Energy Conversion and Management 301 (2024) 118076

Contents lists available at ScienceDirect

Energy Conversion and Management

journal homepage: www.elsevier.com/locate/enconman

Faults detection and diagnosis of PV systems based on machine learning approach using random forest classifier

Ahmed Faris Amiri [a,b,*], Houcine Oudira [b], Aissa Chouder [c], Sofiane Kichou [d]

[a] Laboratory of Signal and System Analysis (LASS), Electronic Department, University of M'sila, PO Box166 Ichebilia, 28000 M'sila, Algeria
[b] Laboratory of Electrical Engineering (LGE), Electronic Department, University of M'sila, PO Box166 Ichebilia, 28000 M'sila, Algeria
[c] Laboratory of Electrical Engineering (LGE), Electrical Engineering Department, University of M'sila, PO Box166 Ichebilia, 28000 M'sila, Algeria
[d] Czech Technical University in Prague, University Centre for Energy Efficient Buildings, 1024 Třinecká St., 27343 Buštěhrad, Czech Republic

ARTICLE INFO

Keywords:
MGWO algorithm
Random forest classifier
Parameter extraction
Dynamic MPP model
Fault detection
Co-simulation

ABSTRACT

Accurate and reliable fault detection procedures are crucial for optimizing photovoltaic (PV) system performance. Establishing a trustworthy PV array model is the primary step and a vital tool for monitoring and diagnosing PV systems. This paper outlines a two-step approach for creating a reliable PV array model and implementing a fault detection procedure using Random Forest Classifiers (RFCs).

Firstly, we extracted the five unknown parameters of the one-diode model (ODM) by combining the current–voltage translation method to predict the reference curve and employing the modified grey wolf optimization (MGWO) algorithm. In the second step, we simulated the PV array to obtain maximum power point (MPP) coordinates and construct operational databases through co-simulations in PSIM/MATLAB. We developed two RFCs: one for fault detection (a binary classifier) and another for fault diagnosis (a multiclass classifier).

Our results confirmed the accuracy of the PV array modeling approach. We achieved a root mean square error (RMSE) value of 0.0122 for the ODM parameter extraction and RMSEs lower than 0.3 in dynamic PV array output current simulations under cloudy conditions. Regarding the fault detection procedure, our results demonstrate exceptional classification accuracy rates of 99.4% for both fault detection and diagnosis, surpassing other tested models like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Neural Networks (MLP Classifier), Decision Trees (DT), and Stochastic Gradient Descent (SGDC).

1. Introduction

In recent years, worldwide energy policies have focused on reducing carbon footprints and moving towards more sustainable energy sources. This is reflected by an increased adoption of renewable energy sources (RES) to ensure a greener future. Among RES, solar photovoltaics (PV) is identified as a key energy source, addressing environmental concerns at very competitive costs [1]. As a result, global PV capacity surpassed the terawatt threshold in early 2022, accounting for two-thirds of the projected increase in global renewable capacity by 2023 [2].

PV systems are designed to operate under harsh external conditions, including extreme weather situations, wind-induced vibrations, and exposure to ultraviolet radiation [3,4]. In these demanding environments, various malfunctions and failures may occur, potentially shortening the lifespan of PV modules, reducing the overall system's energy yields, compromising system availability, and posing safety risks to personnel involved in their operation and maintenance [5]. Hence, the early detection and diagnosis of faults is paramount for ensuring the long-term reliability and sustainable operation of the entire PV system.

Multiple methods for detecting and diagnosing faults in PV systems have emerged over the last decade. Model-based approach procedures involve simulating the performance of the actual PV installation and comparing the simulated output power with the monitored one [6,7]. Chouder and Silvester introduced a fault detection methodology for PV systems based on power loss analysis, categorizing identified faults into a faulty string, faulty module, and partial shading through detailed analysis of simulated and measured output ratios [8]. Silvestre et al. presented an automated procedure for fault detection in grid-connected PV systems centered on current and voltage indicators [9]. The method involves setting thresholds based on typical operational behavior,

[*] Corresponding author at: Laboratory of Signal and System Analysis (LASS), Laboratory of Electrical Engineering (LGE), Electronic Department, University of M'sila, PO Box166 Ichebilia, 28000 M'sila, Algeria.
E-mail address: ahmedfaris.amiri@univ-msila.dz (A.F. Amiri).

https://doi.org/10.1016/j.enconman.2024.118076
Received 24 October 2023; Received in revised form 3 January 2024; Accepted 5 January 2024
0196-8904/© 2024 Elsevier Ltd. All rights reserved.

triggering a fault signal when surpassed, and identifying faults by analyzing current and voltage ratios. Drews et al. employed a fault detection method by setting a power residual threshold using weather satellite data for irradiance and temperature instead of on-site sensors [10]. While this obviates the need for additional on-site sensors, it may compromise accuracy due to potentially more significant margins of error in weather data. In contrast, Garoudja et al.'s method sets a threshold on the exponentially weighted moving average of current, voltage, and power residuals, integrating historical data for fault detection rather than relying solely on the most recent observation [11]. Overall, the fault detection techniques mentioned above are straightforward to implement. Nevertheless, the primary challenge lies in precisely selecting suitable thresholds to ensure their reliability.

In recent years, diverse artificial intelligence (AI) techniques encompassing Machine Learning (ML) and Deep Learning (DL) have been incorporated as the core methodologies of PV fault detection and diagnosis due to their excellent capabilities in addressing feature extraction and classification problems. Several ML techniques were developed for fault detection and diagnosis in PV systems [12–15]. Among these techniques, artificial neural network (ANN), support vector machine (SVM), and Random Forest (RF) are the most common approaches. Bendary et al. proposed two adaptive neuro-fuzzy inference system-based controllers (ANFIS) to address cleaning, tracking, and faulty issues in PV systems [16]. The method is based on associating the actual measured values of current and voltage with respect to the trained historical values for this parameter while considering the ambient changes in conditions, including irradiation and temperature. Madeti and Singh introduced an algorithm based on k-nearest neighbors (KNN) for real-time fault detection in PV systems, capable of detecting and classifying open circuit, line-to-line, and partial shading faults [17]. However, it's important to note that the method's accuracy is not flawless compared to its computational efficiency. Eskandari et al. proposed an ensemble learning method that combines three algorithms—Support Vector Machine (SVM), Naïve Bayes (NB), and KNN [18]. The selected classifiers exhibited impressive performance with an accuracy rate of up to 99.5 %. Nonetheless, it's worth noting that this method was specifically developed for detecting line-to-line faults. Similarly, Kapucu et al. explored an ensemble learning approach that integrates quadratic discriminant analysis (QDA), extra trees with entropy (ETent), and decision trees (DT) [19]. Their investigation focused on identifying two PV faults — partial shading and short circuit — the method achieved an initial accuracy rate of 97.46 %, which increased to 97.67 % after optimization. Likewise, Adhya et al. utilized a diagnostic approach comprising the light gradient boosting method (LGBM), categorical boosting (CatBoost), and extreme gradient boosting (XGBoost) to identify faults in PV systems [20]. This combination of diverse ML algorithms resulted in an impressive accuracy of approximately 99 %. However, despite these promising outcomes, the approach remains intricate, prompting the need for further refinements and enhancements. Akram et al. proposed a monitoring method for the DC side of PV arrays, employing the Probabilistic Neural Network (PNN). Their approach demonstrated a good classification accuracy, reaching 98.53 %. However, the method was specifically tested for detecting and classifying short- and open-circuit faults [21]. Chen et al. utilized a RF to classify partial shading, degradation, open circuit, and short circuit faults, employing only high-frequency current and voltage measurements in parallel circuit substrings [22]. This work showed good results. However, it was based on a limited range of operating weather conditions. Likewise, Gong et al. utilized the classification regression tree to address the issue of photovoltaic array fault diagnosis [23]. The method is based on I-V curves generated under specific working conditions, and the obtained classification accuracy was 97.9 %. Mellit et al. developed an embedded system for remote monitoring and fault diagnosis of PV systems based on two conventional neural network models [24]. The first ANN is used for fault detection, while the second deals with fault diagnosis. Both ML algorithms showed good accuracy when embedded into a low-cost edge device for real-time diagnosis of a PV array. On the other hand, the emergence of DL algorithms represents a transformative leap in machine learning, gaining considerable attention for their prowess in pattern recognition, data mining, and knowledge discovery. A notable contribution in this domain comes from Gao et al. [25], where they introduce a DL approach that integrates a stacked autoencoder (SAE) with a multi-grained cascade forest for diagnosing PV faults – associated with partial shading, open circuit, and short circuit faults – without needing weather data or I-V curves as inputs. In this approach, the SAE extracts the fault features automatically from normalized sequence waveforms of string current and voltage, while the multi-grained cascade forest is responsible for diagnosis. In parallel endeavors, Liu et al. introduced a fault diagnosis method for a PV array utilizing SAE and clustering [26]. This approach mines inherent I-V characteristics, enabling automatic feature extraction and fault diagnosis. In addition, Chen et al. presented an innovative deep residual network (ResNet) for intelligent fault detection and diagnosis. Leveraging output, I-V characteristic curves, and input ambient condition data, this novel approach adds depth to fault analysis [27].

Despite the effectiveness of the AI-based fault detection and diagnostics procedures, their accuracy is compromised by the data used in their training stage. Actual measurement data are not enough to train AI models, so to conceive a trustful training database, developing an accurate model of the PV system is crucial. Efficient models are essential to fully replicate the PV systems operation considering various faults and outdoor conditions. Furthermore, data processing is another crucial approach to consider in deploying AI-based machine learning procedures for PV diagnosis. As Wang et al. have underscored the data processing importance in increasing the accuracy of ML-based algorithms to categorize complex faults in the range of 81 %-99 % [28].

The present work's contributions involve developing a robust PV model that is the foundation for monitoring and fault detection — whether AI-based or conventional model-based — techniques. Additionally, it introduces a fault detection procedure based on Random Forest Classifiers, optimized through a grid-search algorithm for hyperparameter tuning. The adopted methodology unfolds in two crucial steps:

- In the first step, we focus on accurately identifying the unknown parameters of the One-Diode Model (ODM) of the PV array operating under outdoor conditions. This is achieved through a novel application of the translation technique designed to correct randomly measured current–voltage (I-V) curves to reference standard test conditions (STC). The translation technique employs analytical formulations to derive these parameters across various operating conditions, accounting for variations in irradiance and temperature [29]. To determine the five unknown parameters of the ODM at STC, we utilize an optimization algorithm based on the Modified Grey Wolf Optimization (MGWO), an approach initially introduced by Mirjalili et al. in 2014 [30]. The MGWO algorithm's innovative position updating concept enables more efficient searching and exploitation capabilities while maintaining rapid convergence speed. Subsequently, based on the identified parameters, we derived and simulated the evolution of the maximum power point (MPP) model using actual dynamic measurements from a grid-connected PV system in Algeria.
- The second step in our approach involves simulating the PV array to extract MPP coordinates and construct its operational databases through PSIM/MATLAB co-simulations. Additionally, we implement an efficient fault detection and diagnosis process by leveraging the Random Forest Classifier (RFC). This entails the development of two RFCs: the first for fault detection (a binary classifier) and the second for fault diagnosis (a multiclass classifier). Finally, we comprehensively compare our approach with other machine-learning techniques for detecting and diagnosing faults in the considered grid-connected PV system. The testing phase encompasses five

operating scenarios: a healthy system, three short-circuited modules in one string, a line-to-line fault, a string disconnected from the array, and the shading effects on three panels.

The remainder of this paper is organized as follows: Section 2 comprehensively describes the PV system utilized to validate the proposed methods. Section 3 explores the novel approach to PV modeling and parameter extraction. Section 4 is dedicated to explaining the developed fault detection approach. Section 5 presents the results obtained, accompanied by in-depth discussions to elucidate the methodology's performance and effectiveness. Finally, Section 6 summarizes the conclusions drawn from this study.

