Monitoring of photovoltaic systems using improved kernel-based learning schemes

Fouzi Harrou, Member, IEEE, Ahmed Saidi, Ying Sun and Sofiane Khadraoui

Abstract—Data-based procedures for monitoring the operating performance of a PV system are proposed in this paper. The only information required to apply the procedures is the availability of system measurements, which are routinely on-line collected via sensors. Here, kernel-based machine learning methods, including support vector regression (SVR) and Gaussian process regression (GPR), are used to model multivariate data from the PV system for fault detection due to their flexibility and capability to nonlinear approximation. Essentially, the SVR and GPR models are adopted to obtain residuals to detect and identify occurred faults. Then, residuals are passed through an exponential smoothing filter to reduce noise and improve data quality. In this work, a monitoring scheme based on kernel density estimation is used to sense faults by examining the generated residuals. Several different scenarios of faults were considered in this study, including PV string fault, partial shading, PV modules short-circuited, module degradation, and line-line faults on the PV array. Using data from a 20 MWp grid-connected PV system, the considered faults were successfully traced using the developed procedures. Also, it has been demonstrated that GPR-based monitoring procedures achieve better detection performance over SVRs to monitor PV systems.

Index Terms—Temporary shading, Photovoltaic systems, Kernel-based methods, KDE, Nonlinear regression methods, Fault detection.

I. INTRODUCTION

RENEWABLE energy has become essential day-to-day to reduce the issue of fossil energy depletion and mitigate issues caused by CO₂ emissions such as environmental pollution, greenhouse effect, and climate change. Over the last years, the energy generated by photovoltaic systems has received considerable attention and has become a highly recommended energy source in residential, commercial, and industrial applications [1]. According to the International Renewable Energy Agency (IRENA), the total renewable energy generation capacity reached 2,351 GW at the end of 2018 [2].

Faults and failures in PV systems that can greatly affect their performance and increase power losses all the time are essentially unavoidable. Different factors can affect the performance and lifetime of PV systems if they are not detected on time, including maximum power point tracking error, shading, dust accumulation in PV modules, and failure in other PV elements. Presently, numerous research carried out over the years investigating different faults [3], [4], such as line-line (LL) fault which can be generated by an accidental short circuit between two points in the PV array [5], line-ground (LG) [6], and arc faults [7], open circuit faults, hotspots, connection failures, and PV array degradation faults [8], [9]. Faults can occur in the AC and DC sides of the PV system. Authors in [10] compared some of the advanced monitoring techniques and reviewed types of faults, sensor requirements, procedural complexity, and level of protection. Even though advanced PV systems are reliable, the consequences of some challenging failures are too much costly and may create critical safety problems [11], [12]. As an illustration of severe PV systems failures, two PV systems in the US (a 383 KWp PV array in Bakersfield, CA, and a 1.208 MWp power plant in Mount Holly, NC) burned in 2009 and 2011, respectively [11]. Thus, modern PV systems require an efficient monitoring scheme to increase efficiency and productivity and guarantee their implementation in real-time. Essentially, the main objective of the monitoring system is to inspect the operating conditions of a PV system over time to verify they are remaining within the desired specifications.

The success of PV systems operation depends considerably on the effectiveness and sensitivity of the selected monitoring system [13]. Over the last decade, there has been an increasing interest in fault detection and isolation methods in both academic research and industrial applications [14], [15]. Numerous fault detection techniques have been designed to improve PV system monitoring, which could generally be classified into model-based and data-based methods. Model-based fault detection methods are based on the use of analytical models of the system under investigation, which are generally designed using some fundamental knowledge of the system when operating correctly. Traditionally, fault detection within model-based is established into two main steps: residual generation and residual assessment. Thus, faults are uncovered by evaluating the residuals, which are the disparity between measured values and their estimates computed using the analytical model. Specifically, residuals are compared to a prefixed detection threshold, and if the threshold is surpassed, then a fault is declared. There are many model-based approaches to PV system modeling and monitoring based on the one-diode model, and overviews on this topic are readily available in the literature [16], [17]. For instance, in [18], a linear regression model has been used to generate residuals, and then the Support Vector Machine is applied to residuals for fault
detection in a PV system. However, linear regression cannot capture nonlinear features and can lose pertinent information for fault detection. In [19], an online fault detection procedure for PV systems has been proposed by combining the one-diode model and the 3-sigma monitoring scheme (also called Shewhart scheme). However, the Shewhart scheme used to evaluate residuals is not suited to detect small and moderate changes because its decision is based on actual observation alone. In [14], fault detection and isolation approach has been proposed to detect and identify the type of faults in the DC side of a PV system. At first, this approach generates residuals using a simulation model based on the one-diode model. Model parameters are determined using optimization methods by minimizing the deviation between measured data and the model’s estimated data and considering weather conditions and the datasheet from the PV module manufacturer [14]. Then, residuals of MPP current, voltage, and power are verified by the univariate and multivariate exponentially weighted moving average (EWMA) schemes for uncovering faults and partial shading [20]. In [21], the single one diode model and the cumulative sum (CUSUM) scheme are amalgamated to monitor the DC side of the PV system. However, detection thresholds in the previously discussed methods are computed based on the Gaussian assumption of residuals’ distribution. To bypass this limitation, in [9], nonparametric monitoring approaches based on k-nearest neighbors (kNN) algorithm are introduced. The attractive features of these distribution-free approaches are their ability to handle data with any kind of distribution.

In data-driven methods (also called process history-based approaches), only past data gathered when the inspected system is running within the desired specification is needed to design an empirical reference model that reflects the healthy operating conditions. Then, fault detection procedures are performed accordingly. In general, the model is first obtained, and then fault detection procedures are performed accordingly. The residual space estimation, which is crucial in fault detection, depends on the appropriate process modeling. Data-driven fault detection methods comprise computational intelligence and machine learning methods. Recently, several machine learning procedures have been applied for fault detection and diagnosis in PV systems including neural networks [25], convolutional neural network [26], neuro-fuzzy [27], fuzzy logic [28], probabilistic neural network [29], and decision trees [30]. Recently in [8], a data-driven method based on principal component analysis (PCA) has been proposed to detect faults impacting the performance of PV systems. This approach uses PCA to generate residuals and apply multivariate monitoring, including $T^2$ and squared predicted error for fault detection. The principal characteristic of this approach is its flexibility and ability to build an empirical model of the system, which is then used to detect anomalies in future data. However, the PCA model ignores the time dependence on PV systems measurements and cannot capture nonlinearity.

