Real-time fault detection in PV systems under MPPT using PMU and high-frequency multi-sensor data through online PCA-KDE-based multivariate KL Divergence

Azzeddine Bakdi *[a], Wahiba Bounoua[b], Amar Guichi[c], Saad Mekhilef[d,e]

bkdaznsun@gmail.com*, wb.bounoua@gmail.com, guichi.omar@gmail.com, saad@um.edu.my

[a] Department of Mathematics, University of Oslo, 0851 Oslo, Norway. * Corresponding author.

[b] Signals and Systems Laboratory, Institute of Electrical and Electronics Engineering, University M'Hamed Bougara of Boumerdes, Avenue of independence, 35000 Boumerdès, Algeria.

[c] Department of Electronic, University Mohamed Boudiaf, BP 166, 28000 M'Sila, Algeria.

[d] Power Electronics and Renewable Energy Research Laboratory (PEARL), Department of Electrical Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia.

[e] School of Software and Electrical Engineering, Swinburne, Victoria, Australia - email: smekhilef@swin.edu.au

Abstract

This paper considers data-based real-time adaptive Fault Detection (FD) in Grid-connected PV (GPV) systems under Power Point Tracking (PPT) modes during large variations. Faults under PPT modes remain undetected for longer periods introducing new protection challenges and threats to the system. An intelligent FD algorithm is developed through real-time multi-sensor measurements and virtual estimations from Micro Phasor Measurement Unit (Micro-PMU). The high-dimensional high-frequency multivariate characteristics are nonlinear time-varying where computational efficiency becomes crucial to realize online adaptive FD. The adaptive assumption-free method is developed through Principal Component Analysis (PCA) for dimension reduction and feature extraction with reduced complexity. Novel fault indicators $D_x(t)$ and discrimination index $AD(t)$ are developed using Kullback–Leibler Divergence (KLD) for an accurate evaluation of Transformed Components (TCs) through recursive Smooth Kernel Density Estimation (KDE). The algorithm is developed through extensive data with 2.2×10^6 measurements from a GPV system under Maximum PPT (MPPT) and Intermediate PPT (IPPT) switching modes. The validation scenarios include seven faults: open circuit, voltage sags, partial shading, inverter, current feedback sensor, and MPPT/IPPT controller in boost converter faults. The adaptive algorithm is proved computationally efficient and very accurate for successful FD under large temperature and irradiance variations with noisy measurements.

Keywords

Grid-connected PV Systems; Power Point Tracking; Kullback-Leibler Divergence; Principal Component Analysis; Advanced Monitoring; Phasor Measurement Unit.

* Corresponding author. *E-mail address:* bkdaznsun@gmail.com. (Azzeddine Bakdi).

Nomenclature

I-V	Current-voltage (curve)	n	Number of samples
P-V	Power-voltage (curve)	m	Data set dimension
$\boldsymbol{V}_{PV}$	Output voltage	$\boldsymbol{X}_r$	Reference healthy data matrix
$\boldsymbol{I}_{PV}$	Output current	n_r	Number of samples in $\boldsymbol{X}_r$
I_{irr}	Photocurrent	$\boldsymbol{X}_{tst}$	Online testing data matrix
I_0	Diode saturation current	n_{tst}	Number of samples in $\boldsymbol{X}_t$
R_s	Series resistance	H_0	Null hypothesis
V_{therm}	Cell thermal voltage	H_1	Alternative hypothesis
$\mathscr{n}$	Ideality factor of the cell	D_{KL}	Kullback-Leibler divergence
R_{sh}	Shunt resistance	ε_{safe}	Control limit
κ	Boltzmann constant	N	Normal distribution
q	The electron’s charge	$\boldsymbol{\mu}$	Mean vector
T	Actual cell temperature	$\boldsymbol{\Sigma}$	Covariance matrix
G	Actual solar irradiance	tr	Trace of a matrix
G_{st}	Std. cond. irradiance $1000\ W/m^2$	g	Density ratio
T_{st}	Std. cond. cell temperature $25\ ℃$	K	Kernel function
K_I	Relative temperature coefficient	σ	Kernel width
N_s	Number of cells in series	h	Kernel smoothing factor
N_p	Number of cells in parallel	$\boldsymbol{J}$	Objective function
T_s	Sampling Time	$\boldsymbol{\theta}$	Parameter vector
$\boldsymbol{\mathscr{Y}}$	Real time measured data matrix	$\Re$	Set of real numbers
$\boldsymbol{\mathscr{y}}$	Real time measured signal	$\overline{\boldsymbol{X}}$	Auto-scaled data matrix
$\boldsymbol{V}_{dc}$	DC voltage	$\boldsymbol{S}$	Estimated covariance matrix
$\boldsymbol{I}$	Current	$\boldsymbol{P}$	Loadings matrix
$\boldsymbol{f}_I$	Current frequency	$\boldsymbol{\Lambda}$	Eigenvalues diagonal matrix
$\boldsymbol{V}$	Voltage	λ	Eigenvalue
$\boldsymbol{f}_V$	Voltage frequency	t	Time
$\boldsymbol{U}_{abc}$	3-phase voltages	n_t	Sliding window size
$\boldsymbol{I}_{abc}$	3-phase currents	$\boldsymbol{c}_k^*$	k^{th} reference TC
P_t	Distribution of window samples at t	p_t	PDF of window samples at t
$\boldsymbol{x}$	Smoothed version of a noisy signal $\boldsymbol{v}$	$\boldsymbol{c}_{k,t}$	k^{th} online TC at time t
r	Weighting factor	p_k^*	Probability density function of $\boldsymbol{c}_k^*$
w	Window length	$p_{k,t}$	Probability density function of $\boldsymbol{c}_{k,t}$
P_{ref}	Reference distribution	$\hat{p}$	Density function estimation
P_{tst}	Testing distribution	E	Expectation
p_{ref}	Reference probability density function	D_k	k^{th} fault indicator
p_{tst}	Testing probability density function	AD	Discrimination index
$\boldsymbol{X}$	Data matrix	CL_{Dk}	Control limit of D_k
		CL_{AD}	Control limit of AD

Abbreviations

AC	Alternating Current	MW	Megawatt
AI	Artificial Intelligence	PC	Principle Component
ANN	Artificial Neural Networks	PCA	Principal Component Analysis
ARL	Average Run Length	PI	Proportional Integral
DC	Direct Current	PLL	Phase Lock Loop
FD	Fault Detection	MPPT	Maximum Power Point Tracking
FL	Fuzzy Logic	PMU	Measurement Unit
FNR	Fault to Noise Ratio	PSO	Particle Swarm Optimization
GPV	Grid-connected PV	VOC	Voltage Oriented Control
GW	Gigawatt	PV	Photovoltaic
IEC	International Electrochemical Commission	RF	Random Forest
IPPT	Intermediate Power Point Tracking	SFR	Signal to Fault Ratio
KDE	Kernel Density Estimation	SNR	Signal to Noise Ratio
KLD	Kullback–Leibler Divergence	SVD	Singular Value Decomposition
kWh	Kilowatt-hour	SVM	Support Vector Machines
LPF	Low Pass Filter	SVPWM	Space Vector Pulse Width Modulation
MIMO	Multi-Input Multi-Output	TC	Transformed Component
MISE	Mean Integrated Squared Error	T^2	Hotelling T^2

1. Introduction

The record low solar prices that were achieved in 2016 had caught many energy experts by surprise. That year, bids awarded in several tenders were below the 3 US cent per kWh level (2.95 US cents for an 800 MW project in Dubai, 2.91 US cents for a power supply contract in Chile, 2.42 US cents for the ‘winter’ supply part of the 1.18 GW plant PPA in Abu Dhabi) [1]. The largest increments in 2017 were recorded in China (53 GW) and the US (11 GW), together accounting for two-thirds of the growth in global solar capacity. Japan provided the third largest addition (7 GW). China also leads in terms of cumulative installed capacity (130 GW), with one-third of the global total. The US (51 GW) and Japan (49 GW) are in second and third with Germany (42 GW) now in fourth [2]. The cumulative installed solar PV power capacity grew by 32% to 404.5 GW by the end of 2017, up from 306.4 GW in 2016. In only ten years, the world’s total PV capacity increased by over 4,300% (43 times) – from 9.2 GW in 2007; and under optimal conditions, the capacity could reach the terawatt level by the end of 2022 [1].

Solar power costs will continue to decrease due to technical improvements. One key factor of reducing the costs of photovoltaic systems is to increase the reliability and the service lifetime of the PV modules [3]. PV systems are also vulnerable to several anomalies that should be diagnosed as early as possible before any deviations from the designed nominal conditions. Preventive actions must then be implemented to avoid deteriorating the performance and drastically hindering efficiency, reliability, and safety. The ultimate objective is to meet the international protection standards of the International Electrochemical Commission (IEC) [4, 5, 6]. The early detection of potential anomalies in PV systems is crucial to the good performance to avoid small deviations from nominal conditions and to match the predicted energy yield [7] and the desired power quality. Depending on the operative functioning of various components and grid regulation, the availability factor [8] of a PV system is also improved through fast FD techniques by avoiding and/or minimizing the downtime. Besides, an accurate monitoring PV system increases its efficiency while reducing maintenance costs and maximizing the profit during the system lifetime [9]. In addition to their nonlinear time-varying characteristics and high dependency on environmental factors (temperature and irradiance), the properties of electrical systems are naturally inherited in PV systems which have very fast dynamics while abrupt changes have even faster dynamics. This justifies the need for high-frequency and multiple sensors measurements to monitor the GPV system and its faults dynamics. However, these facts introduce the bottleneck problem of computational complexity.

The detection of sensor faults in GPV systems was considered in [10], the authors also proposed the optimal location of current and voltage sensors to limit the increase in cost due to the redundancy of sensory devices. Sensor-based analysis was also implemented in [11] to detect partial shading. Related approaches are based on observing a local signal such as the string current which is then compared to its known patterns to detect local faults [12, 13]. A comprehensive review of metaheuristic tools was provided in [14]. In PV systems, reported model-based FD techniques incorporate mathematical (analytical) models such as state observers [15], parameter identification [14], and impedance-based models [16]. Model-based methods were theoretically proved useful in simulations but their major drawback in practice is the robustness to measurement noise and model uncertainties. Analytical model-based techniques fail to address the broad sense of FD in real PV systems which are complex and cannot be accurately described by a closed-form mathematical model, especially under different conditions. Similarly, FD in PV systems can be achieved through artificial models such as Fuzzy Logic (FL) [17, 18], and Artificial Neural Networks (ANN) [19, 20]. Artificial models indeed provide a better approximation to the nonlinear behavior of a PV system, however, they are computationally demanding and cannot address the time-varying behavior through adaptive learning. Other common PV system FD tools are heavily based on classification methods such as Support Vector Machines (SVM) [21]. The FD task is reduced to classifying different measurements into normal/ faulty operations. The main drawbacks of artificial models are the requirement of labelled data, that's sufficient measurements during real faults labelled by solar experts. Another issue to consider is the multi-classes of normal behavior due to changes in power point and temperature/ irradiance variations.

A comprehensive review of fault diagnosis and protection challenges in PV systems were provided in [22], system faults were classified into physical, electrical, and environmental. Protection devices in the DC-side protect against over-current faults, grounding faults, and arcing faults [23]. It was mentioned in [22] that the mentioned protection devices have failed to detect their corresponding faults in the PV array due to: (i) Lower fault current magnitudes, (ii) Presence of MPPT and (iii) Non-linear PV characteristics and the colossal dependency on the insolation levels. Faults on the DC side have catastrophic effects on the system outputs and may cause the whole system to burn even though it is equipped with protective devices [24]. A critical review of AC Microgrid protection issues was also provided in [25], while their respective protection schemes were classified into protection for only grid-connected mode, protection for only islanded mode, and protection for both modes. In the current digital era, the new aspect of cyber risk introduces emerging challenges for the detection and diagnosis of cyber-attacks in wide-area power systems as highlighted in [64]. The three conditions (i, ii, iii) are experimentally considered in this work for the Grid-connected PV (GPV) system for which traditional techniques are developed to consider the nonlinear time-varying characteristics of the system and its faults.

Among the emerging methods based on a statistical analysis of time-series sensor data [26], Principal Component Analysis (PCA) is a common multivariate data analysis and dimensionality reduction technique [27, 28]. PCA plays a major role in solar engineering for various applications such as the analysis of big time-series data such as the satellite-derived irradiance data and string-level measurements from a utility-scale PV system. PCA is also used in PV systems for power forecasting and monitoring [29, 30]. Despite its paramount advantages in handling big data and reducing computational complexity, PCA theory relies on three heavy assumptions: (a1) multivariate Gaussian distribution of data, (a2) stationarity of the process assuming a fixed operating point of a system, and (a3) linear correlations assuming a linear time-invariant system. Due to these shortcomings, PCA applications for effective FD in PV systems is limited to simulation studies [31, 32] only. Unfortunately, these assumptions cannot be tolerated when considering the practical conditions under which all PV systems operate. [33] emphasized the physical adequacy of a power generation system under long-term conditions. On the other hand, variations of PV module parameters with irradiance and temperature were highlighted in [34, 35], and the influence of increased temperature on energy production was considered in [36]. Compared to the existing FD methods and considering the nonlinear time-varying characteristics, the developed adaptive algorithm updates its model and parameters (in a computationally efficient manner) to the prevailing power point through a novel discrimination index $AD(t)$ that distinguishes the controlled changes of power point and triggers updates when necessary to classify faults from evolving normal behavior.

