RSO based Optimization of Random Forest Classifier for Fault Detection and Classification in Photovoltaic Arrays

Khaled Baradieh
Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia
Khalid.baradia@outlook.com

Mohd Zainuri
Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia
ammirrulatiqi@ukm.edu.my

Mohamed Kamari
Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia
azwank@ukm.edu.my

Yushaizad Yusof
Faculty of Engineering and Built Environment, Universiti Kebangsaan, Malaysia
yushaizad@ukm.edu.my

Huda Abdullah
Faculty of Engineering and Built Environment, Universiti Kebangsaan, Malaysia
huda.abdullah@ukm.edu.my

Mohd Zaman
Faculty of Engineering and Built Environment, Universiti Kebangsaan, Malaysia
hairizaman@ukm.edu.my

Mohd Zulkifley
Faculty of Engineering and Built Environment, Universiti Kebangsaan, Malaysia
asyraf.zulkifley@ukm.edu.my

Abstract: *Fault detection and classification in photovoltaic arrays are critical for increasing grid reliability and reducing the power losses. This paper assesses twelve machine learning classifiers for their effectiveness in detecting and classifying faults in Photovoltaic (PV) systems. Multiple validation methods were used for the algorithm evaluation, including K-fold, stratified K-fold, leave-one-out, and random split cross-validation approaches to ensure robust performance measures. The applied selection criterion of the top performing classifier are the accuracy, precision, recall, and computing efficiency. The utilized dataset, comprising samples with various fault kinds under diverse environmental conditions, received thorough preprocessing to enhance model training and assure generalizability. A large dataset of roughly 10,000 samples was utilized in this research for the model training and to run multiple random tests on new and unseen data. This dataset provides a fair representation of multiple fault types such as the healthy, Line to Line (LL), Line to Ground (LG), Partial Shading (PS), and Complete Shading faults (CS). The data preprocessing comprised normalization, handling of missing values by taking the average, and applying multiple statical analysis approaches to reduce the size of the features matrix and to improve the dependability of the model's predictions across varying operational circumstances. The results illustrate the best performance utilizing the optimized version of the Random Forest classifier, reaching an average fault detection accuracy of 100% and fault classification accuracy of 94.7%, the hyperparameters of the classifier was optimized using Random Search Optimization algorithm (RSO).*

Keywords: *Classifier, diagnosis, fault classification, fault detection, machine learning, optimization, photovoltaic, pv, random forest.*

Received January 15, 2024; accepted July 4, 2024
https://doi.org/10.34028/iajit/21/4/8

1. Introduction

Since the oil crisis of the 1970s, the solar Photovoltaic (PV) industry has developed rapidly, especially in recent years. The additions to renewable capacity expanded by more than 45% in 2020 over 2019 and broke another record. According to the International Energy Agency (IEA), the high rate of renewable energy capacity expansions is reaching 280 GW in 2022, with solar PV counts for 60% of this rise [24]. This growth exceeds the annual capacity record of 2017-2019 by more than 50%, implying that renewables are accounting for 90% of all global power capacity expansions in 2022, a rate comparable to that of the semiconductor and computer industries. For the next several decades, the PV industry can maintain a double-digit annual growth. According to the solar roadmap published by the IEA [25], PV power will supply around 11% of worldwide energy consumption by 2050 and reduce 2.3 Gigatons (Gt) of CO2 emissions per year.

However, PV power generation is affected by many factors that disturb the energy production process. These factors include the condition of the PV arrays and their wire connections, environmental conditions such as temperature and solar radiation, and faults that may occur during operation [39, 40]. Faults are one of the

major issues that increase grid susceptibility and power losses. Such faults can arise for a variety of reasons, including module faults at the manufacturing or installation level, faults on PV systems caused by external factors, and faults caused by internal component failures.

To have a better understanding, PV faults might happen in the DC or AC stages, which form the power flow chain. DC stage faults can arise because of different reasons, including partial shading, hotspots, bypass diode failure, module cracks, maximum power point algorithm failure, and converter switching failure at the DC-DC converter level [26]. On the other hand, AC stage faults are the sort of faults that happen on the distribution side of the PV system, and they generally consist of inverter faults such as the open circuit switches, the short circuit switches, filter failure, and gating failure in the inverter, among other things [33]. These faults may cause incoherency in the PV system operation, the aging of PV arrays, and a lowering of the system efficiency [33, 39]. However, among these faults, Line-Line (LL) and Line-Ground (LG) faults have a disastrous impact on the entire system and are known to be the primary causes of catastrophic failures such as electrical fires [2, 38].

On the other hand, PV system installations worldwide adhere to the protection standards outlined in the National Electric Code (NEC) or the International Electro-Technical Commission (IEC) by employing Over-Current Protection Devices (OCPDs) and Ground Fault Protection Devices (GFPDs). Usually, OCPDs, such as fuses, and GFPDs are the conventional fault detection and protection methods that are used to protect PV components from large fault currents. However, it has been shown that OCPD and GFPD may not be able to clear or detect certain faults in PV arrays due to the non-linear characteristics of PV arrays, high fault impedances, PV current-limiting nature, PV grounding schemes, low irradiance conditions, or MPPT of PV inverters [15, 22, 39]. Such situations bring "blind spots" in the protection schemes, resulting in reduced system efficiency, accelerated system aging, DC arcs, and fire hazards such as the reported cases in [16]. Therefore, advances in the detection and classification of faults in the PV facilitate an improvement in the system's efficiency through a reduction in the downtimes and power losses realized by the early and accurate diagnosis of faults. Improved detection of faults increases safety by preventing potential hazards associated with undetected faults, like electrical fires or damage to equipment. Moreover, early identification of faults means a decrease in maintenance cost and an extension of the operating lifespan of the equipment. These combined advances enable steady and more efficient energy output, thus requiring more accurate fault detection and diagnostics to maintain high-performance PV systems. The paper using advanced ML techniques to develop a robust framework for fault detection and classification. The specifications of overcoming the limitations of traditional methods are discussed in this paper. With its ability to draw upon a comprehensive dataset, including all types of samples for different kinds of faults under various environmental conditions, the goal is to increase the reliability and efficiency in the diagnosis of PV systems. The structure of this paper will form a base for a detailed discussion on the types of faults, the shortcomings associated with conventional methods of detection, and how one could apply the principles of machine learning to address them. The following sections will explain the methodology, data preparation, model training, and validation and conclude with the results of the comparative analysis of twelve ML classifiers.

The following subsections discuss the literature of faults in PV systems and the utilized fault monitoring techniques.

1.1. Faults in PV Arrays

Among the different types of faults that occurred during the operation of the PV array, Line-to-Line (LL), Line-to-Ground (LG), and Arc faults were reported to be catastrophic. These faults might result in deep and long-term failures, as well as the risk of electric fires. This section will discuss in detail the expected reasons for these faults, their effect on the electrical behavior of the PV array, and the challenges in detecting such faults. Figure 1 shows the different types of PV array faults.

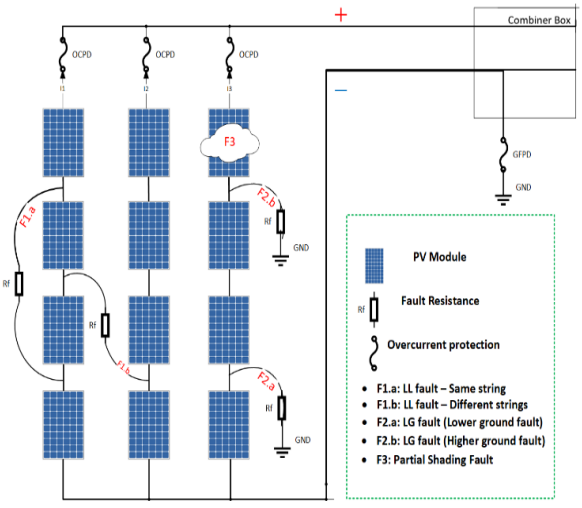


Figure 1. Schematic diagram of various types of faults in the PV array.

1.1.1. Line to Line Faults (LL)

This fault type, shown in Figure 1 (F1.a, and F1.b), is caused by an unintentional connection between two points in the same string or different strings, which will result in a low impedance current path between the connected points or create a reverse current through the affected strings. The amplitude of the fault is determined by the voltage difference between the connection points before the fault, where the fault

current is directly proportional to the potential difference between the two points. The OCPDs (fuses) are generally used to clear such types of faults once the fault current exceeds 156% of the string current, which, in turn, will be able to melt the OCPD and result in an open circuit [3]. However, under low irradiation, such as in cloudy weather or during the day-to-night or night-to-day transition, the amount of the faulty string current will be insufficient to melt the OCPD, and the fault may remain undetected. Furthermore, the inverter's MPPT may shift the operating point to a different position on the I–V curve, causing the fault current amplitude to drop. In addition, while the fault current is smaller than the melting threshold of the OCPD, the fault will remain undetected for an indefinite time, the MPPT will relocate the MPP to another position, and the PV array will appear to be operating normally but with less output power. In addition, the behavior of the PV array will be closer to the healthy case with the increase in the LL fault resistance [59].

On the other hand, the blocking diodes can be optionally utilized in the same PV configuration. Blocking diodes prevent the current from flowing in the reverse direction, which, in turn, will block the fault current (reverse current), i.e., it will not be detected by the OCPD. This behavior will increase the PV power [6] and will make the fault detection task challenging. However, any failure in the blocking diodes with the presence of undetected LL fault is dangerous, and it might cause an electric fire due to the high fault current [2, 38]. Moreover, the blocking diodes with the presence of the low fault resistance will cause different voltage peaks on the I-V characteristics curve, which would make it very similar to the I-V curve of the open circuit and partial shading faults [39].

1.1.2. Line to Ground Faults (LG)

Usually, the PV arrays include different non-Current-Carrying (NCC) parts during normal operation, such as the mounting racks, enclosures, module frames, etc. However, due to an accidental short circuit such as a contact between the junction box and the ground or water corrosion, etc., an electrical connection might be established between the NCC and the Current-Carrying Conductors (CCC) [20], which is known as the ground fault, shown in Figure 1 (F2.a, F2.b).

Fault location and fault impedance are the main factors determining the magnitude of the fault current. Normally, the low fault impedance will generate a high fault current, which will be enough to trigger the Ground Fault Protection Device (GFPD), so the system will be protected. However, not all ground faults have low impedance, in which the GFPD will not be able to detect the fault due to the low amount of generated fault current, and the inverter will not be disconnected [48]. This fault type might result in fire hazards because of the formed DC arc if there is no proper way to clear the fault.

There are two types of ground faults in PV arrays: the lower ground fault (Figure1, F2.a) and the upper ground fault (Figure 1, F2.b). The lower ground fault is in the last two modules in the string, so there will be a difference in voltage and uneven flow of current between the faulty and normal strings. The MPP of the array will be shifted gradually because of the fault, where the MPPT will detect the drop in the output power and shift the MPP by reducing the array voltage to optimize the output power. However, in the case of a lower ground fault, there will be no back-fed current to the faulty string which will be mismatched with the other strings and no longer be able to operate at its real MPP [33]. On the other hand, the upper ground fault happened in the upper PV modules, which will make a high fault current at low or no impedance, the faulty module will have a larger current than the other modules because of the fault current and the back-fed current. In this situation, the GFPD could be triggered and clear the faulty string by disconnecting the faulty path, so the negative CCC will not be grounded anymore. However, when the GFPD is unable to detect the fault, then the inverter's MPPT will shift the MPP to reduce the power loss and fault current by reducing the array voltage. Moreover, high back-fed current in the faulty module might damage the cables and modules. The faulty string in this fault type will operate as a load at the beginning of the fault, then the MPPT will help in finding a new MPP for the string, which might generate power at a lower point.

1.1.3. Partial Shading Faults (PS)

Partial shading occurs in the PV arrays due to various reasons, such as the movement of clouds, dust, trees, snow, and more. This fault type causes an output power loss and might create a local heating problem (hotspot), which will affect the system's safety and reliability, and might damage the PV generator at high temperatures (more than 150 °C) [6, 9]. Usually, the PV systems are protected against the hotspot effect by utilizing the bypass diodes, which are connected in parallel with the PV module at opposite polarity [16]. The MPP of the PV array will be reduced with the increase in the shading percentage, and different local and global MPP peaks will appear based on the number and percentage of the shaded modules. The inverter's MPPT will push the system to continue working on the new global peak to guarantee more output power by reducing the output voltage, which might affect the inverter's lifetime. In addition, at complete shading of the PV array (CS), there will be no local peaks, but the whole output power will be reduced.