The accurate modeling of PV system is critical to developing a monitoring system for PV system control and management. Generally speaking, the success of monitoring systems depends mainly on the accuracy of the used modeling approach and the sensitivity of the anomaly detection scheme. Recently kernel-based machine learning methods, such as support vector regression (SVR) and gaussian process regression (GPR), have gained particular attention in modeling multivariate input-output data due to their capacity and superior nonlinear approximation [31]. Essentially, SVR and GPR models are flexible and efficient tools to describe implicit correlations between a set of variables, which make them incredibly valuable to handle challenging nonlinear regression problems. A particularly attractive characteristic of GPR and SVR is their assumption-free and do not require assumptions on the distribution underlying the data. In this paper, GPR and SVR models were adopted to model multivariate data from a PV system. An innovative fault detection technique based on the kernel-based models (GPR and SVR) and a kernel density estimation (KDE)-based monitoring scheme is introduced for monitoring PV systems. The design of this fault detection scheme is performed in two phases. In the first phase, prediction models (GPR and SVR) of PV array are built on fault-free historical data gathered when the PV system was operating conform to the desired specifications. Here, solar irradiance and ambient temperature are used as input for prediction models and generate the residuals for fault detection. Residuals reflect the operating conditions of the inspected photovoltaic array; in the absence of failures, the residuals are close to zero, while residuals become larger in the case of faulty conditions. The second phase consists of applying the designed prediction models to generated residuals based on testing data and apply the KDE-based monitoring scheme to residuals to check the presence of possible failures and inspect if the PV array follows the desired performance. Because the measured data are often assumed to be noisy, data smoothing through data averaging is the primary statistical tool for removing random noise. Here, we also apply an exponentially-weighted moving average (EWMA) filter to the residual data to reduce high-frequency noise. This also permits us to incorporate information from past data as EWMA exponentially smooths the residuals. However, amalgamating these kernel-based models with a monitoring scheme based on the Gaussian distribution could deteriorate their detection performance. To alleviate this limitation, in this paper, KDE has been used to set a detection threshold for GPR and SVR-based PV system monitoring, which extends the flexibility of the developed methods. To verify the efficiency of the proposed schemes, we are using data collected from an actual operating 20 MWp grid-connected PV system located in the province of Adrar, Algeria. The detection performance of the resulting GPR and SVR-based KDE monitoring schemes were then compared with the GPR, whose detection threshold is based on the Gaussian distribution. Results obtained showed that
the GPR-KDE monitoring scheme produces better monitoring performance than SVR-KDE with different kernels.

The present paper is organized as follows. Section II describes briefly the two kernel-based models (GPR and SVR) and the used data. In Section III, PV system modeling results based on GPR and SVR models are presented. Section IV discusses the monitoring results and Section V concludes this study.

II. MATERIALS AND METHODS

In this section, we briefly describe the basic concept of two important Kernel-based models, namely GPR and SVR regression models, and for more details, the reader details can refer to [32], [33], [34]. We then introduce the measurements of effectiveness adopted to evaluate the investigated models. Here, the real data from the Adrar PV system used to assess the considered methods are described.

A. Methods

1) Gaussian Process Regression (GPR): GPR is an important kernel-based learning algorithm that has attracted attention in modeling multivariate data [31], [35], [36]. Essentially, GPR is an efficient tool for exploring implicit relationships between multiple process variables based on training data, which makes GPR particularly helpful in dealing with challenging nonlinear regression. It should be noted that the most attractive feature of the GPR is its distribution-free learning, which makes it powerful in handling different kinds of data that do not necessarily follow Gaussian distribution [37]. GPR is a non-parametric Bayesian approach that can be used to solve nonlinear regression problems [38]. For a regression problem, in GPR, the response $y$ of a function $f$ at the input $x$ is expressed as

$$ y_i = f(x_i) + \epsilon_i. $$

where $\epsilon \sim \mathcal{N}(0, \sigma^2_{\epsilon})$. It should be noted that this is similar to the hypothesis in ordinary linear regression, where the output (dependent variable $y$) consists of independent inputs $x$ contaminated with additive noise $\epsilon$.

GPR is a Bayesian nonparametric regression characterized by its flexibility and its capacity to offer uncertainty estimates by assuming a Gaussian Process (GP) prior to the regression functions, $f(x)$ [39], [40]. Assume $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ is a given input-output data and $f(\cdot)$ to be approximated is supposed to follow a Gaussian process. The prior GP distribution is expressed as:

$$ f(x) \sim GP(m(x), k(x, x')). $$

GP is fully characterized by its mean function, $m(x)$, and kernel or covariance function, $k(x, x')$.

$$ m(x) = \mathbb{E}[f(x)] $$

$$ k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] $$

Notice that within the mean function and kernel function, the regression process can be adjusted and managed for modeling the evolution of the physical process at hand. Mainly, the kernel function is used to describe the data structure and capture the correlation between data points in a data set.

Assume that the observed $y_i$ values $[y_1, y_2, \ldots, y_n]^T$ are finite values of the function $f(\cdot)$ tainted with noises. Thus, $y_i$'s follow a joint Gaussian distribution [40]:

$$ y = [y_1, y_2, \ldots, y_n]^T \sim \mathcal{N}(m(x), \Sigma), $$

where $m(x) = [m(x_1), m(x_2), \ldots, m(x_n)]^T$ refers to the vector of mean $m(\cdot)$ which frequently are zero when the process is under control, $\Sigma$ is the $n \times n$ covariance matrix with $(i, j)^{th}$ element $\Sigma_{ij} = k(x_i, x_j)$. For a GPR model, $k(x, x)$ is commonly called a kernel function, which defines the generalization capabilities of the model [32]. Generally speaking, there is no automatic way to select a suitable kernel. The selection of a suitable kernel is generally based on hypotheses like smoothness and likely patterns to be suspected in the data. Two popular used kernel functions are given as follows.