More advanced FD approaches are generally required to detect PV system faults in the presence of MPPT and IPPT controllers. Advanced MPPT search algorithms such as dynamic leader based collective intelligence [61] and memetic salp swarm algorithm [62] are reported effective and faster in reducing power losses under partial shading conditions making it very difficult to detect faults for two main reasons: MPPT search algorithms mask the symptoms of mismatch faults, especially faults at low-current, whereas faults have disparate characteristics due to MPPT controllers. A recent review of MPPT algorithms is provided in [63] while [22] highlighted the adverse effects of MPPT controllers on FD strategies in PV systems. This work takes this challenge into account in the design of the novel FD strategy. Real data are collected during real faults from a GPV system with MPPT and IPPT controllers which are based on the common Particle Swarm Optimization (PSO) technique. PSO is increasingly preferred and prompts researchers in recent studies [49] and it is commonly used for MPPT applications in PV systems. PSO algorithm [49] is used as an experimental verification case for its popularity in the literature, whereas advanced MPPT search algorithms [61,62] pose the same challenges for FD strategy design in PV systems.

It was highlighted in [64, 65, 66] that full modelling and analysis of spatiotemporal dependencies are crucial for a reliable fault detection and diagnosis and for a full comprehension of physical and cyber risk assessment in power distribution systems. In the presented work, spatial and temporal dependencies are respectively addressed through PCA and sequential analysis. In a typical MPPT/IPPT controlled GPV system under input variations, real data is far from a multivariate Gaussian distribution for which traditional PCA assumptions of data normality are not satisfied herein. This motivates the current work by proposing a novel non-parametric (distribution-free) indices. The resulting Transformed Components (TCs) are evaluated recursively using novel fault indicators named the $D_x(t)$ indices which are developed through nonparametric Kullback–Leibler Divergence (KLD) instead of the traditional Q and T^2 statistics of PCA [37, 38]. In consequence, another challenge rises due to the high-dimensional high-frequency data acquired from grid-connected PV systems at a sampling time of $100\ \mu s$, multivariate KLD approach [39,40] is known for its high computational complexity as reported in [41, 42], this prevents the realization of real-time online FD. A common state-of-the-art solution to realize FD is through parametrized KLD approaches assuming a Gaussian distribution [43, 44, 45] or Gamma distribution [46]. FD is reduced to simple monitoring of statistical parameters such as mean vector and statistical dispersion for which approximate parametrized KLD approaches are less accurate. While computational complexity is reduced through PCA, accurate FD and discrimination indices $D_x(t)$ and $AD(t)$ are obtained through nonparametric KLD through recursive smooth Kernel Density Estimation (KDE) without any assumptions on GPV system behavior or data distributions. This article validates the significance of the proposed algorithm through several experiments in which validation scenarios include real faults in a GPV system. Realistic faults have different levels of severity and are injected across different parts in the entire energy conversion system including array faults such as open circuit and partial shading, MPPT/IPPT controller faults in a boost converter, inverter fault in form of single IGBT failure, current feedback sensor fault, and grid anomalies such as voltage sags. Compared to the existing literature in solar engineering, this work is of practical novelty that considers (I) a wide-range of realistic faults in GPV systems and (II) further examining their online detectability and detection performance under MPPT/IPPT conditions with variation in temperature and solar insolation, such setups have not been reported before. The presented methodology is novel with advantages of (III) escaping theoretical PCA's assumptions, (IV) greatly reducing the KLD complexity to match the online application, and (V) improving the accuracy compared to parametrized PCA/KLD approaches for successful FD in GPV systems in practice. The algorithm was implemented in real-time online FD in the GPV system where the obtained results reflect its potential applications in practice as it outperforms state-of-the-art methods.

The rest of this article is organized as follows. Section 2 conducts a short description of the GPV system, its nonlinear time-varying characteristics, measured and virtually estimated signals, and the examined real faults. In Section 3, the main contribution of this work, the proposed algorithm is developed, justified, and its steps are detailed while highlighting novel contributions. Experimental results are then discussed in section 4 while comparing with several methods for computational time and memory complexity as well as the results of FD accuracy and robustness performance. Finally, potential applications are summarized and important conclusions and recommendations are drawn in section 5.

2. System description and data preprocessing

In this article, a lab implemented typical grid-connected PV system is used to validate the FD performance of data-driven methods against real faults under MPPT/IPPT modes and practical conditions. In this work, statistical methods incorporate the general knowledge about the system functionality in order to construct a reliable and effective FD algorithm.

2.1 Grid-connected PV system

This section highlights the theoretical particularities of PV systems in which the data-driven algorithm in section 3 is developed to solve. The nonlinear time-varying behavior of PV systems can be theoretically highlighted according to ideal one-diode model [47] relating the output voltage $\boldsymbol{V}_{PV}$ to the output current $\boldsymbol{I}_{PV}$:

$$\boldsymbol{I}_{PV} = I_{irr} - I_0 \left[exp\left(\frac{\boldsymbol{V}_{PV} + R_s \boldsymbol{I}_{PV}}{V_{therm} \mathcal{n}}\right) - 1 \right] - \frac{\boldsymbol{V}_{PV} + R_s \boldsymbol{I}_{PV}}{R_{sh}} \tag{1}$$

where $\mathcal{n}$ is the ideality factor of the cell and $V_{therm} = \kappa T/q$ is the cell thermal voltage. κ, T and q are respectively the Boltzmann constant, the temperature of the p-n junction, and the electron charge. R_{sh} and R_s are the shunt resistance and the series resistance respectively. This system is nonlinear and time-variant since the diode saturation current I_0 depends on the temperature of the cell, the photocurrent I_{irr} is also linearly related to the irradiance level and the temperature of the cell [48]:

$$I_{irr} = I_{irr,st} \left(\frac{G}{G_{st}}\right) [1 + K_I (T - T_{st})] \tag{2}$$

where $I_{irr,st}$, T_{st} and G_{st} are respectively the photocurrent, cell temperature, and solar irradiance under the standard test conditions ($T_{st} \ = \ 25$ °C and $G_{st} = 1000\ W/m^2$); G and T are respectively the actual solar irradiance and the actual cell temperature; and K_I is the relative temperature coefficient of the short-circuit current.

A PV panel of N_s cells in series and N_p cells in parallel have the following I-V relation:

$$\boldsymbol{I}_{PV} = N_p I_{irr} - N_p I_0 \left(exp \left[\frac{1}{V_{therm} n} \left(\frac{\boldsymbol{V}_{PV}}{N_s} + \frac{R_s \boldsymbol{I}_{PV}}{N_p} \right) \right] \right), \quad (3)$$

in addition to nonlinear time-variant behavior, PV systems exhibit two known peculiarities [22]: (i) voltage and current are limited and highly dependent on solar insolation G and temperature T (Eq.(3)), and (ii) the presence of MPPTs/ IPPTs.

In this article, the grid-connected PV system is implemented as shown in Fig.1 [48]. The PV array output is generated through the programmable Chroma 62150H-1000S solar array emulator that allows varying effects of environmental conditions (G and T). The programmable AC source Chroma 61511 is used as a grid emulator. The control algorithm was implemented on a DSpace 1104 environment, which is also used for data acquisition. Voltage Oriented Control (VOC) technique is used in combination with Space Vector Pulse Width Modulation (SVPWM) to control the active and reactive power based on the grid-side signals. The output voltage is synchronized with the grid voltage through the Phase Lock Loop (PLL). The AC load in this work is used for protection purposes while injecting real faults.

This system controller is based on Particle Swarm Optimization (PSO) technique to ensure Maximum Power Point Tracking (MPPT) when the available power level is lower than the rated power $P_{Available} \leq P_{Limit}$ and Intermediate Power Point Tracking (IPPT) [48] mode if $P_{Available} > P_{Limit}$ [48]. This system is used to generate and collect real faulty data for experimental validation of real-time online FD, we refer interested readers to [48] for more details on the control structure, Energy Management System (EMS), communication, and settings of this system.

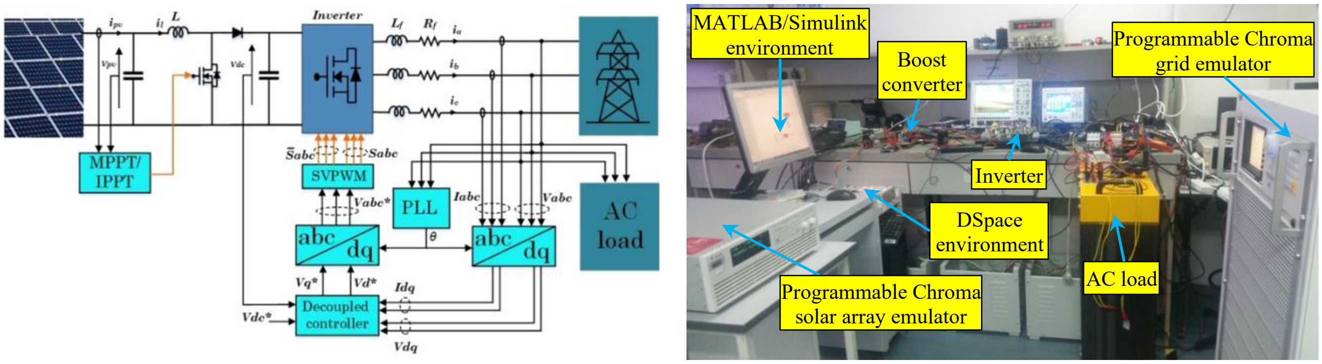


Figure 1. Overview of the implemented grid-connected PV system.

As demonstrated in Fig.2, the power point location varies along P-V curves for different temperature and irradiance levels Eq.(1-3). Collected GPV system data exhibit a varying covariance structure and disparate fault characteristics at various power points. The FD algorithm must update to such variations and distinguish the evolving normal behavior from faults.

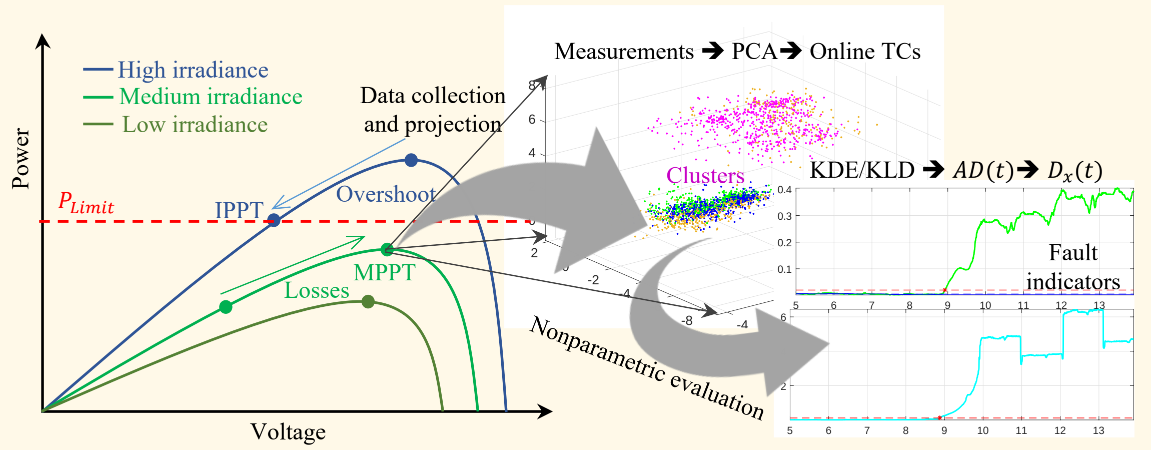


Figure 2. Overview of real-time online adaptive PCA-KDE-KLD FD algorithm in GPV.

2.2 Measurements, estimations, and faults

While the previous theoretical model is simple and cannot be extended for practical FD, the presented algorithm is fully data-driven based on system data and free from assumptions. Real-time measured signals are the PV array voltage $\boldsymbol{V}_{PV}$ and current $\boldsymbol{I}_{PV}$ and DC voltage $\boldsymbol{V}_{dc}$, as shown in Fig.1, these are acquired with a sampling time of $T_s = 100\,\mu s$. The 3-phase grid voltages $\boldsymbol{U}_{abc}$ and currents $\boldsymbol{I}_{abc}$ have distorted periodic patterns and their skewed multimodal distributions do not contribute good quality information in the training stage. Magnitude and frequency in addition to the phase shift (which is regulated for synchronization with the grid) are online estimated from the measured periodic signals using PLL positive-sequence PMU (IEEE Std C37.118.1-2011) [49] as depicted in Fig.3. This work is designed and validated based on a small-scale microgrid application since the process of injecting real faults and collecting real data is hazardous, costly, and impractical in a large-scale system. Data are collected from sensor measurements and a virtual PMU [67] is used to extract the positive-sequence components from three-phase signals. In the case of multi-source wide-area applications, micro-PMUs should be used; moreover, locations of PMUs are of great importance and should be optimized for complete observability [68]. Wide-area systems are also subject to inter-area oscillations and stability issues. Moreover, cyber risk emerges as a new challenge in intelligent digital powers systems where the detection and diagnosis of software failures and cyber-attacks become crucial. However, the broad scope of this work considers the various common physical faults in GPV systems which have a direct impact on the microgrid.

The PMU-estimated quantities are more sensitive to detect faults due to their unimodal distributions and they are more significant to the analysis of the system performance. The 3-phase currents and their PMU-estimated quantities are shown in Fig.4 over a short time window of $100\ ms$. The estimation is based on the simulated model of PMU [49] in Fig.3 where the reporting rate is set to 64 while input data are interpolated and output data are extrapolated to observe the positive-sequence components at the inherited sampling time T_s. Phasor computations are based on Fourier analysis performed using a running average window of one cycle, full details are explained in the standard documentation [49].