1.2. PV Fault Monitoring Techniques

According to the literature, fault monitoring techniques can be divided into two categories:

1.2.1. Signal Processing-Based Methods

These methods are based on sensing real-time data such as temperature and irradiance, and measuring the PV system data such as voltage, current, and power. These methods utilize a predefined threshold to help in comparing the measured data with the expected results [4, 5, 39, 54, 56], or by generating the faulty signal by further analysis of the system output using various signal processing techniques such as wavelet transformation and Fourier transformation, etc., [19, 27, 37]. The advantage of this approach is its quick diagnosis and low computation time. On the other hand, the drawbacks include the need for some of them to update the threshold regularly to mitigate the influence of array aging, while others have a sophisticated structure to alter the threshold adaptively. Furthermore, many threshold approaches can only identify a limited number of PV fault types. In addition, models need to be updated constantly to account for changes in the way the system works due to seasonal changes or short-term environmental effects like low irradiance or different shading conditions, as in [13], where the diagnostics accuracy reached 98%.

1.2.2. Artificial Intelligence-Based Methods and Hybrid Techniques

Machine learning methods that demonstrated high fault detection accuracy under different fault types can be trained to differentiate between normal and faulty operating conditions by finding distinctive features or signatures in the signals [6, 8]. These methods use a large data sample to create fault detection and diagnosis models, taking advantage of current computer technology advancements. While signal processing methods rely on a pre-defined mathematical technique to analyze and interpret signals in PV systems, they may not be optimal for detecting faults under varying operating conditions or in complex systems. Machine learning-based algorithms, on the other hand, can learn from data and adapt to changing conditions, making them better suited for detecting failures in complex and dynamic systems like PV systems. Machine learning algorithms can examine massive datasets of PV system signals to detect trends and anomalies that indicate faults. Machine learning-based solutions also have the advantage of detecting previously unknown faults that may lack a clear signature or trend. By training a machine learning model on a collection of labeled data, the model can learn to recognize different kinds of faults. Furthermore, machine learning models can be tailored to detect faults with high accuracy while limiting false alarms, which can cause expensive maintenance and downtime, and lowering the overall efficiency of the system. The following Table 1 summarizes the main differences between both techniques in PV fault detection and classification [19, 20, 22, 35].

Table 1. Comparison of machine learning and signal processing techniques for fault detection and classification in PV arrays.

Criteria	Machine learning techniques	Signal processing techniques
Fault detection accuracy	High accuracy, especially with complex and nonlinear relationships.	Effective for detecting periodic and transient faults but may struggle with complex signals.
Adaptability to changing conditions	High adaptability; can learn and adjust to new patterns with sufficient training data.	Limited adaptability; requires manual threshold adjustments for new conditions.
Computing complexity	Can be computationally intensive, especially for deep learning models.	Generally low computational complexity, suitable for real-time processing.
Simplicity of use	Requires significant expertise for model training, tuning, and deployment.	Relatively simple to implement, but effectiveness depends on accurate threshold settings.
Data requirements	Requires large, high-quality datasets for training and validation.	Can work with smaller datasets but may require detailed domain knowledge for setup.
Real-time processing	Challenging due to high computational demands; often requires powerful hardware.	Suited for real-time applications due to lower computational needs.
Maintenance	Requires regular updates and retraining to incorporate new data and fault types.	Minimal maintenance, primarily involves updating thresholds as needed.

To date, various efforts have utilized ML techniques to build reliable and robust algorithms to overcome signal processing approach shortages. Artificial intelligence-based approaches such as probabilistic neural networks [1], random forest learning [14], stagewise additive modeling employing a multiclass exponential loss function based on a classification and regression tree [23], conventional neural networks [8, 28], and Genetic algorithm [32], were utilized to diagnose faults in PV systems. Despite this, most of these studies were unable to offer an accurate model for recognizing fault patterns as they did not pay enough attention to the LL and LG faults at low mismatch or high impedance levels. However, some research works have attempted to address the stated issues using various machine learning classifiers, such as decision trees [61], kernel-based extreme learning machine algorithms [15], graph-based semi-supervised learning algorithms [58], fuzzy inference systems [51], and the two-stage Support Vector Machine (SVM) [52]. Harrou *et al*. [21] proposed a data-base procedure for monitoring the operating performance of a PV system using kernel-based machine learning methods for fault detection, specifically Support Vector Regression (SVR) and Gaussian Process Regression (GPR). The procedures only require the availability of system measurements collected via sensors. The developed monitoring scheme based on kernel density estimation successfully detects and identifies different faults in a 20 MWp grid-connected PV system. GPR-based monitoring procedures achieved better detection performance than SVRs for monitoring PV systems, and GPR-based KDE monitoring schemes outperform SVR-based schemes in all cases. Moreover, Taghezouit *et al*. [44] have suggested a monitoring method for photovoltaic systems based on parametric models and double exponentially smoothing. The method used empirical models to obtain residuals and detect faults, and a double exponentially smoothing scheme to sense faults

by examining the generated residuals. The flexibility of the approach is extended with a non-parametric detection threshold computed via kernel density estimation. The method is tested on several fault scenarios and shown to successfully trace faults using real data from a 9.54KWp photovoltaic system. Same Taghezouit *et al*. [43] have proposed a multivariate statistical monitoring of photovoltaic plant operation. This research proposed a simple and efficient monitoring methodology for detecting anomalies in PhotoVoltaic (PV) systems using a principal component analysis model and multivariate monitoring schemes. The research work aimed to design assumption-free principal component analysis-based schemes and proposed a nonparametric approach using kernel density estimation to set thresholds for decision statistics. Real measurements from an actual 9.54 KWP grid-connected PV system were used to illustrate the performance of the proposed approach, and six case studies were investigated to evaluate the fault detection capabilities of the proposed approach. The results highlighted the efficiency of the proposed method in monitoring a PV system and its greater flexibility when using non-parametric detection thresholds. Moreover, Tyagi *et al*. [47] proposed a method to detect the faults in PV arrays by predicting the output power of the PV modules by three machine learning models, the average accuracy of the proposed method was around 85%.

Nonetheless, these fault detection models have a multitude of challenges, such as the need for a large dataset in the learning process, low detection accuracy for faults with a low-percentage mismatch or high impedance, the unreliability in some of these methods because of the use of only one classifier, the absence of hyperparameters optimization methods that remarkably affect the classifier performance, the proposed PCA-based anomaly detection approaches were suitable for one scale (time scale) and may not be suited for detecting anomalies at several scales, and the lack of comparison approaches to select the best classifier.

Lately, Eskandari *et al*. [18], published a ML-based detection and classification method using a hierarchical classification technique. Different mismatches and resistance levels were applied, and the data was tested using three types of classifiers, namely: LR, NB, and SVM. However, this study was interested in detecting the LL and LG faults under different mismatch levels and temperature conditions, despite the partial shading effect. Different levels of classification were used, which caused the proposed algorithm to take even longer. Badr *et al*. [6] have suggested a two stage fault detection and diagnostics algorithm to detect and diagnose four types of faults in the PV system under different shading levels and temperature conditions, namely: LL, MPP, open circuit, and Arc faults. The proposed algorithm has compared the behavior of three ML classifiers: DT, KNN, and SVM. In addition, the optimal classifiers' parameters were chosen based on the Bayesian optimization method. This study has developed two modules, fault detection, and fault diagnosis modules. The required parameters by the developed method include irradiance, temperature, voltage, and current. The accuracy of the proposed model was 100% for fault detection and 89.84% for the fault diagnosis models based on the SVM classifier utilizing the Cubic Kernel.

From the previous studies, it can be noticed that the two-stage models are consuming more time in the detection and classification processes, in addition to the need for a method to differentiate between the faults caused by LL and LG faults. Therefore, this paper proposes a new and accurate machine learning-based fault detection and diagnosis tool. The proposed tool will be utilized to detect different major faults, namely: LL, LG, partial shading, and complete shading faults under different mismatches, fault resistances, and shading levels.

Moreover, when a fault occurs in a PV system, the current will deviate from its expected behavior. By analyzing the current signals, it is possible to detect these deviations and identify the location and type of fault. Voltage and power signals are also important in fault detection, but they can be affected by factors such as temperature and weather conditions, or installation conditions [31], which can make it more difficult to identify faults. Furthermore, in PV systems, the current is typically measured using a shunt resistor or a Hall effect sensor, which is relatively simple and cost-effective compared to measuring voltage and power. This makes current measurements a practical and reliable way to detect faults in PV systems. Therefore, the PV array current will be the only required input to the developed model. On the other hand, and to speed up the algorithm's response time, different statistical calculations will be applied on the model's input, which will in turn extract the distinguishing features for each fault type under the applied environmental and technical conditions. The contribution of this work is to develop a novel and accurate machine learning-based-fault detection and diagnosis tool. The proposed tool can detect major faults in photovoltaic arrays, including LL, LG, Complete shading, and partial shading faults under different environmental and technical conditions. The model utilizing only the PV array current through various statistical calculations to extract the distinguishing features for each fault type, those features will be the input of the developed ML model instead of the current signal itself, which will speed up the proposed algorithm. Twelve machine learning classifiers will be applied, and the hyperparameters of the top-performing classifier will be optimized to provide the most efficient and precise results. The proposed tool offers significant potential for improving the performance and reliability of PV systems. The main contributions of this paper are summarized as follows:

- Studying the behavior of the PV arrays under different faults affecting the DC side of the PV system, including LL, LG, and partial shading faults with different irradiance, temperature, and mismatch conditions. This study will show the possible causes of these faults and their effect on the system's performance. The complete shading that might happen temporarily because of clouds or snow will be distinguished from partial shading that happens permanently.
- Develop a supervised machine learning model with only one data input, the array current, the developed model will test twelve different types of ML classifiers to choose the most efficient and top-performing classifier in terms of speed and accuracy. The utilized classifiers include K-Nearest Neighbor classifier (KNN); Random Forest classifier (RF); Support Vector Machine classifier (SVM); Naive Bayes classifier (NB); Decision Tree classifier (DT); Gradient Boosting Classifier (GBC); Multi-layer Perceptron classifier (MLP); Gaussian Process classifier (GPC), Extra Trees Classifier (ETC), AdaBoost Classifier, Quadratic Discriminant Analysis (QDA), and stochastic gradient descent (SGD) classifier.
- Optimize the accuracy and the processing time of the selected classifier by integrating the Random Search Optimization algorithm (RSO) to tune the hyperparameters of the top performing classifier, Random Forest Classifier (RFC), as it will be shown later.

It is worth mentioning that the proposed fault detection and classification technique is based on several critical assumptions and limitations that would limit the real-life applicability; the assumptions of this research are briefly described next:

- Firstly, it assumes a high-quality input signal free from significant noise or interference. Outliers or missing data can have substantial effects on the performance of machine learning models. Incomplete or noisy datasets might severely downgrade ML model performance and increase unreliable results in fault detection and diagnosis.
- Besides, machine learning models need to be adequately trained with appropriately large and diverse datasets so that different fault conditions and normal operations can be correctly captured.
- It is assuming that all the PV modules in the array are homogenous and working under similar conditions of irradiation, temperature, thermal stress, and humidity, it is assumed that they will degrade similarly over time.
- Most ML models assume that different types of faults are independent-when one fault happens, it will not dramatically influence the detection of another.

The limitations in the developed model can be summarized as follows:

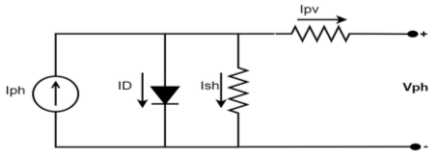
- The need for updating the ML models continuously to incorporate new scenarios of faults that may arise. During training, ML models are trained based on historical data, which, in general, represents known fault conditions. If the PV system undergoes a specific type of fault that the model has not seen in the training dataset, it could fail to identify correctly.
- Machine learning models are susceptible to hyperparameters; therefore, they need to be very well tuned and optimized.
- ML-based models' efficiency is dependent on the quality of data preprocessing and feature extraction, where the selection of the most relevant attributes representing characteristics of the fault is highly affecting the accuracy of the model.