- A squared exponential (SE) function

$$ k_{SE}(x_i, x_j) = \theta_1 \exp \left( \frac{(x_i - x_j)^2}{\theta_2} \right). $$

- A linear kernel function

$$ k_{Lin}(x_i, x_j) = \theta_1^2 + \theta_2^2 (x_i - \theta_3)(x_j - \theta_3). $$

At the training the model step, the optimal values of the kernel parameters (called hyper-parameters) are obtained by maximizing the negative log marginalized likelihood, $L(\theta)$.

$$ \theta_{opt} = \arg \max_{\theta} L(\theta) $$

where $\theta = [\theta_1, \theta_2, \ldots]$ denotes kernel parameters, the mean values $m(\cdot)$ are taken to be zero, and

$$ L(\theta) = \log[\mathbb{P}(y|X, \theta)], $$

$$ = - \frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_{y}| - \frac{1}{2} y^\top \Sigma_{y}^{-1} y, $$

where $\Sigma_{yy} = K_{xy} + \sigma^2 I$ denotes the covariance matrix of the noisy output $y$, and $|.|$ refers to the determinant of a matrix. The marginal log-likelihood can be regarded as a penalized fit metric composed of three terms. The term, $-\frac{1}{2} y^\top \Sigma_{y}^{-1} y$ quantifies how well the current kernel parametrization can explain the dependent variable, and $-\frac{1}{2} \log |\Sigma_{y}|$ refers to a complexity penalization term, and $-\frac{1}{2} \log(2\pi)$ denotes a normalization constant. Hyper-parameters can be computed by maximizing the marginal likelihood in (10) with respect to $\theta$ using any general gradient-based optimization algorithms, such as gradient descent [39]. The gradient of $L(\theta)$ with respect to $\theta$ is expressed as

$$ \frac{\partial L(\theta)}{\partial \theta_j} = - \frac{1}{2} y^\top K_{xy}^{-1} y - \frac{1}{2} y^\top (K_{xy}^{-1} \partial K_{xy} / \partial \theta_j), $$

$$ = - \frac{1}{2} y^\top \left( \alpha \alpha^\top - K_{xy}^{-1} \frac{\partial K_{xy}}{\partial \theta_j} \right) $$

where $\alpha = K_{xy}^{-1} y$. Recently, a fully Bayesian-based estimation framework, which provides uncertainty estimates for the
Assume that we observed \( D_t = X_t, y_t \), we need making prediction for new inputs \( X_s \) by drawing \( f_s \) from the posterior distribution \( p(f|D_t) \). The joint distribution of the observed data \( y \) and the latent noise-free function on the test points \( f_s = f(X_s) \) is expressed as:

\[
\begin{bmatrix} y \\ f_s \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mu(X) \\ \mu(X_s) \end{bmatrix}, \begin{bmatrix} K + \sigma_n^2 I & K_s \\ K_s^T & K_{ss} \end{bmatrix} \right)
\]

(12)

where \( K = k(X, X) \in \mathbb{R}^{n \times n} \) refers to the covariance matrix between all data points, \( K_s = k(X, X_s) \in \mathbb{R}^{n \times n_s} \) denotes the sub-matrix of the covariances assessed at all pairs of training and testing observations, and \( K_{ss} = k(X_s, X_s) \in \mathbb{R}^{n_s \times n_s} \) denotes the sub-matrix of covariances assessed at all pairs of testing observations.

The GP posterior \( p(f|X, y) \) can be obtained by using the GP prior \( p(f|X) \) and after observing some data points \( y \). The posterior could then be employed for predictions \( f_s \) given new input \( X_s \) [39], [42]:

\[
p(f_s|X_s, X, y) = \int p(f_s|X_s, f)p(f|X, y) \, df = \mathcal{N}(f_s|\mu_s, \Sigma_s)
\]

(13)

\[
\mu_s = \mu(X_s) + K_s^T(K + \sigma_n^2 I)^{-1}(y - \mu(X))
\]

(14)

\[
\Sigma_s = K_{ss} - K_s^T(K + \sigma_n^2 I)^{-1}K_s
\]

(15)

Overall, the main steps performed for GPR-based prediction are as follows.

- **Step 1**: Estimate the hyper-parameters (i.e., those of the used kernel and variance of the noise) in the GP model by maximizing the negative log marginalized likelihood in (10) based on training data \( (X, y) \).
- **Step 2**: Compute the predictive posterior distribution of GPR for the new inputs \( X_s \) by using Eq. (14) and (15). The mean of the predictive distribution in (14) is used as the predicted value.

More details about GPR model can be found in [32], [42].

2) **Support vector regression (SVR) approach**: This section is devoted to another important kernel-based learning model termed SVR, which is designed based on the principle of structural risk minimization [33], [34], [43]. Similar to GPR, the SVR model provides superior learning capability by using kernel tricks and does not require the specification of data distribution. Furthermore, SVR possesses good capacity in dealing with limited samples [34]. Essentially, SVR models are designed based on the principle of structural risk minimization. This approach has been widely applied in different applications, such as solar irradiance modeling [35], wind power prediction [36], and anomaly detection [44]. Also, SVR has another important characteristic that ensures a globally optimal solution for a given training data. In ordinary regression modeling, the goal is to find out the relationship between independent and dependent variables by minimizing the error, while the main objective in SVR is to fit the error within a certain threshold. Essentially, in the \( \epsilon \)-SVR, the aim is to model a function \( f(x) \) that represents the relation between inputs and outputs. Indeed, in SVR, \( f(x) \) is determined so that it is close to the current target \( y_i \) by \( \epsilon \) for the whole training data, and it is as smooth as possible [34]. Here, the errors are acceptable if they are less than \( \epsilon \); however, any divergence larger than \( \epsilon \) is not accepted. Of course, a precise model of \( f(x) \) enables predicting the output for many possible input values.