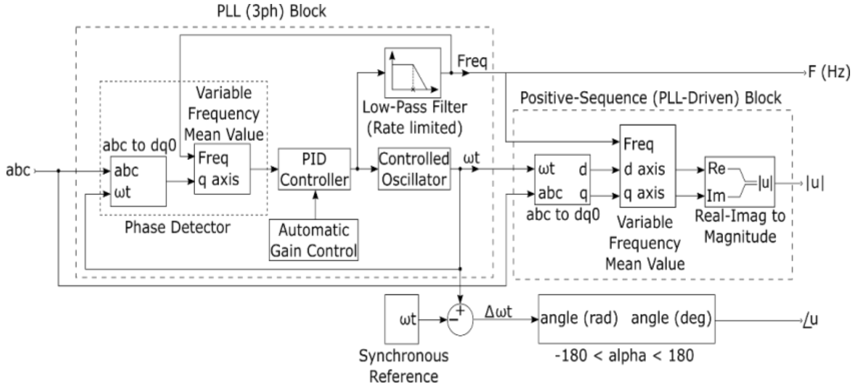


Figure 3. MATLAB block diagram of PMU for estimating 3-phase signals [49].

Without loss of generality, micro PMUs can be used in industrial applications to collect such measurements directly since they provide micro-second resolution with milli-degree accuracy, they are mainly advantageous in local applications to study the grid penetration of renewables and they are well-suited to the current PV microgrid application. The minimum set of fault-relevant variables $\{\boldsymbol{I}_{PV},\ \boldsymbol{V}_{PV},\ \boldsymbol{V}_{dc},\ |\boldsymbol{I}|,\ \boldsymbol{f}_I,\ |\boldsymbol{V}|,\ \boldsymbol{f}_V\}$ is used in this work for monitoring the GPV system. The real-time measured and estimated signals form a data matrix $\boldsymbol{\mathcal{Y}}$ of seven columns:

$$\boldsymbol{\mathcal{Y}} = [\boldsymbol{I}_{PV}, \boldsymbol{V}_{PV}, \boldsymbol{V}_{dc}, |\boldsymbol{I}|, \boldsymbol{f}_I, |\boldsymbol{V}|, \boldsymbol{f}_V]^T, \tag{4}$$

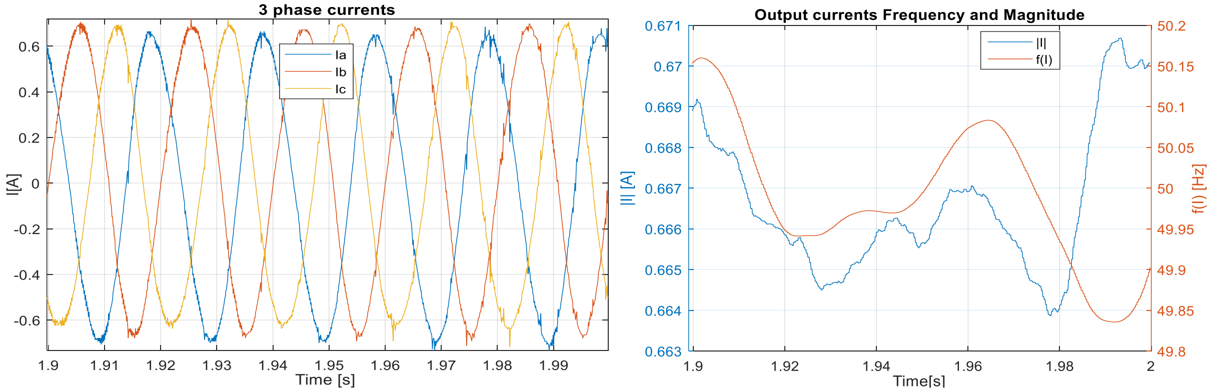


Figure 4. Three-phase currents and their PMU-estimated quantities

The measurements are highly corrupted by noise and hence a preprocessing stage is performed for signal de-noising. The major drawback of using a Low Pass Filter (LPF) is the destruction of the main details which increases fault detection delay and conceals the symptoms of intermittent and low-impact faults. In this direction, an exponential filter is used for smoothing the measured signals using an exponential window function. The j^{th} sample of smoothed version $\boldsymbol{x}_j$ is obtained from noisy signals as:

$$\boldsymbol{x}_j = \frac{1}{\sum_{i=1}^{w} r^i} \sum_{i=1}^{w} r^i \, \boldsymbol{y}_{j-w+i} \tag{5}$$

where $\boldsymbol{y}_i = \boldsymbol{\mathcal{Y}}_{ij}$ for $j = 1,2,\cdots,7$ represents the i^{th} row vector measurement of all variables, r is a weighting factor that controls the smoothing versus the memory of the filter, and w is the window length. This filter is applied to the 7 measured and estimated variables. For example, the measured PV outputs and their filtered versions are depicted in Fig.5.

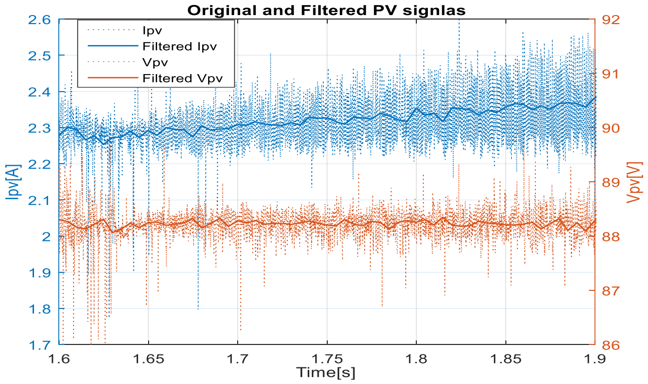


Figure. 5. Measured and filtered signals at the DC side.

Table 1. Realistic injected fault in the GPV system.

Fault	Type	Description
F1	Inverter fault	Complete failure in one of the six IGBTs
F2	Feedback Sensor fault	One phase sensor fault 20%
F3	Grid anomaly	Intermittent voltage sags
F4	PV array mismatch	10 to 20% nonhomogeneous partial shading
F5	PV array mismatch	15% open circuit in PV array
F6	MPPT/IPPT controller fault	-20% gain parameter of PI controller in MPPT/ IPPT controller of the boost converter
F7	Boost converter controller fault	+20% in time constant parameter of PI controller in MPPT/IPPT controller of the boost converter

This article considers the detection of seven realistic faults that are listed in Table 1 and injected in the GPV system of Fig.1. These faults are of various types and locations to ensure a complete analysis. All faults are injected manually in several successive independent experiments, each experiment runs around 10 to 15 seconds where the fault is introduced around the 7th to 8th second. The sampling time for fault-free and faulty data acquisition is $T_s = 100\ \mu s$. Unlike simulation studies, the exact fault occurrence timestamp is unknown for the algorithm. PV array mismatches such as F4 and F5 are challenging to detect due to the large variability in sensor data at the DC-side; Fortunately, these faults are of lower severity levels causing only power losses. Faults F1 and F3 occurring in the grid side of the grid-connected PV system are easy to detect since they affect only the AC side where data exhibits very small variability as demonstrated in Fig.4. Due to their severity, however, these faults must be detected at their early stages within a limited fault time. This work also investigates parametric faults F6 and F7 in MPPT/IPPT Proportional Integral (PI) controller in the DC side, in addition to a feedback current sensor fault, F2. Controller fault F7 indicates an increased time-constant parameter whereas F6 is a biased gain in the PI controller which results in a reduced MPPT/ IPPT trajectory tracking performance without affecting the stability of the closed-loop system. These faults are widely common in practice, their impact on GPV systems, theoretical description, and I-V characteristics are well-detailed in a comprehensive review in [22]. It was stated in [22] that detecting faults of 20% to 40% mismatch levels is difficult, whereas challenging for mismatch levels below 20%. Moreover, the presence of MPPT/ IPPT controllers poses adverse effects on the detection of faults that have disparate characteristics and hidden symptoms due to MPPT/IPPT search algorithms. Incipient faults in PV systems may occur because of corrosion, cells degradation, and partial damage in interconnections, these degradation faults are not severe and they are generally avoided through regularly scheduled preventive maintenance. Degradation faults are not considered in this work, their detection requires long-term data at large sampling intervals.

3. Proposed FD algorithms

Novel data-driven algorithms are designed in this framework for FD in grid-connected PV systems under practical conditions and time-varying parameters. The algorithm models the system behavior under its nonlinear evolving characteristics using its huge data from its high temporal resolution sensors. The effectiveness in detecting various types of faults is improved by accurate statistical modeling, online adaptation to prevailing conditions, and precise (assumption-free) evaluation. Novel fault indicators and a new discrimination index are proposed to detect faults and distinguish model updates. Since it only depends on the available data, the method is cost-efficient, however, major developments are implemented for this algorithm to match online realization. The computational efficiency is improved by extracting the few most sensitive features while monitoring the PV system.

3.1 Kullback–Leibler Divergence

The Kullback–Leibler Divergence (KLD), also called the relative entropy, is the most common of the *f*-divergence family [39] and widely used in practice. The KLD is a discriminant function between two probability distributions, a reference distribution P_{ref} and a test distribution P_{tst}, defined on the same probability space, the KL divergence from P_{tst} to P_{ref} is defined for discrete distributions [40] to be:

$$D_{KL}(P_{ref} \parallel P_{tst}) = \sum_{i} P_{ref}(i) \log \frac{P_{ref}(i)}{P_{tst}(i)} \tag{6}$$

and for continuous distributions as the integral [41]:

$$D_{KL}\left(P_{ref} \parallel P_{tst}\right) = \int_{-\infty}^{+\infty} p_{ref}(\boldsymbol{x}) \log \frac{p_{ref}(\boldsymbol{x})}{p_{tst}(\boldsymbol{x})} d\boldsymbol{x} \tag{7}$$

where p_{ref} and p_{tst} are the probability densities of P_{ref} and P_{tst}, respectively.

The KLD represents the expectation over a reference distribution P_{ref} of the logarithmic difference between the probabilities. Let n measurements data $\boldsymbol{X} = [\boldsymbol{x}_1, \ \boldsymbol{x}_2, \ \cdots, \ \boldsymbol{x}_n] \in \Re^{n \times m}$ be a sample of m-variate random vectors drawn from a common distribution where m is seven for this PV system and n can exceed 10^5. Suppose $\boldsymbol{X}_r$ is a reference data recorded during normal operation and it contains n_r samples described by P_r distribution, and $\boldsymbol{X}_t$ is an online measured data with n_t samples following P_t distribution. The KLD is widely used as a scalar monitoring index [43, 44] which quantifies the deviation between two m-variate time-series data-sets $\boldsymbol{X}_r$ and $\boldsymbol{X}_t$. It is zero if and only if the two distributions are equal, and it is positive and far from zero if the distributions are different. The GPV system is therefore considered to be under statistical control based on the following hypothesis test:

$$\begin{cases} H_0: & P_t = P_r \\ H_1: & P_t \neq P_r \end{cases} \tag{8}$$

According to Eq. (6,7), $D_{KL}(P_r \parallel P_t)$ is zero under the null hypothesis H_0 which represents a fault-free operation of the PV system. Also, $D_{KL}(P_r \parallel P_t)$ is different (greater than) zero under the alternative hypothesis that represents a faulty operation. However, in practice, the measurements in both data-sets are not perfect and their samples are generally noisy, the distributions at different intervals are never identical. A critical region is hence defined to reject the null hypothesis through an upper control limit ε_{safe} which is small but different from zero:

$$\begin{cases} H_0: & D_{KL}(P_r \parallel P_t) \leq \varepsilon_{safe} \\ H_1: & D_{KL}(P_r \parallel P_t) > \varepsilon_{safe} \end{cases} \tag{9}$$