This paper is organized as follows: Section 2 presents the modeling of the employed PV array and its characteristics. Section 3 explains different PV faults and investigates their effect on the PV array. Section 4 explains the methods used for data acquisition and pre-processing. In section 5, the developed ML-based fault detection and classification algorithm is presented. It also discusses different types of ML classifiers, their hyperparameters, and the integrated optimization technique. Simulation results and model evaluation are presented in section 6. The conclusions and recommendations are in section 7.

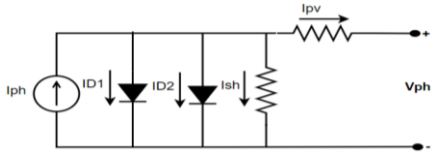
2. Modeling of PV Array

2.1. PV Cell Equivalent Circuit

Because of the non-linear I-V characteristics of the solar cells, modeling them as a constant voltage or constant current source is not appropriate. To describe the electrical characteristics of solar cells, the one-diode and double-diode models are the most common. Figure 2 illustrates the analogous circuits for the one-diode and the double-diode models [31, 59]. However, the one-diode solar cell model has several advantages over the double-diode model, including the high accuracy for the steady-state and fault analysis at the system level, the availability of the data for most PV modules on the market, and the rapid responses in the simulation environment [36]. Therefore, the one-diode model will be used in this research to build the PV array. By incorporating the one-diode model into the simulations, realistic current-voltage (I-V) curves that reflect the actual performance of the PV system can be generated. These simulated I-V curves serve as the basis for creating the statistical characteristics necessary for training and validating the machine learning-based fault detection and classification algorithms [17].



a) The one-diode model.



b) The double-diode model.

Figure 2. PV cell equivalent circuit.

2.2. Mathematical Representation

The one-diode model and its I-V characteristic can be described by the following Equations [37, 56]:

$$I_{PV} = I_{ph} - I_d - I_{sh} \tag{1}$$

where the I_{ph} is the photocurrent (A), I_d is the diode current (A), and I_{sh} is the shunt current (A) and it can be represented as follows:

$$I_{ph} = \left[I_{sc,STC} K_I \left(T_{cell} - T_{ref}\right)\right] \frac{G}{G_{ref}} \tag{2}$$

$$I_d = I_o \left(e^{\left(\frac{V_d + IR_s}{aV_t}\right)} - 1\right) \tag{3}$$

$$I_{sh} = \frac{V_{PV} + R_s I_{PV}}{R_{sh}} \tag{4}$$

where *Io* is the diode reverse saturation current, and V_d is the diode voltage. Both parameters are represented in Equations (5) and (6).

$$I_o = \frac{I_{sc,STC} + K_I \left(T_{cell} - T_{ref}\right)}{e^{\left(\frac{V_{oc,TC} + K_v \left(T_{cell} - T_{ref}\right)}{a_{vt}}\right)} - 1} \tag{5}$$

$$V_d = V_{PV} + R_s I_{PV}) \tag{6}$$

where G (W/m^2) is the solar irradiation, G_{ref} (W/m^2) is the illumination reference=1000W/m^2, T_{cell} (^oK) is the cell temperature, T_{ref} (^oK) is the reference temperature, K_I ($A/^oK$) is the short-circuit current temperature, (*a)* is the diode ideality factor, R_s (Ω) is the PV series resistance, R_{sh} (Ω) is the PV shunt resistance, and V_t is the thermal voltage. However, the diode ideality factor represents the divergence of the diode from ideal behavior. It affects the recombination processes in PV cells, hence affects the I-V properties. The variation in the ideality factor can change the curvature of the I-V curve, and without being well accommodated into the model, making it harder to distinguish normal and different faulty conditions if not properly accounted in the model [41]. On the other hand, thermal voltage affects the voltage drop across the diode and is directly related to the temperature of the PV cells. Since the I-V curve moves with temperature and thermal voltage changes, the shifts can hide or mask the fault conditions, and that complicates the fault detection process [1].

In contrast, the solar module is several solar cells connected in series and parallel to provide the required output voltage and power. The characteristic equation for the equivalent circuit of the photovoltaic module arranged in N_s series and N_p parallel cells is described as in the following Equation [45]:

$$I^M = N_p I_{ph} - N_p I_o \left(e^{\left(\frac{\frac{V^M}{N_s} + \frac{I^M R_s}{N_p}}{AV_t}\right)} - 1 \right) - \frac{\frac{N_p V^M}{N_s} + I^M R_s}{R_{sh}} \tag{7}$$

To first model the PV system under different operating conditions covering normal and faulty scenarios including Line-to-Line (LL) faults, Line-to-Ground (LG) faults, and partial shading faults, a full Simulink model with two parallel strings and three modules per string was created, as shown in Figure 3.

Moreover, identical PV system setups and fault situations were applied on the built model using PSIM^{TM} software to verify the SIMULINK model, as shown in Figure 4. Both models were exposed to the same environmental and technical parameters, such as the temperature and irradiance. The main metrics compared were the output power, current, and voltage at different PV system configurations. To assess the consistency between the two models, their output datasets were compared. As a result, high consistency was found in the Simulink and PSIM dataset analysis, with relatively little variations explained by simulation artefacts and model-specific subtleties. A sample results of the I-V curve from both models of the healthy setup under STC is shown in Figure 5, while the P-V curve generated by the same models is shown in Figure 6.

On the other hand, to guarantee the precision and robustness of the developed model, the simulation data was cross-checked with the datasheets provided by different PV manufacturers. The comparison criteria are by matching the simulation parameters such as the Maximum Power Point (MPP), open-circuit voltage (V_{oc}), short-circuit current (I_{sc}), and temperature coefficients with the actual values listed in the datasheets. Through this comparison, the reliability of the developed model's results was increased by ensuring that the simulation model faithfully captured the real-world behavior of the PV components across a variety of environmental variables and fault situations. An example of simulation results for the PV module (KC130GT) is shown in Figure 7.

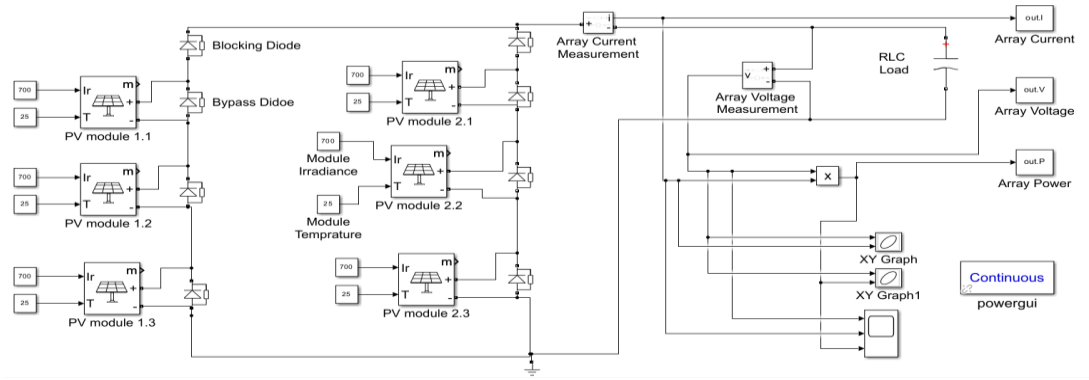


Figure 3. SIMULINK model of the PV array.

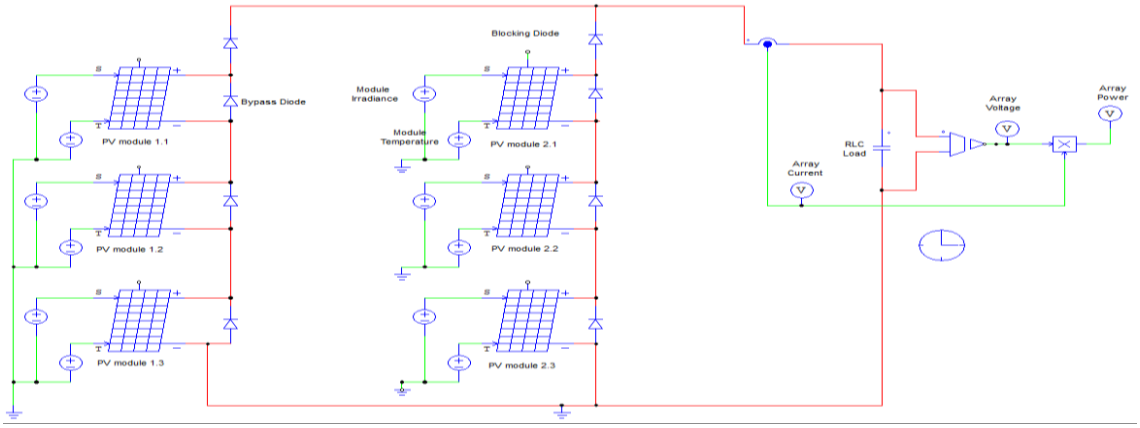


Figure 4. PSIM model of the PV array.

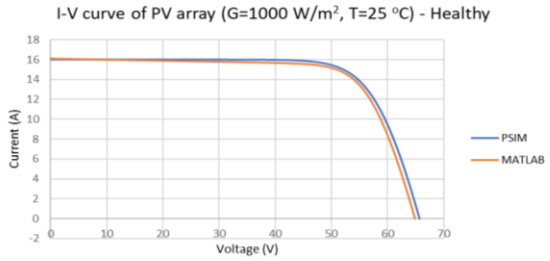


Figure 5. I-V curve of the designed model using MATLAB and PSIM.

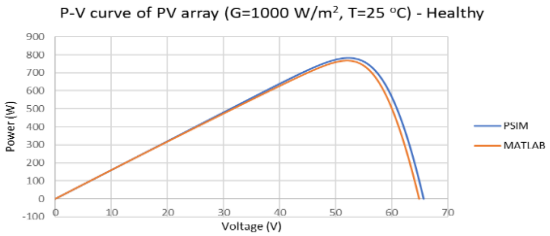


Figure 6. P-V curve of the designed model using MATLAB and PSIM.

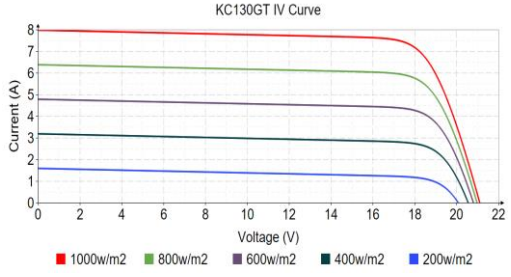


Figure 7. The I-V curve of the simulated module (KC130GT) under different irradiance levels.

3. Data Acquisitions and Processing

Machine learning will be used to address the above faults, and since the input and output are both known, the supervised machine learning technique will be utilized, where the output current of the PV array will be the model's input, without the need for any other attributes or live measurements. A set of statistical milestones will be extracted from the current signature, which, in comparison with the PV voltage and power, can be easily measured, including the important features that are the best representatives of the PV situation.

3.1. PV Data Generation

Under different environmental and electric conditions, various cases will be simulated, as shown in Table 2. The designed PV model generates three types of row data, namely, current, voltage, and power. Moreover, wide range of variable parameters were considered, such as temperature, irradiance, number of shaded modules, and different levels of fault resistance to ensure the consistency of the proposed model with the variable environmental conditions shown in Table 2. It can be noticed from the table that the situations when the PV array experiences no fault are the "healthy and complete shading situations", where the data are extracted at a varying temperature between 5°C and 50°C with 5°C increment, while the irradiance is varying between 200W/m^2 and 1000W/m^2 with 200W/m^2 increment. In addition, the generated data for different faults is

simulated under the same conditions of temperature and irradiance. Moreover, the LL and LG faults are studied under varying resistance between pure short circuit (0Ω) and 50Ω with 10Ω increment. In the partial shading case, a random distribution of the shading levels is considered, including a combination of one, two, and three partially shaded modules. However, it was noticed that the PV current carries the most important information about the health of the PV system. Part of the generated data will be used for ML training and testing, while the other part will be kept for algorithm validation and testing at a later stage. The collected PV current datasets are reformulated so that each dataset will be a combination of two parts; the attribute, which comprises the classification features, and the label (class), which is the status or category of the collected PV current. This class will be identified as a healthy condition or any other fault type from the faults of focus in this research.

Table 2. Training data generation under different environmental and technical conditions.