2. Experimental setup description

To rigorously assess the accuracy of the proposed fault detection methodology and the new procedure for extracting the PV model's unknown parameters, monitored data from a grid-connected PV system were used. The proposed PV system is located in Algiers, Algeria, at coordinates 36°43′N latitude and 3°15′E longitude. This PV installation boasts a total capacity of 9.54 kW, organized into three sub-arrays, each with a capacity of 3.18 kW. Each sub-array comprises 30 Isofoton 106–12 panels arranged in two parallel strings of 15 modules in series. These PV modules are connected to a 2.5 kW single-phase inverter (IG30 Fronius).

The PV plant's tilted and horizontal irradiance levels are monitored using a Kipp & Zonen CM11 thermoelectric pyranometer. Additionally, temperature measurements of the PV modules are conducted using K-type thermocouples. Meteorological and electrical variables are systematically recorded using a data logger (Agilent 34970). The data, including weather (Solar irradiance (*G*), module temperature (*T*), and PV output (*Impp*, *Vmpp*, *Pmpp*) parameters at the Maximum Power Point (MPP), were collected at a sampling rate of 1 min.

The main specifications of the selected PV array used in this work are listed in Table 1, while further details of the whole PV installation can be found in [31].

Table 2 summarizes the key electrical parameters for the Isofoton 106–12 PV module under Standard Test Conditions (STC), characterized by a temperature of 25 °C and an irradiance level of 1000 W/m^2.

3. Developed approach for PV modeling

The basis for our photovoltaic (PV) modeling approach is the widely adopted one-diode, five-parameter solar cell model [32]. This model is a popular choice in PV module modeling for various technologies, encompassing crystalline and thin-film designs. It is favored for striking a balance between model complexity and predictive accuracy. The solar cell I–V characteristic is described by the implicit and nonlinear expression given in Eq. (1).

$$I = I_{ph} - I_o\left[\exp\left(\frac{q(V+R_sI)}{nkT}\right) - 1\right] - \frac{V+R_sI}{R_{sh}} \tag{1}$$

where I_0 is the diode saturation current (A). I_{ph} represents the photocurrent in (A). n is the diode ideality factor. The Boltzmann constant (1.38x10^{-23}JK^{-1}) is defined by k. T is the cell temperature in (K). The parameter q is the electrical charge (1.602 x10^{-19}C). V$_t$(V) is the thermal voltage expressed as $V_t = kT/q$. Finally, R_{sh} and R_s represent shunt and series resistances (Ω). For an in-depth understanding of this model, including the requisite equations to extend its applicability from a single solar cell to an entire PV array, an extensive description is referenced in [33].

Table 1
Main specifications of the selected PV array.

Parameter	Description
Module technology	Mono-crystalline (mc-Si)
PV array nominal power	3.18 kWp
Inverter type and size	IG30 Fronius single-phase, 2.5 kW
Modules per inverter	30
Modules in series (*Ns*)	15
Strings in parallel (*Np*)	2
Tilt - Azimuth	35° – 10° West

Table 2
Electrical characteristics of the considered PV module.

Parameter	Value
P_{mp} (W)	106
I_{SC} (A)	6.54
V_{OC} (V)	21.6
I_{mp} (A)	6.10
V_{mp} (V)	17.4
βV_{OC} (%/°C)	−0.36
αI_{SC} (%/°C)	0.06

The five parameters, namely I_{ph}, I_o, n, R_{sh}, and R_s, are typically not explicitly provided by PV module manufacturers. Previous investigation revealed that the extracted actual values of these parameters often deviate from calculated ones based on nominal data provided in the datasheet specified at the STC [34]. Consequently, achieving a precise alignment between the PV model outputs defined by Eq. (1) and real-world monitored data is essential for accurate simulation and fault detection. Therefore, the necessity of using an effective parameter identification procedure is crucial.

3.1. Current-voltage translation to reference conditions

The parameters of PV cells are notably influenced by weather conditions, making it inaccurate to assume their constancy. Additionally, the mathematical expressions employed in these models depend on access to reference parameters. However, replicating the standard test conditions proves challenging under typical outdoor conditions. To address this challenge, we present an efficient translation method inspired by a technique introduced in [29] initially employed for analyzing the degradation of amorphous silicon-based modules. This method transforms three arbitrary I-V curves, each obtained under varying temperature and irradiance conditions, into a reference curve.

It is important to note that many translation methods in literature often necessitate prior knowledge of additional parameters. In contrast, our innovative approach requires no prior information about temperature coefficients or internal parameters. It solely relies on data obtained from three measured I-V curves (Curves 1,2 and 3) defined as:

- Curve 1: $(V_1[i], I_1[i])$ where i = 1,…,n$_1$, measured at an irradiance G$_1$ and a cell temperature T$_1$
- Curve 2: $(V_2[j], I_2[j])$ where j = 1,…,n$_2$, measured at an irradiance G$_2$ and a cell temperature T$_2$
- Curve 3: $(V_3[k], I_3[k])$ where k = 1,…,n$_3$, measured at an irradiance G$_3$ and a cell temperature T$_3$

The proposed methodology is rooted in the derivation of a new Curve 0, denoted by $(V_0[i], I_0[i])$, aligning with the desired conditions G_0 and T_0 at standard test conditions (STC). An intermediate curve is introduced to achieve this, initiating an interpolation process denoted as Curve 4, governed by the operating conditions G_4 and T_4. Initially, Curve 4 is extracted from Curve 1 and Curve 2. Subsequently, Curve 3 and Curve 4 are employed to attain the target Curve 0. The interpolation process begins within the irradiance/temperature plane, as illustrated in Fig. 1, and is subsequently carried out in the voltage/current space, employing identical parameters as elaborated below.

The values of G_4 and T_4 are established based on combinations of G_1 and G_2, and T_1 and T_2, respectively, as depicted in Eqs. (2) and (3),

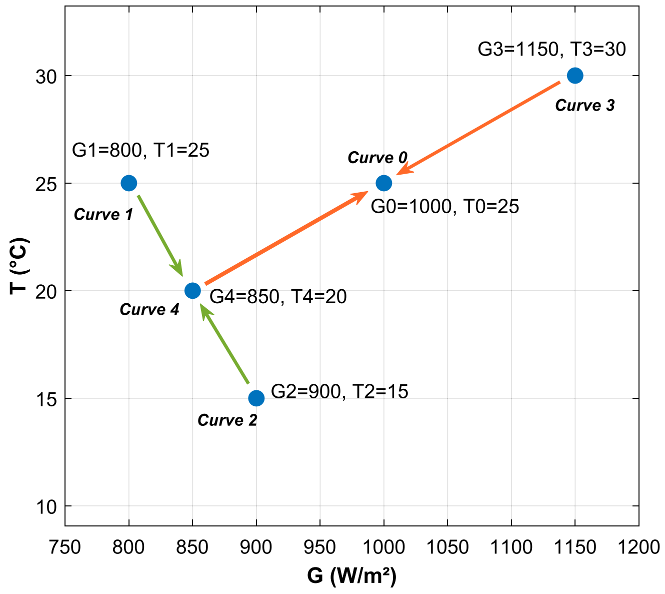

Fig. 1. The operating conditions of curves 1, 2, and 3 are interpolated to obtain the operating conditions of Curves 4 and 0.

wherein the parameter α will be determined. Additionally, as demonstrated in Eqs. (4) and (5), the desired irradiance G_0 and temperature T_0 are inferred from G_3 and G_4, as well as T_3 and T_4, respectively, with the incorporation of another unknown parameter $\varnothing$. This configuration yields a system of four equations and four unknowns (G_4, T_4, α, and $\varnothing$).

$$G_4 = G_1 + \propto (G_2 - G_1) \tag{2}$$

$$T_4 = T_1 + \propto (T_2 - T_1) \tag{3}$$

$$G_0 = G_3 + \varnothing (G_4 - G_3) \tag{4}$$

$$T_4 = T_3 + \varnothing (T_4 + T_3) \tag{5}$$

The set of equations specified in the standard conditions has been streamlined by introducing a novel translation parameter, denoted as ω, defined as the product of $\varnothing$ and α. Additionally, the values of G_4 and T_4 have been integrated into Eqs. (4) and (5), resulting in the formulation of Eqs. (6) and (7), which can be straightforwardly computed.

$$G_0 - G_3 = (G_1 - G_3).\varnothing + (G_2 - G_1).\omega \tag{6}$$

$$T_0 - T_3 = (T_1 - T_3).\varnothing + (T_2 - T_1).\omega \tag{7}$$

The next step is intended to find the I-V curves. It has been assumed that I_{SC1} and I_{SC2} are the short-circuit currents of Curve 1 and Curve 2, respectively. For each point of Curve 1 $(\mathrm{V}_1[\mathrm{i}], \mathrm{I}_1[\mathrm{i}])$, its partner $(\mathrm{V}_2[\mathrm{j}], \mathrm{I}_2[\mathrm{j}])$ is sought in Curve 2 so that the next condition is satisfied: $\mathrm{I}_2[\mathrm{j}] - \mathrm{I}_1[\mathrm{i}] = \mathrm{I}_{\mathrm{sc}2} - \mathrm{I}_{\mathrm{sc}1}$. Then, a new point $(\mathrm{V}_4[\mathrm{i}], \mathrm{I}_4[\mathrm{i}])$ of Curve 4 is obtained by applying Eqs. (8) and (9). By the same manner, for each point of Curve 3 $(\mathrm{V}_3[\mathrm{i}], \mathrm{I}_3[\mathrm{i}])$, the best matching point $(\mathrm{V}_4[\mathrm{j}], \mathrm{I}_4[\mathrm{j}])$ of Curve 4 is selected fulfilling $\mathrm{I}_4[\mathrm{j}] - \mathrm{I}_3[\mathrm{i}] = \mathrm{I}_{\mathrm{sc}4} - \mathrm{I}_{\mathrm{sc}3}$ and the point $(\mathrm{V}_0[\mathrm{i}], \mathrm{I}_0[\mathrm{i}])$ of Curve 0 is produced based on Eqs. (10) and (11).

$$V_4[i] = V_4[i] + \propto (V_1[i] - V_2[j]) \tag{8}$$

$$I_4[i] = V_4[i] + \propto (I_1[i] - I_2[j]) \tag{9}$$

$$V_0[i] = V_3[i] + \varnothing (V_3[i] - V_4[j]) \tag{10}$$

$$I_0[i] = V_3[i] + \varnothing (I_3[i] - I_4[j]) \tag{11}$$

3.2. Parameter extraction based on modified grey wolf optimization

In this section, we introduce an offline optimization method for parameter identification. The reason for opting for the optimization approach is that the characteristic equations, as defined in Eq. (1), have an implicit form making the direct parameters identification challenging. The parameter identification process can be likened to an optimization problem, and we tackle this challenge using the Modified Grey Wolf Optimization (MGWO) algorithm. This method effectively optimizes the unknown parameters to reconcile the implicit characteristic equations, enabling us to precisely determine the desired values based on actual measurement data. In this approach, we focus on quantifying the disparity between the outputs derived from Eq. (1) and the data obtained from the current–voltage translation methodology described above (section 3.1). We employ the root mean square error (RMSE) as a key criterion to measure this difference. For each set of experimental values (I, V), the RMSE is computed according to the following formula:

$$RMSE = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N} (f(V,I,x))^2\right)} \tag{12}$$

$$f(V,I,x) = I - \left(I_{ph} - I_o\left[\exp\left(\frac{q(V + R_s I)}{nkT}\right) - 1\right] - \frac{V + R_S I}{R_{sh}}\right) \tag{13}$$

where, $x = \left[I_{ph,ref}, I_{o,ref}, R_{sh,ref}, R_{s,ref}, n_{ref}\right]$, and N represents the data points quantity.