Unfortunately, this technique cannot be implemented for FD in real-time PV systems since the estimation of the joint distribution of m-variate data is highly challenging as reported in [42]. A linear increase in the dimension m results in an exponential increase in the number of required samples, the estimated parameters, and also the computation time and complexity. To comfort this challenge, previous works assumed all samples in $\boldsymbol{X}$ to follow a multivariate Gaussian distribution, $P_r(\boldsymbol{x}) \sim N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ and $P_t(\boldsymbol{x}) \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$. The KLD is then given as:

$$D_{KL}(P_r \parallel P_t) = \frac{1}{2}\left\{(\boldsymbol{\mu}_t - \boldsymbol{\mu}_r)^T \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_r) + ln\frac{|\boldsymbol{\Sigma}_t|}{|\boldsymbol{\Sigma}_r|} + tr(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}_r) - m\right\} \quad (10)$$

where the problem is reduced to the estimation of the mean vectors $\boldsymbol{\mu}_{ref}$ and $\boldsymbol{\mu}_{test}$ of dimension m and the covariance matrices $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_t$ of dimension $m \times m$. Notice here that the KLD index is reduced to monitoring changes in the process mean and statistical dispersion, and therefore its FD sensitivity is highly deteriorated. Besides, the assumption of normality does not hold in practical applications and especially during a faulty operation. In recent FD applications, the KLD is estimated in [50,51] using the direct importance estimation [52], where the importance [53] represents the density ratio $p_r(\boldsymbol{x}) \,/\, p_t(\boldsymbol{x})$, and it is given as a statistical model:

$$\frac{p_r(\boldsymbol{x})}{p_t(\boldsymbol{x})} = g(\boldsymbol{x};\ \boldsymbol{\theta}) = \sum_{i=1}^{n_t} \theta_i K(\boldsymbol{x}, \boldsymbol{x}_t(i)) \quad (11)$$

where $K(\boldsymbol{x}, \boldsymbol{x}')$ is a kernel function, Gaussian Kernel is widely used in practice with a parameter σ as a kernel width:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right) \quad (12)$$

and θ_i are the n_t parameters to be learned by minimizing the KLD of $p_t(\boldsymbol{x})$ with respect to its estimate $\widetilde{p_t}(\boldsymbol{x})$ $D_{KL}(p_t(\boldsymbol{x}), \widetilde{p_t}(\boldsymbol{x}))$:

$$D_{KL}\big(p_t(\boldsymbol{x}), \widetilde{p_t}(\boldsymbol{x})\big) = \int_{-\infty}^{+\infty} p_t(\boldsymbol{x}) \log \frac{p_t(\boldsymbol{x})}{\tilde{g}(\boldsymbol{x};\ \boldsymbol{\theta})\, p_r(\boldsymbol{x})} d\boldsymbol{x}$$

$$= \int_{-\infty}^{+\infty} p_t(\boldsymbol{x}) \log \frac{p_t(\boldsymbol{x})}{p_r(\boldsymbol{x})} d\boldsymbol{x} - \int_{-\infty}^{+\infty} p_t(\boldsymbol{x}) \log \tilde{g}(\boldsymbol{x};\ \boldsymbol{\theta})\ d\boldsymbol{x} \quad (13)$$

since the first term is constant, the density is approximated by maximizing the following objective function

$$\boldsymbol{J} \coloneqq \int_{-\infty}^{+\infty} p_t(\boldsymbol{x}) \log \tilde{g}(\boldsymbol{x};\ \boldsymbol{\theta})\ d\boldsymbol{x}$$

$$\approx \frac{1}{n_t}\sum_{j=1}^{n_t} \log \tilde{g}(\boldsymbol{x_t}(j);\ \boldsymbol{\theta}) = \frac{1}{n_t}\sum_{j=1}^{n_t} \log\left(\sum_{i=1}^{n_t} \theta_i K(\boldsymbol{x_t}(j), \boldsymbol{x}_t(i))\right) \quad (14)$$

with respect to the parameter vector $\boldsymbol{\theta}$:

$$\widetilde{\boldsymbol{\theta}} = \underset{\theta_1^{n_t}}{\arg\max}\ (\boldsymbol{J}) \quad (15)$$

The computational complexity of this approach is still high especially for high dimension data of GPV systems (n_t is large), another shortcoming of this approach is the density ratio divergence problem [42].

3.2 Dimension reduction

The KLD measure is very sensitive to anomalous behaviors, however, its computation is not feasible for large dimension multivariate data, especially in adaptive approaches. To comfort the curse of dimensionality, PCA is used in this framework to de-correlate the system variables and obtain the transformed components [54] at the prevailing power point. These features are more sensitive to faults because they capture the correlation among variables. More importantly, the estimation and evaluation tasks will be faster and efficient when reducing the dimensions of the estimated parameters, computation time, and the required number of samples.

Suppose $\boldsymbol{X}_r = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n_r}] \in \Re^{n_r \times m}$ is a reference data matrix of n_r samples, this descriptive set is independent of time. The seven variables in Eq.(4) are observed with different units and scales and hence are auto-scaled:

$$\overline{\boldsymbol{X}_r} = [\overline{\boldsymbol{x}_1}, \overline{\boldsymbol{x}_2}, \cdots, \overline{\boldsymbol{x}_{n_r}}], \overline{\boldsymbol{x}_\imath} = [\boldsymbol{x}_i - \boldsymbol{\mu}_r]\, \boldsymbol{\Sigma}_r{}^{-1} \tag{16}$$

$\boldsymbol{\mu}_r$ and $\boldsymbol{\Sigma}_r$ are the reference standardization parameters, they respectively represent the mean vector and the reference standard deviation matrix:

$$\boldsymbol{\mu}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} \boldsymbol{x}_i = [\mu_{r\,1}, \cdots, \mu_{r\,m}]^T \tag{17}$$

$$\boldsymbol{\Sigma}_r = \mathrm{diag}\{\sigma_{r\,1}, \cdots, \sigma_{r\,m}\}, \qquad \sigma_{r\,k}{}^2 = \frac{1}{n_r - 1} \sum_{i=1}^{n_r} [\boldsymbol{X}_{r\,i,k} - \mu_{r\,k}]^2 \tag{18}$$

Take a reference data matrix $\overline{\boldsymbol{X}_r}$ constructed from the PV system variables given in Eq.(4) and illustrated in Fig.2. These signals are filtered using Eq.(5) and down-sampled to $T_s = 1ms$ then auto-scaled as given by Eq.(16). Suppose an informative and descriptive data-set is collected, the covariance structure of the reference data matrix is approximated as follows:

$$\mathbf{cov}(\overline{\boldsymbol{X}_r}) \approx \boldsymbol{S} = \frac{1}{n_r - 1} \overline{\boldsymbol{X}_r}^T \overline{\boldsymbol{X}_r} \tag{19}$$

using Singular Value Decomposition (SVD), the covariance matrix is decomposed into:

$$\mathbf{S} = \boldsymbol{P\Lambda P}^T, \qquad \boldsymbol{P} = [\boldsymbol{P_1}, \cdots, \boldsymbol{P_m}]^T, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{bmatrix} \tag{20}$$

The orthonormal eigenvectors $\boldsymbol{P_i} \in \Re^{m \times 1}$ are the m loadings which represent the independent directions of variability within the original data. The m eigenvalues λ_i represent the amount of variation per each direction, the eigenvalues are in descending order such that $\boldsymbol{P_1}$ and λ_1 is the direction and amount of maximum variation. Conventional is based on defining two subspaces called the principal subspace and residual subspace:

$$\boldsymbol{P} = \left[\widehat{\boldsymbol{P}}|\widetilde{\boldsymbol{P}}\right]^T, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \widehat{\boldsymbol{\Lambda}} & \boldsymbol{0} \\ \boldsymbol{0} & \widetilde{\boldsymbol{\Lambda}} \end{bmatrix} \tag{21}$$

where $\widehat{\boldsymbol{P}} \in \mathfrak{R}^{m\times l}$ and $\widehat{\boldsymbol{\Lambda}} \in \mathfrak{R}^{l\times l}$ are the projection operators on the principal subspace, and they contain the first l principal (significant) features which are interpreted as the useful information that represents the natural variability. $\widetilde{\boldsymbol{P}} \in \mathfrak{R}^{m\times(m-l)}$ and $\widetilde{\boldsymbol{\Lambda}} \in \mathfrak{R}^{(m-l)\times(m-l)}$ are the projection operators on the residual subspace which is generally interpreted as noise. However, the determination of the appropriate number of Principal Components (PCs) is not clear even for a particular system. Data-sets generated from different systems exhibit different covariance structures and signal to noise ratios.

PCA based FD is simply the measure of deviation within each subspace using the T^2 and the Q statistics. Based on the squared-distance these statistics are less accurate in detecting system faults. While relying on particular distribution assumptions, these statistics are also less robust to noise and outliers. Furthermore, the choice of l is primarily related to overfitting or underfitting the constructed model which consequently controls the degree of a tradeoff between different measures of fairness. In this article, all the components are used to monitor all the changes in an online operation. The components are analyzed through a moving window using exact KLD.

The transformed components (TCs) are obtained as linear combinations of the original variables, and they represent the projection of the online measurements on the orthogonal directions obtained in the offline stage. The TCs are obtained through this mapping:

$$TC^* = \overline{\boldsymbol{X}_r}\, \boldsymbol{P} = [\boldsymbol{c}^*{}_1, \cdots, \boldsymbol{c}^*{}_m]^T \tag{22}$$

$\boldsymbol{c}^*{}_k \in \mathfrak{R}^{n_r}$ (for $k = 1, \cdots, m$) are the m reference uncorrelated TCs obtained by projecting the auto-scaled reference data on the orthonormal loadings.

Let $\boldsymbol{X}_t(t) = [\boldsymbol{x}(t - n_t + 1), \cdots, \boldsymbol{x}(t)] \in \mathfrak{R}^{n_t\times m}$ is an online measured data-set of the most recent n_t samples. These samples are scaled in the same manner but using updated parameters:

$$\overline{\boldsymbol{X}}_t(t) = [\overline{\boldsymbol{x}}(t - n_t + 1), \cdots, \overline{\boldsymbol{x}}(t)], \qquad \overline{\boldsymbol{x}}(i) = [\boldsymbol{x}(i) - \boldsymbol{\mu}(i)]\, \boldsymbol{\Sigma}^{-1}(i) \tag{23}$$

where $\boldsymbol{x}(i)$ is a vector measurement at time instance i, $\boldsymbol{\mu}(i)$ and $\boldsymbol{\Sigma}(i)$ represent the most recent updated scaling parameters where reference parameters are used for initial standardization. The new TCs are obtained by projecting the new scaled measurements:

$$\boldsymbol{tc}(t) = \overline{\boldsymbol{x}}(t)\boldsymbol{P} \tag{24}$$

3.3 Novel detection indices

The PV system measurements are online projected on the orthonormal loadings of the constructed model. The TCs are observed through a sliding window of size n_t, they are given at a time t as:

$$\boldsymbol{TC}(t) = [\boldsymbol{tc}(t - n_t + 1), \cdots, \boldsymbol{tc}(t)] = \left[\boldsymbol{c}_{1,t}, \cdots, \boldsymbol{c}_{m,t}\right]^T \tag{25}$$

$\boldsymbol{c}_{k,t} \in \Re^{n_r}$ (for $k = 1, \cdots, m$) are the m online TCs at a time t. These are successively updating by removing the oldest values and augmenting the values of the new projections. The transformed components are analyzed in this framework. It is known in PCA theory that the first (principal) components are the most sensitive to incipient changes such as an evolving normal behavior (changing conditions) or incipient faults (not considered in this work). Distance-based statistics, such as the Hotelling T^2, are however inefficient in detecting those changes within the principle subspace due to its large variability [55]. On the other hand, the residual components are more sensitive to small and abrupt shifts, but the distance-based Q statistic is very limited to measure the deviation within this subspace.

The resulting TCs are orthogonal and have different sensitivities to faults; the most sensitive are hence independently analyzed through a fast and efficient methodology. Let p_k^* denote the density function of the k^{th} reference TC $\boldsymbol{c}_k^*$, and $p_{k,t}$ the density of the k^{th} TC $\boldsymbol{c}_{k,t}$ at time t. These densities are estimated through smooth Kernel Density Estimation (KDE) as:

$$\hat{p}_k^*(c; h) = \frac{1}{n_r} \sum_{i=1}^{n_r} K_h(c, \boldsymbol{c}^*{}_k(i)) \quad \text{for} \quad k = 1,2, \cdots, 7 \tag{26}$$

$$\hat{p}_{k,t}\ (c; h) = \frac{1}{n_t} \sum_{i=1}^{n_t} K_h(c, \boldsymbol{c}_{k,t}(i)) \quad \text{for} \quad k = 1,2, \cdots, 7 \tag{27}$$

using the following kernel:

$$K_h(c, c') = \frac{1}{\sqrt{2\pi}h} \exp - \left(\frac{1}{2} \left(\frac{c - c'}{h} \right)^2 \right) \tag{28}$$

with a smoothing factor h [56], that minimizes the Mean Integrated Squared Error (MISE):

$$MISE(h) = \mathrm{E}\left[\int \left(\tilde{p}(c; h) - p(c; h) \right)^2 dc \right] \tag{29}$$

The KLD of some measurements at time t along the k^{th} TC is hence obtained according to Eq.(7) as:

$$D_k(t) = KLD\left(\hat{p}_k^* \parallel \hat{p}_{k,t}\right) = \int \hat{p}_k^*(c)\, log \frac{\hat{p}_k^*(c)}{\hat{p}_{k,t}\ (c)}\, dc\ , \qquad \text{for k} = 1,2, \cdots, 7 \tag{30}$$

Seven different deviation measures are hence obtained. In this paper, three indices are used D_1, D_7, and AD . The first two indices are referred to as the detection indices and used to detect various types of anomalies. Since they measure any deviation from the reference densities, these indices are super sensitive to any change including small and slow variations of the operation mode. D_1 is indeed the most sensitive to monitor the evolving normal behavior since it is associated with the first dominant TC reflecting the natural variability of the system. The adaptation index AD is hence obtained from D_1 through the Xbar chart [57] as follows:

$$AD(t) = \overline{D_1(t)} = \sum_{i=t-w_{AD}+1}^{t} \frac{D_1(i)}{w_{AD}} \tag{31}$$

The AD index is used to monitor the local mean divergence over the dominant transformed component. w_{AD} is the length of the window and it is taken as half the number of the reference samples $n_r/2$. This index is the most sensitive to normal behavior evolution and used to trigger model updates. Using these three indices, the hypothesis test of Eq (9) is generalized to decisions on the PV system operation to be made at a time t following these three conditions:

$$\begin{cases} \blacksquare C1\colon\ D_1(t) > CL_{D1}\ OR\ D_7(t) > CL_{D7} \\ \ \Rightarrow Fault, Trigger\ alarm, hold\ AD(t) = AD(t-1) \\ \blacksquare C2\colon\ D_1(t) < CL_{D1}, D_7(t) < CL_{D7}, AD(t) < CL_{AD} \\ \ \Rightarrow Fault\_free\ operation \\ \blacksquare C3\colon\ D_1(t) < CL_{D1}, D_7(t) < CL_{D7}, AD(t) > CL_{AD} \\ \ \Rightarrow Change\ operating\ point, Trigger\ update. \end{cases} \tag{32}$$