Assume that we are concerned about the relationship between a response variable \( y \) and a predictor \( x \). The SVR model maps the input vector \( (x_i \in \mathbb{R}^n) \) to a higher-dimensional space called a feature space. This mapping is performed by nonlinear mapping \( \phi \), and then linear regression is applied in this feature space. The regression function is expressed as:

\[
f(x) = w^T\phi(x) + b
\]

(16)

The weight vector \( w \) and the bias \( b \) are computed by minimizing the following regularized risk function [34].

\[
\min_w \frac{1}{2}\|w]\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*)
\]

(17)

s.t. \[
\begin{align*}
& y_i - (w^T\phi(x_i) - b) \leq \epsilon + \xi_i, \\
& w^T\phi(x_i) + b - y_i \leq \epsilon + \xi_i^*, \\
& \xi_i, \xi_i^* \geq 0, i = 1, \ldots, l.
\end{align*}
\]

where \( C > 0 \) is a constant defining the trade-off between the flatness of \( f \) and the margin of tolerance. In other words, it has an impact on the degree of penalty loss when an error happens in training. In this convex optimization problem, the decision function is determined by expecting that almost all of the data are within the zone between \( f(x) + \epsilon \) and \( f(x) - \epsilon \). However, when a data pair \( (x_i; y_i) \) is located outside of this zone, the slack variables \( \xi_i \) and \( \xi_i^* \) minimize the objective function. Notably, the slack variables for measuring the disparity between the point and marginal hyperplane. In other words, they are adopted for quantifying the level of error. Only the observations that are the interval of tolerance are taken to obtain an appropriate fitting model. Of course, the minimization of Equation 17 enables minimizing the learning machine confidence interval, \( \frac{1}{2}\|w\|^2 \), and the empirical risk, \( C \sum_{i=1}^l (\xi_i + \xi_i^*) \), respectively. In [45], least squares support vector regression has been introduced for solving the quadratic programming problem in (17) by solving a set of linear equations.

The dual of Equation (17) would be computed as [34],

\[
\begin{align*}
\max & \quad \frac{1}{2} \sum_{i,j=1}^m (\alpha_i - \alpha_j^*)(\alpha_i - \alpha_j^*)K(x_i, x_j) \\
& \quad - \epsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*) + \sum_{i=1}^m y_i(\alpha_i - \alpha_i^*)
\end{align*}
\]

(18)

such that \( \sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0, 0 \leq \alpha_i, \alpha_i^* \leq C \). Here, \( \alpha_i \) and \( \alpha_i^* \) denotes Lagrange multipliers [34].

Lastly, by using the kernel function, the equation (16) is formulated as:

\[
f(x) = \sum_{i=1}^N (\alpha_i - \alpha_i^*)k(x_i, x_j) + b
\]

(19)

where \( K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j) \) is the kernel function applied to transform a lower-dimensional data inputs into a higher-
dimensional space. Mapping input variables into a feature space offers considerably more flexibility and allows one to model functions of any shape. The radial basis function kernel is one of the commonly used kernels, which is expressed as: RBF kernel:

$$K(x_i, x_j) = \exp\left(\frac{||x_i - x_j||^2}{2\sigma^2}\right),$$  

(20)

where $\sigma$ is the width of Gaussian kernel. For more details about SVR models, refer to [46], [34].

These two kernel-based learning models (GPR and SVR) will be used for input-output data of a PV system and generate residuals that will be used for anomaly detection purposes.

3) Evaluation metrics: The prediction performance of each of the learning models is evaluated in terms of the following indexes: coefficient of determination ($R^2$), Root Mean Square Error (RMSE), mean absolute error (MAE), explained variance (EV), and mean absolute percentage error (MAPE). Where

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$  

(21)

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n}(\hat{y}_t - y_t)^2},$$  

(22)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|,$$  

(23)

$$EV = 1 - \frac{\text{Var}(\hat{y})}{\text{Var}(y)},$$  

(24)

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{\hat{y}_t - y_t}{y_t} \right| \%	ext{,}$$  

(25)

$y_i$ are the actual values, $\hat{y}_i$ are the corresponding estimated values, and $n$ is the number of measurements. Lower RMSE, MAE, or MAPE values represent more accurate prediction performances.

B. Dataset

The proposed fault detection strategy has been experimentally tested using practical data gathered from an actual grid-connected PV system of 20 MWp located at the province of Adrar. This city is a south-eastern Saharian province ranked as the second-largest Algerian province with a total area of 427,386 km$^2$ [47]. This province has a huge solar potential, an average of 5.7 kWh/m$^2$/day [48], that could satisfy the total energetic request of the entire city if it is efficiently exploited.

The PV system under study has started operating from December 10th, 2015, while its output generated power is directly injected to the grid utility. The PV system output DC voltage is 480-850 V while its output AC voltage is 315 V. Adrar PV plant is a huge system that includes 20 PV arrays of 1 MW each. Each array is composed of 93 sub-arrays of 44 YINGLI (YL245P-29b) application class (A) PV modules each. The sub-arrays outputs are connected to a 2*550 KW PV grid-connected inverter SUNGROW (5SG 500MX). There is a step-up transformer (SUNTEN ZBW10A-1250/30/0.315-0.315) with 1250 KVA of nominal capacity and 50HZ nominal frequency between inverter and the grid. The irradiance measurements are gathered according to a Kipp & Zonen CMP21 pyranometer with expected daily uncertainty $<2\%$. While the temperature level is measured by a J-type thermocouple, with $\pm1.1^\circ\text{C}$ accuracy. The system is also equipped with a wind speed sensor type WE-100 sonic anemometer with 0.05m/s resolution. The used data is gathered from January 2017 to March 2017 with 10 min temporal resolution.

Lastly, DC and AC electrical measurements from the inverter and meteorological data from the used sensors are then transmitted to a sophisticated Supervisory Control and Data Acquisition (SCADA) control system, which utilizes computers, graphical user interfaces and communication tools of the networked data to supervise the system performance. Table I presents the PV module electrical parameters measured at the standard test condition (STC), i.e., a temperature level of 25$^\circ\text{C}$ and irradiance measurement of 1000 W/m$^2$. The proposed fault detection methods are applied to monitor the PV array that contains 44 YINGLI (YL245P-29b) PV modules. This PV array is composed of two parallel PV strings of 22 PV modules in series each.

TABLE I: The main electrical parameters of YINGLI (YL245P-29b) PV module at STC.

