems. The principal contributions of this article are threefold.

1) The first contribution of this paper aims to develop an effective EL models for interval valued data with KNN, SVM and DT classifiers using bagging, boosting and random subspace combination tools. The developed paradigms are so-called interval EL (IEL)-based centers and ranges (IEL_{CR}), and upper and lower bounds (IEL_{UL}). The objective behind these proposed methods is to show the impact of using interval-valued data instead of single-valued data to improve the fault diagnosis abilities. The main idea of the developed techniques is to represent the interval-valued data matrix using centers and ranges or upper and lower bounds approaches. Then, the feature matrices

are constructed and introduced to the proposed EL classifier for fault classification purposes. In this study, we use two methods based on interval-valued data to further assess the effectiveness of using model uncertainties. The developed techniques achieve higher classification accuracy compared to the single-valued data EL approaches. However, the two proposed techniques suffer from some limitations due to the direct use of interval-valued features at the nodes.

2) To surmount this problem, two interval-valued data kernel PCA (IKPCA) methods are applied to extract features by transforming the single-valued data set into interval-valued latent variables. The IKPCA$_{CR}$ consists first to compute the new numerical matrix by the concatenation of center and range matrices and then to perform KPCA model on the new matrix. The second method, the IKPCA$_{UL}$, aims to fit two KPCA models on the lower and upper bounds of the interval values. Next, to enhance the diagnosis effectiveness, it is important to select the most significant and relevant features before doing the classification task. Finally, the faults are classified using the EL model. Second, in order to further improve the classification accuracy, EL-based IKPCA methods are proposed. The proposed EL-IKPCA schemes are addressed such that the interval kernel PCA (IKPCA$_{CR}$ and IKPCA$_{UL}$) techniques are developed for features extraction and selection. Then, the more relevant features are fed to the EL for classification purposes.
3) An improved IKPCA technique, called interval reduced KPCA (IRKPCA) is proposed. This proposal aims to overcome the problem of computation time and storage cost. The improved IKPCA technique consists of reducing the number of observations in the training data set using Hierarchical K-means (H-K-means) clustering method.

To summarize, six multi-class (MC) classifiers called IEL$_{CR}$, IEL$_{UL}$, IKPCA$_{CR}$-based EL, IKPCA$_{UL}$-based EL, IRKPCA$_{CR}$-based EL, IRKPCA$_{UL}$-based EL are used. The main goals behind the proposed methods are to show the efficiency of using interval-valued data, features extraction and selection, and data size reduction step by step. The MC classifiers consist of classifying instances into one or more classes. To further improve the classification performances of the developed classifiers, a set of one-class (OC) classifiers is proposed. To do that, a bank of OC IEL$_{CR}$, IEL$_{UL}$, IKPCA$_{CR}$-based EL and IKPCA$_{UL}$-based EL, IRKPCA$_{CR}$-based EL and IRKPCA$_{UL}$-based EL classifiers are developed (there are as many classifiers as classes). An emulated PV system is applied to demonstrate the effectiveness of the proposed diagnosis methods.

The rest of the work is presented as follows. Section II presents the GCPV system description and data collection. A brief description of machine based ensemble learning techniques is given in Section III. Section IV presents the proposed paradigms. The performance of the proposed methods is evaluated in Section V. At last, some conclusions are drawn in Section VI.

II. PV IMPLEMENTATION AND DATA COLLECTION

Figure 1 shows the synoptic of the GCPV system under study, where PV and grid emulators are used to emulate the operation of PV panels and a 3-phase grid respectively (under different operating modes). 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.

The faults were emulated at different system stages (common coupling point, inverter, sensors, emulated PV arrays, ...) to ensure a comprehensive analysis [30], [32]. 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.

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

The healthy and faulty operation modes are showed in Table 2.

III. MACHINE LEARNING ALGORITHMS

A. CLASSIFICATION TECHNIQUES

In this study, we use SVM, KNN, and DT models to construct different ensemble classifiers. The main advantage of SVM technique is that it is able to handle high dimensional data without overfitting problems. Moreover, the kernel trick is a real strength of SVM in which one can solve any complex problem [36]. However, SVM does not perform very well when the data set has more noise which affects the final decision. The KNN model is a very efficient classifier in terms of improvisation for random modeling on available data [37]. A tree model is very useful for solving decision-related problems and it can work well even if the assumptions are somewhat violated by the dataset from which the data is extracted [38]. Therefore, in this work, we propose three well-used machine learning algorithms, each of which differs in its

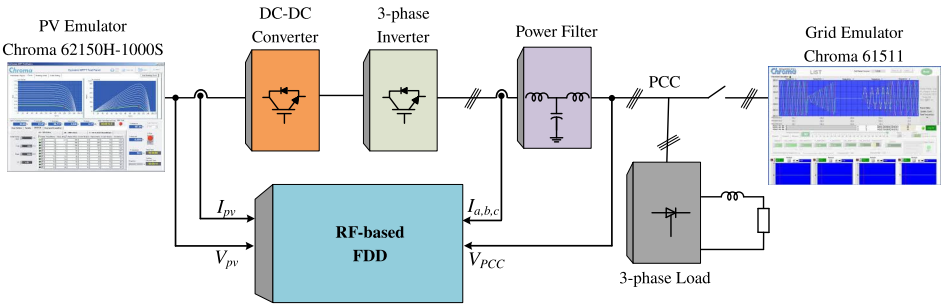


FIGURE 1. Synoptic of the grid-connected PV system under study.

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

TABLE 2. Construction of database for fault diagnosis system.

Class	Mode	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

way of training from the other, to overcome the shortcomings that result from the use of a single classifier. Thus, they work in an integrated way.

1) SUPPORT VECTOR MACHINES (SVM)

SVM has been first introduced by Vapnik [39]. There are two main categories for SVM: support vector classification (SVC) and support vector regression (SVR). In this study, an overview of the basic ideas underlying support vector (SV) machines for classification is presented. For a considered training data set with N samples $\{x_k, y_k\}_{k=1}^{N}$, with input data $x_k \in \mathbb{R}^m$ and output $y_k \in \{-1\ 1\}$ which represents a set of labeled training features. The SVM for classification is presented as following:

$$y_k = f(x_k) = w^T x_k + b \tag{1}$$

where $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$.

2) DECISION TREE (DT)

Decision Tree (DT) is a well-known technique that has been applied to real-world problems [40]. DT is a symbolic learning technique that organizes information extracted from a training dataset in a hierarchical structure composed of nodes and ramifications. The main advantage of using DT algorithms is that they involve minimal requirements for data preparation and are robust on large datasets.