The classical Grey Wolf Optimizer (GWO) algorithm was introduced in 2014 by Mirjalili et al. [30], and its mathematical social behavior model is represented as follows:

$$\overrightarrow{\mathrm{D}} = \left|\overrightarrow{\mathrm{C}}.\overrightarrow{\mathrm{X}_{\mathrm{p}}}(\mathrm{t}) - \overrightarrow{\mathrm{X}}(\mathrm{t})\right| \tag{14}$$

$$\overrightarrow{\mathrm{X}}(\mathrm{t}+1) = \overrightarrow{\mathrm{X}_{\mathrm{p}}}(\mathrm{t}) - \overrightarrow{\mathrm{A}}.(\overrightarrow{\mathrm{D}}) \tag{15}$$

where t is the current iteration, $\overrightarrow{\mathrm{X}_{\mathrm{p}}}$ is the position vector of the prey, $\overrightarrow{\mathrm{X}}$ is the position vector of the hail wolf, and $\overrightarrow{\mathrm{A}}$ and $\overrightarrow{\mathrm{C}}$ are coefficient vectors, calculated as follows:

$$\begin{cases} \overrightarrow{\mathrm{A}} = 2\,\overrightarrow{\mathrm{a}}.\overrightarrow{\mathrm{r}_1} - \overrightarrow{\mathrm{a}} \\ \overrightarrow{\mathrm{C}} = 2.\overrightarrow{\mathrm{r}_2} \end{cases} \tag{16}$$

where the components of $\overrightarrow{\mathrm{a}}$ decrease linearly from 2 to 0 over the course of the iterations and $\overrightarrow{\mathrm{r}_1}$, $\overrightarrow{\mathrm{r}_2}$ are random numbers in [0,1]. The equation for position update is shown below.

$$\begin{cases} \overrightarrow{\mathrm{D}_\alpha} = \left|\overrightarrow{\mathrm{C}_1}.\overrightarrow{\mathrm{X}_\alpha} - \overrightarrow{\mathrm{X}}\right| \\ \overrightarrow{\mathrm{D}_\beta} = \left|\overrightarrow{\mathrm{C}_2}.\overrightarrow{\mathrm{X}_\beta} - \overrightarrow{\mathrm{X}}\right| \\ \overrightarrow{\mathrm{D}_\delta} = \left|\overrightarrow{\mathrm{C}_3}.\overrightarrow{\mathrm{X}_\delta} - \overrightarrow{\mathrm{X}}\right| \end{cases} \tag{17}$$

$$\begin{cases} \overrightarrow{\mathrm{X}_1} = \overrightarrow{\mathrm{X}_\alpha} - \overrightarrow{\mathrm{A}_1}.(\overrightarrow{\mathrm{D}_\alpha}) \\ \overrightarrow{\mathrm{X}_2} = \overrightarrow{\mathrm{X}_\beta} - \overrightarrow{\mathrm{A}_2}.(\overrightarrow{\mathrm{D}_\beta}) \\ \overrightarrow{\mathrm{X}_3} = \overrightarrow{\mathrm{X}_\delta} - \overrightarrow{\mathrm{A}_3}.(\overrightarrow{\mathrm{D}_\delta}) \end{cases} \tag{18}$$

Each wolf in the pack updates its position following the positions of $\overrightarrow{\mathrm{X}_1}$, $\overrightarrow{\mathrm{X}_2}$, and $\overrightarrow{\mathrm{X}_3}$ which stand for the top three solutions thus far in the iteration process.

$$\overrightarrow{\mathrm{X}}(\mathrm{t}+1) = \frac{\overrightarrow{\mathrm{X}_1} + \overrightarrow{\mathrm{X}_2} + \overrightarrow{\mathrm{X}_3}}{3} \tag{19}$$

This article introduces an adaptable method that leverages the GWO algorithm, with a minor modification in the selection phase. As depicted in Fig. 2, the diagram outlines the steps of the proposed Modified GWO (MGWO) technique, which closely aligns with a method previously employed in a prior study [35]. This technique determines alpha, beta, and delta members by evaluating the fitness function for individual positions, specifically the five unknown parameters. Other agents adjust their positions accordingly.

A novel approach to position updating is integrated into GWO, enhancing both search and exploitation capabilities while ensuring rapid convergence. This novel concept draws inspiration from the competitive exclusion method found in genetic algorithms [36]. In this approach, only positions from the current iteration of search agents (wolves) that exhibit higher fitness compared to positions from previous iterations are replaced. Only the top positions are considered during the final phase for selecting new alpha, beta, and delta members. The process iterates to update search agent positions based on these selections, repeating as necessary to reach the maximum number of iterations [37]. The MGWO with an additional phase can search for fully optimal results without using any parameters like conventional methods would.

3.3. Prediction of the PV outputs under actual outdoor conditions

Using fully analytical formulas and reference parameters obtained through the MGWO algorithm, the next crucial step involves establishing the values of the unknown parameters within real operational contexts. Eqs. (20) to (25) encompass the analytical expressions that enable the calculation of the five parameters in question as functions of temperature and irradiance [38–40].

$$n(T) = n_{ref}\left(T/T_{ref}\right) \tag{20}$$

$$I_{ph}(G,T) = \frac{G}{G_{ref}}\left[I_{ph,ref} + \alpha\left(T - T_{ref}\right)\right] \tag{21}$$

$$R_{sh}(G) = R_{sh,ref}\left(G_{ref}/G\right) \tag{22}$$

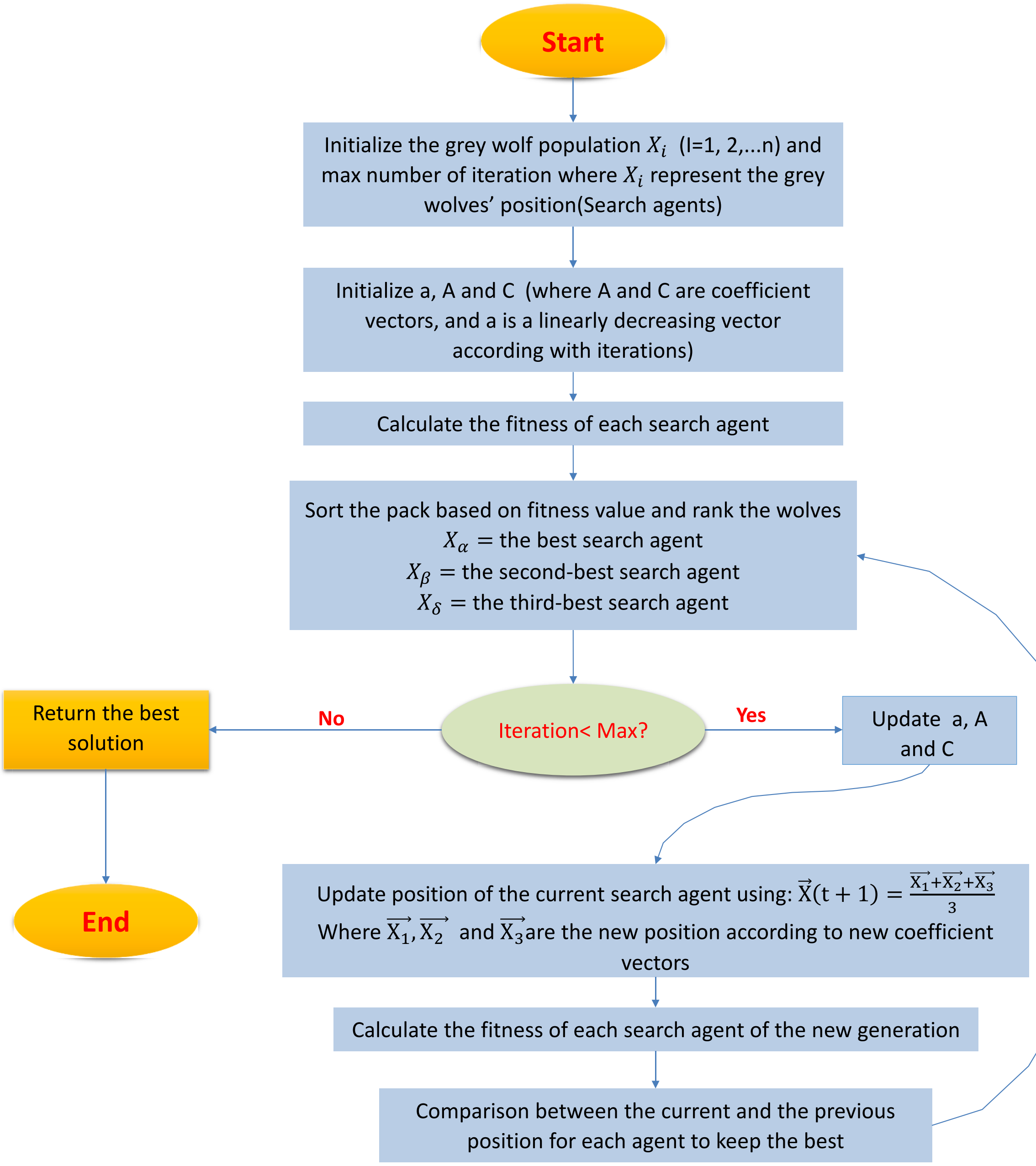
Start
Initialize the grey wolf population X_i (I=1, 2,...n) and max number of iteration where X_i represent the grey wolves' position(Search agents)
Initialize a, A and C (where A and C are coefficient vectors, and a is a linearly decreasing vector according with iterations)
Calculate the fitness of each search agent
Sort the pack based on fitness value and rank the wolves
X_α = the best search agent
X_β = the second-best search agent
X_δ = the third-best search agent
Iteration< Max?
No
Yes
Return the best solution
End
Update a, A and C
Update position of the current search agent using: $\vec{X}(t+1) = \frac{\vec{X_1}+\vec{X_2}+\vec{X_3}}{3}$
Where $\vec{X_1}, \vec{X_2}$ and $\vec{X_3}$are the new position according to new coefficient vectors
Calculate the fitness of each search agent of the new generation
Comparison between the current and the previous position for each agent to keep the best

Fig. 2. MGWO algorithm flowchart.

$$E_g = E_{g,ref}\left[1 - 0.0002677(T - T_{ref})\right] \tag{23}$$

$$R_s(G,T) = R_{s,ref}\left(\frac{T}{T_{ref}}\right)\left[1 - \beta ln\left(G/G_{ref}\right)\right] \tag{24}$$