<table>
<thead>
<tr>
<th>Peak power (W)</th>
<th>Voltage at maximum power point (V)</th>
<th>Current at maximum power point (A)</th>
<th>Open circuit voltage (V)</th>
<th>Short circuit current (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>245</td>
<td>29.6</td>
<td>8.28</td>
<td>37.5</td>
<td>8.83</td>
</tr>
</tbody>
</table>

The collected data contains peak power as a response variable and two observed predictors, namely ambient temperature and solar irradiance. Figure 1 displays the heatmap of the lower triangular half of the Pearson correlation matrix for the training dataset. It allows for understanding the relationship between the input and output variables [49]. By definition, the Pearson’s correlation coefficient between two random variables $X_i$ and $X_j$ is computed as follows:

$$\rho_{X_i,X_j} = \frac{E[(X_j - \mu_{X_j})(X_j - \mu_{X_j})^T]}{\sigma_{X_i}\sigma_{X_j}},$$  

(26)

where $\mu_{X_i}$ and $\mu_{X_j}$ are the mean of $X_i$ and $X_j$, respectively. The $m \times m$ matrix of all pairwise correlation coefficients is called the correlation matrix of $X$. Figure 1 indicates that there is a moderate positive correlation between ambient temperature and solar irradiance. This is because the ambient temperature is directly related to solar radiation; a rise in solar radiation increases the ambient temperature. Also, It can be observed that there is a moderate positive correlation between ambient temperature and solar irradiance. As expected, there is a high and positive correlation between solar irradiance and the peak power. Figure 1 suggests that solar irradiance has a significant influence on peak power compared to the ambient temperature.
This study compares two data-driven methods, namely SVR and GPR models and the baseline Ordinary Least Squared (OLS) model [18], to monitor a PV system. Here, the SVR model with six different kernels and GPR with four different kernels are considered (Table II). In total, eleven models are used for modeling nominal operating conditions, and the outputs are compared. The models in Table II are used to model the training data based on 5-fold cross-validation. Figure 2(a) shows the prediction results of the DC power using the SVR and GPR models. Figure 2(b) shows the zoom on one day of prediction data for each model. Figure 2(b) indicates that the GPR models provide a better prediction of the power data compared to the SVR models and the OLS model.

### TABLE II: Considered models.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Kernel function</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVR_C</td>
<td>SVR with the Cubic kernel</td>
<td>$(1 + x_i^T x_j)^3$</td>
</tr>
<tr>
<td>SVR_CG</td>
<td>SVR with the Coarse Gaussian kernel</td>
<td>$\exp(-\frac{1}{\sigma^2}</td>
</tr>
<tr>
<td>SVR_FG</td>
<td>SVR with the Fine Gaussian kernel</td>
<td>$\exp(-\frac{1}{\sigma^2}</td>
</tr>
<tr>
<td>SVR_L</td>
<td>SVR with the Linear kernel</td>
<td>$x_i^T x_j$</td>
</tr>
<tr>
<td>SVR_MG</td>
<td>SVR with the Medium Gaussian kernel</td>
<td>$\exp(-\frac{1}{\sigma^2}</td>
</tr>
<tr>
<td>SVR_Q</td>
<td>SVR with the Quadratic kernel</td>
<td>$(1 + x_i^T x_j)^2$</td>
</tr>
<tr>
<td>GP_RQ</td>
<td>GPR with the Rational Quadratic kernel</td>
<td>$\sigma_i^2 \left(1 + \frac{x_i^2}{\lambda_i^2}\right)^{m_i}$</td>
</tr>
<tr>
<td>GP_SE</td>
<td>GPR with the Squared Exponential kernel</td>
<td>$\sigma_i^2 \exp\left(\frac{x_i^2}{\lambda_i^2}\right)$</td>
</tr>
<tr>
<td>GP_M52</td>
<td>GPR with the Matern 5/2 kernel</td>
<td>$\sigma_i^2 \left(1 + \frac{x_i^2}{\lambda_i^2} + \frac{5x_i^4}{3\lambda_i^4}\right) \exp\left(\frac{x_i^2}{\lambda_i^2}\right)$</td>
</tr>
<tr>
<td>GP_Exp</td>
<td>GPR with the Exponential kernel</td>
<td>$\sigma_i^2 \exp\left(\frac{x_i^2}{\lambda_i^2}\right)$</td>
</tr>
</tbody>
</table>

* $\lambda = \sqrt{(x_i - x_j) \cdot (x_i - x_j)}$ is the GPR kernel function.

Figure 3 displays the boxplots of the prediction errors of the ten models based on training data. From these boxplots, the distribution of the prediction errors for each model can be assessed. Essentially, in comparing boxplots, large boxes with larger ranges indicate wider distribution, that is, more variable data. While compact boxes with a central line (median) around zero indicate that prediction error is small. From Figure 3, the GPR models achieve improved prediction performance in comparison to the SVR models. We can also see that the SVR_CG and SVR_L models possess the widest variability (i.e., largest prediction errors) in comparison to the other models. Hence, results in Figure 3 confirm the superior prediction performance of GPR models.

To visually display the prediction performance obtained with GPR and SVR models, Figure 4 presents the Empirical cumulative distribution function (ECDF) of the prediction errors. It confirms that the GPR method’s effectiveness for predicting peak power is promising and can be adopted as a reference for fault detection.

IV. GPR AND SVR-BASED FAULT DETECTION IN PV SYSTEM
This section provides the general framework of the proposed kernel-based machine learning methods (GPR and SVR) used to detect anomalies in the DC output of a PV system. Essentially, the proposed GPR and SVR-based PV system monitoring algorithms are performed into two main phases (Figure 5). First, GPR and SVR models are constructed based on training input-output data, \((X, y)\). Before model training, the input-output data have been scaled to zero mean and variance one. Here, the input data contains the solar irradiance and ambient temperature, and the peak power is the response variable. After constructing these reference models, for testing unseen input data, the built model is used to predict the input variable (peak power), \(\hat{y}\). In other words, the GPR, and SVR models are designed based on anomaly-free data and then adopted for monitoring new data. It is worth pointing out that the use of the smoothed residuals in fault detection permits incorporating information from past data in the detection process.

There are several schemes in the literature for monitoring univariate time-series data, such as Shewhart, exponential smoothing, and cumulative sum monitoring schemes. However, these conventional schemes are derived by assuming the normality distribution of data [50]. When the Gaussian assumption is not verified, the monitoring results would be inappropriate. Here, to alleviate this limitation, a nonparametric approach using kernel density estimation (KDE) has been used to set a threshold based on residuals for fault detection.

### A. Residuals denoising: exponential smoothing

Real data collected using sensors can be easily contaminated by noise and outliers. Usually, residuals are filtered or smoothed to reduce noise, remove outliers, and enhance fault detection quality by reducing the false alarm rate [50]. Numerous filters have been proposed and developed in the literature for time series denoising. In this paper, an EWMA filter is applied to the residuals to smooth the data and remove outliers. The EWMA filter takes the form of a first-order process and can be expressed as

\[
\begin{align*}
    s_t &= \nu x_t + (1 - \nu) s_{t-1} \\
    s_0 &= \mu_0
\end{align*}
\]

(27)

where \(x_t\) is the observation at time \(t\), \(S_t\) is the filter output at time \(t\), and \(\nu \in [0, 1]\) is the smoothing parameter that defines the depth of the memory of EWMA. It should be noted that the use of the smoothed residuals in fault detection permits incorporating information from past data in the detection process.