3) K-NEAREST NEIGHBORS (KNN)

KNN is a non-parametric algorithm, which means it does not make any assumption on underlying data [41]. The main step of KNN technique is to classify samples from the available data based on similarity. Therefore, when new data appears then it can be easily classified into a good suite category by using K- NN method [42]. The Euclidean distance is used to compute the KNN class as follows,

For a given known class $X = [x_1, x_2, \ldots, x_k]$ and a data to be classified $Y = [y_1, y_2, \ldots, y_k]$. So, the distance is given by

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_k - y_k)^2} \tag{2}$$

Then a class is assigned to which the distance defined as in Eq. 2 is minimum.

B. ENSEMBLE LEARNING TECHNIQUES

Ensemble technique is a machine learning technique that combines the decisions from multiple models in order to generate one optimal predictive model and to enhance global performance. The main idea behind ensemble techniques is to improve predictability in models and decrease bias and variance to boost the accuracy of models [43]. Boosting, bagging, and random subspace are the most popular ensemble methods. Next, we discuss the three advanced ensemble methods.

1) BOOSTING

Boosting is one of the most popular ensemble techniques. The main objective behind boosting algorithm is to combine many weak learners into strong learners [44]. So, it learns from previous predictor mistakes to make improved predictions

in the future. Therefore, it significantly improved the predictability of models [45]. The main steps of boosting technique are threefold: i) Bias the training data towards those examples which are difficult to predict, ii) add assembly members to correct predictions from previous models, and iii) combine predictions using a weighted average of the models. Some commonly boosting algorithms are adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost) and gradient boosting.

2) BAGGING

Bagging, also called bootstrap aggregating, is an ensemble learning method that decreases the variance and improves the accuracy of different models to form one ensemble model [44]. The first step of the bagging technique is to create multiple models. Then, the created models are generated based on the actual method with random sub-samples of the dataset which are constructed from the original dataset randomly using bootstrap sampling technique [45].

3) RANDOM SUBSPACE

Random subspace (RS) is similar to the bagging method but the variables are randomly sampled, with replacement, for each learner [46]. By training the estimators on random samples of characteristics instead of all characteristics set, RS aims to decrease the correlation between models. RS outperforms other ensemble techniques in terms of computational cost thanks to the use of random subsets [20].

IV. PROPOSED TECHNIQUES

The main contributions are threefold. First, two alternative and effective interval-valued learning methods (interval EL_{CR} and interval EL_{UL}) based on the direct use of variables measured with uncertainties are presented. In this study, we used three classification algorithms and three ensemble techniques. The used classification algorithms are SVM, DT, and KNN. The used EL techniques are Bagging, Boosting, and Random sub-space. The main steps of interval-valued raw data-based EL (IEL) techniques are illustrated in Figure 2. Then, in order to further improve the efficiency of the developed IEL methods, two additional intervals KPCA (IKPCA)-based FDD techniques are developed, where the most relevant characteristics are first extracted and selected from the original data then the final features are fed to the proposed EL model for classification purposes. Once the samples representing the healthy and different possible faulty scenarios in the process are available, the IKPCA models are constructed using only the healthy data. The built models are applied to extract and select the most significant features. However, the main disadvantage of IKPCA-EL is the computational cost which is proportional to the number of measurements. To overcome this challenge, an improved IKPCA technique based on a data reduction scheme using H-K-means clustering is proposed. The first objective behind this proposed technique is to reduce the number of samples. The improved IRKPCA-EL not only decreases the computation time and storage cost but also keeps the diagnosis capacity. Next, some arbitrary groups of selected features are applied to train the EL classifier. Finally, to make efficient decisions, we compare the EL output results using the different picked arbitrary groups.

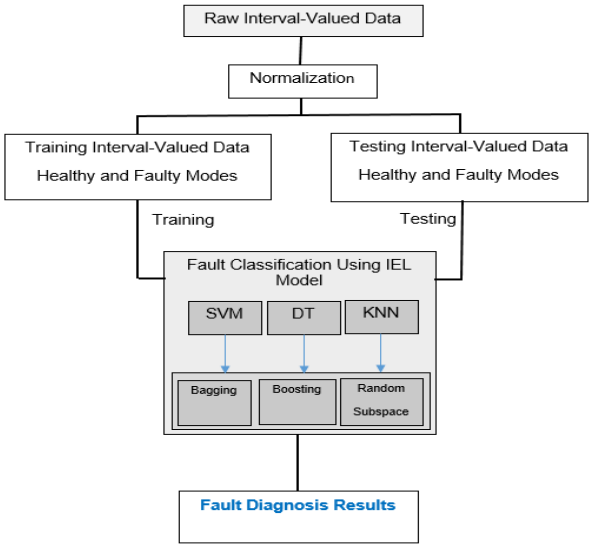


FIGURE 2. Schematic diagram of the interval RF (IRF)-based interval-valued raw data.

A. INTERVAL-VALUED Data

In order to keep the variable information, it is more relevant to present these measurements by interval values instead of single values. Given that x_{ij}, $i = 1, \ldots, N$ and $j = 1, \ldots, m$, is an observation is the $i-$th sample of the $j-$th variable, the interval representation of the data measurement x_{ij} is given by,

$$[x_{ij}] = [\underline{x_{ij}}, \overline{x_{ij}}] \tag{3}$$

where $\underline{x_{ij}}$ and $\overline{x_{ij}}$ are the lower bound and upper bound of the interval,respectively. The interval-valued matrix $[X]$ is defined as follows:

$$[X] = \begin{pmatrix} [\underline{x}_{11}, \overline{x}_{11}] & . & . & [\underline{x}_{1m}, \overline{x}_{1m}] \\ . & . & & . \\ . & & . & . \\ [\underline{x}_{N1}, \overline{x}_{N1}] & . & . & [\underline{x}_{Nm}, \overline{x}_{Nm}] \end{pmatrix} = ([x_1], \ldots, [x_N])^T \tag{4}$$

where $[x_k] = \left([\underline{x}_{1k}, \overline{x}_{1k}], \ldots, [\underline{x}_{mk}, \overline{x}_{mk}]\right)$.

The generic interval $[x_{jk}]$ can be also represented as a couple $\{x^c_{jk}, x^r_{jk}\}$. The center x^c_{jk} of the interval is given as [47], [48]

$$x^c_j(k) = \frac{1}{2}(\overline{x}_{jk} + \underline{x}_{jk}) \tag{5}$$

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

$$x^r_{jk} = \frac{1}{2}(\overline{x}_{jk} - \underline{x}_{jk}) \tag{6}$$

Usually, data is composed of variables that belong to different physical quantities with different scales and spreads. To deal with this problem, the data matrix is scaled to zero mean and unit variance. Thus, the pre-processing step is very important and it is recommended before applying any model in order to enhance the simulation results.