The densities are estimated for univariate TCs through sufficient samples (n_r and n_t are large), the smooth KDE yields a very accurate approximation for the actual densities. Accordingly, the proposed indices are very robust to individual outliers and very sensitive to real changes. Therefore, the control limits for the detection indices are set empirically in the training stage without using confidence intervals and without any assumption on their distributions. The control limits CL_{D1} and CL_{D7} measure the highest acceptable divergence attributed to noise and inaccuracies without faults and condition variations. Validation fault-free data are collected separately from training data without condition variations and used in the validation stage to tune the control limits. The validation divergence indices $D_1(t)$ and $D_7(t)$ during the validation stage correspond to fault-free, variation-free, and they are independent of training data. Validation divergence in various directions is negligible and accounts for imperfectness aspects only. CL_{D1} and CL_{D7} are therefore tuned empirically just above the maximum values of their respective validation divergence $D_1(t)$ and $D_7(t)$ to ensure a minimum false alarms rate in fault-free conditions. Since model and reference parameters adaptation accounts for varying conditions, the control limits are constant once they are tuned in the validation stage. The control limit for the adaptation index is empirically set to be $CL_{AD} = CL_{D1}/4$. The adaptation index AD is derived from D_1 ($AD(t) = \overline{D_1(t)}$) based on an Xbar chart such that $AD(t)$ measures the variation in the mean value of the divergence within the first TC. Since the Xbar chart has a longer Average Run Length (ARL) [58], AD is slower than $D_1(t)$ in detecting fast changes (due to faults). However, since $CL_{AD} = CL_{D1}/4$, AD is more accurate than $D_1(t)$ in triggering slow and small mean changes (varying conditions). Due to its longer ARL but smaller control limit, the adaptation index is triggered in cases of small, persistent, and slow variations only. In case of a fault, the detection of indices trigger alarms ($D_1(t)$>CL_{D1} or $D_1(t)$>CL_{D1}) before the adaptation index is triggered ($AD(t) > CL_{AD}$) due to the fact that PV system faults have much faster dynamics compared to the slow variation speed of temperature and irradiance. On the contrary, in case of continuous variations of temperature and/or irradiance, the mean D_1 will slowly increase by small steps (<CL_{D1}) where the adaptation index triggers an update ($AD(t) > CL_{AD}$).

The proposed algorithm is hence based on local data analysis around the current operating point. This point is in an unpredictable continuous change due to environmental inputs. Assuming changes due to anomalies are faster than changes in the operating point, this methodology is very efficient in discriminating faults from normal changes. The assumption is correct without a loss of generality since the PV system is known for faster dynamics compared to the natural slow variation of temperature and irradiance during normal operation (excluding abrupt partial shading which is regarded as a fault herein). The adaptation index is based on the change in the mean divergence across the first component, and the changes of the operating point are tracked through the changes in the local mean value and local statistical dispersion. Under the last condition (C3) in Eq.(32) an update is triggered so the normalization parameters are updated as follows:

$$\boldsymbol{\mu}(t) = \frac{1}{n_t} \sum_{i=t-n_t+1}^{t} \boldsymbol{x}(i) = [\mu_1(t), \cdots, \mu_m(t)]^T \tag{33}$$

$$\boldsymbol{\Sigma}(t) = \mathrm{diag}\{\sigma_1(t), \cdots, \sigma_m(t)\}, \qquad {\sigma_k}^2(t) = \frac{1}{n_t - 1} \sum_{i=t-n_t+1}^{t} \left[\boldsymbol{X}_{i,k}(t) - \mu_k(t)\right]^2 \tag{34}$$

Otherwise:

$$\boldsymbol{\mu}(t) = \boldsymbol{\mu}(t - T_s), \text{and } \boldsymbol{\Sigma}(t) = \boldsymbol{\Sigma}(t - T_s) \tag{35}$$

using the initial scaling parameters given at Eq.(11, 17) where $T_s = 1\ ms$ is the sampling time.

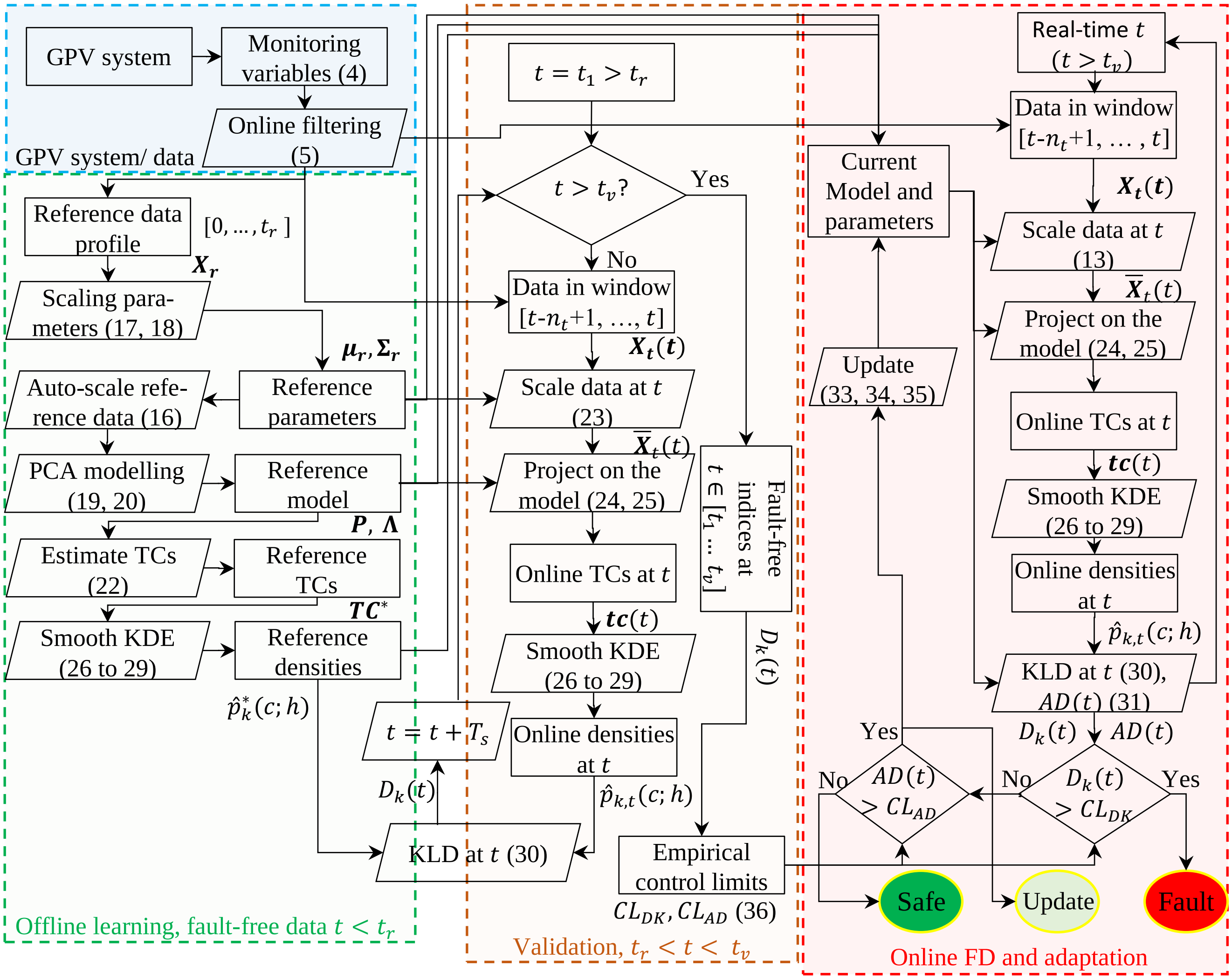


Figure 6. A simplified flowchart for the overall algorithm. Between brackets are the corresponding equations for each statement.

The overall algorithm is summarized in Fig.6. For simplicity purposes, this figure lists the steps of the entire procedure in one operation mode only, while the procedure is based on two parallel models. The developed algorithm, in fact, switches its monitoring models and their respective parameters once the real operation is changed from MPPT to IPPT and vice-versa. The versatile semi-supervised algorithm is optimally trained in an offline stage which yields one initial PCA model for each mode, initial scaling parameters, and reference densities. Outcomes of this training stage are first validated using independent measurements which are also offline collected and selected, but the validation stage interprets the data in an online-like manner to check if the obtained settings are applicable. After the validation stage, the control limits are calculated empirically from validation samples

$$CL_{Dk} = \max\big(D_k(t)\big) + \varepsilon, \qquad s.t. \quad t_r < t < t_v \tag{36}$$

where ε is a small number. The sensitivity of the max operator to noise in original data is not an issue since $D_k(t)$ is obtained through a moving-window sequential analysis and they it is robust to noise. Once the settings are validated, the overall algorithm is put for online application with no prior knowledge about the class (faulty, fault-free, changed mode / operating point). The established models and their parameters are only updated once the discrimination index triggers an update and faults are detected through the D indices.

4. Results and discussion

This section describes the fault injection procedure and the collected data sets. The overall computational complexity is first compared across different methods. The sensitivity of different approaches in tracking the nonlinear time-varying behavior is then examined and compared. The detection performance of contemporary methods is then demonstrated and compared for each GPV system fault.

4.1. Complexity analysis and comparisons

Real-time data-sets are acquired at a high-frequency rate ($T_s = 100\ \mu s$) during the experiments using the dSpace 1104 environment as illustrated in Fig.1, several tests were independently experimented in real-time (Table 1) as demonstrated in Fig.7 below, each test runs for around 15 seconds as described in subsection 2.2. It is worth mentioning again that the exact fault occurrence is unknown in this work because of a high sampling rate and the realistic experimental setup that reflects practical applications. The traditional theoretical FD performance assessment tools such as false alarm rate, detection rate, and detection delay are not accurate in this realistic setup. FD performance comparison is introduced through practical aspects. The reference training data is collected for both MPPT and IPPT modes where each set spans an initial operation period of 3.3 seconds ($n_r = 3.3 \times 10^4$), $\boldsymbol{X_r} \in \Re^{33000\times 7}$ during normal operation conditions while test experiments are of 10 to 15 seconds ($\boldsymbol{X} \in \Re^{100\,000\times 7}$ to $\boldsymbol{X} \in \Re^{150\,000\times 7}$). This algorithm is hence developed, tested and evaluated based on such extensive data measurements. The online TCs $\boldsymbol{c}_{k,t}$ (for $k = 1, \cdots, m$) are observed within a sliding window of $n_t = 1.1 \times 10^4$ in order to estimate their prevailing densities at each time instance and measure the online divergence with respect to their reference values $\boldsymbol{c}^*{}_k$.

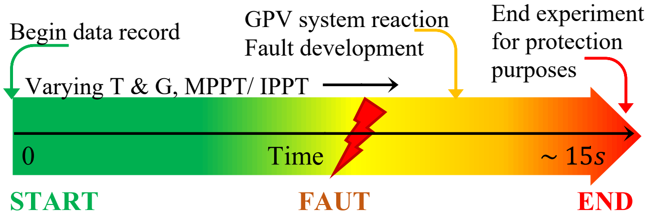
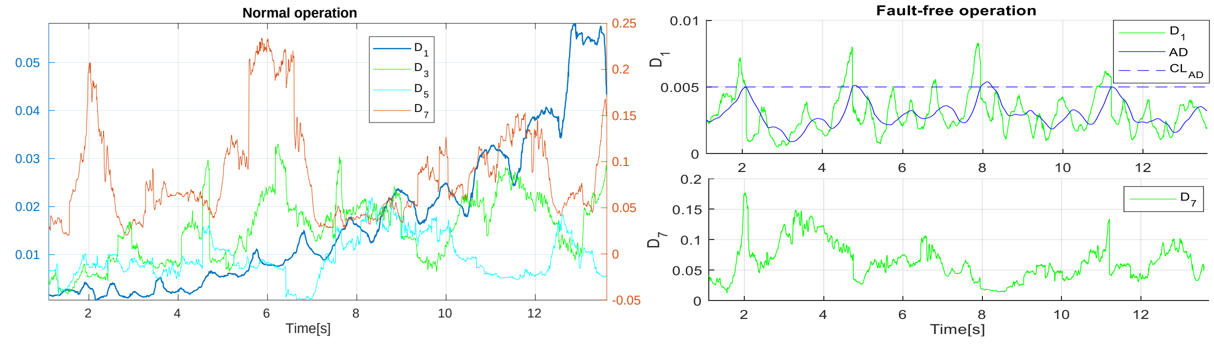


Figure 7. Real fault experiment in the GPV system.

It is known that kernel density estimation methods are very powerful for nonparametric analysis but also computationally highly expensive [59]. In an m-variate data the direct evaluation of kernel density estimates at n_p evaluation points given n_t input sample points require a quadratic $\mathcal{O}\left(\left(n_t \times n_p\right)^m\right)$ operations in a single window only. This dictates the need for huge memory and computation requirements. These requirements are reduced to the m decorrelated features through $\mathcal{O}\left(m \times n_t \times n_p\right)$ which are a huge improvement for accurate density and KLD evaluations ($m = 7$, $n_t = 1.1 \times 10^4$, and $n_p = 100$).