### B. Kernel density estimation-based monitoring scheme

Kernel density estimation is a nonparametric procedure for estimating the probability density function of a set of random samples [51]. Generally, it is employed when a parametric form of data distribution is not available or to avoid making assumptions about distribution underlying the data. If the data distribution has a parametric form, it can be estimated by parametric methods, such as the maximum likelihood estimation approach [52]. It has been widely exploited in many applications, including outlier detection and quality control [53], [54], [55]. In this paper, a nonparametric monitoring scheme based on KDE is presented and applied to monitor a PV system. To this end, at first, the probability density function (PDF) of residuals from the prediction model (e.g., GPR or SVR) is estimated based on fault-free data since residuals are univariate. Then, the detection threshold based on KDE could be determined by taking the percentile of the estimated distribution.

Assume that \(e = [e_1, e_2, \ldots, e_n]\) are residuals generated from prediction model (e.g., GPR or SVR). For any real model, \(\hat{y}\) from the PV array using measured temperature and irradiance. The residuals \(E = y - \hat{y}\) offer pertinent information about the presence of abnormal conditions. Residuals reflect the operating conditions of the inspected photovoltaic array, in the absence of failures, the residuals are close to zero, while residuals become larger in the case of faulty conditions.
values of $e$, KDE nonparametrically estimates the PDF of the residuals by the following formula [51]:

$$\hat{f}(e) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{e - e_i}{h}\right),$$

(28)

where $e_i$ refers to residuals, $n$ is the number of samples. $K$ is the kernel function and $h$ denotes the kernel bandwidth parameter. Among a lot of kernel functions, the Gaussian kernel is usually used:

$$K(e) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{e^2}{2}\right).$$

(29)

$h$ represents the smoothing bandwidth factor that determines the probability estimation quality. Small values of $h$ can lead to quite a noisy estimation and under-smoothed PDFs. On the other hand, large values over-smooth them. Various procedures have been developed in the literature to compute the adequate parameter $h$ [56]. Silverman’s rule of thumb calculates an optimal $h$ based on the assumption that the data is normally distributed [57]. This rule is usually used if the data is unimodal and close to normal distribution. When a large amount of data is available, and the data is far away from normal or multimodal, the approach called Improved Sheather Jones can be used to make the approximation of data distribution [57]. As shown in [58], [57], for $n$ observations with the standard deviation $\sigma$, the optimal choice of $h$ (i.e., that minimizes the mean integrated squared error) can be obtained by the following expression:

$$H = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}} \approx 1.06 \sigma n^{-1/5}.$$ 

(30)

For fitting long-tailed distributions and outliers, a robust estimation of $\sigma$ could be computed as $\sigma = \text{median}(|X_i - \hat{X}|)/0.6745$, where $\hat{X}$ represents the median of the data [59]. After estimating the distribution of the monitoring statistics based on KDE, the detection threshold is computed by the 100$(1 - \alpha)$th quantile of $\hat{f}(e)$.

The detection threshold, $\text{DT}(\alpha)$ can then be obtained from the PDF of the residuals, $e$, for a given probability false alarm, $\alpha$ as follows:

$$\int_{-\infty}^{\text{DT}(\alpha)} \hat{f}(e)de = 1 - \alpha$$

(31)

$$\int_{\text{DT}(\alpha)}^{+\infty} \hat{f}(e)de = \alpha$$

(32)

The detection threshold, $\text{DT}(\alpha)$, is computed as the $(1 - \alpha)^{\text{th}}$ quantile of the distribution $\hat{F}(e)$ of the residuals.

$$\text{DT}(\alpha) = F^{-1}(1 - \alpha).$$

(33)

where $F^{-1}$ denotes the inverse CDF (the quantile function) of the residuals based on fault-free data.

The core of the described fault detection algorithm with the KDE detection threshold can be described as follows:

- **Step 1**: Generate the residuals vector, $e = [e_1, e_2, \ldots, e_n]$, based on the reference model (i.e., GPR or SVR).
- **Step 2**: Select the kernel function, here Gaussian kernel is used.
- **Step 3**: Compute the PDF of residuals as given in Equation (28).
- **Step 4**: Calculate the CDF of residuals using the following equation:

$$\hat{F}(e) = \int_{-\infty}^{\text{DT}(\alpha)} \hat{f}(e)de.$$ 

(34)

- **Step 5**: From the estimated cumulative distribution function (CDF) of residuals, the detection threshold, $\text{DT}(\alpha)$, is computed by taking the $(1 - \alpha)^{\text{th}}$ quantile of the estimated CDF (see Equation (33)).
- **Step 6**: An anomaly is reported if the residuals surpass the detection threshold.

V. EXPERIMENTAL RESULTS

This section is devoted to evaluating the performance of GPR and SVR-based procedures to sense faults in the DC sides of a PV system. In this study, six types of anomalies are examined: (1) string fault, (2) partial shading, (3) module degradation, (4) line-line faults, (5) four PV modules short-circuit, and (6) eight modules short-circuited on the PV array (Figure 7). Experimental data collected from the PV system described in Section II.
In this work, the commonly used metrics to verify the performance of anomaly detection, including true positive rate (TPR), false-positive rate (FPR), accuracy, and area under the curve (AUC, are used [60]). Also, the EER (Equal Error Rate) measure is used to evaluate the performance of the proposed method. The lower the EER, the more accurate is the fault detection. It is worth pointing out that the investigated models are constructed based only on daytime data and without considering night-time data.

A. Scenarios with string faults:

This scenario shows the GPR and SVR-based fault detection methods when applied to the PV system with open-circuit faults. Importantly, this kind of fault happened if DC protection has deteriorated or detached between PV modules in series [61]. To this end, one string disconnected fault is established by completely disconnecting the first string of the PV array (F1 in Figure 7). As discussed above, at first, we generate residuals based on the constructed GPR and SVR models. We applied the conventional Shewhart scheme to the residuals for fault detection. Also, we applied the KDE-based scheme to filtered DC power residuals with the EWMA filter to sense the presence of energy losses. Figure 8(a-b) illustrates, respectively, the performance of the GPR and SVR-based Shewhart and KDE methods applied to the faulty dataset over time.