B. ENSEMBLE LEARNING FOR INTERVAL-VALUED DATA (IEL) METHOD

In this section, EL techniques based on interval centers and ranges IEL_{CR} and interval upper and lower bounds IEL_{UL} are presented.

1) EL BASED ON INTERVAL CENTERS AND RANGES (IEL_{CR})

In this method, the Center and Range (CR) approach is used. The CR technique is one of the most used models for analyzing interval-valued data. Let X be the training data sets, where m is the number of variables and N is the number of observations.

In the CR technique, the interval-valued data matrix is first transformed into center and range matrices as:

$$X^c = \frac{1}{2}\begin{pmatrix} \underline{x}_{11} + \overline{x}_{11} & . & . & \underline{x}_{1m} + \overline{x}_{1m} \\ . & . & & . \\ . & & . & . \\ \underline{x}_{N1} + \overline{x}_{N1} & . & . & \underline{x}_{Nm} + \overline{x}_{Nm} \end{pmatrix} \tag{7}$$

$$X^r = \frac{1}{2}\begin{pmatrix} \overline{x}_{11} - \underline{x}_{11} & . & . & \overline{x}_{1m} - \underline{x}_{1m} \\ . & . & & . \\ . & & . & . \\ \overline{x}_{N1} - \underline{x}_{N1} & . & . & \overline{x}_{Nm} - \underline{x}_{Nm} \end{pmatrix} \tag{8}$$

Then, the obtained data matrix is constructed by the concatenation of center and range data matrices. Thus, the new input $\mathbf{X}^{CR}$ data matrix is presented as:

$$X^{CR} = \left[X^c \; X^r\right] \in \mathbf{R}^{N\times 2m} \tag{9}$$

2) EL BASED ON INTERVAL UPPER AND LOWER BOUNDS (IEL_{UL})

For the interval EL_{LU} method, an upper-lower approach is considered to classify the data. Let X^L and X^U be the lower and upper bounds of the input matrices, respectively.

$$X^L = \begin{pmatrix} \underline{x}_{11} & . & . & \underline{x}_{1m} \\ . & . & & . \\ . & & . & . \\ \underline{x}_{N1} & . & . & \underline{x}_{Nm} \end{pmatrix} \tag{10}$$

$$X^U = \begin{pmatrix} \overline{x}_{11} & . & . & \overline{x}_{1m} \\ . & . & & . \\ . & & . & . \\ \overline{x}_{N1} & . & . & \overline{x}_{Nm} \end{pmatrix} \tag{11}$$

The upper and lower matrices can be considered at the same time. According to the above definitions, let X^{LU} be the upper-lower value that can be represented by:

$$x_{ij}^{LU} = \gamma \underline{x}_{ij} + (1-\gamma)\overline{x}_{ij} \tag{12}$$

where, $\gamma \in [0, 1]$, γ can be used as the adjustment weight of interval-valued data unit, which is used to balance the relationship between the upper and lower bounds of the interval-valued data unit. The upper and lower matrix is given by:

$$X^{LU} = \begin{pmatrix} x_{11}^{LU} & . & . & x_{1m}^{LU} \\ . & . & & . \\ . & & . & . \\ x_{N1}^{LU} & . & . & x_{Nm}^{LU} \end{pmatrix} \tag{13}$$

When $\gamma = 1$, it can be studied as a lower scheme with one feature. If $\gamma = 0$, then, it can be represented by an upper bound that contains the size information of $\overline{x}$.

The next section proposes two EL algorithms based on IKPCA models. In the proposed IKPCA-EL methods, only the most informative extracted features from the dataset are selected and applied to the EL algorithm for classification in the diagnosis problem.

C. ENSEMBLE LEARNING BASED INTERVAL KPCA METHODS

The main idea behind the proposed IKPCA-EL methods is to extract and select the most pertinent nonlinear features from interval-valued data using two IKPCA models. Then, the selected pertinent nonlinear features are fed to the EL to address the fault classification problem. The feature extraction and selection steps are used to retain only the most relevant and effective measurements in order to better present any system under different operating modes. IKPCA method consists of transforming the interval-valued dataset on a numerical dataset and then a KPCA is applied to the created numerical dataset. Besides, it aims to calculate the interval kernel principal components (IKPCs) in the characteristics space using nonlinear kernel functions and integral operators [49]. Let us consider three data matrices $X^{CR} \in \mathbf{R}^{N\times 2m}$, $X^L \in \mathbf{R}^{N\times m}$, and $X^U \in \mathbf{R}^{N\times m}$, which represent the center and range matrix, the lower matrix and the upper matrix, respectively. IKPCA technique consist of applying the KPCA model in the given interval data matrices.

1) FEATURE EXTRACTION USING IKPCA

Given a training interval data matrix $[X]$. The matrix regrouping the mapped interval vectors is arranged as follows: $[\mathcal{X}] = \left[\phi([x_1])\; \phi([x_2]) \ldots \phi([x_1])\right]^T \in \mathbb{R}^{N\times h}$, where $h \gg m$ is the dimension of the characteristic space. Using the kernel trick, we can compute the kernel principal components (KPC_s) using eigenvector expression as follows:

$$\lambda\alpha = K\alpha \tag{14}$$

where λ and α are the eigenvector and eigenvalue of the gram matrix K.

The interval kernel matrix K can be expressed as follows:

$$[K] = [\mathcal{X}]\left[\mathcal{X}^T\right] = \begin{bmatrix} k([x_1]),([x_1]) & \dots & k([x_1]),([x_N]) \\ \vdots & \dots & \vdots \\ k([x_N]),([x_1]) & \dots & k([x_N]),([x_N]) \end{bmatrix} \quad (15)$$

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

$$k([x]) = (k([x_1]),[x],\dots,k([x_N]),([x]))^T \quad (16)$$

2) FEATURE SELECTION USING IKPCA

Let be consider the eigenvector of the kernel matrix in the feature space $v = \lambda^{-1}\left[\mathcal{X}^T\right]\boldsymbol{\alpha}$ [34]. The matrix with the ℓ leading eigenvectors is computed as,

$$P = \left[{\lambda_1}^{-1}\left[\mathcal{X}^T\right]\boldsymbol{\alpha}_1,\dots,{\lambda_\ell}^{-1}\left[\mathcal{X}^T\right]\boldsymbol{\alpha}_\ell\right] \quad (17)$$

where $\Lambda = \text{diag}\{\lambda_1,\dots,\lambda_\ell\}$ is the ℓ largest eigenvalues of the matrix $[K]$.