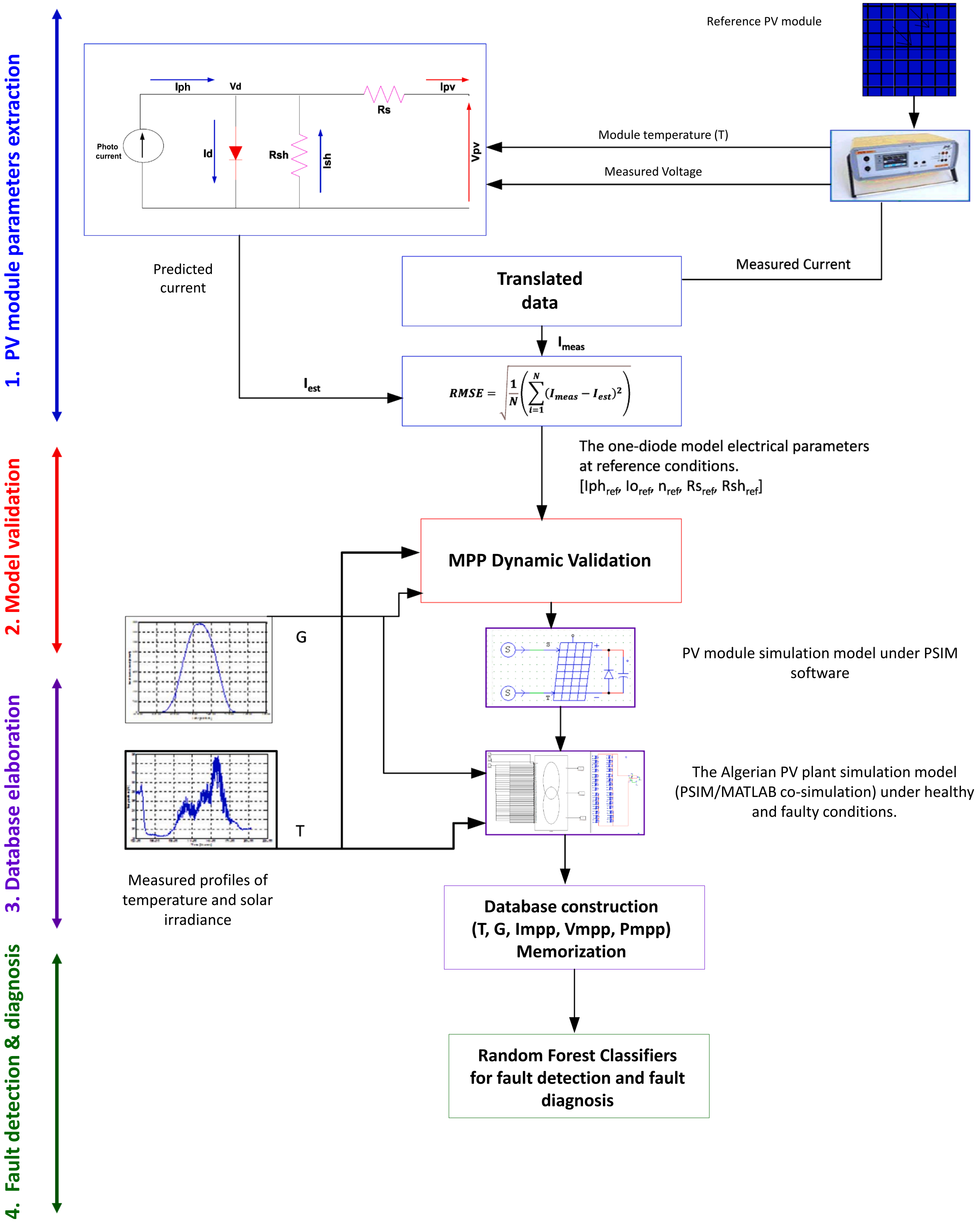
1. PV module parameters extraction
2. Model validation
3. Database elaboration
4. Fault detection & diagnosis
Reference PV module
Iph
Vd
Rs
Ipv
Photo current
Id
Rsh
Ish
Vpv
Module temperature (T)
Measured Voltage
Predicted current
Measured Current
Translated data
I_{meas}
I_{est}
$RMSE = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N}(I_{meas} - I_{est})^2\right)}$
The one-diode model electrical parameters at reference conditions.
[Iph_{ref}, Io_{ref}, n_{ref}, Rs_{ref}, Rsh_{ref}]
MPP Dynamic Validation
G
T
PV module simulation model under PSIM software
The Algerian PV plant simulation model (PSIM/MATLAB co-simulation) under healthy and faulty conditions.
Measured profiles of temperature and solar irradiance
Database construction (T, G, Impp, Vmpp, Pmpp) Memorization
Random Forest Classifiers for fault detection and fault diagnosis

Fig. 3. Flowchart of the proposed fault detection and diagnosis strategy.

$$I_o(G,T) = I_{o,ref}\left(\frac{T}{T_{ref}}\right)^3 e^{\left(\frac{q}{nK_B}\left(\frac{E_{g,ref}}{T_{ref}}-\frac{E_g}{T}\right)\right)} \quad (25)$$

E_g is the semiconductor's band gap energy, and $E_{g,ref}$ is the band gap energy for reference conditions. I_{ph}, I_o, n, R_s, and R_{sh} are the five parameters at actual operating conditions. In contrast, $I_{ph,ref}$, $I_{o,ref}$, n_{ref}, $R_{s,ref}$, and $R_{sh,ref}$ are the five unknown parameters at the reference conditions found by the extraction method application.

4. Faults detection and diagnosis strategy

Operating a PV system during certain types of failures can lead to complete insecurity, catastrophic damages, and safety risks. The primary objective of this work is to establish a robust and reliable fault detection procedure using Random Forest Classifiers to detect anomalies within a PV system and pinpoint their root causes. To accomplish this, conceiving a high-quality database that clearly delineates the characteristics of each class of fault is imperative. Therefore, having a reliable simulation model that accurately represents the behavior of a PV system in both its healthy and faulty states is the best course of action to handle this case. Fig. 3 provides a comprehensive flowchart outlining the steps to develop the proposed strategy.

The validated PV system model, as described in the preceding section, serves as the foundation for constructing databases that capture the performance of the PV system under actual outdoor conditions. This PV model is harnessed to produce datasets comprising optimal operation and intentionally simulated defects, utilizing daily solar irradiance and module temperature profiles. To achieve this, the physical model of the grid-connected PV system under consideration was implemented within the PSIM^{TM} software platform. Subsequently, the values of the unknown parameters extracted under reference conditions are incorporated into the physical PV array model.

In this work, the simulated healthy/faulty scenarios, encompassing the most prevalent issues encountered in grid-connected PV systems, are described below and depicted in Fig. 4.

a) A healthy system: This scenario mirrors the operation of the PV system without any anomalies.
b) Three short-circuited modules: Represents the case of one string in the PV system with fewer PV panels in operation.
c) Open circuit faults: This scenario simulates where one string within the PV system becomes non-functional.
d) Line-to-line fault: This is the case of a short-circuit between two PV strings.
e) Three PV modules shaded: This scenario replicates the effects of partial shading experienced by PV systems due to factors such as cloud movement or the presence of nearby objects for a specific duration.

The resulting databases contain five key attributes - Irradiance, Temperature, and the output Current, Voltage, and Power at Maximum Power Point (MPP) - extracted from each simulated operational scenario. An illustration of the simulated faults and their impact on the output power of the grid-connected PV system, based on clear-sky weather data for a typical day, is presented in Fig. 5.

The concluding phase of the proposed fault detection strategy involves the deployment of two Random Forest Classifiers (RFCs). The first RFC is dedicated to identifying anomalies within the PV system, while the second RFC is responsible for diagnosing the specific faults that have been detected.

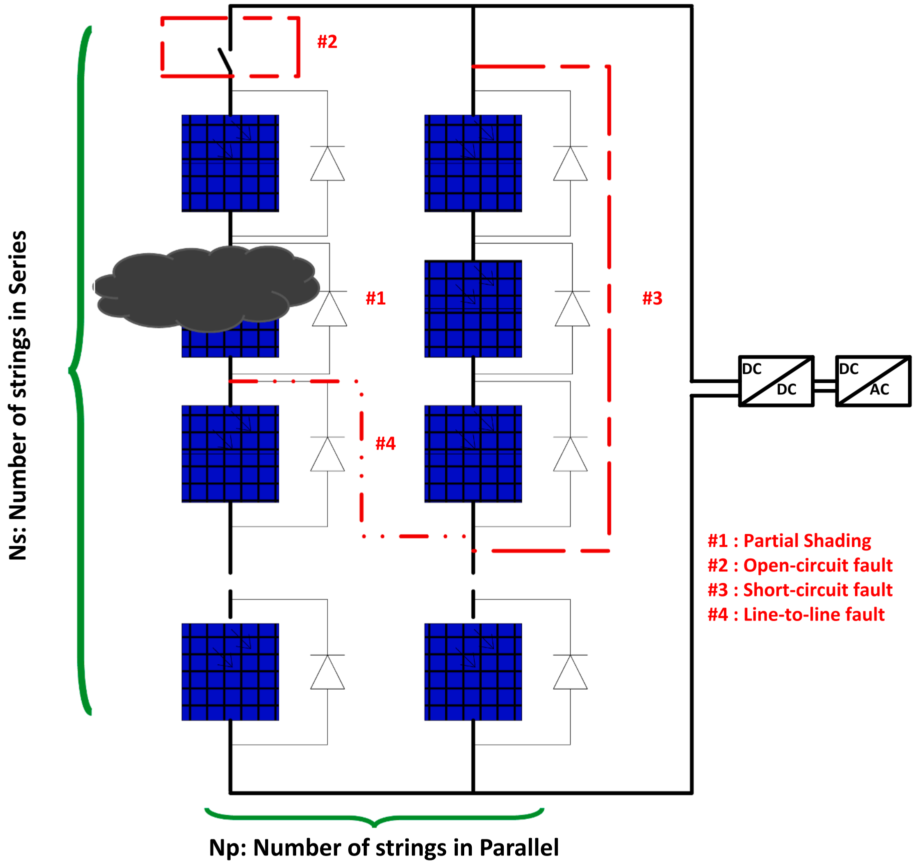

Fig. 4. Failure types considered in the proposed methodology (#1 partial shading, open-circuit fault#2, #3 short-circuit fault, and #4 Line-to-Line fault).

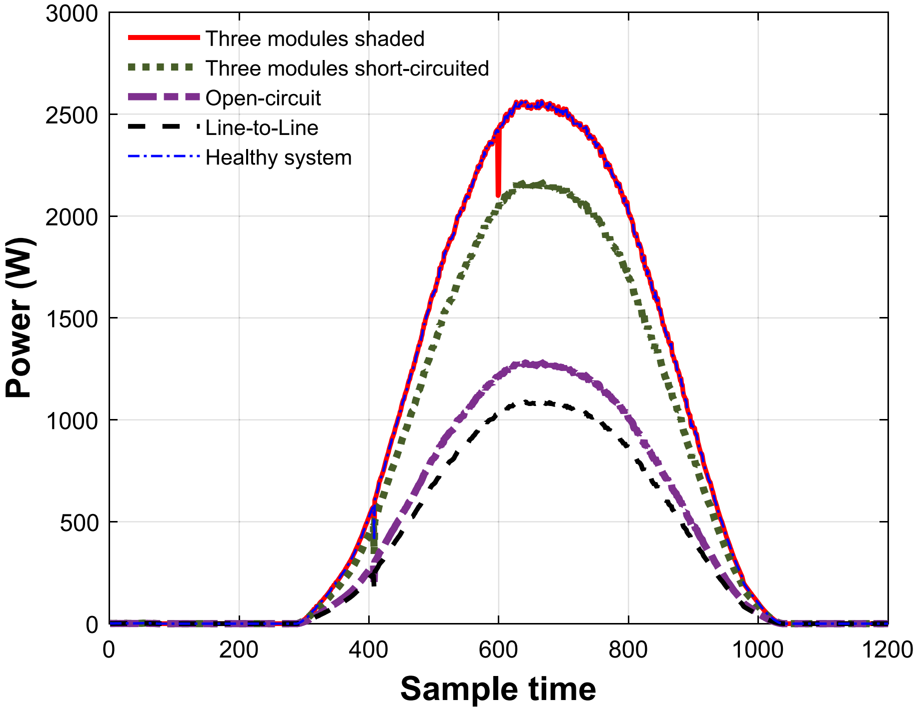

Fig. 5. DC output power of the grid-connected PV system within various fault scenarios.

4.1. Random forest classifier

The Random Forest (RF) algorithm is a supervised machine-learning technique widely employed for Classification and Regression tasks. It operates on the principle of ensemble learning, which involves combining multiple decision trees on different subsets of the input data to enhance predictive accuracy. As a fundamental concept in machine learning, Random Forest's effectiveness and problem-solving capabilities increase with the number of trees it encompasses.