PV Status	Temperature (°C)	Irradiance (W/m²)	Fault Resistance (Ω)	Remarks
Healthy	[5:5:50]	1000	NA	-
Cloud shading	[5:5:50]	[200:200:800]	NA	-
Partial shading	[5:5:50]	850 for all modules except modules 1.1 (300) and 2.2 (500), 2.3 (350)	NA	This case to be runs for module 1.1 alone, then 1.1 and 2.2, then 1.1, 2.2 and 2.3 together
LL fault	[5:10:50]	[400:200:1000]	[0:10:50]	-
LG fault	[5:10:50]	[400:200:1000]	[0:10:50]	-

The utilized methodology in generating the current data for different situations of the PV array is shown in Figure 8, where Rf_{LL}, and Rf_{LG} are the LL and LG fault resistances, respectively. The flag "i" represents the column number where the newly generated current row will be saved. The outcome of this step will be the input to the data preprocessing and statistical analysis phase. As shown in the flowchart, the simulation parameters are first initialized. These parameters include temperature (T), irradiance (I_r), resistance values for line-to-line (Rf_{LL}) and line-to-ground (Rf_{LG}) faults, and a counter (i). The initial temperature is 5°C, the initial irradiance is 200 W/m², and both Rf_{LL} and Rf_{LG} are configured to 10Ω. The Simulation retrieves the PV current and stores it in the i^{th} column of its dataset and increment the counter *i*. The simulation would then check whether the temperature is over 50°C or the irradiance is over 1000W/m², where they will be resetted to their initial conditions and the resistance values for the LL and LG faults will be increased by 10Ω. If the resistance value for Rf_{LL} exceeds 50Ω, the LL fault is cleared by setting Rf_{LL} to infinity, and similarly for the LG fault with Rf_{LG}. Once all faults are applied, the simulation proceeds to apply partial shading faults (PS_1M, PS_2M, PS_3M) and Complete Shading faults (CS), generating data for each fault situation.

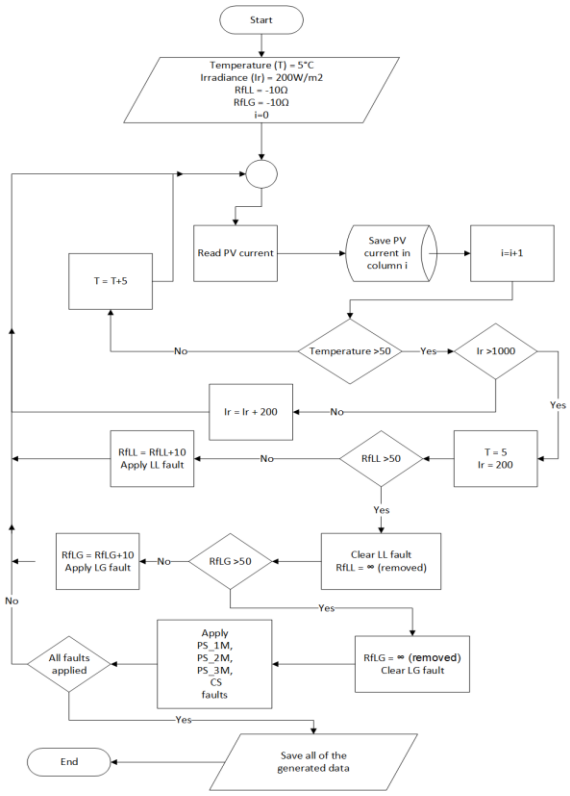


Figure 8. Data generation flowchart.

3.2. Features Extraction

The main goal of data processing and features extraction is to convert the raw data into a numerical feature without losing any important data from the original dataset. Different statistical calculations will be applied to the collected data to get the milestone features and reduce the data size, which in turn will speed up the algorithm and better visualize the PV faults. In addition, converting the row current into a statistical scalar will help in getting a meaningful input format and substituting any missing instance of data. Therefore, the model's input will be the statistical features of each current signature instead of the current row itself. The applied statistical calculations include the maximum value, minimum Value, mean value, standard deviation, Root Mean Square (RMS), skewness, kurtosis, crest factor, and form factor.

1) Peak value (Xp)

The peak value of the current vector *x* including the observations $x_i(i=1to\ N)$ in a dataset made up of *N* scalar observations, is defined by:

$$x_p = \max|x_i| \tag{8}$$

2) Minimum value

The minimum value of the current vector *x* including the observations $x_i(i=1to\ N)$ in a dataset made up of *N* scalar observations, is defined by:

$$x_{min} = min|x_i| \tag{9}$$

3) Mean value (μ)

The mean value of the current vector *x* including the observations $x_i(i=1to\ N)$ in a dataset made up of N scalar observations, is defined by:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{10}$$

4) Standard Deviation (STD)

The STD of the current vector *x*, including the observations $x_i(i=1to\ N)$ in a dataset made up of *N* scalar observations, is defined by:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}|x_i-\mu|^2} \tag{11}$$

5) Root Mean Square (RMS)

The RMS of the current vector x including the observations $x_i(i=1to\ N)$ in a dataset made up of N scalar observations is defined by:

$$x_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} {x_i}^2} \tag{12}$$

6) Kurtosis

Which is a measurement tool to describe the difference in the current vector distribution from the normal distribution. The faults might increase the value of the kurtosis because of the increase in the number of outliers in the current vector. The Kurtosis of the current vector *x* including the observations $x_i(i=1to\ N)$ in a dataset made up of *N* scalar observations, is defined by:

$$x_{kurt} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\right]^2} \tag{13}$$

7) Skewness

A measurement for the asymmetry of current waveform. In general, faults might affect the symmetrical distribution of the signal, i.e., the level of skewness will be increased based on the fault size and type. The skewness of the current vector *x* including the observations $x_i(i=1to\ N)$ in a dataset made up of N scalar observations, is defined by [32]:

$$x_{skew} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\right]^{\frac{3}{2}}} \tag{14}$$

8) Form factor

Is the ratio between the RMS value and the average value of the output current, the form factor will give an accurate measurement for the RMS value of the sinusoidal waveform, especially if there is any distortion in the signal, such as the distortion caused by faults. The form factor of the current signal *x* including the observations $x_i(i=1to\ N)$ in a dataset made up of *N* scalar observations, is defined by:

$$x_{form} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}}{\frac{1}{N}\sum_{i=1}^{N} x_i} \tag{15}$$

9) Crest factor

Which is the ratio between the peak value of the current and its effective value. It is represented by dividing the peak value of the current row (x_p) by its RMS. Faults generally appear as a change in the peak of the current signal before appearing in the energy, which is represented by the root mean square of the signal. The crest factor provides an early warning of the faults once they develop in the PV system. The crest factor of the current signal *x* including the observations $x_i(i=1to\ N)$ in a dataset made up of *N* scalar observations, is defined by:

$$x_{crest} = \frac{x_p}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}} \tag{16}$$

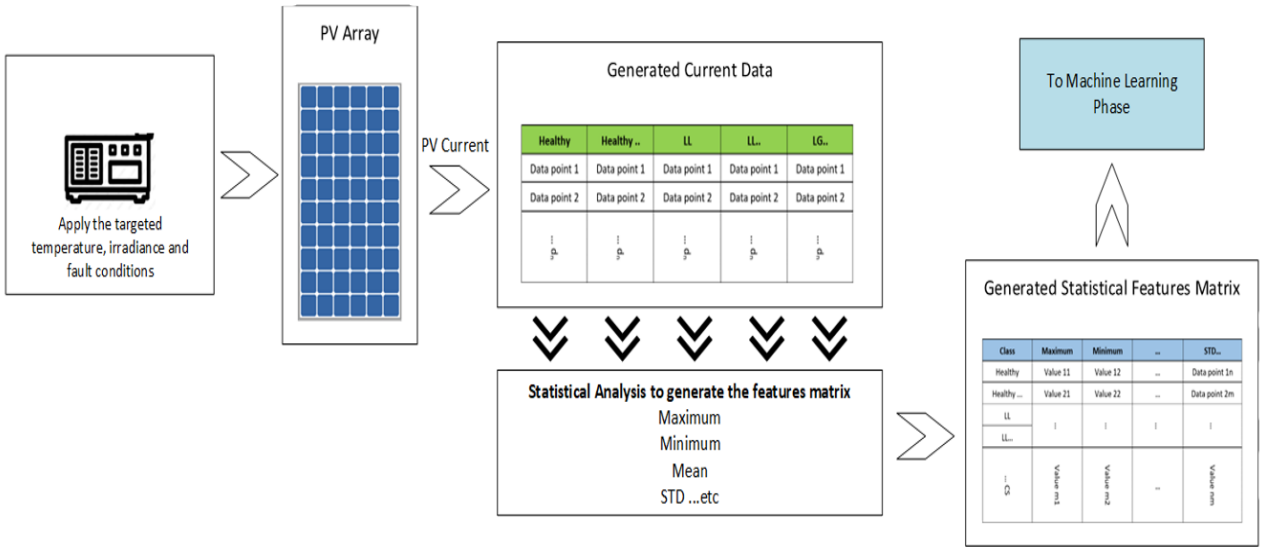


Figure 9. Statistical features matrix generation.

The flowchart in Figure 9 depicts the process of constructing a statistical characteristics matrix from PV current data under various fault scenarios using the developed algorithm. Initially, temperature, irradiance, and fault conditions are applied to the PV array to simulate several operational scenarios, including normal (healthy) and faulty states. The PV array, working under these controlled conditions, generates the current data that captures the electrical characteristics of the system. This data is then formatted into columns with labels corresponding to the operational condition of the system being Healthy, Complete Shading (CS), Line-to-Line (LL) faults, Line-to-Ground (LG) faults, and Partial Shading (PS) with 1, 2, or 3 shaded modules (PS_1M, PS_2M, PS_3M). Each column of data is subsequently analyzed to extract its statistical properties, which include maximum, minimum, mean, standard deviation, and other metrics. These extracted statistical features for each instance are reformatted into a row format, where each row includes the fault label and the eight statistical values.

This reformatting guarantees that the data is organized in a structured manner, simplifying subsequent process.

4. Machine Learning Based Fault Detection and Classification Algorithm

Machine learning is a sub-area of artificial intelligence; it involves taking automatic decisions about the class of the input data based on the learned knowledge from the provided training dataset. Each dataset includes two parts; the attribute, which is the feature of the instance that determines its properties or classification, and the label, or the class to which each dataset belongs [9, 46, 60]. In this research, the attributes are collected from the statistical analysis of the PV current, which is labeled based on the fault type as “normal behavior” (Healthy), or any of the targeted fault types, namely, Complete Shading (CS), Line to Line fault (LL), Line to Ground fault (LG), and Partial Shading with one, two, or three shaded modules (PS_1M, PS_2M, and PS_3M, respectively). The datasets are arranged in the feature matrix of dimension $dx(n+1)$, where d represents the sequence number of different fault cases under various temperature, irradiance, and fault resistance levels. While each instance of the training row data is annotated in the form $\{y\_i, x_{i\_1}, x_{i\_2}, \ldots, x_{i\_n}\}$, where yi is the category label, $i \in [1,d]$, d represents the number of rows, n represents the number of the used statistical features as listed in the previous section, and $(n+1)$ represents the complete row data including both label and statistical features (number of columns.

In general, machine learning is categorized into three main categories:

1. Supervised learning, which uses a labeled dataset for training.
2. Unsupervised learning, which uses unlabeled dataset.
3. Semi-supervised learning, which is a combination of the two types, i.e., using both labeled and unlabeled data for training.

However, since this is a classification algorithm where input and output are both known, the supervised machine learning method will be used.

On the other hand, to allow the algorithm to directly retrieve the most up-to-date data from the PV array and to ensure a precise and rapid detection and classification of faults, Figure 10 illustrates the proposed position of the developed algorithm into the DC side of the PV system. This position is critical as it will perform continuous monitoring of the PV current on a regular basis. Moreover, the algorithm undertakes different critical jobs to ensure the system's operating integrity. First, the pre-processing of the current data, including cleaning of the data and removing any noise or anomalies that can bias the analysis, ensuring consistency by normalizing the data, and interpolating missing values to preserve a complete dataset. Once the data is pre-processed, the algorithm starts with the statistical analysis. The results of the statistical analysis will represent the attributes of the ML model to identify the health of the PV system by using the ML classification processes. By using the algorithm at the DC side, it ensures that any PV fault can be recognized as soon it occurs, allowing for rapid operations, therefore, the suggested method and position is crucial for ensuring the reliability and efficiency of the PV system for real-time analysis.