Then, the kernel principal components are defined as, [34],

$$t = \Lambda^{-1/2}P^T k([x]) \quad (18)$$

Additional to the ℓ first KPCs, IRKPCA based features extraction is performed using the Hotelling's T^2, squared prediction error (Q) and combined φ statistics which are used to select the optimal features [50]. The statistical features are calculated as follows:

$$T^2 = k([x])^T P\,\Lambda^{-1}P^T k([x]) \quad (19)$$

$$SPE = k([x],[x]) - k^T([x])Ck([x]) \quad (20)$$

$$\varphi = \frac{SPE}{\tau_\alpha^{SPE}} + \frac{T_{CR}^2}{\tau_\alpha^{T^2}} \quad (21)$$

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

$$\tau_\alpha^{T^2} = \frac{\ell(N_r-1)(N_r+1)}{N_r(N_r-\ell)}F_\alpha(\ell,N_r-\ell) \quad (22)$$

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

$$\tau_\alpha^{SPE} = g_{SPE}\chi^2_{h_{SPE},\alpha} \quad (23)$$

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. For the IKPCA based on upper and lower bounds, new interval squared prediction error (ISPE) index is given by:

$$ISPE = \gamma\underline{SPE} + (1-\gamma)\overline{SPE} \quad (24)$$

where $\gamma \in [0,1]$, γ is the weight that defines the trade-off between the upper and lower bounds.

In the same way, the interval Hotteling's IT^2 statistic is given by:

$$IT^2 = \gamma\underline{T^2} + (1-\gamma)\overline{T^2} \quad (25)$$

where, $\underline{SPE}$ and $\underline{T^2}$ are the statistical characteristics for lower bound and $\overline{SPE}$ and $\overline{T^2}$ are the statistical characteristics for upper bound of interval-valued data. The interval combined index $I\varphi$ is given by,

$$I\varphi = \frac{ISPE}{\tau_\alpha^{ISPE}} + \frac{IT^2}{\tau_\alpha^{IT^2}} \quad (26)$$

where $\tau_\alpha^{IT^2}$ and τ_α^{ISPE} represent control limits of IT^2 and $ISPE$ at the confidence level $\alpha = 95\%$, respectively.

$$\tau_\alpha^{IT^2} = \frac{\ell(N_r-1)(N_r+1)}{N_r(N_r-\ell)}F_\alpha(\ell,N_r-\ell) \quad (27)$$

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

$$\tau_\alpha^{ISPE} = g_{ISPE}\chi^2_{h_{ISPE},\alpha} \quad (28)$$

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

The variance D^2, mean m, kurtosis K and skewness S of the first ℓ retained KPCs $t = [t_1,\dots,t_N]^T$, where $t_k = [t_{k1},\dots,t_{k\ell}]$; $k = 1,\dots,N$ are calculated by [34],

$$m_j = \frac{1}{\ell}\sum_{i=1}^{\ell} t_{ji} \quad (29)$$

$$D_j^2 = \frac{1}{\ell}\sum_{i=1}^{\ell}(t_{ji}-m_j)^2 \quad (30)$$

$$K_j = \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\frac{t_{ji}-m_j}{D_j^2}\right)^4 \quad (31)$$

$$S_j = \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\frac{t_{ji}-m_j}{D_j^2}\right)^3 \quad (32)$$

D. EL BASED INTERVAL REDUCED KPCA METHODS ($IRKPCA_{HKMEANS}$)

1) HIERARCHICAL CLUSTERING

Hierarchical clustering aims to group similar objects into groups called clusters [51]. We can compute the distance between two clusters as following,

- **Single Linkage**: Compute the minimum distance $d(s,c)$ between any single data point in the two S and C clusters:

$$D(S,c) = min(d(s,c)) \quad s \in S, \quad c \in C \quad (33)$$

- **Complete Linkage**: Compute the maximum distance between a S and C:

$$D(S,C) = max(d(S,C)) \quad s \in S, \quad c \in C \quad (34)$$

- **Ward's linkage**: Regroup the clusters in which the inertial losses within clusters $\Delta(S,C)$ at each step are decreased.

$$\Delta I(S,C) = \frac{m_S\, m_C}{m_S+m_C}\, d^2(g_S\, g_C) \quad (35)$$

where m_S and m_C are the total weight of the observations, g_S and g_C are the center of gravity of S and C,respectively, and $d^2(g_S\ g_C)$ is the Euclidean distance between g_S and g_C.

In this work, we use the Ward's linkage distance method [31]. Let consider the the original matrix $X = \left[x_1\ x_2 \cdots x_N\right]^T \in \mathbb{R}^{m\times N}\ i = 1, \ldots, N$. Using the Agglomerative hierarchical clustering [52], N_r clusters are obtained $\{C'_1, C'_2, \ldots, C'_{N_r}\}$ where $x_j \in C'_i\ j = 1, \ldots, n'_i,\ \ i = 1, \ldots, N_r$ with n'_i is the number of samples in C'_i.

2) K-MEANS CLUSTERING

K-means clustering is one of the simplest and popular machine learning techniques [53]. The main idea behind K-means clustering is to attributes samples to the cluster with a smallest distance between samples to centroid cluster. The objective of using K-means clustering is to improve the quality of the clusters result obtained using Agglomerative hierarchical clustering. It compute the squared distances between the data and centroids, and attributes data to the nearest centroid. We purpose to enhance the N_r clusters $\{C'_1, C'_2, \ldots, C'_{N_r}\}$ and we classify into N_r disjoint subsets $\{C_1, C_2, \ldots, C_{N_r}\}$ each containing n_i observations, where $x_j \in C_i\ j = 1, \ldots, n_i,\ \ i = 1, \ldots, N_r$ by the reduction of the mean-square-error cost function

$$E1 = \sum_{i=1}^{N_r} \sum_{x_j \in C_i} ||x_j - M_i||^2, \tag{36}$$

The resulting input data set obtained using H-K-means is given as,

$$X^r = \left\{x_1^r\ x_2^r \cdots x_{N_r}^r\right\} \tag{37}$$

where

$$x^r(i) = \frac{1}{n_i} \sum_{x_j \in C_i} x_j, \quad i = 1, \ldots, N_r \tag{38}$$

with $x_j \in \mathbf{R}^m,\ j = 1, \ldots, n_i$ and $N_r = \ell + 1, \ldots, N$.