The RF model used in this study is characterized by the decision tree algorithm's benefits linked to speed and high accuracy. The developed model's structure involves two key steps: first, selecting a sampling method to create a data subset, and second, constructing a decision tree, as illustrated in Fig. 6. Notably, four hyperparameters must be considered, including the minimum number of samples for leaf nodes, the minimum number of samples for internal node splitting, the maximum number of selections, and the maximum depth of the decision tree. It's important to note that the RFC's performance is influenced by various

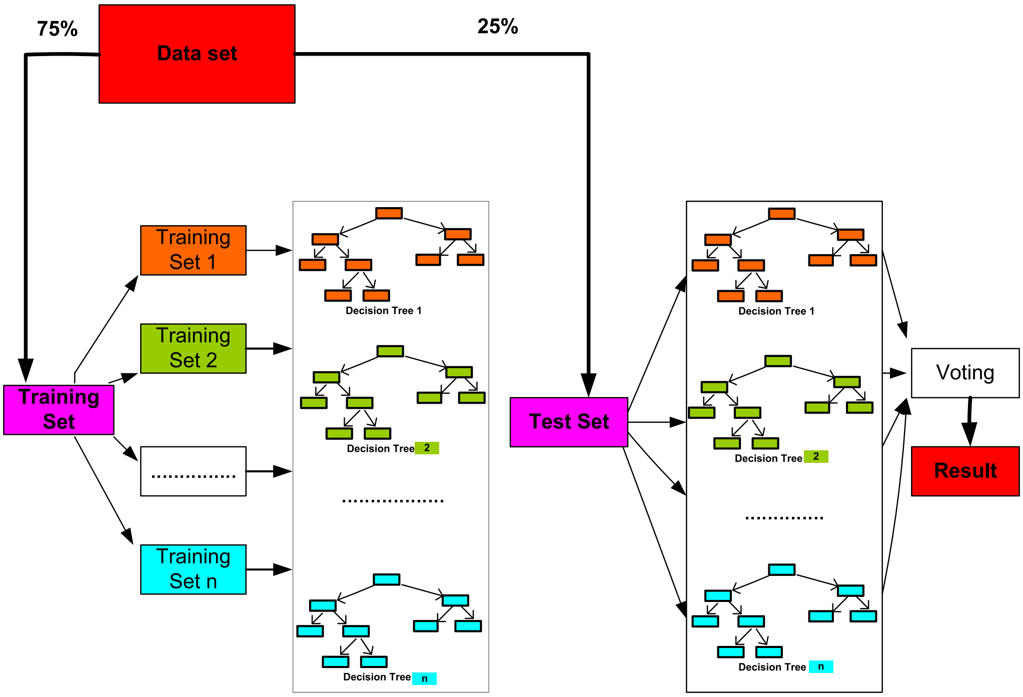

Fig. 6. The general structure of the deployed RF model.

hyperparameters such as splitting criteria, the minimum number of samples for leaf nodes, and internal node splitting, and this study delves into optimizing their combination for enhanced results [41].

After establishing the RF model, the test set samples are input into the model. Each individual decision tree evaluates the classification results for each sample. Once this evaluation is complete, the class that receives the most votes from all decision trees is assigned as the classification for the sample [42]. This is achieved by employing a voting mechanism that combines the results of all the decision trees. A grid search optimization method is utilized to further optimize the RF algorithm's parameters, as illustrated in Fig. 7. This approach aids in identifying the most influential parameter combinations for the RF model, enhancing its classification performance. First, the Cartesian product is applied to the value set of each hyperparameter to generate the hyperparameter configuration space (the box on the left side of Fig. 7), which contains all potential hyperparameter combinations. The grid search algorithm then trains a model for each hyperparameter combination in the configuration space. As seen in the box on the right side of Fig. 7, the experiment with the best validation set error is picked as having discovered the optimal hyperparameters [43].

As mentioned earlier, this study incorporates two RFCs, each with a specific role. The first classifier is designed to identify any indications of faults within the PV system, while the second classifier's task is to pinpoint the particular type of fault.

The diagnosis model aims to provide output that categorizes the specific fault cases (fault #1, fault #2, fault #3, fault #4), as visualized in Fig. 4. The detection model has five input parameters (*T*, *G*, *Impp*, *Vmpp*, and *Pmpp*), a data processing module, the Random Forest/Grid search optimization method application, and two outputs (healthy state, faulty state). This setup enables the models to effectively detect and diagnose faults within the PV system, offering a comprehensive view of the specific fault conditions.

4.2. Data preparation for learning and testing stages

The preprocessing phase of the raw data is imperative to enhance problem-solving capabilities and achieve higher accuracy. In this context, the 'sklearn' library offers a suite of functions for handling missing values, allowing us to identify and address them effectively using the 'isnull' function. To unveil relationships within the data, Pearson's correlation coefficient is employed. This metric yields values ranging from −1 (indicating a perfect negative correlation) to + 1 (indicating a perfect positive correlation), quantifying the strength of linear relationships. Notably, this measure is distinct from correlations between variables [44]. We employed a normalization process based on a calibration technique to facilitate a meaningful comparative analysis of information across attributes in the dataset. This technique centers values around the mean and utilizes a unit standard deviation. Additionally, to provide context for the recorded data, we assigned appropriate class labels, facilitating the creation of well-defined data samples. The defined classes with their corresponding fault type are given in Table 3.

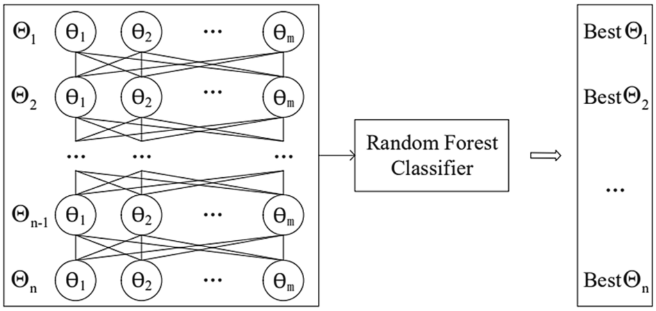

Fig. 7. The grid search algorithm's principle.

Table 3
Defined classes and their corresponding fault type.

Phase	Class	Corresponding fault type
Detection	0	Healthy
	1	Faulty
Diagnosis	0	#2: Open-circuit fault
	1	#1: Partial Shading
	3	#3: Short-circuit fault
	9	#4: Line-to-line fault

For the training and evaluating the two RFCs, we utilized a dataset that includes monitored data from selected 60 days throughout the year, covering all seasonal variations. 75 % of the complete data samples were randomly chosen for training purposes. Subsequently, the remaining 25 % of the data samples were used as an independent set of unknown data to assess performance in each scenario. Considering the data preprocessing of the original data, the dataset resulted in 242,890 data samples designated for detection and 194,400 data samples for diagnosis. As explained in the above section, the classifiers under consideration were fed with both learning and testing datasets, comprising five key attributes (*T*, *G*, *Impp*, *Vmpp*, and *Pmpp*). These attributes serve as input features, and the resulting outputs correspond to the estimated class labels for each data point. The specifics of the constructed detection and diagnosis databases are detailed in Table 4.

A performance evaluation of the classifiers was conducted, employing the confusion matrix as a critical assessment tool to gauge their effectiveness. The confusion matrix provides insights into the accuracy of the classifier's predictions and reveals areas where it made errors. In this matrix, the rows represent the actual labels, and the columns depict the predicted labels. The diagonal values indicate how often the predicted label aligns with the actual label, demonstrating correct predictions. Values in the remaining cells reflect instances where the classifier incorrectly assigned labels to observations, with columns indicating what the classifier predicted and rows showing the actual correct labels.

To comprehensively evaluate our proposed system, we employ metrics such as accuracy, *Precision*, *Recall*, and $F1_{score}$, as expressed in the following equations [45]. These metrics are instrumental in assessing our models' overall performance and reliability.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{25}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{26}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{27}$$

$$F1_{score} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{28}$$

where *TP* signifies the number of samples correctly classified into class "*x*" as they should have been. Conversely, *FN* represents the count of samples that were incorrectly classified, as they should have belonged to class "*x*" but were placed in another class by the classifier. On the other hand, *TN* corresponds to the True Negatives, denoting the number of

Table 4
Details of the detection and diagnosis database construction.

Phase	Class	Test data set (25 %)	Train data set (75 %)	Total
Detection	0	12,145	36,433	*242,890*
	1	48,578	145,734	
Diagnosis	0	12,145	36,433	*194,312*
	1	12,144	36,434	
	3	12,145	36,433	
	9	12,144	36,434	

samples that were correctly classified as not belonging to class "*x*." These samples were placed in a different class per the classifier's judgment. Finally, *FP* stands for False Positives, signifying the number of samples that were incorrectly labeled as belonging to category "*x*," even though they should not have been according to the classifier's assessment. Furthermore, for better interpretability in the case of multi-class classification, we adopt averaging methods, and the macro and weighted average of *Precision*, *Recall*, and $F1_{score}$ is calculated. Macro average (Macro avg) is calculated using the unweighted mean that can penalize the model if the performance in minority classes is poor. On the other hand, weighted average (weighted avg) considers the number of true instances in each class to cope with class imbalance and consequently favors the majority class.

5. Results and discussion

This section demonstrates the validation of the newly developed PV array modeling approach. Subsequently, the effectiveness of the proposed automatic fault detection system is assessed under various weather conditions, considering different faulty patterns in PV array operation. Finally, the fault detection method based on Random Forest Classifier (RFC) is evaluated through benchmarking against other established machine learning techniques. It's worth noting that these validations rely on data collected from the monitored PV system described earlier.

5.1. PV modeling and parameter estimation approach validation

The newly developed procedure for Current-Voltage translation to STC has been validated using three measured curves denoted as Curve 1, Curve 2, and Curve 3, as shown in Fig. 8. The methodology involves an intermediate step where Curve 4 is determined based on the operating conditions (*T* and *G*) of Curve 1 and Curve 2. Subsequently, the target curve, which represents the operation of the PV array at STC (referred to as Curve 0), is predicted from Curve 4 and Curve 3.

Following the extraction of the reference I-V curve for the PV array, the unknown parameters of the one-diode model (ODM) have been extracted through a parameter extraction technique utilizing the Modified Grey Wolf Optimization (MGWO). The optimization algorithm has demonstrated high accuracy, with an RMSE value of 0.0122, and the resulting parameters are presented in Table 5.

The proposed methodology for modeling the PV array has undergone extensive validation, employing the extracted parameters to simulate the PV array under varying irradiance (*G*) and temperature (*T*) conditions, as described in equations (19–24). A comparison was made between the experimental I-V and P-V curves and the simulated data to evaluate the model's accuracy under static conditions. The results are depicted in Fig. 9, revealing a noteworthy agreement between the measurements and the simulated values. This observation is further corroborated by the RMSE indicator values, which stand at 0.0266 and 0.1024, respectively.

Table 5
Extracted ODM parameters at STC.

Parameter	Value
R_p (Ω)	42.9633
R_S(Ω)	0.2212
I_o (A)	4.344 10^{-7}
n	45.1606
I_{ph} (A)	6.8378
RMSE	0.0122

Dynamic validation of the PV array model was conducted using an adapted co-simulation model that integrates the MATLAB and PSIM environments. This dynamic validation incorporated daily temperature and irradiation profiles, along with measured MPP output profiles from a real PV system located in Algiers, under three distinct weather conditions: a) clear sky, b) semi-cloudy, and c) cloudy day. Fig. 10 illustrates the temporal evolution of the developed model's simulated PV array output current. The results demonstrate a strong agreement between the measured and estimated values of the maximum power point current, as indicated by the RMSE values (RMSE = 0.1416, 0.216, and 0.2971, respectively). This substantiates the efficacy of the identification process and the robustness of the proposed approach.

5.2. Evaluation of the proposed fault detection and diagnosis strategy

The automated fault detection procedure developed using Random Forest (RF) is implemented with the Python library scikit-learn. It leverages the "Random Forest Classifier" class from the "ensemble" module. The method is entirely built in Python, with key libraries such as scikit-learn, NumPy, SciPy, seaborn, matplotlib, and the open-source machine learning library dlib [46]. Scikit-learn primarily handles random forest implementation, while dlib is employed for automatic error detection and diagnosis. The computational environment used for this work comprised a personal computer equipped with an Intel Core i7 processor (2.50 GHz), 16 GB of RAM, and a GTX 1060 GPU with a 6 GB of memory.