The reduced complexity of the proposed algorithm is a remarkable advantage since the computational time is crucial for online real-time application for GPV system FD in face of its high-dimensional high-frequency data. The multivariate KLD change-point detection approaches [42, 50, 51] are implemented offline with a very rough approximation of multivariate density ratio estimation around only 10 points in each direction (10^7 points in total). For the seven-dimensional data of GPV system, approximate KLD approaches [42, 50, 52] take in average 170 to 190 seconds to verify the statistical control hypothesis for one single measurement only. These approaches are extremely far from realization in online practical applications since sensor measurements arrive at a rate of $T_s = 100\,\mu s$. These approaches are limited in the literature for univariate processes [50, 51] and cannot be extended to high dimensional problems except for very slow processes [42], they are hence expelled from comparisons in this work. On the contrary, the presented nonparametric PCA-DKE-KLD takes on average 0.65×10^{-4} seconds only to evaluate each new measurement. This advantage is due to the fact that the correlation among the original attributes is already captured through PCA TCs which are orthogonal and vary in a one-dimensional space where only two univariate independent evaluations apply (Eq.(26-29)) without any loss of precision.



(a) (b)

Figure 8. FD performance under fault-free time-varying power point: (a) $D_k(t)$-sensitivity to evolving normal behavior without model update, and (b) $AD(t)$-detected and triggered updates.

4.2. Performance analysis and comparisons

It is known in the literature that the principal subspace (first TCs) is more sensitive to incipient changes such as incipient faults but incipient changes can be an evolving normal behavior; whereas the residual subspace is more sensitive to small changes and small faults but the smallest changes can also be attributed to noise. The theoretically-proved temperature and insolation-dependent time-variant behavior (recall Eq.(1-3)) is examined in Fig.8 during a normal (fault-free) GPV operation test where temperature and irradiance are gradually altered: Time-development of different fault indicators ($D_k(t)\ for\ k = 1, 3, 5, 7$ of Eq.(30)) without any model update are demonstrated in Fig.8(a); Notice that all the indicators have a negligible value (around-zero) that is attributed to noise only; However $D_1(t)$ exhibits an evolving pattern due to the normal evolving behavior. $D_1(t)$ measures the divergence across the first TC which is commonly known as the most sensitive to normal evolving behavior, this fact supports the choice of this indicator as the basis for the discrimination index $AD(t)$ of Eq.(31); In Fig.8(a) $D_1(t)$, $D_3(t)$, and $D_5(t)$ are plotted versus the left axis, while $D_7(t)$ (divergence across the last TC) is plotted versus the right axis since it exhibits higher values, this justifies the selection of $D_1(t)$ and $D_7(t)$ as the main indicators. The two fault indicators are derived from the first and the last TCs which respectively measure the highest and lowest variabilities and are respectively sensitive to different types of controlled and uncontrolled variabilities in the GPV system. While Fig.8(a) shows the development of the indicators during varying power point without any statistical model update, Fig.8(b) shows the two indicators with their control limits for the same experiment under models updates which are triggered through the discrimination index $AD(t)$ when it reaches its established threshold (Eq.(32)), all the indices are now under control limits during normal operations and remain within a negligible range as the system operation is safe. Fig.9 demonstrates the detection performance of several methods for fault F6 which stands for a biased controller gain in the PI controller of MPPT/IPPT unit of the boost converter controller (Fig.1) where gain parameters are deliberately biased after few seconds (around 11s) of normal operation as described in Table 1 in subsection 2.2. This fault does not imply a risk to the system but it may damage the converter and cause power losses if it remains undetected for a long time. Fig.9 also reflects the successful detection of this fault by the proposed method (Fig.9(f)) versus the complete failure of recent methods to show any fault symptoms. Fig.9(a) and (b) show PCA performance for decorrelating its huge input data into uncorrelated PCs (Eq.(17-21)) for dimensionality reduction. However, the four relevant clusters corresponding to fault-free (training and testing) are not classified from faulty data (during and after fault F6). Recall in section 2.2 that the fault was implemented in the second half of the period of each test and remains until the end of the experiment, the traditional PCA's Q and T^2 statistics [27, 28] fail to show any sensitivity to this fault as demonstrated in Fig.9(c). These approaches completely fail to detect fault F6 where Q and T^2 statistics remain within their respective thresholds Q_α and ${T^2}_\alpha$ established with a significance level of $\alpha = 0.01$ [37, 38]. Fig.9(d) shows the distributions of PC_1 at different instances, even PC_1 of training data does not follow a parametric distribution and it is instead multimodal, a fact that violates a heavy assumption of most statistical methods. The other plots of Fig.9(d) also show the smoothed KDE of this PC (Eq.(26-29)) at pre-fault (P_{ref}, $P_{test}Tr$, $P_{test}Ts$), during fault ($P_{test}DF$), and post-fault ($P_{test}AF$) which are fairly far from a Gaussian distribution. The change of such density estimate during pri-fault and post-fault is hardly distinguished from the change during pre-fault situations, the last change is attributed to measurement noise and varying power point of the GPV system. Fig.9(e) shows the performance of the I index [43] based on a parametrized KLD approach that assumes a Gaussian distribution [44, 45]. The divergence measured by this conventional index is fairly high even before the fault is introduced, this implies that the Gaussian approximation is not correct.

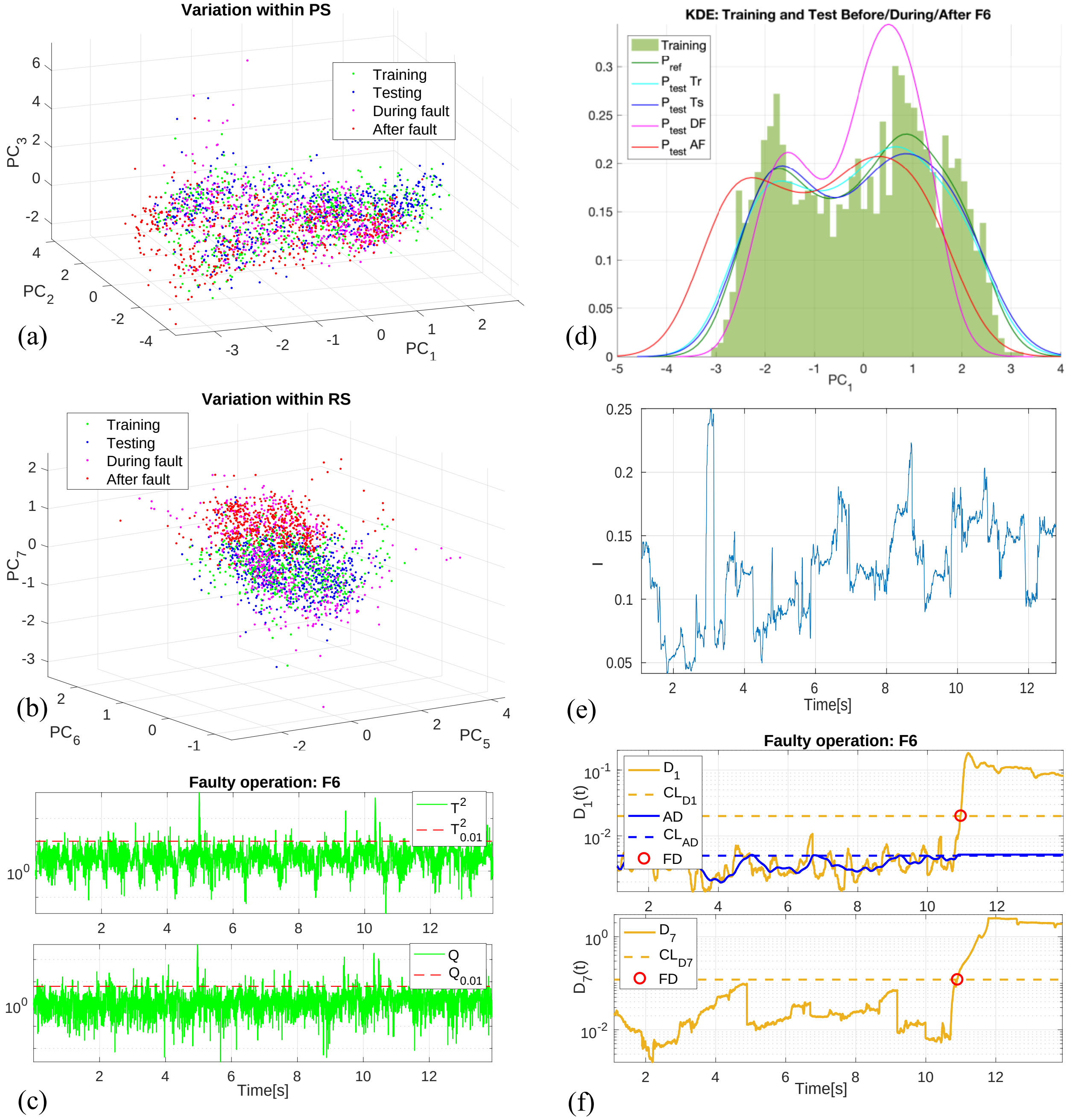


Figure 9. Comparison of recent methods for the detectability of fault F6 (biased PI MPPT/IPPT boost converter controller).

The failure of traditional methods to detect fault F6, as seen in Fig.9(a, b, c, e), is explained in Fig.9(d) where such theoretical assumptions lead to biased estimations of statistical parameters (mean and variance). Moreover, these parameters are changing during a normal behavior due to measurement noise and varying power point of the GPV system, notice the mean is not always zero, and the variance is changing. On the contrary, Fig.9(f) shows the successful detection of fault F6 using the proposed algorithm through both fault indicators $D_1(t)$ and $D_7(t)$ obtained through Eq.(25-30). Notice first that these indicators are very accurate as they have negligible values (<< 0.05) during normal operating conditions compared to the I index, $D_1(t)$ and $D_7(t)$ remain clearly under their respective control limits CL_{D1} and CL_{D7} (Eq.(32)). Both indicators increase considerably (>>1) above their control limits upon the fault occurrence and generate fault alarms (FA) a few seconds before the experiment ended due to protection purposes. Fig.9(f) also demonstrates the sensitivity of the adaptive discrimination index $AD(t)$ (Eq.(31)) which discriminates an evolving normal behavior from GPV system faults where several updates have been triggered each time $AD(t)$ reaches its upper control limit CL_{AD} before $D_1(t)$ reaches its CL_{D1} and triggers a false alarm.

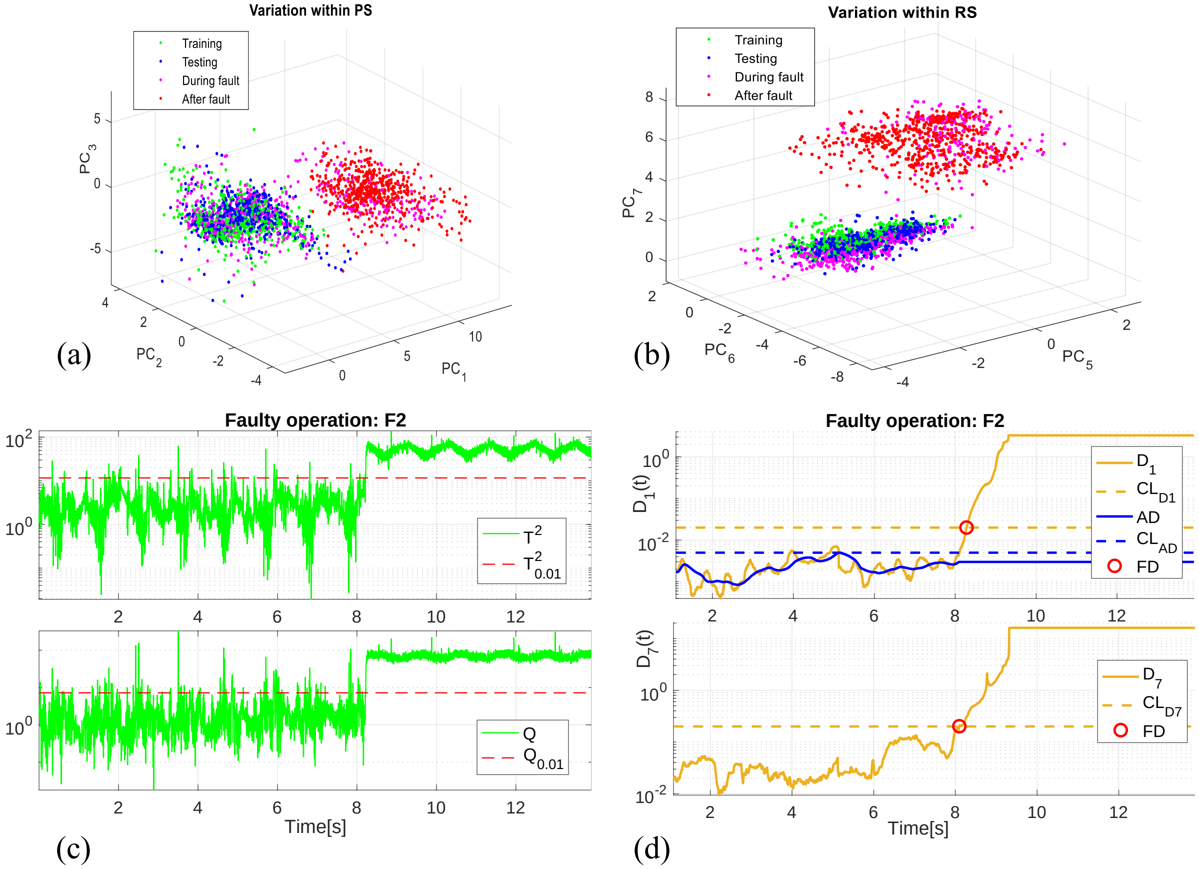


Figure 10. Comparison of detection of F2 (current feedback sensor fault 20%).