3) FEATURE EXTRACTION AND SELECTION USING IRKPCA$_{HKMEANS}$

Let be consider a mapped interval valued data X^r defined as,

$$\mathcal{X}^r = \left[\phi(x^r(1))\ \phi(x^r(2))\ \cdots\ \phi(x^r(N_r))\right]^T \in \mathbb{R}^{N_r\times h} \tag{39}$$

The reduced kernel $K_r \in \mathbf{R}^{N_r\times N_r}$ is constructed as follows:

$$\begin{aligned} K^r &= \mathcal{X}^r(\mathcal{X}^r)^T \\ &= \begin{bmatrix} k(x_1^r, x_1^r) & \cdots & k(x_1^r, x_{N_r}^r) \\ \vdots & \ddots & \vdots \\ k(x^r(N_r), x_1^r) & \cdots & k(x_{N_r}^r, x_{N_r}^r) \end{bmatrix} \end{aligned} \tag{40}$$

The eigenvector λ^r and the corresponding eigenvalue α^r of the new reduced kernel matrix K^r are determined by solving the following equation:

$$\lambda^r \alpha^r = K^r \alpha^r \tag{41}$$

Next we extract and select the most significant features from the reduced interval valued data using IKPCA methods as given in section IV-C2.

E. FAULT CLASSIFICATION METHODS

During the classification stage, once the global characteristics are extracted and selected using the four proposed methods IKPCA$_{UL}$, IKPCA$_{CR}$,IRKPCA$_{UL}$, and IRKPCA$_{CR}$, they are used as input data for the proposed EL technique. Finally, to make efficient decisions, we compare the EL output results and choose the best one. The main steps of the proposed techniques are illustrated in Algorithm 1.

Algorithm 1 IRKPCA-EL Algorithm

Input: Collect the normal $N \times m$ interval data matrix X.

Offline phase

1. Normalize X to zero mean and unit variance,
2. Compute the reduced interval data matrix X',
3. Calculate the kernel matrix K,
4. Extract characteristics using IRKPCA technique,
5. Select the pertinent characteristics using IRKPCA method,
6. Present the selected characteristics as input to the EL model,
7. Classify the faults using EL model,
8. Determine the classification task,

Online phase

1. standardize a new samples using the mean and the variance computed from the training data,
2.Determine the kernel vector $k(x)$,
3. Extract the characteristics using IRKPCA technique,
4. Select the efficient characteristics using IRKPCA method,
5. Present the selected features as input to the EL classifier,
6. Classify the faults using EL model,
7. Determine the prediction model,
8. Compute the fault diagnosis results.

V. RESULTS AND DISCUSSIONS

This section presents the results and discussions of our experimental.

A. EVALUATION PARAMETERS

In this section, a set of emulated PV system data is used to assess the effectiveness of the proposed methods. The adopted criteria are: Normalized Classification Accuracy (NCA), which represents the number of correct predictions divided by the total number of input samples. Normalized Recall (NR), which represents the number of correct positive results divided by the number of all pertinent samples. Normalized Precision (NP), which represents the number of correct positive results divided by the number of positive results predicted by the classifier. Computation time (CT (s)) which represents the time needed to execute the algorithm.

TABLE 3. Global performances using EL, IEL_{CR} and IEL_{UL} methods.

Methods	Global Performances		
	Accuracy	Recall	Precision
Ensemble learning (EL)	0.48.	0.48	0.48
IEL_{CR}	0.72	0.71	0.72
IEL_{UL}	0.74	0.74	0.73

B. MULTI-CLASS (MC) CLASSIFICATION RESULTS

In this study, we use the minimum root mean-square error (RMSE) as a selection criterion for different ML classifiers. The 10-fold cross-validation approach was used to obtain the classification accuracy and to illustrate the efficiency of the proposed techniques for FDD purposes. For the proposed ensemble learning techniques, the DT was tested with 50 trees, the K and C parameters for SVM are selected with the lowest RMSE value and the K value for KNN is equal to 1, 3, and 5. For the FFNN, MNN, GRNN, CFNN, PNN, NN, RNN, and CNN classifiers, the number of hidden layers chosen is ten and the number of hidden neurons in the hidden layer is equal to 50.

The first step of this work aims to compare the performance of the presented interval IEL_{CR} and interval IEL_{UL} techniques, the results are compared to EL for single valued-data. The number of variables m equals to 9 and the number of samples N equals to 1501 for both IEL_{CR} and IEL_{UL} techniques. The results of the multi-class classification are summarized in the Table 3 where it can be distinctly noticed that the classification metrics obtained using the two proposed methods is higher than the one obtained using the EL for single-valued data. It is easy to conclude that the use of interval representation instead of a single value representation enhances the fault classification performance.

At the second phase, in order to further improve the performance of the proposed IEL-based techniques, novel EL-based frameworks ($IKPCA_{CR}$, $IKPCA_{UL}$, $IRKPCA_{CR}$ and $IRKPCA_{UL}$) were proposed. Firstly, the data set is normalized under normal operating modes. Secondly, the interval KPCA (IKPCA) and interval reduced KPCA (IRKPCA) models are constructed in which the cumulative percent variance (CPV) criterion is equal to 95% as confidence level. CPV criterion is adopted to retain the number of first kernel component ℓ. The reduced datasets (number of samples N) obtained through H-K-means equal to 806 and 800 have fed to respectively $IRKPCA_{CR}$ and $IRKPCA_{UL}$ techniques. The IKPCA models are structured by 31 interval kernel principal components (IKPCs) while the selected number of IKPCs using CPV criterion is equal to 18 and 17 using $IRKPCA_{CR}$ and $IRKPCA_{UL}$ models, respectively. To generate the simulation data 6 operating modes are considered. The operating modes include one healthy referred to class C0 and 5 faulty modes ($F_1 - F_5$) assigned to classes C1-C5 (Table 2). In this study, 5 groups of features are extracted and then we select the best one from them. Table 4 shows the performed groups of features.

TABLE 4. 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, kurtosis, variance and skewness of the ℓ retained IKPCs
Group 5	The first ℓ IKPCs

TABLE 5. Accuracies using $IKPCA_{CR}$-EL, $IKPCA_{UL}$-EL, $IRKPCA_{CR}$-EL, $IRKPCA_{UL}$-EL techniques.

Methods	Extracted Features				
	group 1	group 2	group 3	group 4	group 5
$IKPCA_{CR}$-EL	0.35	0.54	0.42	0.88	0.99
$IKPCA_{UL}$-EL	0.37	0.56	0.41	0.89	0.99
$IRKPCA_{CR}$-EL	0.35	0.57	0.43	0.90	1
$IRKPCA_{UL}$-EL	0.33	0.56	0.42	0.91	1

TABLE 6. Global performances using $IKPCA_{CR}$-RF, $IKPCA_{UL}$-RF methods.