As explained in the section above, the grid search algorithm was utilized to optimize the hyperparameters in this study. Table 6 lists the optimal hyperparameters for each RF model.

The output reports generated by the two developed Random Forest Classifiers (RFCs) are summarized in Table 7 and Table 8. Both RFCs

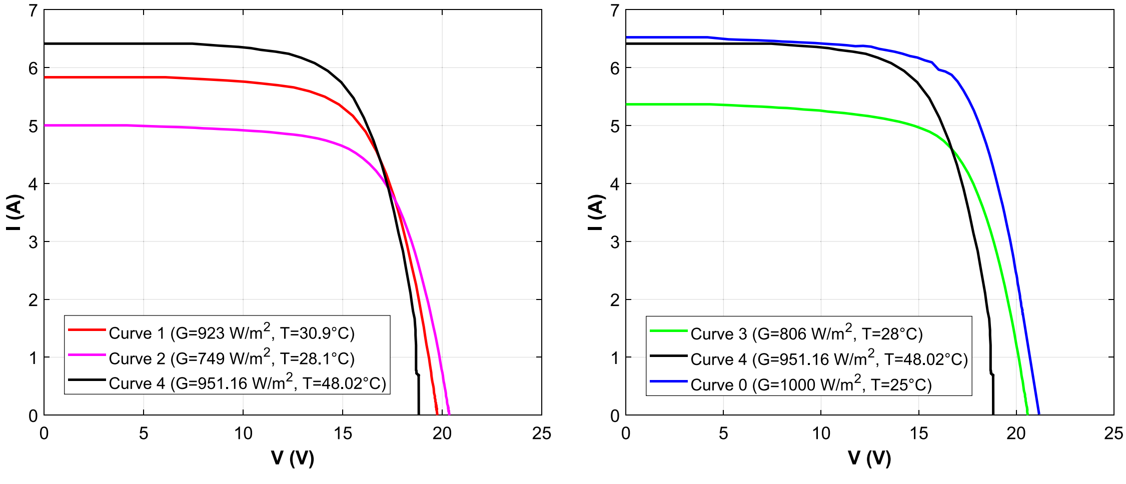

Fig. 8. Predicted I-V curve at STC (Curve 0) using the current–voltage translation method.

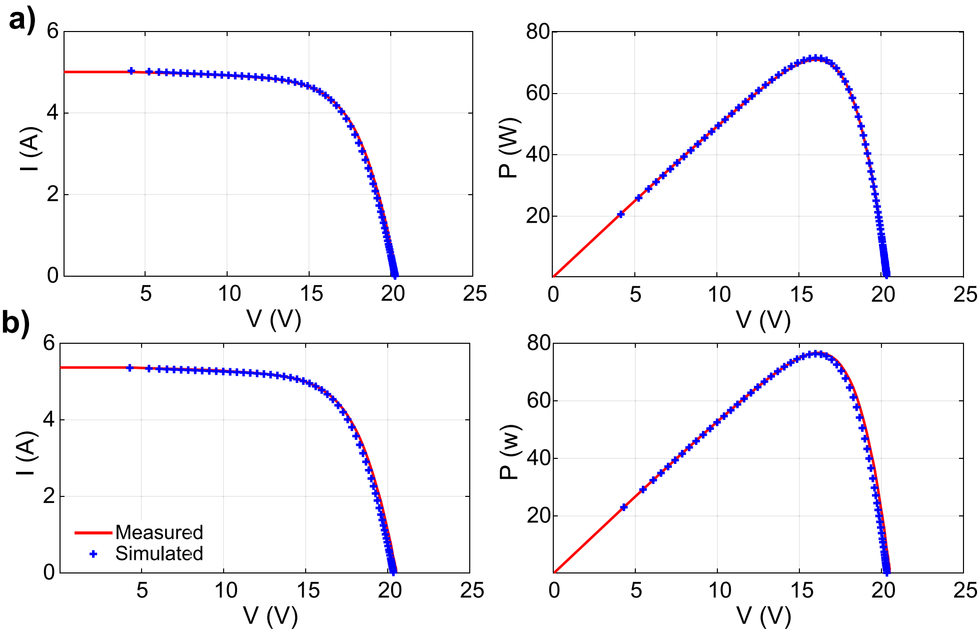

Fig. 9. PV array model validation under a) T = 28.1, G = 749, b) T = 28.2, G = 800.

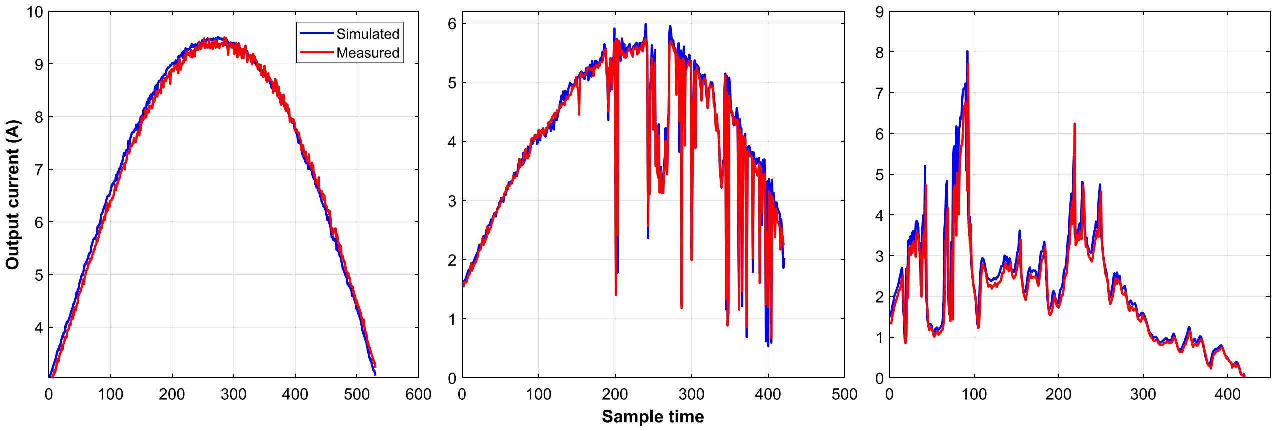

Fig. 10. Dynamic validation of the PV array model under different weather conditions.

Table 6
Optimal hyperparameters.

Hyperparameter	RF Detection model	RF Diagnosis model
max_depth	45	85
n_estimators	65	35
Criterion	gini	entropy
Bootstrap	True	True
Min_samples_leaf	1	1
Min_sample_split	2	2
Max_features	6	6

Table 7
Classification report of RF detection model.

	Precision	Recall	$F1_{score}$	Samples number
Class0	1.00	0.970	0.985	12,145
Class1	0.993	1.000	0.996	48,578
Macro avg	0.996	0.985	0.991	60,723
Weighted avg	0.994	0.994	0.994	60,723
Accuracy (%)	99.4			60,723

Table 8
Classification report of RF diagnosis model.

	Precision	Recall	$F1_{score}$	Samples number
Class0	0.978	1.000	0.989	12,145
Class1	1.000	0.974	0.987	12,144
Class3	0.999	1.000	1.000	12,145
Class9	0.998	1.000	0.999	12,144
Macro avg	0.994	0.994	0.994	48,578
Weighted avg	0.994	0.994	0.994	48,578
Accuracy (%)	99.4			48,578

demonstrate outstanding performance in detecting faults, with an accuracy rate of 99.4 %. Among classification metrics, a lot of emphasis is given to measures of *Precision* and *Recall* (sensitivity) as they are more effective in dealing with imbalanced distributions. However, it is difficult to achieve a trade-off between these two, and the nature of the classification problem often dictates the requirement. For our case, the lowest values of the *Precision* and *Recall* are found during the diagnosis phase and are equal to 0.978 and 0.974, respectively. This can be explained by the model challenges in distinguishing between faults labeled as #1 and #3, representing three partially shaded and three short-circuited PV modules, respectively. These two faults have a similar impact on the PV output power. As seen in Fig. 4, when three PV modules are shaded, the PV system's output power closely resembles the scenario with three short-circuited PVs (represented in green). Despite this complexity, the developed fault detection procedure maintains an overall accuracy of 99.4 %.

The normalized confusion matrices generated by both RF models, one for fault detection and the other for fault diagnosis, are presented in Fig. 11 and Fig. 12, respectively. Examining the data provided in Table 7 and Fig. 11, it becomes evident that the binary classification model performs exceptionally well. In the case of the healthy system, denoted as *Class0*, the model exhibits high precision, with very few false positives, though it has a slightly lower *Recall*, capturing 97 % of cases. This implies that it effectively identifies non-risky instances. In contrast, when dealing with faulty cases labeled as *Class1*, the model excels with high *Precision*, high *Recall,* and an excellent $F1_{score}$, indicating near-perfect performance.

Regarding the diagnosis phase, it is represented by Table 8 and Fig. 12. A detailed explanation of rows and columns of the normalized confusion matrix of the RF diagnosis model is given in the following bullet points:

- Row 1 (*Class0*): The first row corresponds to *fault #2*. The value 1.00 in the top-left cell means that the model correctly identifies *Class0* instances almost perfectly. The remaining values in this row are zeros, indicating that the model rarely misclassifies *Class0* as any other class. This demonstrates that the model is highly accurate in recognizing open circuit faults.
- Row 2 (*Class1*): The second row represents *fault #1*. The value 0.974 in the second cell (from the left) means that the model correctly identifies *Class1* instances with a true positive rate of about 97.4 %. The value 0.023 in the third cell suggests that the model occasionally misclassifies *Class1* as *Class3*. The value 0.003 in the first cell suggests that the model occasionally misclassifies *Class1* as *Class0*. The remaining values in this row are zeros, indicating rare misclassifications into other classes. This shows that the model is highly effective at identifying *Class1* but may occasionally confuse it with *Class3*.
- Row 3 (*Class3*): The third row corresponds to *fault #3*. A value of 1.00 in the third cell (from the left) signifies that the model accurately recognizes *Class3* instances, achieving a true positive rate of 100 %. The values in the remaining rows are all zero, indicating that the model consistently avoids misclassifying instances from other classes. This underscores the model's reliability in identifying *Class3*.
- Row 4 (*Class9*): The fourth row represents *fault #4*. The value 1.0 in the last cell indicates a perfect true positive rate, meaning the model accurately identifies all *Class9* instances. The values in this row suggest that the model never misclassifies *Class9* as any other class, emphasizing the model's exceptional performance in recognizing line-to-line faults.

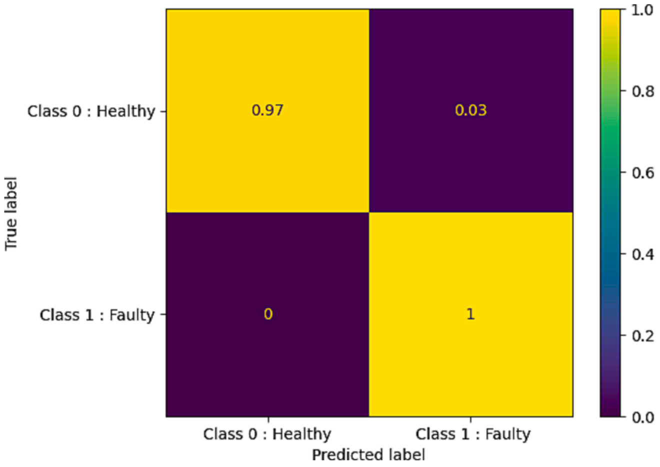

Fig. 11. Normalized Confusion matrix of RF detection model.