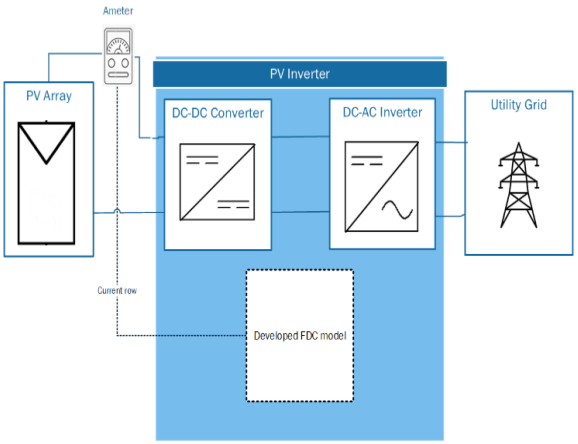


Figure 10. Overview of the proposed fault detection and classification tool position.

However, one significant challenge in the proposed system is data acquisition, which requires high-quality, consistent data from reliable sensors mounted across the PV array. Ensuring data quality and limiting noise are crucial for accurate model performance. For model deployment, the integration of machine learning algorithms into existing PV monitoring systems demands scalable computer resources and real-time data processing capabilities. Strategies such as leveraging edge computing and cloud-based platforms can promote

efficient model deployment. Maintenance is another key feature, as the model needs regular updates and retraining to include new data and respond to developing fault circumstances. Implementing automatic maintenance methods and continuous monitoring systems can assist sustain the model's effectiveness over time. These concerns are critical for transferring from a theoretical framework to real, robust applications in PV systems.

The following subsections explain the flow of the algorithm.

4.1. Dataset Splitting and Cross Validation

In ML, there are different splitting methods, such as the holdout, cross validation (*K-fold*), stratified cross validation (*sK-fold*), and Leave-One-Out Cross Validation (LOOCV). The holdout is the basic method of data splitting and model validation, where the dataset is divided into two sets: the training set, which is what the model is trained on, and the testing set, which is used to test the performance of the model with unseen data. The cross-validation method, on the other hand, is a statistical approach used to measure the overall accuracy of the model based on the random splitting of the dataset into "*K*" groups or "Folds", where one of the groups utilized for testing the model, while other "*K-1*" groups are for training. The process will be repeated until all the groups are used as testing sets, which means it allows the model to be trained on several portions of the dataset [30].

Stratified cross validation is comparable to *K-fold* cross-validation with a few minor differences. This method is based on the stratification principle, which is the process of reorganizing data to ensure that each fold or group is a good representation of the entire dataset. It is one of the finest ways for dealing with bias and variation. The leave one out cross-validation method requires creating a model for each dataset with one datapoint reserved from each dataset. The model is trained on the remaining dataset, and the process is repeated for all datapoints in the dataset, which means all datapoints are used to test the model. It is a computationally costly process, which makes it inappropriate for a large dataset, but it yields an accurate and unbiased measure of model performance.

In this research, the four types of data splitting and model validation methods will be tested to choose the best splitting method in terms of accuracy and speed (time). In the Hold-out method, the datasets will be distributed into 80% for training and 20% for testing. On the other hand, different folds will be tested in *K-fold* and *sK-fold* cross validation methods ranging from 2 to 13 folds, and the best-performing number of folds for each of them will be selected and compared with other methods. The LOOCV method will be tested since the developed model is using the statistical dataset as its input, which will not be as large as the original current row.

Figure 11 represents the general flowchart of data splitting. The process begins with a chosen subset of the training data for each tree which is used to generate individual trees inside an ensemble approach, such as a Random Forest algorithm. The method examines if the stopping condition holds for each tree based on parameters such as maximum tree depth or a minimum number of samples necessary to divide a node. If the stopping condition is not met, the process then builds the next split. The method proceeds to calculate the prediction error, which measures the difference between the expected and actual outputs. The algorithm compares this error against a predefined threshold. If the error is below the threshold, the process finishes, indicating that the tree has reached adequate precision. If the error exceeds the threshold, the procedure continues, and further splits are examined to increase the tree's accuracy. Moreover, the algorithm calculates a quality metric (such as Gini impurity or information gain) for each potential split point to quantify how well the split separates the data into distinct classes. The optimal split point, which maximizes the separation of the data into homogenous classes, is then picked. By iterating through these processes, the decision tree is formed in a manner that optimizes the splits at each node, boosting the model's capacity to effectively categorize input. This methodology assures that the final model is both accurate and efficient in detecting and categorizing defects in PV systems, ultimately contributing to increased reliability and performance.

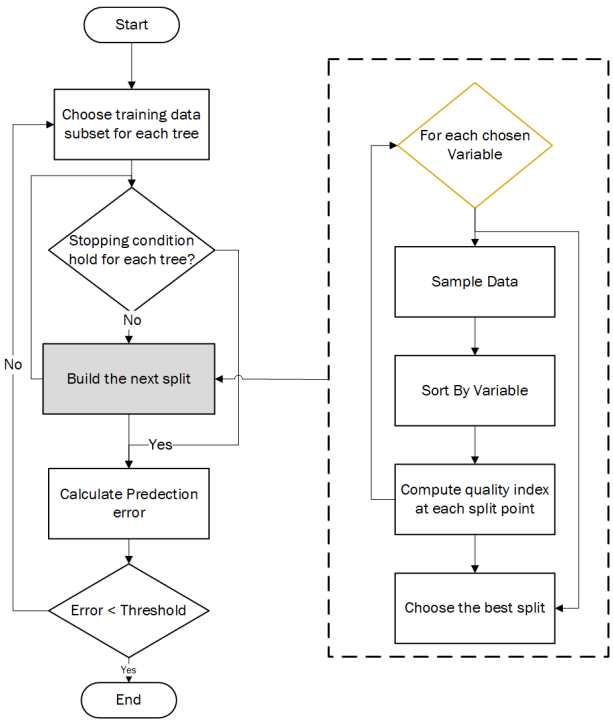


Figure 11. Data splitting general flowchart.

4.2. ML Training and Classification

The process of determining the class or category of a

new dataset is known as classification. In this study, twelve machine learning classifiers will be applied to detect and classify PV faults. However, different classifiers' configurations result in a new ML model. The utilized classifiers include KNN; Random Forest classifier (RF); SVM classifier; Naive Bayes classifier (NB); DT classifier; Gradient Boosting classifier (GBC); Multi-layer Perceptron classifier (MLP); Gaussian Process classifier (GPC), Extra Trees Classifier (ETC), AdaBoost Classifier, Quadratic Discriminant Analysis (QDA), and Stochastic Gradient Descent (SGD) classifier. Table 3 compares the twelve classifiers based on the pros, cons, average speed, and the processing time for each one of the classifiers.

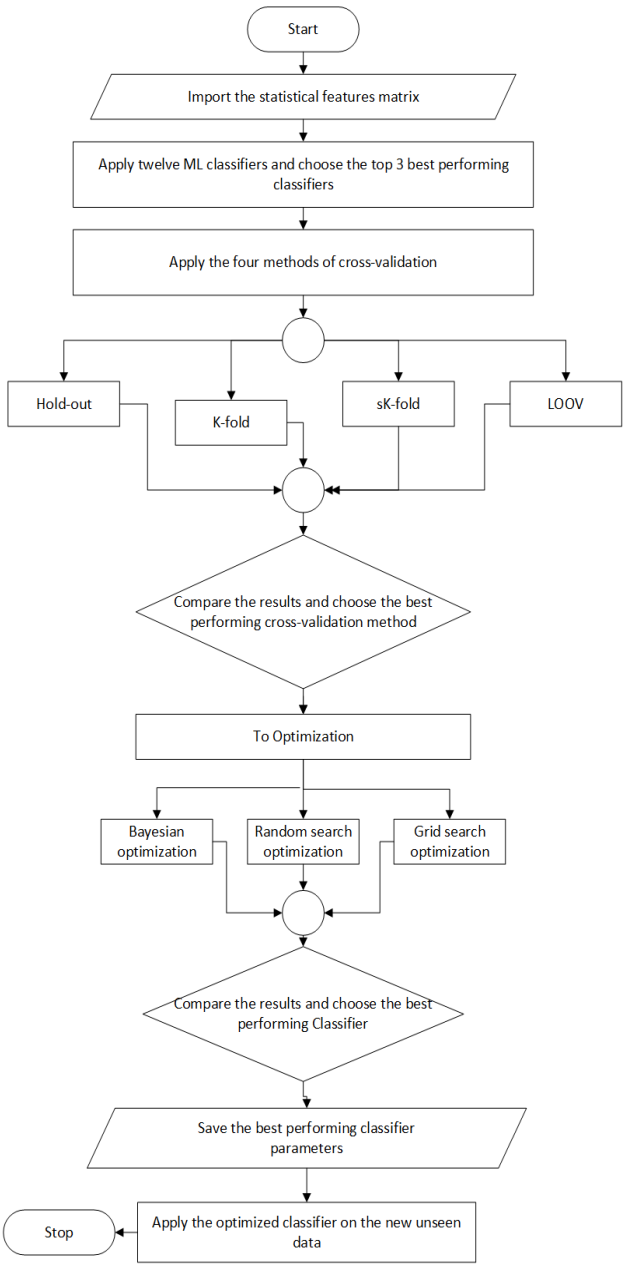


Figure 12. ML Classifier selection and optimization process.

The flowchart shown in Figure 12 illustrates the selection and optimization of the top-performing classifier. The process is initiated by importing the statistical characteristics of the PV current (features matrix), a total of twelve ML classifiers are then implemented to single out three top performers. The classifiers are evaluated utilizing four cross-validation approaches: Hold-out, *K*-fold, stratified *K*-fold (*sK-fold*), and Leave-One-Out (LOOV). The results of the cross-validation methods are compared to select the top-performing approach. The parameters of the selected classifier are then optimized with the assistance of three optimization techniques: Bayesian optimization, random search optimization, and grid search optimization. The results are compared, and the best-performing classifier is chosen. The Parameters of the optimized classifier are then used on new and unseen data.

Based on the classification accuracy results that will be shown in the later section, RFC has shown the highest accuracy, therefore, its structure and hyperparameters are explained in detail, while an optimization algorithm will be integrated to improve the performance of the classifier for PV fault detection and classification, as in the following subsections.

1) Random Forest Classifier (RFC)

The RF algorithm, originally proposed by Breiman [11], it is a broad category encompassing ensemble approaches that utilize classifiers based on trees. RF constructs a significant quantity of decision trees employing bagging, a meta-algorithm employed to enhance classification and regression models based on their stability and accuracy in classification. These decision trees are built from a sub-dataset derived from a distinct initial training set by using two-thirds of the original dataset for training and one-third for testing [34, 36]. The utilization of bagging techniques in machine learning models effectively mitigates variation and concurrently mitigates the risk of overfitting. The process involves the random selection of cases from the initial training dataset, while the bootstrap sets are utilized to build each of the decision trees in the Random Forest (RF) algorithm. Each tree classifier is referred to as a component predictor. The Random Forest algorithm determines its judgments by aggregating the votes of individual predictors for each class, thereafter, picking the class with the highest number of votes as the winner [50].

The fundamental concept underlying RF classification is a process known as bootstrap sampling [40], which is employed to extract k samples from the original training set. Each sample has the same capacity as the original training set. Subsequently, k Decision Tree (DT) models are constructed for each of the k samples, resulting in k classification outcomes. Ultimately, based on the outcomes of the k classification process, a decision is made to assign a final classification to each individual record using a voting mechanism.

Table 3. A Comparison between the twelve classifiers used for PV fault detection and classification.