Methods	Global Performances		
	NCA	NR	NP
$IKPCA_{CR}$-EL	0.99	0.98	0.98
$IKPCA_{UL}$-EL	0.99	0.99	0.98
$IRKPCA_{CR}$-EL	1	1	1
$IRKPCA_{UL}$-EL	1	1	1

The main goal of this part is to extract and select the most effective characteristics from raw data in order to obtain the best classification results. In the first stage, emulation data is used to collect and label the database in faulty mode. Then, we apply the labeled data as inputs for the proposed techniques. For this purpose, a comparison between five arbitrary groups using the proposed techniques is presented in Table 5. From this table, we can see that the proposed methods based on data reduction scheme can achieve higher accuracy using group 5 of features. Both EL-based methods provide an accuracy of 0.99 (EL-IKPCA) and 1 (EL-IRKPCA) using group 5 of features. As shown in Table 5, the accuracy of the proposed EL-based methods performed better the classification results comparing with IEL techniques. The overall accuracy can improve from about 0.72 using IEL to 1 using EL-based IRKPCA techniques. Additionally, to further evaluate the results, recall and precision classification metrics are used. As shown in Table 6, the proposed EL-based methods present perfect results in all used classification metrics.

Additionally, we used confusion matrix to more demonstrate the diagnosis performance of the proposed methods

TABLE 7. Confusion matrix of IKPCA$_{CR}$-EL and IKPCA$_{UL}$-EL classifiers using group 5.

Conf. Matrix		Predicted process statuses						NR
True classes	C0	1500	1	0	0	0	0	.99
	C1	0	1501	0	0	0	0	1
	C2	0	0	1500	1	0	0	.99
	C3	0	0	0	1501	0	0	1
	C4	0	0	0	0	1500	1	.99
	C5	0	0	0	0	0	1501	1
NP		1	.99	1	1	1	.99	**.99**

TABLE 8. Confusion matrix of IRKPCA$_{CR}$-EL and IRKPCA$_{UL}$-EL 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**

(see Tables 7 and 8). The confusion matrix represent the visualization of the performance of the proposed algorithms. The rows present instances in an actual class while the columns represent the instances in a predicted class. In addition, the confusion matrix represents the correct classified and mis-classified samples for the condition modes (C_0 to C_5). Referring to the results given in Tables 7 and 8, the proposed EL-based methods achieved the highest accuracy correctly identifying 1501 measurements among 1501 during the healthy case (C0). Furthermore, the NP is 1 and its recall is 1 for all different modes using both IRKPCA-EL during all faulty cases with 0 of misclassification. We can conclude from these results that the proposed methods are able to distinguish the six different modes and obtain good classification results.

To further evaluate the effectiveness of the proposed techniques, a comparative study between 14 machine learning (ML) methods is done. The ML techniques include the proposed methods, interval principal components analysis based EL (IPCA-based EL) [30], Feed-Foward Neural Network (FFNN) [54], Multiple Layers (MNN) [55], Generalized Regression Neural Network (GRNN) [56], Cascade Foward Neural Network (CFNN) [55], Probabilistic Neural Network (PNN) [8], Neural Network (NN) [54], Recurrent Neural Network (RNN) [57] and Convolutional Neural Network (CNN) [58]. Table 9 presents the results according to the NCA and computation time (CT). The classification outcomes, given in Table 9, demonstrate that the enhanced ensemble methods using IKPCA and IRKPCA models provide the best results in terms of NCA compared to other techniques. Besides, one can notice from Tables 9 that the results are significantly improved compared to the IPCA-based EL. IPCA-based EL classifier reached quite high performance, with an NCA value of 0.92 and with a misclassification rate equal to 0.08. From Table 9, it is shown that both IKPCA and IRKPCA improve the

TABLE 9. Comparative classification accuracy and computation time results using group 5.

Methods	Global Performance	
	NCA	CT(s)
IRKPCA$_{CR}$-EL	**1**	**105.2**
IRKPCA$_{UL}$-EL	**1**	**111.33**
IKPCA$_{CR}$-EL	**.99**	221.17
IKPCA$_{UL}$-EL	**.99**	233.78
IEL$_{CR}$	.72	59.12
IEL$_{UL}$	.74	59.12
IPCA-EL [54]	.92	**93.14**
FFNN [55]	.83	59.12
MNN [56]	.86	25.87
GRNN [57]	.69	35.96
CFNN [56]	.85	71.16
PNN [8]	.71	31.73
NN [55]	.72	**13.7**
RNN [58]	.84	267.1
CNN [59]	.76	389.16

TABLE 10. Multiple one class classifier logic for fault diagnosis.

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

feature extraction results and outperform the linear IPCA model because they can handle the nonlinearity of the PV system. Also, we can be noticed that the presented IEL classifier makes the performance of fault diagnosis efficient for fault classification. The IEL$_{CR}$ and IEL$_{UL}$ classifiers provide a classification NCA equal to 0.72% and 0.74%. A classification error of 0.28 is achieved using IEL$_{CR}$ and for IEL$_{UL}$, the misclassification is 0.26. The poor NCA using IEL$_{CR}$ and IEL$_{UL}$ are due to the use of measured variables without characteristics extraction and selection steps which indicates the effectiveness of the developed IKPCA-EL and IRKPCA-EL techniques to perform the classification task. In addition, we can conclude from the results summarized in Table 9 that the developed IRKPCA$_{CR}$-EL and IRKPCA$_{UL}$-EL methods afford the best tread-off between NCA and computation time (CT). Therefore, the proposed methods based on characteristics extraction and selection phases and data reduction scheme are considered as good alternatives for faults classification due to their high NCA and reliability. For FFNN, MNN, GRNN, CFNN, PNN, NN, and RNN classifiers, the best results in terms of NCA are obtained using MNN with NCA values of 0.86 and misclassification value of 0.14.

C. ONE-CLASS (OC) CLASSIFICATION RESULTS

To more highlight the effectiveness of the developed techniques a bank one class classifiers is presented. One

TABLE 11. NCA using group 5 with different one class classifiers.

Class	Methods			
	IKPCA$_{CR}$-EL	IKPCA$_{UL}$-EL	IRKPCA$_{CR}$-EL	IRKPCA$_{UL}$-EL
C0	.97	.98	1	1
C1	.96	.96	1	1
C2	.96	.97	1	1
C3	.95	.96	1	1
C4	.94	.96	1	1
C5	.96	.95	1	1
Mean	.95	.96	1	1

class classification is a specific type of classification task done by only instances of one class. In our case study, we apply one healthy and five faulty classes [59]. As shown in Table 10, each one is trained in order to classify a specific class labeled by 1 or -1. The performance of the proposed methods in terms of NCA is presented in Table 11 using the selected features of group 5. Classification results of all classifiers, given in Table 11, demonstrate the effectiveness of the proposed techniques based on feature extraction and selection steps thanks to the high ability of the proposed kernel-based methods to extract and select the most pertinent and significant characteristics from interval raw data.