Among the different approaches in Fig.9, it is only the proposed assumption-free adaptive method in Fig.9(f) that yields a successful evaluation of PC_1 to detect F6. Considering now the closed-loop current feedback sensor fault F2 in Fig.10, the same comparison can be made between traditional PCA's Q and T^2 statistics [27, 28, 37, 38] and the proposed method. This fault is more severe, compared to F6 above, since it yields a non-zero steady-state error and leads to a wrong configuration of feedback controllers. This severe fault is poorly separated from normal data within both principal and residual PCA subspaces as seen in Fig.10(a, b), F2 is also poorly detected using PCA's Q and T^2 statistics [27, 28, 37, 38] in Fig.10(c) with a lot of false alarms in the pre-fault stage. In comparison, the proposed method yields successful and clear detection that is confirmed through fault alarms (FA) of both fault indicators assisted by the discrimination index.

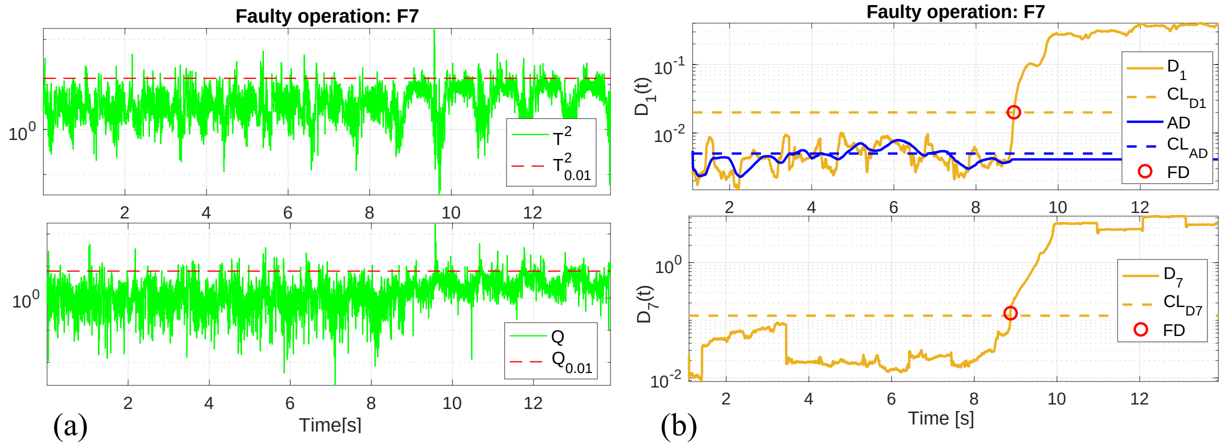


Figure 11. Comparison of detection of F7 (slow-response of PI MPPT/IPPT boost converter controller).

The last comparisons are also made for fault F7 which stands for a slowed control that was introduced in the PI controller of the MPPT/IPPT boost converter controller. Again, the conventional PCA's Q and T^2 completely failed to detect this fault as demonstrated in Fig.11(a), while the fault is successfully detected through both $D_1(t)$ and $D_7(t)$ fault indicators in Fig.11(b).

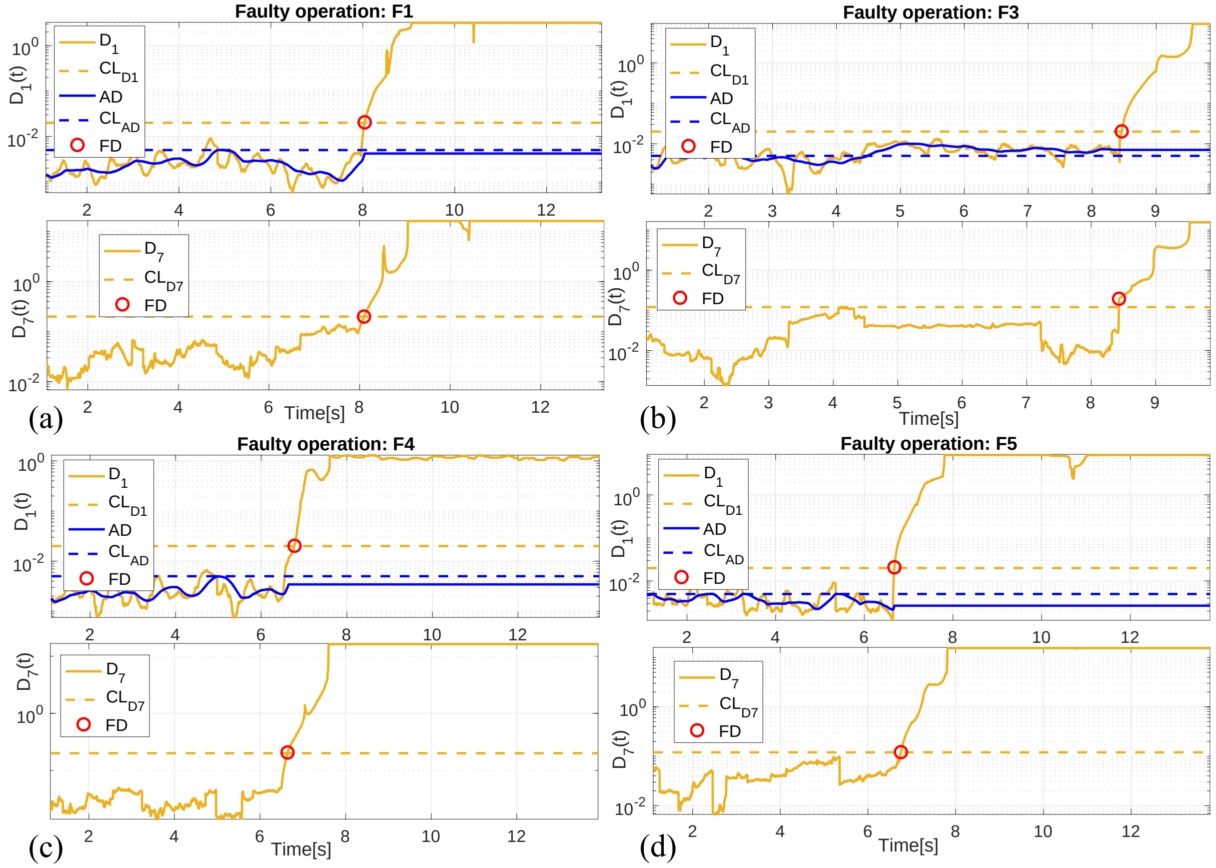


Figure 12. Detection performance: (a) Inverter fault F1, (b) Grid anomaly F3, (c) Partial shading F4, (d) Open circuit F5.

In addition to the previous comparisons across F6, F2, and F7, the successful fault detections (FD) of the remaining four faults of Table 1 are all illustrated in Fig.12. The presented recursive algorithm was hence proved computationally efficient compared to multivariate KLD approaches [42, 50, 51] to match online real-time FD in GPV systems. The adaptive algorithm also considers the nonlinear time-varying behavior of GPV systems by triggering updates upon changes of power point which were successfully sensed through the novel discrimination index $AD(t)$. For these improvements, the obtained results of Fig.9-11 demonstrated superior detection performance compared to traditional PCA's Q and T^2 statistics [27, 28, 37, 38]. Moreover, the developed assumption-free fault indicators $D_1(t)$ and $D_7(t)$ greatly outperformed the parametrized KLD approaches [43-45] as proved in Fig.9. Thus, the adaptive assumption-free PCA-KDE-KLD-based fault indicators and discrimination index are efficient and effective in detecting realistic faults in GPV systems in real-time.

Table 2. Comparison of robustness, detection sensitivity, and computational time.

FD approach	False alarms	Fault detection							Computational time (s)
		F1	F2	F3	F4	F5	F6	F7	
PCA Q, T^2	1%	✓	✓	✗	✗	✗	✗	✗	4.009
I	17.43%	✓	✓	✗	✗	✓	✗	✗	98.474

KLD*	5.94%	✓	✓	✗	✗	✓	✗	✗	21600*
Proposed	<1%	✓	✓	✓	✓	✓	✓	✓	11.809

* sampling frequency, number of evaluation points, and reference sample size are reduced to a rate of 1/10.

Table 2 compares fault detection robustness and sensitivity as well as computational complexity of closely related methods for real-time fault detection performance, (✓) refers to definite detection, whereas (✗) means no prompt detection. The results are obtained through offline performance evaluation using real GPV system data collected during real faults which are injected manually in the system; the timestamp of the physical fault-occurrence is not known precisely. This poses a challenge for the fault detection algorithms which are trained in a semi-supervised approach. Unfortunately, the detection rate and detection delay cannot be assessed precisely without a reference method. Table 2 lists the highest false alarms rate, successful detection of the seven GPV system faults, and average offline computational time on experimental test data collected during 15 seconds. The computational time of a method should be under 15s to allow its online implementation. Recall that lower false alarms rates ensure higher robustness, the more detected faults reflect a higher detection sensitivity, and the shorter computational time indicates computational efficiency. A remarkable advantage of PCA is its dimensionality reduction which yields the fastest data processing to match online application (4.009 << 15s), this aspect is utilized in the proposed approach to reduce computational complexity without a loss of generality. Parametrized KLD approaches [43-46, 60] (referred to as the I index in Fig.9(e)) exhibit a slightly increased complexity where local mean and covariance of a seven-dimensional data of size $n_t = 1.1 \times 10^4$ are updated each timestamp. Multivariate KLD approaches [42, 50, 52] based on direct density (or indirect density-ratio) estimation are impractical for the seven-dimensional data, sample data size and evaluation points are reduced to verify their performance in a finite time. On the other hand, the distance-based Q and T^2 statistics of PCA are ineffective in detecting faults due to their assumptions of normality and stationarity. The parametrized multivariate KLD is slightly better in terms of detection (3/7 faults detected), but it is less robust with 17.43% false alarms since the assumption of normality deteriorates its performance. Despite its reduced parameters to allow a feasible computational time, nonparametric KLD achieves better performance. The proposed approach reduces the dimensionality of the problem using PCA which allows full and nonparametric KLD evaluation without assumptions nor any approximations; Moreover, the highest sensitivity and lower complexity allow for sensing varying conditions and updating the models which greatly improves the FD robustness and detection performance.

Faults in PV systems exhibit disparate characteristics under MPPT/IPPT controllers and varying environmental conditions. Consequently, fault classes exhibit varying characteristics which make it challenging for data-driven fault diagnosis approaches. The presented methods successfully mitigated this challenge for fault detection and they can be extended in future works from fault detection to fault diagnosis with cyber-attacks detection and diagnosis in multi-source wide-area power systems under the penetration of renewables with practical conditions.

5. Conclusions

This article presented an experimental analysis of real-time fault detection in grid-connected PV systems. Realistic-faults were injected in the system from which labelled data sets were collected from several experiments during varying power point through MPPT/IPPT modes and large variations in temperature and solar insolation. Moreover, this work examined the detectability of various types of GPV system faults at different levels and components including PV module mismatches such as open circuit and partial shading, inverter IGBT failure, grid anomalies in form of voltage sags, biased and slowed PI controller in the MPPT/IPPT boost converter controllers, and also current feedback sensor.

This work aimed for the online real-time detection of a set of seven faults injected in a real system. The nonlinear time-varying behavior of the GPV system was successfully treated through the designed intelligent algorithm which is explicitly data-driven. The major issue of high-dimensional high-frequency GPV system data was solved through the decorrelation of such space into the transformed components from which the two most sensitive detectors $D_1(t)$ and $D_7(t)$ were developed and successfully employed for online FD. The presented discrimination index $AD(t)$ was proved a very useful solution to distinguish the evolving normal behavior from faults to avoid false alarms and update statistical models and their parameters upon considerable changes in prevailing conditions only.

On the other hand, the novel fault indicators $D_1(t)$ and $D_7(t)$ were proved superior in detecting GPV system faults during the various tests for their assumption-free approach compared to parametrized KLD approaches and PCA's Q and T^2 statistics. Their higher accuracy is due to the exact KDE-based KLD across decorrelated one-dimensional TCs where the measured divergence is indeed negligible near zero during a pre-fault stage and increases considerably once the GPV system is under a faulty operation.

The novel algorithm of this work was designed, validated, tested, and compared based on extensive amounts of measured data from several tests on a GPV microgrid application. The obtained results proved the reliability of the proposed algorithm for computation-efficient and effective online real-time fault detection in GPV systems.