Model	Description	Pros	Cons	Average speed	Processing time
Rule-based models (Signal processing models)	Uses a set of rules based on a pre-defined thresholds to detect faults	Simple and easy to implement, low computational cost	Limited accuracy, highly dependent on the quality of input data and expert knowledge	Fast	Low
ANNs	Uses a multi-layered neural network to detect faults	High accuracy, can learn complex patterns in data, and can handle non-linear relationships	Can be computationally expensive and requires large amounts of training data	Slow	High
Fuzzy logic-based models	Uses fuzzy sets to capture the imprecision and uncertainty in data	Robust to noise and can handle uncertain and vague information	Difficult to implement and requires expert knowledge to define fuzzy rules	Fast	Low
DL models	Uses multiple layers of artificial neural networks to extract features from data and detect faults	High accuracy, can learn complex patterns in data, and can handle large datasets	Requires significant computational resources and large amounts of training data	Slow	High
SVM	Uses a hyperplane to separate data into fault and non-fault classes	High accuracy, can handle non-linear data, and can work with small datasets	Limited interpretability and high computational cost	Moderate	Moderate
RF	Uses an ensemble of decision trees to detect faults	High accuracy, can handle missing data and noisy features, and can provide feature importance information	Can be overfit with noisy data and requires careful parameter tuning	Fast	Low
GBC	Uses an ensemble of weak models to detect faults	High accuracy, can handle missing data and noisy features, and can provide feature importance information	Can be overfit with noisy data and requires careful parameter tuning	Slow	High
KNN	Uses the distance between data points to detect faults	Simple and easy to implement, can handle noisy data	Computationally expensive for large datasets and can be sensitive to irrelevant features	Fast	Low to Moderate
DT Classifier	Uses a tree-like model of decisions and their possible consequences to detect faults	Easy to interpret, can handle non-linear data and can provide feature importance information	Can be sensitive to noisy data and can be overfit with complex models	Fast	Low
LR Classifiers	Uses a linear model to estimate the probability of a fault occurring	Simple and interpretable, can work with small datasets and can handle binary classification tasks	Limited to linear relationships between features	Fast	Low
GPC	Uses Bayesian inference to model the underlying function and estimate the probability of a fault occurring	Provides uncertainty estimates and can handle non-linear relationships between features	Computationally expensive and requires careful parameter tuning	Slow	High
NB classifier	Uses probabilistic models based on Bayes' theorem to detect faults	Simple and fast, can work with small datasets and handle irrelevant features	Assumes independence between features and can be sensitive to outliers	Fast	Low
MLP Classifier	Uses a multi-layered artificial neural network to detect faults	High accuracy, can handle non-linear relationships and large datasets	Requires significant computational resources and large amounts of training data	Slow	High
ET classifier	Uses an ensemble of decision trees to detect faults	High accuracy, can handle missing data and noisy features, and can provide feature importance information	Can be computationally expensive and requires careful parameter tuning	Fast	Low

On the other hand, the sampling process in RF algorithm is random and involves replacement, resulting in certain duplication of samples within the training subset. This duplication is intended to prevent decision trees in the forest from generating local optimal solutions. Therefore, RF utilizes the bagging technique, which eliminates the need for an additional validation step. The internal validation is conducted by utilizing the testing samples, which consist of one-third of the original dataset. The samples are commonly referred to as the Out of Bag (OOB) samples. Therefore, the validation errors that arise from RF are commonly referred to as out of bag errors. A reduced OOB error rate is indicative of superior performance by the RF model. In general, errors of the RF algorithm typically revolve around two primary factors, the correlation between any pair of decision trees inside the forest, and the individual effectiveness of each decision tree. An increase in the association among trees within a RF might result in a more intricate structure, hence potentially leading to an elevation in the OOB error rate. On the other hand, as the strength of each decision tree increases, there is a corresponding decrease in the OOB error rate. Hence, it is imperative to optimize the RF parameters to achieve optimal strength and minimize correlation.

Fortunately, it has been shown that adjusting the overall number of decision trees in the forest and the number of random features utilized to establish the optimal split at each node can enhance the performance of Random Forests. In general, the conventional practice is to set the number of random features in the initial RF algorithm as the square root of the entire number of features. The selection of the number of trees

in the forest can be determined by considering the OOB error rate.

Moreover, different parameters need to be adjusted to construct an accurate model, such as the maximum number of samples, which is controls the sub-sample size; the number of estimators, which determines the number of trees; the criteria used to measure the quality of splitting, such as the "*entropy*" and "*gini*" functions; the maximum depth of the tree, which is the minimum sample split size; and the minimum number of samples required to split the internal node. In this research, a range of values will be assigned for each parameter to select the optimal parameters based on the classifier accuracy and processing time. The parameters' range includes various number of estimators and the "entropy" and "*gini*" criteria, while the Random Search Optimization algorithm (RSO) will be integrated to achieve the optimal value for each parameter.

2) Hyperparameters Optimization (HPO)

The performance of the ML classifier is highly affected by its hyperparameters that are utilized in the training phase, such as the number of estimators, splitting criteria, splitting size, and much more. The manual isolating and tuning of some hyperparameters regardless of the effect of others will end up with a suboptimal solution, whereas all parameters are dependent on each other. Furthermore, the appropriate hyperparameters might differ tremendously amongst datasets. Therefore, automating the process of tuning these parameters would lessen the need for human effort while improving the performance of the proposed model. The process of hyperparameter optimization is based on finding a global D-dimensional optimum hyperparameter setting (x) that will return the best performance of the validation dataset while minimizing the validation error with learned weights (w) and in the least number of steps, as described in Equation (17) [57].

$$\min_{x\in\mathbb{R}^D} f(x,w,I_{val})\ s.t.\quad w = arg\min_{w} f(x,w,I_{train}) \tag{17}$$

where I_{train} and I_{val} are the training and validation datasets, respectively.

Global optimizations are preferred in the black-box functions, such as the phenomenon of this study. There are three black-box systematic approaches for tuning hyperparameters in ML, namely, Grid Search, Random Search, and Bayesian Optimization methods. Grid search is the most basic HPO method; it depends on a pre-defined range for each hyperparameter and is very efficient in low-dimensional space but suffers from the exponential increase in the search space with the increase in the number of parameters or the parameter range, which will increase the number of the required evaluations for the function [57]. RSO algorithm, on the other hand, is a simple alternative to grid search. It is based on proposing random search points from the hyperparameter space with easier parallelization, which can find a comparable hyperparameter setting to grid search in less time, particularly if the effects of some hyperparameters are more important than others. Random search has no assumptions about the ML algorithm being optimized, and with enough features, it can achieve settings that are equivalent or very close to optimum.

In contrast, Bayesian optimization is the state-of-the-art paradigm for the costly black-box functions; it is not directly targeting HPO but can be generally applied with the newly developed models and kernels based on an iterative process. It consists of two fundamental components: a probabilistic surrogate that fits all the target observations, and an acquisition function that determines which point to examine next based on the predictive distribution of the model [42].

However, it was shown in the literature that RSO is more efficient algorithm for tuning the hyperparameters of the ML classifiers in the case of not all parameters are equally important, such as the case of this study [10]. Therefore, RSO will be used for tuning the parameters of the best performing algorithm. The response time and accuracy of the optimized classifier will be compared with the default classifier, and the best-performing hyperparameter configuration will be chosen as the optimal hyperparameters of the proposed model.

3) Random Search Optimization (RSO)

RSO was initially conceptualized by Brooks [12]. The process involves multiple iterations of sampling from the feasible search space, and often following a uniform sampling distribution. Within the framework of RSO method, every candidate point is generated in a manner that is independent of other points. The update of the parameter's value only occurs if the candidate point demonstrates improvement. The main goal of integrating RSO algorithm in this work is to determine the optimal value for each of the best performing classifier's hyperparameters. This is achieved by iteratively exploring several random directions from the present point to identify a descent direction at each step. The distinguishing feature of the RSO algorithm is in its method of determining the descent direction d^{k-1} during the k^{th} optimization update phase. The algorithm selects a specified quantity of random directions originating from W^{k-1}, assesses each potential update point, and selects the one that yields the lowest evaluation (if it is lower than the assessment of the current point), the updated direction of the parameter will be given by Equation (18) [49].

$$W^k = W^{k-1} + d^{k-1} \tag{18}$$

During the k^{th} iterations, a selection of P random directions is made to be tested. The candidate point ($W_{candidate}$) to assess is generated by adding the P^{th} random

direction d^p to the preceding step W^{k-1}. Thus, $W_{candidate}$ may be expressed as:

$$W_{candidate} = W^{k-1} + d^p \quad (19)$$

Upon thorough review of all P candidate locations, the parameter's point (s) that yields the most minimal evaluation will be selected, as denoted by Equation (20).

$$s = \underset{p=1\ldots p}{\text{ARGMIN}}\, g(W^{k-1} + d^p) \quad (20)$$

If the best point discovered possesses a lower evaluation value compared to the current point, denoted as $W^{k-1} + d^s$, the algorithm proceeds to transition to the new point, alternatively, additional set of P random directions can be explored, and the process will be repeated.

On the other hand, the selection of a set of P random directions can be achieved by employing a certain distribution, such as Gaussian distribution, as utilized in this research. The primary concern associated with this approach pertains to the matter of consistency, as each of the potential orientations would possess varying lengths. To maintain consistency in the directions provided for random candidates, it is necessary to normalize them by adjusting their length to a uniform size, such as a length of one [53, 56]. However, the usage of unit-length directions implies that the norm of each direction vector is always equal to one (i.e., $\|d\|=1$). Consequently, at each iteration of the algorithm, exactly one unit will be traversed, as indicated by the Equation [21]:

$$\|W^k - W^{k-1}\| = \|(W^{k-1} + d) - W^{k-1}\| = \|d\| = 1 \quad (21)$$

The length of each step can be adjusted according to the assumed preference by incorporating a step length parameter (α) into each step, that is enabling a complete control over the distance covered by each step. This step of a more generic kind can be represented as follows:

$$W^k = W^{k-1} + \alpha \quad (22)$$

The magnitude of this step is now precisely equivalent to the step length α. This can be demonstrated by the equation:

$$\|W^k - W^{k-1}\| = \|(W^{k-1} + \alpha d) - W^{k-1}\| = \|\alpha d\| = \alpha \quad (23)$$

Therefore, at the k[th] iteration, P random directions of unit length are considered, each scaled by the step length (α). The direction that results in the largest decrease in the function value will be considered. The pseudo code of selecting the optimal value for each of the classifier's parameters with integrating of RSO is shown in the following algorithm.

Algorithm 1: The Pseudo Code of the Random Search Algorithm

1: input: initial point w^o, maximum number of steps K, number of steps K, number of random samples per step P, a length α.
2: for k=1 : K
Compute P-unit length random directions $\{d\}_{P=1}^{P}$, by sampling and
normalizing N dimensional Gaussian.
3: find $s = \underset{p=1\ldots p}{argmin}\, g(W^{k-1} + \alpha d^p)$
4: set $d^k = d^s$
5: form the new point $W^k = W^{k-1} + \alpha d^k$
6: if $g(W^k) < g(W^{k-1})$
$W^{k-1} \xleftarrow{substitute} W^k$
7: output: history of weights $\{W^k\}_{k=0}^{K}$
and corresponding function evaluations $\{g(W^k)\}_{k=0}^{K}$

4) RFC optimization using RSO

As will be shown in the results section, RFC was showing the highest performance when applied on different PV current datasets, this classifier took less classification time compared to other classifiers and provided the highest classification accuracy with a high size dataset, and even with a missing part the data. To improve the performance of this classifier, its hyperparameters need to be studied, and their ranges need to be determined for the tuning stage. The hyperparameters that can be tuned for better performance include the Number of estimators (*N_estimator*), which is the number of decision trees that is mostly correlated with the dataset size. Criterion, which is the function of measuring the quality of data splits, including "*gini*", and "entropy" functions. Maximum features, which is the maximum features for node splitting, its types are the "*sqrt*" and "*log2*". In addition, there are another hyperparameters that can be tuned including Bootstrap, Minimum sample split, Minimum sample leaf, and maximum leaf node. RSO algorithm was integrated to find the optimal hyperparameters values for RFC [30]. A range of values were examined for each of the classifier's hyperparameters. After each optimization round, a set of new and unseen data was used to evaluate the performance of the optimized classifier by measuring the prediction accuracy and processing time. The flowchart of optimizing the hyperparameters of RF classifier by integrating RSO algorithm is shown in Figure 13.

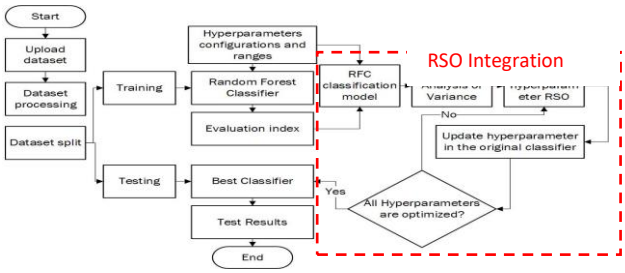


Figure 13. Flowchart of the classification model using optimized RFC utilizing RSO algorithm.