All schemes can easily detect this string open-circuit fault because of the large magnitude of the string faults (approximately 50% of the rated power has been lost). It is worth pointing out that the GPR-based monitoring schemes with exponential kernel sense this fault very clearly compared to the other methods. This may be due to the high prediction quality of GPR with the exponential kernel.

B. PV system with temporary shaded modules
This experiment is aimed at evaluating the ability of GPR and SVR-based monitoring schemes to discover partial shading. Partial shading would negatively affect the production of a PV system. As the power generated by a PV system is related to the received solar irradiance, when partial shading takes place, the power output will decrease proportionally as the shading heaviness increases. In general, shading losses can be caused by different elements, including pylons and trees [61]. Here, we assess the effectiveness of the considered schemes to detect partial shading in four PV modules #5, #6, #7, and #8 in the first string is simulated (F2 in Figure 7). This partial shading has occurred within samples 60 and 105.

Table IV presents the statistical indicator of the monitoring schemes. At first, it can be seen that GPR-based monitoring schemes outperform SVR-based monitoring schemes with different kernels. This result could be interpreted by the fact that the GPR models can better describe the relationship between input-output PV data, as shown in Section III. Also, results show that GPR-based schemes with nonparametric thresholds performed better than the other GPR and SVR-based methods by achieving lower FPR and EER and the highest TPR and AUC (Table IV). Using the denoised residuals by EWMA filter for sensing anomalies improves this scheme’s robustness to measurement noise. This study shows that better performance is achieved when GPR and SVR-based schemes are applied to filtered data than to unfiltered residuals. For instance, GPR(Exp)-based KDE scheme detected reaches an AUC of 0.989, high TPR of 0.978, and without false alarms (FPR=0). The SVR-based methods flag this shading but with some false alarms and missed detection (Table IV). Such results suggest that GPR with KDE-based detection threshold rather than the SVR and GPR-based Shewhart schemes to suitably uncover partial shading in a PV array.

### TABLE IV: Detection results by procedure for PV array with four shaded PV modules.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parametric threshold</th>
<th>KDE-based threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TPR</td>
<td>FPR</td>
</tr>
<tr>
<td>OLS</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SVR(1)</td>
<td>0.644</td>
<td>0</td>
</tr>
<tr>
<td>SVR(k)</td>
<td>0.822</td>
<td>0</td>
</tr>
<tr>
<td>SVR(c)</td>
<td>0.733</td>
<td>0</td>
</tr>
<tr>
<td>SVR(C)</td>
<td>0.000</td>
<td>0</td>
</tr>
<tr>
<td>SVR(g)</td>
<td>0.644</td>
<td>0</td>
</tr>
<tr>
<td>SVR(G)</td>
<td>0.400</td>
<td>0</td>
</tr>
<tr>
<td>GPR(k)</td>
<td>0.978</td>
<td>0.01</td>
</tr>
<tr>
<td>GPR(k2)</td>
<td>0.978</td>
<td>0.01</td>
</tr>
<tr>
<td>GPR(Exp)</td>
<td>0.978</td>
<td>0.03</td>
</tr>
</tbody>
</table>

C. PV system with module degradation

In this scenario, we assess the detection methods’ performance in uncovering power losses caused by PV modules degradation. Namely, a degradation fault can happen because of an increase in the series resistance between the modules due to limited adherence to contacts or corrosion induced by water vapor [62]. Degradation faults can negatively impact the performances of PV systems by causing power losses [63]. Here, a degradation fault has been simulated by introducing a series resistance of R=15 Ω between PV module #1 and PV modules #2 of the first PV string (F3 in Figure 7). Figure 9(a-b) displays the detection results of the GPR and SVR with parametric and nonparametric thresholds. In Figure 9(a), the obtained residuals using the previously constructed GPR and SVR models are evaluated with the Shewhart scheme. In Figure 9(b), the residuals are first denoised with EWMA filter with smoothing parameters $\beta = 0.4$ and then compared to a nonparametric KDE-based detection threshold for anomaly detection. As it can be observed from Figure 9(a-b), the SVR-based schemes cannot correctly recognize this degradation and results in high missed detections. On the other hand, GPR-based schemes can appropriately recognize this PV module degradation and GPR with a nonparametric threshold provides a quite accurate detection compared to its parametric counterpart.

![Fig. 9: Detection performances of (a) the GPR and SVR-based parametric detection threshold and (b) GPR and SVR-based KDE schemes applied to filtered residuals in the presence of one degraded PV module.](image-url)
TABLE V: Detection results by procedure for PV array with one degraded PV module.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>AUC</th>
<th>EER</th>
<th>Accuracy</th>
<th>AUC</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>0.310</td>
<td>0.5</td>
<td>0.690</td>
<td>0.310</td>
<td>0.5</td>
<td>0.690</td>
</tr>
<tr>
<td>SVR_L</td>
<td>0.690</td>
<td>0.5</td>
<td>0.310</td>
<td>0.690</td>
<td>0.5</td>
<td>0.310</td>
</tr>
<tr>
<td>SVR_Q</td>
<td>0.786</td>
<td>0.656</td>
<td>0.214</td>
<td>0.690</td>
<td>0.5</td>
<td>0.310</td>
</tr>
<tr>
<td>SVR_C</td>
<td>0.690</td>
<td>0.5</td>
<td>0.310</td>
<td>0.690</td>
<td>0.5</td>
<td>0.310</td>
</tr>
<tr>
<td>SVR_FG</td>
<td>0.690</td>
<td>0.5</td>
<td>0.310</td>
<td>0.690</td>
<td>0.5</td>
<td>0.310</td>
</tr>
<tr>
<td>SVR_MG</td>
<td>0.690</td>
<td>0.5</td>
<td>0.310</td>
<td>0.690</td>
<td>0.5</td>
<td>0.310</td>
</tr>
<tr>
<td>SVR_CG</td>
<td>0.690</td>
<td>0.5</td>
<td>0.310</td>
<td>0.690</td>
<td>0.5</td>
<td>0.310</td>
</tr>
<tr>
<td>( GP_{RQ} )</td>
<td>0.945</td>
<td>0.923</td>
<td>0.055</td>
<td>0.952</td>
<td>0.922</td>
<td>0.048</td>
</tr>
<tr>
<td>( GP_{SR} )</td>
<td>0.938</td>
<td>0.912</td>
<td>0.062</td>
<td>0.952</td>
<td>0.922</td>
<td>0.048</td>
</tr>
<tr>
<td>( GP_{MS2} )</td>
<td>0.945</td>
<td>0.923</td>
<td>0.055</td>
<td>0.952</td>
<td>0.922</td>
<td>0.048</td>
</tr>
<tr>
<td>( GP_{ESP} )</td>
<td>0.979</td>
<td>0.979</td>
<td>0.021</td>
<td>0.986</td>
<td>0.984</td>
<td>0.014</td>
</tr>
</tbody>
</table>