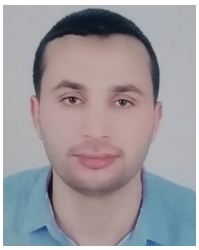
VI. CONCLUSION

New fault detection and diagnosis (FDD) techniques dealing with uncertain Grid-Connected Photovoltaic (PV) systems have been proposed in this paper. The uncertainty was addressed by using the interval-valued data representation. Firstly, two interval-valued ensemble learning (IEL) classifiers based on the direct application of the interval-valued dataset were proposed. Secondly, two enhanced IEL methods based on features extraction, selection, and fault classification steps were developed. For the features extraction and selection steps, two interval KPCA (IKPCA) methods were performed to extract and select the most significant features by transforming the single-valued data set into interval-valued latent variables. Then, the most pertinent characteristics were fed to the proposed EL technique for classification purposes. Finally, in order to further improve the diagnosis results in terms of computation time, an improved IEL techniques based on data reduction and interval KPCA (IRKPCA) were proposed. The proposed methods applied the Hierarchical K-means (H-K-means) clustering measure to remove the irrelevant and redundant samples. The simulation results using a grid-connected PV system under healthy and faulty conditions showed the impact of using interval-valued instead of single value representation and the effectiveness of the proposed techniques for features extraction and selection to provide the best compromise between diagnosis metrics and low computation time.

REFERENCES

[1] 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. 126286–126306, 2021.
[2] Y. Liu, K. Ding, J. Zhang, Y. Li, Z. Yang, W. Zheng, and X. Chen, "Fault diagnosis approach for photovoltaic array based on the stacked auto-encoder and clustering with I-V curves," *Energy Convers. Manage.*, vol. 245, Oct. 2021, Art. no. 114603.
[3] M. Mansouri, M.-F. Harkat, H. N. Nounou, and M. N. Nounou, *Data-Driven Model-Based Methods for Fault Detection Diagnosis*. Amsterdam, The Netherlands: Elsevier, 2020.
[4] V. S. B. Kurukuru, A. Haque, M. A. Khan, S. Sahoo, A. Malik, and F. Blaabjerg, "A review on artificial intelligence applications for grid-connected solar photovoltaic systems," *Energies*, vol. 14, no. 15, p. 4690, Aug. 2021.
[5] 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, 2016.
[6] K.-H. Chao, C.-T. Chen, M.-H. Wang, and C.-F. Wu, "A novel fault diagnosis method based-on modified neural networks for photovoltaic systems," in *Proc. Int. Conf. Swarm Intell.* Springer, 2010, pp. 531–539.
[7] V. S. B. Kurukuru, F. Blaabjerg, M. A. Khan, and A. Haque, "A novel fault classification approach for photovoltaic systems," *Energies*, vol. 13, no. 2, p. 308, Jan. 2020.
[8] B. Basnet, H. Chun, and J. Bang, "An intelligent fault detection model for fault detection in photovoltaic systems," *J. Sensors*, vol. 2020, pp. 1–11, Jun. 2020.
[9] 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.
[10] M. Malinowski, M. Jasinski, and M. P. Kazmierkowski, "Simple direct power control of three-phase PWM rectifier using space-vector modulation (DPC-SVM)," *IEEE Trans. Ind. Electron.*, vol. 51, no. 2, pp. 447–454, Apr. 2004.
[11] M. S. Ramakrishna and S. N. Singh, "Modeling of PV system based on experimental data for fault detection using kNN method," *Sol. Energy*, vol. 173, pp. 139–151, Oct. 2018.
[12] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018.
[13] W. Wu, J. Liu, H. Rong, H. Wang, and M. Xian, "Efficient k-nearest neighbor classification over semantically secure hybrid encrypted cloud database," *IEEE Access*, vol. 6, pp. 41771–41784, 2018.
[14] Y. Zhao, L. Yang, B. Lehman, J.-F. de Palma, J. Mosesian, and R. Lyons, "Decision tree-based fault detection and classification in solar photovoltaic arrays," in *Proc. 27th Annu. IEEE Appl. Power Electron. Conf. Expo. (APEC)*, Feb. 2012, pp. 93–99.
[15] Y. Zhao, J.-F. de Palma, J. Mosesian, R. Lyons, and B. Lehman, "Line–line fault analysis and protection challenges in solar photovoltaic arrays," *IEEE Trans. Ind. Electron.*, vol. 60, no. 9, pp. 3784–3795, Sep. 2013.
[16] 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.
[17] 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.
[18] L. I. Kuncheva, J. J. Rodríguez, C. O. Plumpton, D. E. J. Linden, and S. J. Johnston, "Random subspace ensembles for fMRI classification," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 531–542, Feb. 2010.
[19] M. Skurichina and R. P. W. Duin, "Bagging, boosting and the random subspace method for linear classifiers," *Int. J. Pattern Anal. Appl.*, vol. 5, no. 2, pp. 121–135, 2002.
[20] J. Shin, "Random subspace ensemble learning for functional near-infrared spectroscopy brain-computer interfaces," *Frontiers Hum. Neurosci.*, vol. 14, p. 236, Jul. 2020.
[21] M. Zounemat-Kermani, O. Batelaan, M. Fadaee, and R. Hinkelmann, "Ensemble machine learning paradigms in hydrology: A review," *J. Hydrol.*, vol. 598, Jul. 2021, Art. no. 126266.
[22] Y. Zhu, C. Xie, G.-J. Wang, and X.-G. Yan, "Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance," *Neural Comput. Appl.*, vol. 28, no. S1, pp. 41–50, Dec. 2017.
[23] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," 2021, *arXiv:2104.02395*.
[24] V. Le, X. Yao, C. Miller, and T.-B. Hung, "Arc fault detection in DC distribution using semi-supervised ensemble machine learning," in *Proc. IEEE Energy Convers. Congr. Expo. (ECCE)*, Sep. 2019, pp. 2939–2945.