In summary, it can be observed that for *Class0* (*fault #2*), *Class3* (*fault #3*) and *Class9* (*fault #4*), the model demonstrates a high True Positive rate, accurately predicting instances of these faults. This indicates its effectiveness in identifying these types of defects with precision. However, the model displays high precision for Class1 (*fault #1*), with minimal false positives. This could be attributed to the similarity between this type of fault and *Class3* (*fault #3*), as previously explained. Overall, the model efficiently minimizes misclassifications.

To explain the classification results better, we have created graphical representations of the confusion matrices for both RFC models. These visual summaries are presented in Fig. 13 for the detection stage and Fig. 14 for the diagnosis stage. It can be seen that the graphical outputs align with the data in the confusion matrices. It can be concluded that the RFC models demonstrate robust performance across all fault classes defined in this work. The model's ability to maintain high precision and recall metrics is essential for accurate classification. It effectively reduces misclassifications and establishes its efficacy as a valuable tool for fault detection and diagnosis in grid-connected PV systems.

5.3. Comparative analysis

To underscore the effectiveness of our machine learning-based RFCs in fault detection, we conducted a comparative analysis with various alternative approaches, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Neural Networks (MLP Classifier), Decision Trees (DT), and Stochastic Gradient Descent (SGDC). To ensure a fair and thorough comparison, we followed the same steps as outlined in our study and fine-tuned the internal hyperparameters for each algorithm using a grid search approach. The summarized results are presented in Table 9.

The observed results scores for the detection phases shows that, the SVM accuracy is 84.5 %, and the MLP Classifier achieved a good accuracy of 97.3 %. The SGDC yielded the lowest accuracy at 79.6 %, whereas both KNN and DT algorithms exhibited similar high-performance levels, both achieving an accuracy of 98.3 %. Moving on to the diagnosis stage, Table 9 illustrates that all algorithms demonstrated improved performance. The DT algorithm achieved the highest accuracy value of 98.3 %, and the SGDC exhibited notable progress, achieving an accuracy value of 89.8 %. Notably, our proposed RFCs method outperformed all other methods in both the detection and diagnosis phases, achieving a remarkable overall accuracy of 99.4 %. Our RFC model also excelled in other evaluation metrics, including *Precision*, *Recall,* and $F1_{score}$, surpassing SVM, KNN, DT, SGDC, and MLP Classifier.

The results obtained in this study align well with the findings in the existing literature. Eskandari et al. proposed an SVM-based method specifically for detecting and classifying Line-to-Line faults (LL), achieving average accuracies of 96 % and 97.5 %, respectively [18]. While their accuracies surpass ours, it's important to note that our study encompasses various types of faults, not limited to LL faults alone. Moving on to our K-Nearest Neighbors (KNN) model, it achieved a classification accuracy of 98.3 %, closely matching the results from Madeti and Singh [17], who attained an average fault classification accuracy of 98.70 %, focusing on open-circuit, Line-to-Line, and different short-circuit faults, including those represented by bypass diodes. In the domain of Neural Networks, specifically the Multilayer Perceptron (MLP) Classifier, our model achieved an accuracy of 98.2 %. In comparison, Chine et al. [47] reported a reasonable accuracy of 90.3 %, potentially attributed to the number of faults considered and the absence of optimization in the neural network architecture in their work. Benkercha and Moulahoum presented a fault detection and

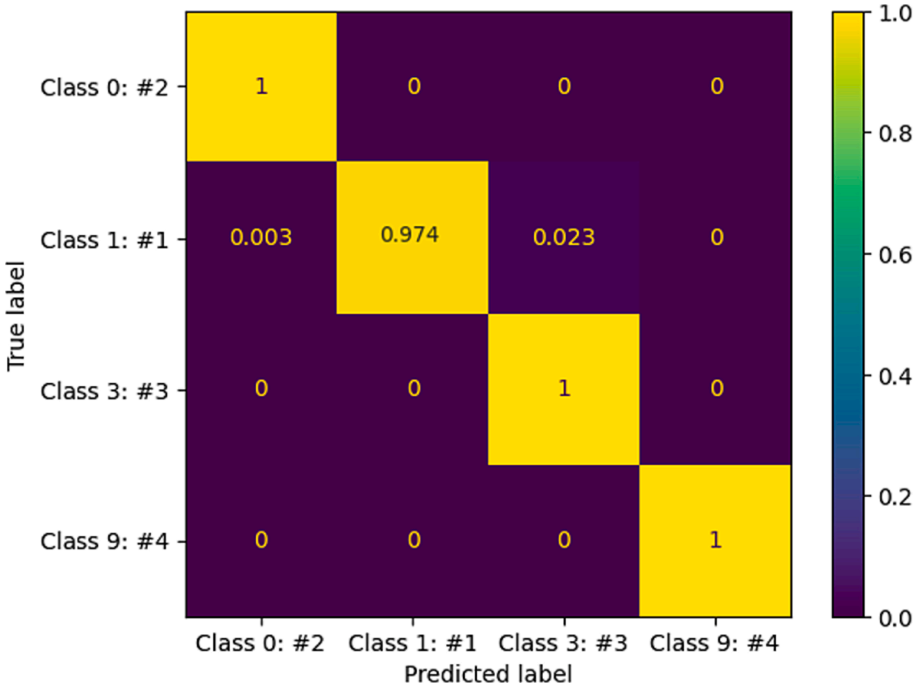

Fig. 12. Normalized Confusion matrix of RF diagnosis model.

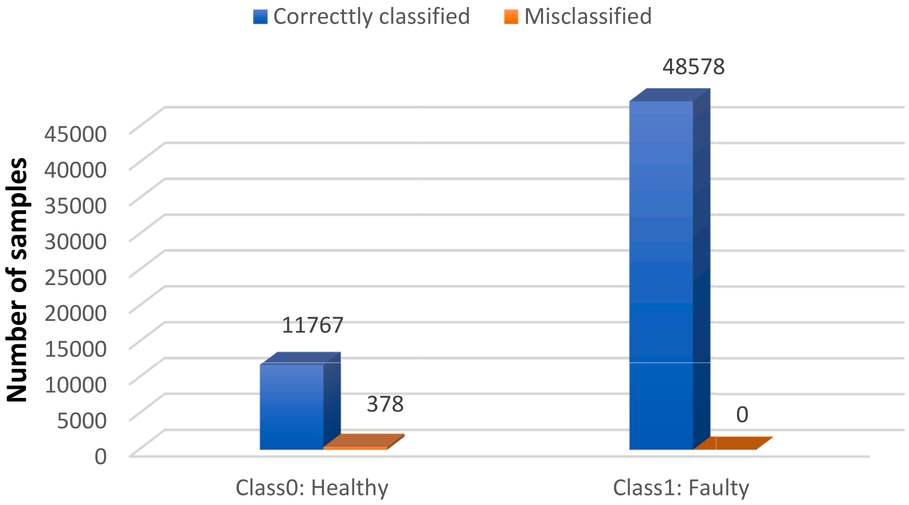

Fig. 13. Fault detection results.

diagnosis technique based on the Decision Trees (DT) algorithm, achieving an overall accuracy of around 99 % [48]. Although slightly higher than our study's accuracy of about 98.3 %, this difference can be justified by considering a broader spectrum of fault types. Notably, Kapucu and Cubukcu [19] reported slightly lower accuracies of 97.46 % and 97.67 % using quadratic discriminant analysis-extra trees-Decision Trees (QDA-ETent-DT) for PV fault detection before and after optimization, respectively. It's worth mentioning that their study focused on partial shading and short-circuit faults without accounting for changes in weather conditions.

6. Conclusion

Photovoltaic systems are continuously exposed to many faults that significantly impact their performance and overall efficiency. These issues, including short circuits, shading, line-to-line problems, and open circuits, can substantially reduce harvested solar energy. In response, this manuscript introduces a robust machine learning (ML) technique that harnesses the Random Forest Classifier (RFC) to effectively detect and monitor PV system performance.

Our approach builds on a precise one-diode (ODM) simulation model, accurately replicating actual PV system behavior. Identifying the unknown parameters of the ODM involves a new application of the current–voltage translation technique combined with the Modified Grey Wolf Optimization (MGWO) algorithm.

The extracted ODM parameters are integrated into the developed physical model of the studied PV system. Trustworthy databases representing normal and abnormal PV system operation are constructed using PSIM and MATLAB software co-simulations.

Following the development of the RFC-based fault detection procedure, our results demonstrate exceptional classification accuracy rates of 99.4 % for both fault detection and diagnosis. This outperforms alternative models like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Neural Networks (MLP Classifier), Decision Trees

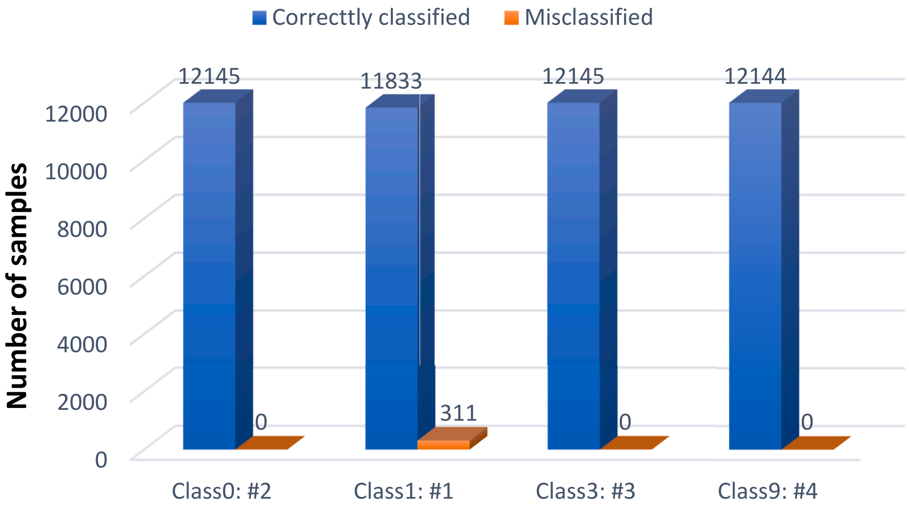

Fig. 14. Fault diagnosis results.

Table 9
Comparative Analysis between SVM, KNN, DT, SGDC, MLP, and RF trained and tested using the same data set.

Phase	Indicator	label	SVM	MLP Classifier	KNN	DT	SGDC	RF
Detection	Precision	0	0.979	0.913	0.979	0.988	0.000	1.000
		1	0.984	0.990	0.984	0.981	0.796	0.993
	Recall	0	0.939	0.960	0.939	0.927	0.000	0.970
		1	0.995	0.977	0.995	0.997	1.000	1.000
	$F1_{score}$	0	0.959	0.936	0.959	0.956	0.000	0.985
		1	0.990	0.983	0.990	0.989	0.886	0.996
	Accuracy (%)		84.5	97.3	98.3	98.3	79.6	99.4
Diagnosis	Precision	0	0.958	0.992	0.923	0.992	0.851	0.978
		1	1.000	1.000	0.996	0.997	0.876	1.000
		3	0.933	0.974	0.998	0.972	0.986	0.999
		9	0.964	0.964	0.997	0.971	0.908	0.998
	Recall	0	0.928	0.975	0.997	0.972	0.967	1.000
		1	0.951	0.972	0.905	0.975	0.930	0.974
		3	0.968	0.981	0.997	0.985	0.697	1.000
		9	1.000	1.000	1.000	0.997	1.000	1.000
	$F1_{score}$	0	0.943	0.983	0.959	0.982	0.905	0.989
		1	0.975	0.986	0.948	0.986	0.902	0.987
		3	0.951	0.978	0.997	0.979	0.816	1.000
		9	0.981	0.982	0.998	0.984	0.952	0.999
	Accuracy (%)		96.1	98.2	97.5	98.3	89.8	99.4

(DT), and Stochastic Gradient Descent (SGDC).