D. PV system with line-line fault

Additionally, we tested the GPR and SVR-based monitoring schemes in the presence of a line-line fault. Basically, a line-line fault in a PV system could happen when a low resistance connection is unintentionally placed within two points of different potential in a string of modules or cabling [64]. Generally speaking, it can be caused by insulation failure of cables, like an animal chewing through cable insulation, mechanical damage, water ingress, or corrosion. In this scenario, we simulate line-line fault by introducing a short circuit between PV #4 and PV #5 of the first string with PV #9 and PV #10 of the second string (see F4 in Figure 7). This line-line fault has occurred within samples 100 till the end of testing data. Since this fault is relatively large, as expected, both GPR and SVR-based schemes achieved good detection performance (Table VI). We also observe that the baseline OLS-based detection approach provides reasonable detection performance with an AUC of 0.911. Again, the GPR-based methods outperform the SVR-based methods by correctly detecting this fault without missed detection and false alarms (Table VI).

TABLE VI: Detection results by procedure for PV array with a line-line fault.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parametric threshold</th>
<th>KDE-based threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>AUC</td>
</tr>
<tr>
<td>OLS</td>
<td>0.945</td>
<td>0.911</td>
</tr>
<tr>
<td>SVR_L</td>
<td>0.986</td>
<td>0.978</td>
</tr>
<tr>
<td>SVR_Q</td>
<td>0.993</td>
<td>0.989</td>
</tr>
<tr>
<td>SVR_C</td>
<td>0.972</td>
<td>0.956</td>
</tr>
<tr>
<td>SVR_FG</td>
<td>0.972</td>
<td>0.956</td>
</tr>
<tr>
<td>SVR_MG</td>
<td>0.972</td>
<td>0.956</td>
</tr>
<tr>
<td>SVR_CG</td>
<td>0.979</td>
<td>0.967</td>
</tr>
<tr>
<td>( GP_{RQ} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( GP_{SR} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( GP_{MS2} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( GP_{ESP} )</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

E. PV system with Short-circuit fault

This case study is dedicated to evaluating the potentials of the proposed techniques to detect short-circuit faults. Short-circuit faults can be caused by bad wiring within PV modules. Also, aging, vibration, and PV modules abrasion represent key sources of short-circuit faults. This kind of fault can cause damage or performance degradation of the PV Modules. Here, two cases are considered to assess the GPR and SVR-based methods; one PV module short-circuited, and eight PV modules short-circuited (F5 and F6 in Figure 7, respectively) between samples 60 and 105 of testing data.

The SVR-based and OLS-based schemes fails to uncover this small fault as illustrated in Figure 10(a-b) and Table VII. Figure 10(a-b) indicates that the GPR-based schemes can sense this fault. The model comparison results confirm that the GPR-based monitoring scheme with exponential kernel significantly outperforms the SVR and GPR with the other kernel function (Table VII).

As expected, SVR and GPR-based monitoring can easily identify the presence of eight short-circuited PV modules in the inspected PV system. This is because the magnitude of this fault is large and can easily exceed the detection thresholds.

The effective detection of faults in PV systems is essential to their efficient operation and management. Two kernel-based monitoring schemes have been designed for monitoring the DC side of the PV system. These charts make decisions about the PV system performance at a given time point by checking the residuals generated by the kernel-based models (i.e., GPR and SVR). The model comparison results show that the GPR model outperforms the SVR model in prediction accuracy. In summary, the GPR model provides a promising way to
Table VII: Detection results by procedure for PV array with one short-circuit PV module.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parametric threshold</th>
<th>KDE-based threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy  AUC  EER</td>
<td>Accuracy  AUC  EER</td>
</tr>
<tr>
<td>OLS</td>
<td>0.310  0.5  0.685</td>
<td>0.310  0.5  0.685</td>
</tr>
<tr>
<td>SVRL</td>
<td>0.690  0.5  0.308</td>
<td>0.690  0.5  0.308</td>
</tr>
<tr>
<td>SVRQ</td>
<td>0.690  0.5  0.308</td>
<td>0.690  0.5  0.308</td>
</tr>
<tr>
<td>SVRc</td>
<td>0.690  0.5  0.308</td>
<td>0.690  0.5  0.308</td>
</tr>
<tr>
<td>SVRFG</td>
<td>0.690  0.5  0.308</td>
<td>0.690  0.5  0.308</td>
</tr>
<tr>
<td>SVRMG</td>
<td>0.690  0.5  0.308</td>
<td>0.690  0.5  0.308</td>
</tr>
<tr>
<td>SVRCG</td>
<td>0.690  0.5  0.308</td>
<td>0.690  0.5  0.308</td>
</tr>
<tr>
<td>GPwQ</td>
<td>0.966  0.944 0.034</td>
<td>0.972  0.956 0.027</td>
</tr>
<tr>
<td>GSpQ</td>
<td>0.959  0.933 0.041</td>
<td>0.966  0.944 0.034</td>
</tr>
<tr>
<td>GPMQ</td>
<td>0.966  0.944 0.034</td>
<td>0.972  0.956 0.027</td>
</tr>
<tr>
<td>GPSP</td>
<td>0.972  0.980 0.028</td>
<td>0.993  0.995 0.007</td>
</tr>
</tbody>
</table>

Despite the satisfying detection performance obtained using the GPR-based monitoring scheme, this approach has been tested only on monitoring the DC side of the PV system. As future work, we plan to extend this scheme to monitor both the DC and AC sides of the PV systems. Besides, we plan to incorporate more input data, including fill factor open-circuit voltage and short circuit current to improve the sensitivity of the proposed monitoring scheme in detecting different types of faults. Also, other types of faults, such as MPPT error, could be investigated in future work. On the other hand, we plan identification of types of faults by applying univariate statistical monitoring charts on the residuals of the DC current and DC voltage.

REFERENCES