References

1. Schmela, M., 2018. Global Market Outlook for Solar Power 2018 – 2022. SolarPower Europe. http://www.solarpowereurope.org/
2. Renewable energy. 2018. BP global. https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy/renewable-energy.html/solar-energy.
3. Gunduz, H., & Jayaweera, D., 2018. Reliability assessment of a power system with cyber-physical interactive operation of photovoltaic systems. International Journal of Electrical Power & Energy Systems, 101, 371-384.
4. International Electrotechnical Commission, 2016. IEC 62548:2016. Photovoltaic (PV) arrays - Design requirements. https://webstore.iec.ch/publication/25949
5. International Electrotechnical Commission, 2017. IEC 61643-32:2017. Low-voltage surge protective devices. Part 32: Surge protective devices connected to the d.c. side of photovoltaic installations – Selection and application principles. https://webstore.iec.ch/publication/30774
6. International Electrotechnical Commission, 2018. IEC TR 63227 ED1. Lightning and surge voltage protection for photovoltaic (PV) power supply systems. https://www.iec.ch/dyn/www/f?p=103:27:12290348286666::::FSP_ORG_ID,FSP_LANG_ID:1274,25
7. de la Parra, I., Muñoz, M., Lorenzo, E., García, M., Marcos, J., Martínez-Moreno, F., 2017. PV performance modelling: A review in the light of quality assurance for large PV plants. Renew. Sust. Energy Rev. 78, 780-797.
8. Kumar, N. M., Dasari, S., Reddy, J. B., 2018. Availability factor of a PV power plant: evaluation based on generation and inverter running periods. Energy Procedia. 147, 71-77.
9. Cai, B., Liu, Y., Ma, Y., Huang, L., Liu, Z., 2015. A framework for the reliability evaluation of grid-connected photovoltaic systems in the presence of intermittent faults. Energy. 93, 1308-1320.
10. Madeti, S. R., Singh, S. N., 2017. Online fault detection and the economic analysis of grid-connected photovoltaic systems. Energy. 134, 121-135.
11. Saha, S., Haque, M. E., Tan, C. P., Mahmud, M. A., Arif, M. T., Lyden, S., & Mendis, N., 2020. Diagnosis and mitigation of voltage and current sensors malfunctioning in a grid connected PV system. International Journal of Electrical Power & Energy Systems, 115, 105381.
12. Naik, J., Dhar, S., & Dash, P. K., 2019. Adaptive differential relay coordination for PV DC microgrid using a new kernel based time-frequency transform. International Journal of Electrical Power & Energy Systems, 106, 56-67.
13. Huka, G. B., Li, W., Chao, P., & Peng, S., 2018. A comprehensive LVRT strategy of two-stage photovoltaic systems under balanced and unbalanced faults. International Journal of Electrical Power & Energy Systems, 103, 288-301.
14. Pillai, D. S., Rajasekar, N., 2018b. Metaheuristic algorithms for PV parameter identification: A comprehensive review with an application to threshold setting for fault detection in PV systems. Renew. Sust. Energy Rev. 82, 3503-3525.
15. Bhagavathy, S., Pearsall, N., Putrus, G., & Walker, S., 2019. Performance of UK Distribution Networks with single-phase PV systems under fault. International Journal of Electrical Power & Energy Systems, 113, 713-725.
16. Dashti, R., Ghasemi, M., Daisy, M., 2018. Fault location in power distribution network with presence of distributed generation resources using impedance based method and applying π line model. Energy. 159, 344-360.
17. Bukhari, S. B. A., Haider, R., Saeed Uz Zaman, M., Oh, Y.-S., Cho, G.-J., & Kim, C.-H., 2018. An interval type-2 fuzzy logic based strategy for microgrid protection. International Journal of Electrical Power & Energy Systems, 98, 209-218.
18. Dhimish, M., Holmes, V., Mehrdadi, B., Dales, M., Mather, P., 2017. Photovoltaic fault detection algorithm based on theoretical curves modelling and fuzzy classification system. Energy. 140, 276-290.
19. Manohar, M., Koley, E., Ghosh, S., Mohanta, D. K., & Bansal, R. C., 2020. Spatio-temporal information based protection scheme for PV integrated microgrid under solar irradiance intermittency using deep convolutional neural network. International Journal of Electrical Power & Energy Systems, 116, 105576.
20. Menke, J.-H., Bornhorst, N., & Braun, M., 2019. Distribution system monitoring for smart power grids with distributed generation using artificial neural networks. International Journal of Electrical Power & Energy Systems, 113, 472-480.
21. Ahmad, M. W., Mourshed, M., Rezgui, Y., 2018. Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. Energy. 164, 465-474.
22. Pillai, D. S., Rajasekar, N., 2018a. A comprehensive review on protection challenges and fault diagnosis in PV systems. Renew. Sust. Energy Rev. 91, 18-40.

23. Lu, S., Phung, B. T., Zhang, D., 2018. A comprehensive review on DC arc faults and their diagnosis methods in photovoltaic systems. Renew. Sust. Energy Rev. 89, 88-98.

24. Brooks, B., 2011. The Bakersfield Fire: A Lesson in Ground-Fault Protection. Sol. Pro. Mag., 4(2), 62-70.

25. Memon, A. A., Kauhaniemi, K., 2015. A critical review of AC Microgrid protection issues and available solutions. Elect. Pow. Syst. Res. 129, 23-31.

26. Wang, G., Xin, H., Wu, D., & Ju, P., 2019. Data-driven probabilistic small signal stability analysis for grid-connected PV systems. International Journal of Electrical Power & Energy Systems, 113, 824-831.

27. Bakdi, A., Kouadri, A., 2018. An improved plant-wide fault detection scheme based on PCA and adaptive threshold for reliable process monitoring: Application on the new revised model of Tennessee Eastman process. J. Chemom. 32(5), e2978.

28. Bakdi, A., Kouadri, A., Mekhilef, S., 2019. A data-driven algorithm for online detection of component and system faults in modern wind turbines at different operating zones. Renew. Sust. Energ. Rev. 103, 546-555.

29. Malvoni, M., De Giorgi, M. G., Congedo, P. M., 2016a. Data on Support Vector Machines (SVM) model to forecast photovoltaic power. Data Brief. 9, 13-16.

30. Chrétien, S., Clarkson, P., & Garcia, M. S., 2018. Application of Robust PCA with a structured outlier matrix to topology estimation in power grids. International Journal of Electrical Power & Energy Systems, 100, 559-564.

31. Mansouri, M., Hajji, M., Trabelsi, M., Harkat, M. F., Al-khazraji, A., Livera, A., Nounou, H., Nounou, M., 2018. An effective statistical fault detection technique for grid connected photovoltaic systems based on an improved generalized likelihood ratio test. Energy. 159, 842-856.

32. Fezai, R., Mansouri, M., Trabelsi, M., Hajji, M., Nounou, H., Nounou, M., 2019. Online reduced kernel GLRT technique for improved fault detection in photovoltaic systems. Energy. 179, 1133-1154.

33. Abadie, L. M., Chamorro, J. M., 2019. Physical adequacy of a power generation system: The case of Spain in the long term. Energy. 166, 637-652.

34. Ibrahim, H., Anani, N., 2017. Variations of PV module parameters with irradiance and temperature. Energy Procedia. 134, 276-285.

35. Reddy, G. S., Reddy, T. B., Kumar, M. V., 2017. A MATLAB based PV Module Models analysis under Conditions of Nonuniform Irradiance. Energy Procedia. 117, 974-983.

36. Poulek, V., Matuška, T., Libra, M., Kachalouski, E., Sedláček, J., 2018. Influence of increased temperature on energy production of roof integrated PV panels. Energ. Buildings. 166, 418-425.

37. Azzeddine Bakdi, Wahiba Bounoua, Saad Mekhilef, and Laith M. Halabi. 2019. Nonparametric Kullback-divergence-PCA for intelligent mismatch detection and power quality monitoring in grid-connected rooftop PV, Energy, 116366.

38. Bakdi, A., Kouadri, A., Bensmail, A., 2017. Fault detection and diagnosis in a cement rotary kiln using PCA with EWMA-based adaptive threshold monitoring scheme. Control Eng. Pract. 66, 64-75.

39. Basseville, M., 2013. Divergence measures for statistical data processing—An annotated bibliography. Signal Process. 93(4), 621-633.

40. MacKay, D. J. C., 2003. Information Theory, Inference and Learning Algorithms, First ed. Cambridge University Press, pp. 34.

41. Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer-Verlag, New York, pp. 55.

42. Hamadouche, A., Kouadri, A., Bakdi, A., 2017. A modified Kullback divergence for direct fault detection in large scale systems. J.Process Contr. 59, 28-36.

43. Chen, H., Jiang, B., Lu, N., 2018. An improved incipient fault detection method based on Kullback-Leibler divergence. ISA Trans. 79, 127-136.

44. Xie, L., Zeng, J., Kruger, U., Wang, X., Geluk, J., 2015. Fault detection in dynamic systems using the Kullback–Leibler divergence. Control Eng. Pract. 43, 39-48.

45. Zeng, J., Kruger, U., Geluk, J., Wang, X., Xie, L., 2014. Detecting abnormal situations using the Kullback–Leibler divergence. Automatica. 50(11), 2777-2786.

46. Delpha, C., Diallo, D., Youssef, A., 2017. Kullback-Leibler Divergence for fault estimation and isolation : Application to Gamma distributed data. Mech. Syst. Signal Process. 93, 118-135.

47. Adhikari, S., Li, F., 2014. Coordinated V-f and P-Q Control of Solar Photovoltaic Generators with MPPT and Battery Storage in Microgrids. IEEE Trans. Smart Grid. 5(3), 1270-1281.

48. Guichi, A., Talha, A., Berkouk, E. M., Mekhilef, S., Gassab, S., 2018. A new method for intermediate power point tracking for PV generator under partially shaded conditions in hybrid system. Sol. Energy. 170, 974-987.

49. 2011. IEEE Std C37.118.1-2011 (Revision of IEEE Std C37.118-2005) IEEE Standard for Synchrophasor Measurements for Power Systems. https://ieeexplore.ieee.org/document/6111219

50. Kawahara, Y., Sugiyama, M., 2012. Sequential change-point detection based on direct density-ratio estimation. Stat. Anal. Data Min: The ASA Data Science Journal. 5(2), 114-127.

51. Liu, S., Yamada, M., Collier, N., Sugiyama, M., 2013. Change-point detection in time-series data by relative density-ratio estimation. Neural Networks. 43, 72-83.

52. Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., Kawanabe, M., 2008. Direct importance estimation for covariate shift adaptation. Ann. Inst. Stat. Math. 60(4), 699-746.

53. Shimodaira, H., 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Stat. Plan. Inference. 90(2), 227-244.

54. Shang, J., Chen, M., Ji, H., Zhou, D., 2017. Recursive transformed component statistical analysis for incipient fault detection. Automatica. 80, 313-327.

55. Harmouche, J., Delpha, C., Diallo, D., 2014. Incipient fault detection and diagnosis based on Kullback–Leibler divergence using Principal Component Analysis: Part I. Signal Process. 94, 278-287.

56. El Heda, K., Louani, D., 2018. Optimal bandwidth selection in kernel density estimation for continuous time dependent processes. Stat. Probabil. Lett. 138, 9-19.

57. Wu, Z., Yang, M., Khoo, M. B. C., Castagliola, P., 2011. What are the best sample sizes for the Xbar and CUSUM charts? Int. J. Prod. Econ. 131(2), 650-662.

58. Weiß, C. H., Steuer, D., Jentsch, C., Testik, M. C., 2018. Guaranteed conditional ARL performance in the presence of autocorrelation. Comput. Stat. Data. Anal. 128, 367-379.

59. Langrené, N., & Warin, X., 2019. Fast and Stable Multivariate Kernel Density Estimation by Fast Sum Updating. Journal of Computational and Graphical Statistics, 28(3), 596-608.

60. Bounoua, W., Benkara, A. B., Kouadri, A., & Bakdi, A. (2019). Online monitoring scheme using principal component analysis through Kullback-Leibler divergence analysis technique for fault detection. Transactions of the Institute of Measurement and Control, 0142331219888370. https://doi.org/10.1177/0142331219888370

61. Yang, B., Yu, T., Zhang, X., Li, H., Shu, H., Sang, Y., & Jiang, L. (2019). Dynamic leader based collective intelligence for maximum power point tracking of PV systems affected by partial shading condition. Energy Conversion and Management, 179, 286-303. https://doi.org/10.1016/j.enconman.2018.10.074

62. Yang, B., Zhong, L., Zhang, X., Shu, H., Yu, T., Li, H., Jiang, L., Sun, L. (2019). Novel bio-inspired memetic salp swarm algorithm and application to MPPT for PV systems considering partial shading condition. Journal of Cleaner Production, 215, 1203-1222. https://doi.org/10.1016/j.jclepro.2019.01.150

63. Mohapatra, A., Nayak, B., Das, P., & Mohanty, K. B. (2017). A review on MPPT techniques of PV system under partial shading condition. Renewable and Sustainable Energy Reviews, 80, 854-867. https://doi.org/10.1016/j.rser.2017.05.083.

64. Cui, M., Wang, J., & Chen, B. (2020). Flexible Machine Learning-Based Cyberattack Detection Using Spatiotemporal Patterns for Distribution Systems. *IEEE Transactions on Smart Grid, 11*(2), 1805-1808. https://doi.org/10.1109/TSG.2020.2965797

65. Sun, C., Wang, X., Zheng, Y., Chen, S., & Yue, Y. (2019). Early warning system for spatiotemporal prediction of fault events in a power transmission system. *IET Generation, Transmission & Distribution, 13*(21), 4888-4899. https://doi.org/10.1049/iet-gtd.2018.6389

66. Dubey, R., Samantaray, S. R., & Panigrahi, B. K. (2017). An spatiotemporal information system based wide-area protection fault identification scheme. *International Journal of Electrical Power & Energy Systems, 89*, 136-145. doi:https://doi.org/10.1016/j.ijepes.2017.02.001

67. Sun, Y., Chen, X., Yang, S., Rusli, Tseng, K. J., & Amaratunga, G. (2017, 12-15 Dec. 2017). *Micro PMU based monitoring system for active distribution networks.* Paper presented at the 2017 IEEE 12th International Conference on Power Electronics and Drive Systems (PEDS). https://doi.org/10.1109/PEDS.2017.8289180

68. Gomathi, V., & Ramachandran, V. (2011, 3-5 Jan. 2011). *Optimal location of PMUs for complete observability of power system network.* Paper presented at the 2011 1st International Conference on Electrical Energy Systems. https://doi.org/10.1109/ICEES.2011.5725349