4.3. Evaluation Metrics

There are several measures that can be used to evaluate the performance of the tested classifiers [46]. Four of these measures will be utilized in this study: accuracy,

precision, recall, and F1-score. Such measures are used to answer the question of how precise the prediction model will be. These measures are derived from the confusion matrix, which is a two-dimensional matrix (*N x N*) summarizing the performance of the multiclass classification algorithm by providing the number of correct and incorrect predictions, which will give a clear idea about the performance of the model and the type of errors. The columns of the confusion matrix represent the predicted classes, while the rows are the actual classes, as shown in Table 4, which shows the correct predictions (TP_{xy}) and the misclassified cases (E_{xy}) at a specific class (C_i). The multi-class confusion matrix includes four types of elements: the *TP*, which is the correctly predicted fault class; the True Negative (TN), which is the correctly predicted non-fault class; the False Positive (FP), which is the incorrectly predicted fault class; and the False Negative (FN), which is the incorrectly predicted non-fault class.

The four evaluation measures are derived from the confusion matrix, where the accuracy is the description of the correctly classified data over all predictions, i.e., the ratio between the true positives and true negatives to all predicted positive and negative cases.

The precision is the description of how many positive predictions are correct; it is the ratio between the correctly classified observations TPs and all true and false positive predictions for a specific class (C_i). The Recall, on the other hand, is a metric of how many true positive cases are correctly predicted by the classifier over all positive cases in the matrix. Both precision and recall measures are combined in the F1-score measure, which provides a single metric weighting the two measures and doesn't require knowing the total number of observations. The utilized metrics for the multiclass confusion matrix are defined in Table 5 [7, 31].

Table 4. Multi-class classification problem confusion matrix.

		Predicted Classes			
		C_1	**C_2**	**…**	**C_N**
Actual Classes	**C_1**	TP_{11}	E_{12}	…	E_{1N}
	C_2	E_{21}	TP_{22}	…	E_{2N}
	⋮	⋮	⋮	…	⋮
	C_N	E_{N1}	E_{N2}	…	TP_{NN}

Table 5. Multi class confusion matrix performance metrics.

Metric	Formula
Accuracy	$Acc = \frac{\sum_{i=1}^{N} TP(C_i) + \sum_{i=1}^{N} TN(C_i)}{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{i,j}}$
Precision (for class Ci)	$Precision\ (C_i) = \frac{TP(C_i)}{TP(C_i) + FP(C_i)}$
Recall (for class Ci)	$Recall\ (C_i) = \frac{TP(C_i)}{TP(C_i) + FN(C_i)}$
F1-score (for class Ci)	$F1\ (C_i) = 2x\frac{Recall\ (C_i).Precision\ (C_i)}{Recall\ (C_i) + Precision\ (C_i)}$

5. Results and Discussion

The performance of the proposed method was evaluated under different conditions, as explained in Table 2. To investigate the behavior of the PV array under the provided fault conditions, the developed technique was applied to a PV array of two parallel strings with three modules each using a one-diode model, and each of them was equipped with a bypass diode. The I-V characteristics of the PV array were extracted using MATLAB/SIMULINK™ and validated by PSIM™ software. The electrical characteristics of the utilized PV module are shown in Table 6. The developed model was prepared and tested in five steps, as follows:

5.1. Step 1: PV Faults Creation and Analysis

For datasets generation, four different faults were generated, namely, LL, LG, PS, and CS. In the LL fault, where unintentional connection between two points in the same string or different strings were created. It was noticed that the inverter's MPPT may shift the operating point to a different position on the I–V curve, causing the fault current amplitude to drop, as shown in Figure 14, which depicts the effect of the LL fault on the behavior of the PV array with one faulty module under different resistance levels. The MPPT will relocate the MPP to another position, and the PV array will appear to be operating normally but with less output power. Moreover, the effect of the LL fault resistance can be seen in Figures 14 and 15, where the behavior of the PV array will be closer to the healthy case with the increase in the fault resistance. On the other hand, the blocking diodes can be optionally utilized in the same PV configuration, as shown in Figure 15. The blocking diodes will cause different voltage peaks on the I-V characteristics curve with the presence of a low fault resistance, which would make it very similar to the I-V curve of the open circuit and partial shading faults.

Table 6. Electrical characteristics of the utilized PV module.

Parameters under STC	**Kyocera solar KC130GT**
Maximum Power (Pmax)	130W
Maximum Power Voltage (V_{mpp})	17.6V
Maximum Power Current (I_{mpp})	7.39A
Open Circuit Voltage (Voc)	21.9V
Short Circuit Current (Isc)	8.02A
Temperature Coefficient of Voc	−8.21×10-2 V/°C
Temperature Coefficient of Isc	3.18×10-3 A/°C

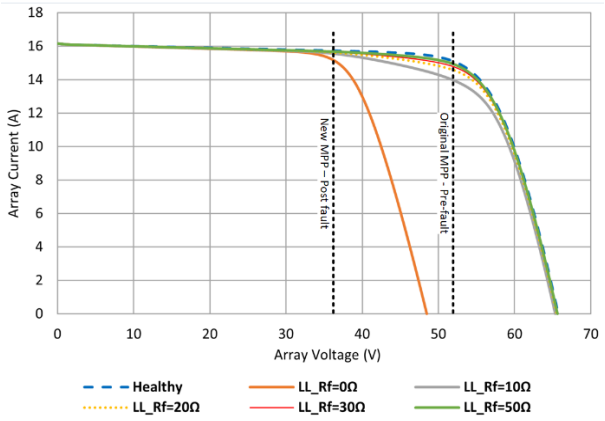


Figure 14. I-V curve of the PV array without blocking diodes under different LL fault resistances in STC (25°C, 1000W/m^2).

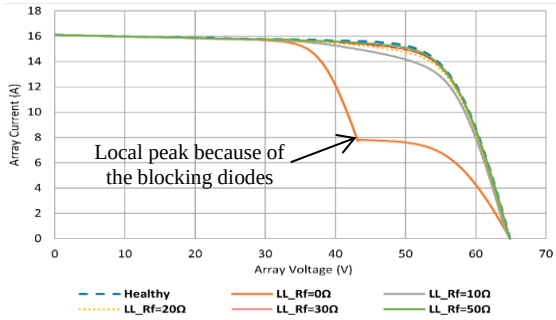


Figure 15. I-V curve of the PV array with blocking diodes under different LL fault resistances in STC (25°C, 1000W/m^2).

In the LG fault, the MPP of the array will be shifted gradually from point "A" to point "B" after applying the fault, where the MPPT will detect the drop in the output power and shift the MPP by reducing the array voltage to optimize the output power, as shown in Figure 16. As can be noticed, in the case of a lower ground fault, there is no back-fed current to the faulty string which will be mismatched with the other strings and no longer be able to operate at its real MPP. On the other hand, the upper ground fault will generate a high fault current at low or no impedance, the faulty module will have a larger current than the other modules because of the fault current and the back-fed current. The inverter's MPPT will shift the MPP from point "A" to point "C" to reduce the power loss and fault current by reducing the array voltage. Figure 17 is showing the effect of the resistance level and the blocking diodes on the behavior of the PV array.

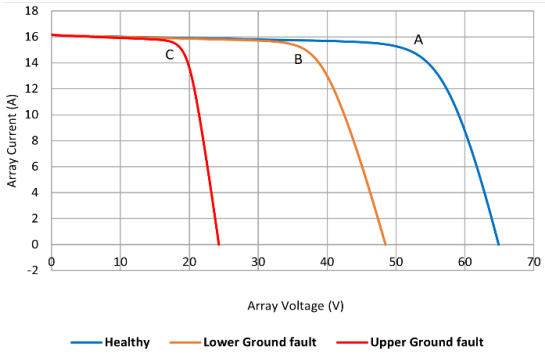


Figure 16. I-V curve of the PV array under lower and upper ground fault.

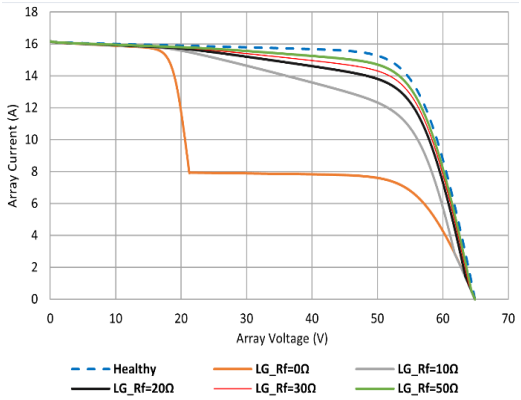


Figure 17. I-V curve of the faulty PV array with upper ground fault at different resistance levels with the presence of the blocking diodes.

In the Partial shading fault, as shown in Figure 18, it can be noticed that the MPP of the PV array has reduced with increasing the partial shading percentage, and different local and global MPP peaks appeared based on the number and percentage of the shaded modules. The inverter's MPPT has pushed the system to continue working on the new global peak to guarantee more output power by reducing the output voltage, which might affect the inverter's lifetime. In addition, it can be observed that at complete shading of the PV array (CS), there will be no local peaks, but the whole output power was reduced.

5.2. Step 2: Data Generation and Preparation

The output current was generated from the PV array as explained in Table 2, and the statistical analysis of the generated current rows under different fault and environmental conditions shows a distinctive sign for each of the studied cases. The results of all cases as shown in will be collated together and separated randomly into two matrices; the first one is the feature matrix, which will be the input of the machine learning classifier for training and testing purposes, while the other matrix will be a new data matrix and unseen by the classifier before, it will be used to validate the developed model as a final step. Figure 19 visualizes the mean value of the statistical features in a line plot for each of the fault types, while Figure 20. Depicts the boxplot of the current rows' statistical features. As it can be noticed from the figure, the healthy class can be distinguished by the maximum, STD, and RMS features. Complete shading, which is the most competitive case compared to partial shading, can be distinguished by the crest factor, which got the lowest value. The LL and LG classes are also competitive, where the maximum, minimum, and skewness for most of the datapoints were very close, but they can be distinguished by the minimum and mean for the LL class, and the kurtosis and crest factor for the LG class. The partial shading classes have a common value for most of the statistical analysis, but it can be noticed that the PS_1M can be distinguished by the form factor, PS_2M can be distinguished by the skewness, and the PS_3M can be distinguished by the kurtosis.

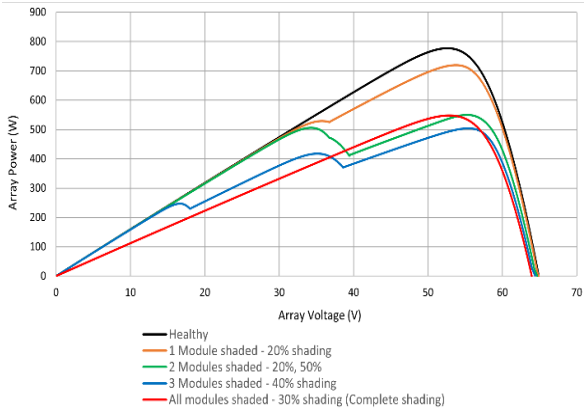


Figure 18. Power curve of the PV array under different shading conditions.

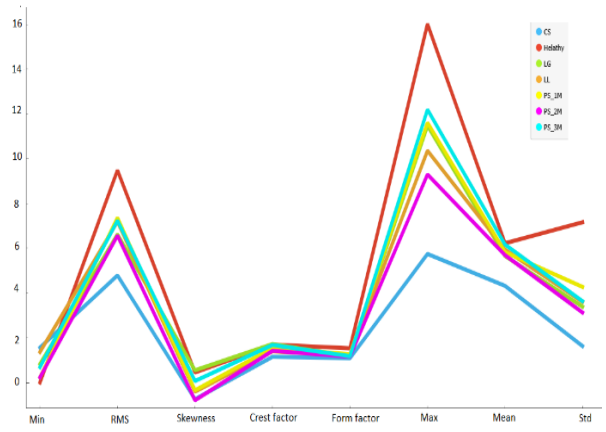


Figure 19. Line graph for the average value of the statistical features.

5.3. Step 3: Classifiers Training

Twelve classifiers with their default hyperparameters were evaluated using four splitting and cross-validation methods as shown in Table 7. Eight classifiers have demonstrated a performance of greater than 85%, while five of them have achieved a performance of greater than 90%.