[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] A. Emami-Naeini, M. M. Akhter, and S. M. Rock, "Effect of model uncertainty on failure detection: The threshold selector," *IEEE Trans. Autom. Control*, vol. 33, no. 12, pp. 1106–1115, Dec. 1988.
[27] V.-L. Nguyen, S. Destercke, and M.-H. Masson, "K-nearest neighbour classification for interval-valued data," in *Proc. Int. Conf. Scalable Uncertainty Manage.* Springer, 2017, pp. 93–106.
[28] L. V. Utkin and F. P. Coolen, "Interval-valued regression and classification models in the framework of machine learning," in *Proc. ISIPTA*, vol. 11, 2011, pp. 371–380.
[29] J. Pan, W. He, Y. Shi, R. Hou, and H. Zhu, "Uncertainty analysis based on non-parametric statistical modelling method for photovoltaic array output and its application in fault diagnosis," *Sol. Energy*, vol. 225, pp. 831–841, Sep. 2021.
[30] 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.
[31] R. Fezai, K. Dhibi, M. Mansouri, M. Trabelsi, M. Hajji, K. Bouzrara, H. Nounou, and M. Nounou, "Effective random forest-based fault detection and diagnosis for wind energy conversion systems," *IEEE Sensors J.*, vol. 21, no. 5, pp. 6914–6921, Mar. 2021.
[32] 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, 2020.
[33] 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.
[34] 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.
[35] 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.
[36] L. Auria and R. A. Moro, "Support vector machines (SVM) as a technique for solvency analysis," 2008.
[37] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.* Springer, 2003, pp. 986–996.
[38] B. Gupta, A. Rawat, A. Jain, A. Arora, and N. Dhami, "Analysis of various decision tree algorithms for classification in data mining," *Int. J. Comput. Appl.*, vol. 163, no. 8, pp. 15–19, Apr. 2017.
[39] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1997.
[40] L. Breiman, *Classification and Regression Trees*. Evanston, IL, USA: Routledge, 2017.
[41] 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, Jun. 2019.
[42] P. Müller, K. Salminen, V. Nieminen, A. Kontunen, M. Karjalainen, P. Isokoski, J. Rantala, M. Savia, J. Väliaho, P. Kallio, and J. Lekkala, "Scent classification by K nearest neighbors using ion-mobility spectrometry measurements," *Expert Syst. Appl.*, vol. 115, pp. 593–606, 2019.
[43] Y. Ren, P. N. Suganthan, and N. Srikanth, "Ensemble methods for wind and solar power forecasting—A state-of-the-art review," *Renew. Sustain. Energy Rev.*, vol. 50, pp. 82–91, Oct. 2015.
[44] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: CRC Press, 2019.
[45] L. I. Kuncheva, *Combining Pattern Classifiers: Methods Algorithms*. Hoboken, NJ, USA: Wiley, 2014.
[46] S. Pathical and G. Serpen, "Comparison of subsampling techniques for random subspace ensembles," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Jul. 2010, pp. 380–385.
[47] M. A. O. Domingues, R. M. C. R. de Souza, and F. J. A. Cysneiros, "A robust method for linear regression of symbolic interval data," *Pattern Recognit. Lett.*, vol. 31, no. 13, pp. 1991–1996, Oct. 2010.
[48] L. Billard, "Dependencies and variation components of symbolic interval-valued data," in *Selected Contributions in Data Analysis and Classification*. Springer, 2007, pp. 3–12.
[49] 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.
[50] 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.
[51] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *Wiley Interdiscipl. Rev. Data Mining Knowl. Discovery*, vol. 2, no. 1, pp. 86–97, 2012.
[52] F. Murtagh and P. Contreras, "Methods of hierarchical clustering," 2011, *arXiv:1105.0121*.
[53] S. Na, L. Xumin, and G. Yong, "Research on K-means clustering algorithm: An improved K-means clustering algorithm," in *Proc. 3rd Int. Symp. Intell. Inf. Technol. Secur. Informat.*, Apr. 2010, pp. 63–67.
[54] C. Lv, Y. Xing, J. Zhang, X. Na, Y. Li, T. Liu, D. Cao, and F.-Y. Wang, "Levenberg–marquardt backpropagation training of multilayer neural networks for state estimation of a safety-critical cyber-physical system," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3436–3446, Aug. 2018.
[55] B. Warsito, R. Santoso, and H. Yasin, "Cascade forward neural network for time series prediction," *J. Phys., Conf.*, vol. 1025, May 2018, Art. no. 012097.
[56] S. Alby and B. Shivakumar, "A prediction model for type 2 diabetes using adaptive neuro-fuzzy interface system," *Biomed. Res.*, 2018.
[57] S. Zhou, M. Mao, L. Zhou, Y. Wan, and X. Xi, "A shadow fault diagnosis method based on the quantitative analysis of photovoltaic output prediction error," *IEEE J. Photovolt.*, vol. 10, no. 4, pp. 1158–1165, Jul. 2020.
[58] 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.
[59] N. Seliya, A. A. Zadeh, and T. M. Khoshgoftaar, "A literature review on one-class classification and its potential applications in big data," *J. Big Data*, vol. 8, no. 1, pp. 1–31, Dec. 2021.

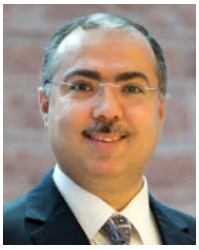


KHALED DHIBI received the Ph.D. degree in electronics and microelectronics engineering from the Faculty of Sciences of Monastir (FSM), University of Monastir, Monastir, Tunisia. His work focuses on the implementation of data-driven techniques for fault detection and diagnosis of industrial processes.

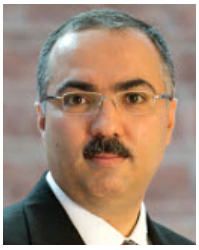
MAJDI MANSOURI (Senior Member, IEEE) received the degree in electrical engineering from SUPCOM, Tunis, Tunisia, in 2006, the M.Sc. degree in electrical engineering from ENSEIRB, Bordeaux, France, in 2008, the Ph.D. degree in electrical engineering from UTT Troyes, France, in 2011, and the H.D.R. (Accreditation to Supervise Research) degree in electrical engineering from the University of Orleans, France, in 2019. He joined the Electrical 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.

KAMALELDIN ABODAYEH received the M.Sc. degree in functional analysis from University College Dublin and the Ph.D. degree from University College Cork, Ireland, in 1997. He had his Postdoctoral Research with the Department of Process Engineering, University College Cork. Since 2001, he has been with Prince Sultan University, Saudi Arabia. He has published more than 60 articles in various areas of pure and applied mathematics. His research interests include functional analysis, theoretical physics, discrete potential theory, fixed point theory, quality monitoring, and statistical hypothesis testing.

KAIS BOUZRARA is currently a Professor in 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 the area of systems engineering and control, with emphasis on process modeling, monitoring, and estimation.



HAZEM NOUNOU (Senior Member, IEEE) is currently a Professor in electrical and computer engineering with Texas A&M University at Qatar. He has more than 19 years of academic and industrial experience. He has served as an associate editor and on the technical committees of several international journals and conferences. He has significant experience in research on control systems, databased 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 journals and conference papers and book chapters.



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