In conclusion, the RF algorithm emerges as a robust tool for fault diagnosis, offering higher accuracy and efficiency, particularly in cases of partial shading. While these promising results are encouraging, it's essential to acknowledge the complexity of PV systems, with potential challenges in fault detection. Future research will explore more advanced techniques, potentially using deep learning methods, to precisely locate faults within PV systems. These advancements aim to enhance the reliability and efficiency of PV systems for a more sustainable solar energy future. Furthermore, it is noteworthy that the proposed technique eliminates the need to install any additional sensors beyond those already present in a standard PV installation. This adaptability enables its application across various PV systems, making the suggested approach straightforward to implement.

CRediT authorship contribution statement

Ahmed Faris Amiri: . **Houcine Oudira:** Writing – review & editing, Supervision, Resources, Methodology, Formal analysis, Conceptualization. **Aissa Chouder:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization. **Sofiane Kichou:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

[1] IEA – International Energy Agency - IEA n.d. https://www.iea.org/reports/worldenergy-outlook-2023. (accessed December 4, 2023).
[2] Solar Power EU. Global market outlook for solar power focus on southeast asia. 2023.
[3] Marion B, Schaefer R, Caine H, Sanchez G. Measured and modeled photovoltaic system energy losses from snow for Colorado and Wisconsin locations. Sol Energy 2013;97:112–21. https://doi.org/10.1016/J.SOLENER.2013.07.029.
[4] Potnuru SR, Pattabiraman D, Ganesan SI, Chilakapati N. Positioning of PV panels for reduction in line losses and mismatch losses in PV array. Renew Energy 2015;78:264–75. https://doi.org/10.1016/J.RENENE.2014.12.055.
[5] Pillai DS, Rajasekar N. Metaheuristic algorithms for PV parameter identification: A comprehensive review with an application to threshold setting for fault detection in PV systems. Renew Sustain Energy Rev 2018;82:3503–25. https://doi.org/10.1016/J.RSER.2017.10.107.

[6] Daliento S, Chouder A, Guerriero P, Pavan AM, Mellit A, Moeini R, et al. Monitoring, diagnosis, and power forecasting for photovoltaic fields: A review. Int J Photoenergy 2017;2017. https://doi.org/10.1155/2017/1356851.

[7] Hariharan R, Chakkarapani M, Saravana Ilango G, Nagamani C. A Method to detect photovoltaic array faults and partial shading in PV systems. IEEE J Photovoltaics 2016;6:1278–85. https://doi.org/10.1109/JPHOTOV.2016.2581478.

[8] Chouder A, Silvestre S. Automatic supervision and fault detection of PV systems based on power losses analysis. Energy Convers Manag 2010;51:1929–37. https://doi.org/10.1016/J.ENCONMAN.2010.02.025.

[9] Silvestre S, Kichou S, Chouder A, Nofuentes G, Karatepe E. Analysis of current and voltage indicators in grid connected PV (photovoltaic) systems working in faulty and partial shading conditions. Energy 2015;86:42–50. https://doi.org/10.1016/J.ENERGY.2015.03.123.

[10] Drews A, de Keizer AC, Beyer HG, Lorenz E, Betcke J, van Sark WGJHM, et al. Monitoring and remote failure detection of grid-connected PV systems based on satellite observations. Sol Energy 2007;81:548–64. https://doi.org/10.1016/J.SOLENER.2006.06.019.

[11] Garoudja E, Harrou F, Sun Y, Kara K, Chouder A, Silvestre S. Statistical fault detection in photovoltaic systems. Sol Energy 2017;150:485–99. https://doi.org/10.1016/J.SOLENER.2017.04.043.

[12] Dhimish M, Holmes V, Mehrdadi B, Dales M. Comparing mamdani sugeno fuzzy logic and RBF ANN network for PV fault detection. Renew Energy 2018;117:257–74. https://doi.org/10.1016/J.RENENE.2017.10.066.

[13] Momeni H, Sadoogi N, Farrokhifar M, Gharibeh HF. Fault diagnosis in photovoltaic arrays using GBSSL method and proposing a fault correction system. IEEE Trans Ind Informatics 2020;16:5300–8. https://doi.org/10.1109/TII.2019.2908992.

[14] Yi Z, Etemadi AH. Fault detection for photovoltaic systems based on multi-resolution signal decomposition and fuzzy inference systems. IEEE Trans Smart Grid 2017;8:1274–83. https://doi.org/10.1109/TSG.2016.2587244.

[15] Leva S, Mussetta M, Ogliari E. PV module fault diagnosis based on microconverters and day-ahead forecast. IEEE Trans Ind Electron 2019;66:3928–37. https://doi.org/10.1109/TIE.2018.2879284.

[16] Bendary AF, Abdelaziz AY, Ismail MM, Mahmoud K, Lehtonen M, Darwish MMF. Proposed ANFIS based approach for fault tracking, detection, clearing and rearrangement for photovoltaic system. Sensors 2021, Vol 21, Page 2269 2021;21:2269. 10.3390/S21072269.

[17] Madeti SR, Singh SN. Modeling of PV system based on experimental data for fault detection using kNN method. Sol Energy 2018;173:139–51. https://doi.org/10.1016/J.SOLENER.2018.07.038.

[18] Eskandari A, Milimonfared J, Aghaei M. Line-line fault detection and classification for photovoltaic systems using ensemble learning model based on I-V characteristics. Sol Energy 2020;211:354–65. https://doi.org/10.1016/J.SOLENER.2020.09.071.

[19] Kapucu C, Cubukcu M. A supervised ensemble learning method for fault diagnosis in photovoltaic strings. Energy 2021;227:120463. https://doi.org/10.1016/J.ENERGY.2021.120463.

[20] Adhya D, Chatterjee S, Chakraborty AK. Performance assessment of selective machine learning techniques for improved PV array fault diagnosis. Sustain Energy, Grids Networks 2022;29:100582. https://doi.org/10.1016/J.SEGAN.2021.100582.

[21] Akram MN, Lotfifard S. Modeling and health monitoring of DC side of photovoltaic array. IEEE Trans Sustain Energy 2015;6:1245–53. https://doi.org/10.1109/TSTE.2015.2425791.

[22] Chen Z, Han F, Wu L, Yu J, Cheng S, Lin P, et al. Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. Energy Convers Manag 2018;178:250–64. https://doi.org/10.1016/J.ENCONMAN.2018.10.040.

[23] Gong S, Wu X, Zhang Z. Fault diagnosis method of photovoltaic array based on random forest algorithm. Chinese Control Conf CCC 2020;2020-July:4249–54. 10.23919/CCC50068.2020.9189016.

[24] Mellit A, Benghanem M, Kalogirou S, Massi PA. An embedded system for remote monitoring and fault diagnosis of photovoltaic arrays using machine learning and the internet of things. Renew Energy 2023;208:399–408. https://doi.org/10.1016/J.RENENE.2023.03.096.

[25] Gao W, Wai RJ, Chen SQ. Novel PV fault diagnoses via SAE and improved Multi-Grained Cascade Forest with string voltage and currents measures. IEEE Access 2020;8:133144–60. https://doi.org/10.1109/ACCESS.2020.3010233.

[26] Liu Y, Ding K, Zhang J, Li Y, Yang Z, Zheng W, et al. Fault diagnosis approach for photovoltaic array based on the stacked auto-encoder and clustering with I-V curves. Energy Convers Manag 2021;245:114603. https://doi.org/10.1016/J.ENCONMAN.2021.114603.

[27] Chen Z, Chen Y, Wu L, Cheng S, Lin P. Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions. Energy Convers Manag 2019;198:111793. https://doi.org/10.1016/J.ENCONMAN.2019.111793.

[28] Wang M, Xu X, Yan Z. Online fault diagnosis of PV array considering label errors based on distributionally robust logistic regression. Renew Energy 2023;203:68–80. https://doi.org/10.1016/J.RENENE.2022.11.126.

[29] Piliougine M, Sánchez-Friera P, Petrone G, Sánchez-Pacheco FJ, Spagnuolo G, Sidrach-de-Cardona M. Analysis of the degradation of amorphous silicon-based modules after 11 years of exposure by means of IEC60891:2021 procedure 3. Prog Photovoltaics Res Appl 2022;30:1176–87. https://doi.org/10.1002/PIP.3567.

[30] Mirjalili S, Mirjalili SM, Lewis A. Grey Wolf Optimizer. Adv Eng Softw 2014;69:46–61. https://doi.org/10.1016/J.ADVENGSOFT.2013.12.007.

[31] Garoudja E, Chouder A, Kara K, Silvestre S. An enhanced machine learning based approach for failures detection and diagnosis of PV systems. Energy Convers Manag 2017;151:496–513. https://doi.org/10.1016/J.ENCONMAN.2017.09.019.

[32] Castañer L, Silvestre S. Modelling photovoltaic systems using PSpice 2002:358.

[33] Kichou S, Silvestre S, Guglielminotti L, Mora-López L, Muñoz-Cerón E. Comparison of two PV array models for the simulation of PV systems using five different algorithms for the parameters identification. Renew Energy 2016;99:270–9. https://doi.org/10.1016/J.RENENE.2016.07.002.

[34] Kichou S, Wolf P, Wolf P. Approach for Simulating outputs of PV module/array of different technologies with high accuracy. EUPVSEC proceedings 2018. https://doi.org/10.4229/35thEUPVSEC20182018-1CV.4.2.

[35] Amiri AF, Oudira H, Chouder A. Faults detection of PV systems based on extracted parameters using Modified Grey Wolf algorithm. In: Proc 2022 Int Conf Adv Technol Electron Electr Eng ICATEEE; 2022 2022.. https://doi.org/10.1109/ICATEEE57445.2022.10093747.

[36] F. h.. Tackling real-coded genetic algorithms : operators and tools for behavioral analysis. Artif Intell Rev 1998;12:265–319.

[37] Ali M, El-Hameed MA, Farahat MA. Effective parameters' identification for polymer electrolyte membrane fuel cell models using grey wolf optimizer. Renew Energy 2017;111:455–62. https://doi.org/10.1016/J.RENENE.2017.04.036.

[38] Chouder A, Silvestre S, Sadaoui N, Rahmani L. Modeling and simulation of a grid connected PV system based on the evaluation of main PV module parameters. Simul Model Pract Theory 2012;20:46–58. https://doi.org/10.1016/J.SIMPAT.2011.08.011.

[39] Elkholy A, Abou El-Ela AA. Optimal parameters estimation and modelling of photovoltaic modules using analytical method. e02137 Heliyon 2017. https://doi.org/10.1016/j.heliyon.2019.e02137.

[40] Tossa AK, Soro YM, Azoumah Y, Yamegueu D. A new approach to estimate the performance and energy productivity of photovoltaic modules in real operating conditions. Sol Energy 2014;110:543–60. https://doi.org/10.1016/J.SOLENER.2014.09.043.

[41] Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9:e1301.

[42] Resende PAA, Drummond AC. A Survey of random forest based methods for intrusion detection systems. ACM Comput Surv 2018;51. https://doi.org/10.1145/3178582.

[43] Zhang A, Lipton ZC, Li MU, Smola AJ. Dive into Deep Learning. ArXiv 2021, arXiv:2106.11342.

[44] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al. Scikit-learn: Machine learning in Python. J. Mach Learn Res 2011;2825-2830–12.

[45] Mellit A, Kalogirou S. Assessment of machine learning and ensemble methods for fault diagnosis of photovoltaic systems. Renew Energy 2022;184:1074–90. https://doi.org/10.1016/J.RENENE.2021.11.125.

[46] King DE. Dlib-ml: A Machine Learning Toolkit. J Mach Learn Res 2009;10:1755–8.

[47] Chine W, Mellit A, Lughi V, Malek A, Sulligoi G, Massi PA. A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. Renew Energy 2016;90:501–12. https://doi.org/10.1016/J.RENENE.2016.01.036.

[48] Benkercha R, Moulahoum S. Fault detection and diagnosis based on C4.5 decision tree algorithm for grid connected PV system. Sol Energy 2018;173:610–34. https://doi.org/10.1016/J.SOLENER.2018.07.089.