It can be noticed from the same table that the Gradient Boosting Classifier (GBC) yielded a high degree of accuracy with (91.6%) using K-fold cross validation method, while the Ada-Boost classifier scored (92.25%), whose scores were identical to those obtained with the sK-fold cross validation method. Moreover, the extra trees classifier achieved a score of (91.92%) using the LOOCV cross-validation method. However, The Random Forest Classifier (RFC) showed the highest performance utilizing the holdout splitting method; its accuracy was (93.2%), therefore, the holdout splitting validation method will be used in the optimization process along with the Random Forest classifier, which is the classifier with the best performance among all the other classifiers.

5.4. Step 4: RFC Hyper Parameters Optimization using RSO

In this research, the Random Search Optimization Algorithm was integrated to fine-tune the Random Forest classifier's hyperparameters. This method allowed for the efficient examination of a broad range of hyperparameter configurations, resulting in the optimization of the model's performance while minimizing the computational resources required for an exhaustive search. By randomly sampling the predefined search spaces for the RFC parameters: Criterion, Max features, and N_estimators, using a Gaussian distribution, the accuracy of the classifier was systematically enhanced, where the performance of the optimized model has increased from 93.2% to 94.7%. The estimated duration for the complete prediction of the PV array status was 314ms. Table 8 shows a comparison between the scores of the default RFC and the optimized RFC according to the integration of RSO algorithm. Consequently, the developed fault detection and classification model will use the "Random Forest Classifier" as the top-performing classifier with the highest classification accuracy (94.7%), and prediction time of 314ms.

The developed model used the holdout splitting method, and a random search optimization algorithm. The optimal values for RFC hyperparameters are the "entropy" criterion, 13 estimators, and Max features of $\log_2$.

Table 7. Twelve classifiers training accuracy results under four splitting methods.

Classifier	Accuracy (%)			
	Holdout (Train-test-split)	K-Fold	sK-Fold	LOOCV
DT	88.1	88.2	87.87	87.87
SVM	88.1	76.6	74.74	77.1
NB	59.3	54.9	57.2	57.91
RFC	93.2	91.2	91.92	91.25
KNN	91.5	88.2	88.55	88.2
GB	91.5	91.6	91.24	90.57
MLP	48.3	30.7	46.1	46.8
GP	36.7	34.0	36.37	36.36
ET	90.0	64.3	91.44	91.92
ABC	91.5	92.25	92.25	91.58
QDA	80.0	44.74	88.93	89.56
SGD	74.6	78.01	70.0	68.35

5.5. Step 5: Real Test on Unseen Data

To test the performance of the developed model based on the optimized RFC, new datasets were generated for different faults under random environmental and technical conditions as shown in Table 9. The new datasets included current rows resulting from healthy, faulty modules with LL and LG faults at different fault resistances, in addition to a partially shaded modules with one, two, and three shaded modules, and a completely shaded modules at random shading percentages. As it can be noticed from the confusion matrix depicted in Figure 21, the model was able to perfectly detect the fault in the unseen with a percentage of 100%, i.e., the healthy cases were classified as healthy all the time, and none of the faulty cases or completely shaded cases were classified as a healthy case. On the other hand, the average accuracy of the fault classification of unseen data was 96.64%. The confusion matrix shows that the accuracy of detecting the LG, PS_1M, PS_2M, and PS_3M was 100%, the accuracy of detecting the LL fault was 94.7%, and the CS detection accuracy was 81.8%. The average precision for all faults was 95.5%, the average recall was 94.7%, and the average F1_Score was 94.5%. A sample of the tested data is shown in Table 10.

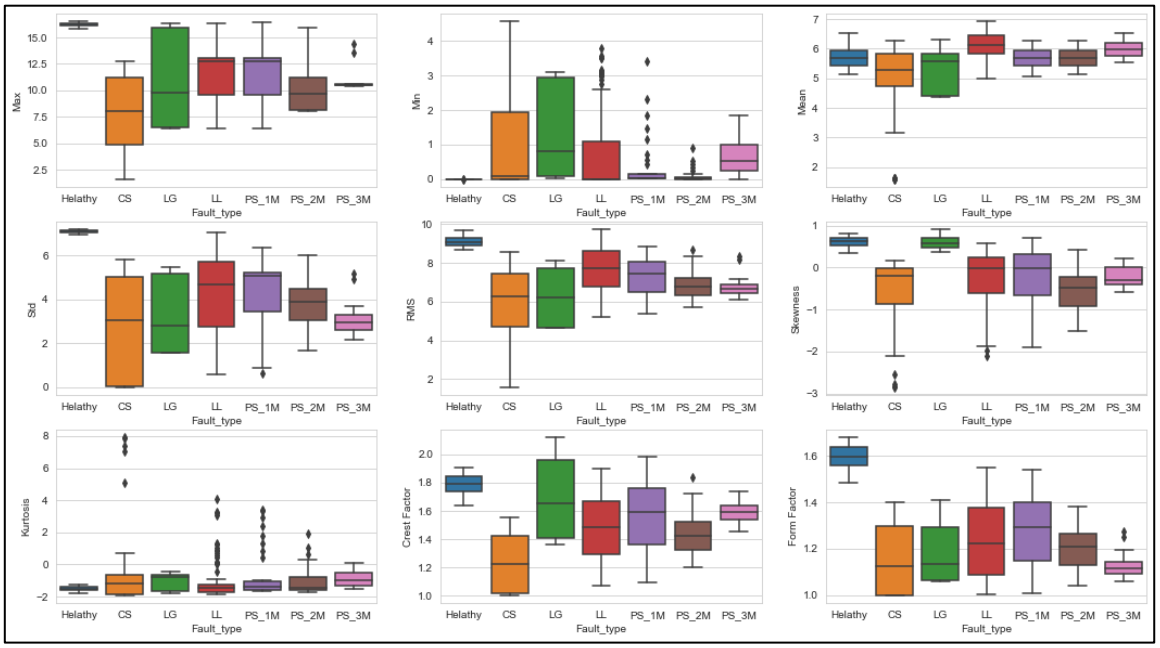


Figure 20. Boxplot of the processed data.

Table 8. Comparison between the accuracy scores of the default RFC and the optimized RFC.

Classifier	Parameters	Accuracy score (%)	Prediction time (*ms*)
Default RFC	Default parameters	93.2	419
Optimized RFC using RSO	Criterion: entropy Max features: log2 N_estimators: 13	94.7	314

Table 9. Sample of testing data gendered under various conditions for real test.

PV Status(No. cases)	Temperature (°C)	Irradiance (W/m^2)	Fault Resistance (Ω)
Healthy (3)	Random between 5 to50	1000	NA
Complete shading (9)	Random between 5 to50	Random between 200 and 800	NA
Partial Shading (18)	Random between 5 to50	850 for all modules except modules 1.1 (random) and 2.2 (random), 2.3 (random)	NA
LL fault (18)	Random between 5 to50	Random between 400 and 1000	Random between 0 and 50
LG fault (9)	Random between 5 to50	Random between 400 and 1000	Random between 0 and 50

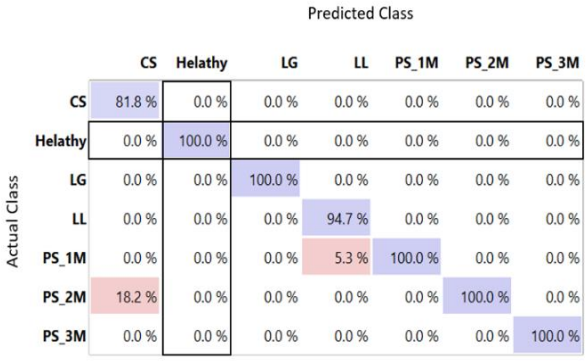


Figure 21. The confusion matrix for the test on unseen data.

6. Conclusions

This paper presented an accurate and fast-responding fault detection and classification model for the PV array using machine learning. The model can classify the PV system's health condition based on its output current. According to the developed model, the PV system will be classified as either healthy, completely shaded, faulty with LL or LG, and partially shaded with 1, 2, or 3 modules. The model's input attributes included nine statistical features derived from the generated PV current. The proposed model has been tested under a variety of environmental and technical conditions. The study compared the performance of twelve ML classifiers. Four different cross-validation techniques were applied to the classifiers in use. The hyperparameters of the top-performing classifier were optimized using RSO algorithm. The chosen classifier was nominated based on four performance metrics: accuracy, precision, F1 score, and recall score. A new set of unseen data generated under random environmental and technical conditions was used to test the performance of the proposed model. Using the optimized RFC, this model has achieved a performance of 100% for fault detection, 94.7% for fault classification of the training data, and 96.6% for fault classification of the testing on new unseen data. The average prediction time was 314 milliseconds. The holdout splitting method was used to improve the performance of the final model. The optimal parameters of the optimized RFC were the “entropy” criterion, 13 estimators, and “max features of log2.”

The proposed method included a few potential limitations and complexities to be considered, including:

1. Signal quality: the accuracy of fault detection and classification depends heavily on the quality of the input signal. If the signal is noisy, incomplete, or contains outliers, it can negatively impact the performance of the statistical analysis and ML classification.
2. Statistical analysis: the effectiveness of the statistical analysis methods utilized to extract features from the existing data plays a key role. While the statistical features considered for this study have proven beneficial, the selection procedure must be continually updated to adapt to varied fault scenarios and environmental variables.
3. Model hyperparameters: each ML classifier has its own hyperparameters that must be tuned for optimal performance.

The model performance is greatly dependent on the selection of hyperparameters. Optimization is a challenging procedure that can considerably influence the model capability to effectively detect and categorize problems.

Table 10. Step 3: a sample of testing unseen data utilizing the developed model.

Case No.	Actual Class of Unseen data	Predicted Class	Case No.	Actual Class of Unseen data	Predicted Class
1	Healthy	Healthy	**30**	LL	LL
2	Healthy	Healthy	**31**	LL	LL
3	Healthy	Healthy	**32**	LL	LL
4	CS	CS	**33**	LL	LL
5	CS	CS	**34**	LL	LL
6	CS	CS	**35**	LL	LL
7	CS	CS	**36**	LL	LL
8	CS	CS	**37**	LL	LL
9	CS	CS	**38**	LL	LL
10	CS	CS	**39**	LL	LL
11	CS	CS	**40**	PS_1M	PS_1M
12	CS	CS	**41**	***PS_1M***	***LL***
13	LG	LG	**42**	PS_1M	PS_1M
14	LG	LG	**43**	PS_1M	PS_1M
15	LG	LG	**44**	PS_1M	PS_1M
16	LG	LG	**45**	PS_1M	PS_1M
17	LG	LG	**46**	PS_2M	PS_2M
18	LG	LG	**47**	***PS_2M***	***CS***
19	LG	LG	**48**	PS_2M	PS_2M
20	LG	LG	**49**	PS_2M	PS_2M
21	LG	LG	**50**	PS_2M	PS_2M
22	LL	LL	**51**	***PS_2M***	***CS***
23	LL	LL	**52**	PS_3M	PS_3M
24	LL	LL	**53**	PS_3M	PS_3M
25	LL	LL	**54**	PS_3M	PS_3M
26	LL	LL	**55**	PS_3M	PS_3M
27	LL	LL	**56**	PS_3M	PS_3M
28	LL	LL	**57**	PS_3M	PS_3M
29	LL	LL			

Acknowledgement

This research work is supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2020/TK0/UKM/02/11) and under Universiti Kebangsaan Malaysia Geran Universiti Penyelidikan (GUP-2022-024).

References

[1] Akram M. and Lotfifard S., "Modeling and Health Monitoring of DC Side of Photovoltaic Array," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 4, pp. 1245-1253, 2015. Doi:10.1109/TSTE.2015.2425791

[2] Alam M., Khan F., Johnson J., and Flicker J., "A Comprehensive Review of Catastrophic Faults in PV Arrays: Types, Detection, and Mitigation Techniques," *IEEE Journal of Photovoltaics*, vol. 5, no. 3, pp. 982-997, 2015. doi:10.1109/JPHOTOV.2015.2397599

[3] Alam M., Khan F., Johnson J., and Flicker J., "A Comprehensive Review of Catastrophic Faults in Mitigation Techniques," *IEEE Journal of Photovoltaics*, vol. 5, no. 3, pp. 982-997, 2015. Doi:10.1109/JPHOTOV.2015.2397599

[4] Ali M., Rabhi A., El-Hajjaji A., and Tina G., "Real Time Fault Detection in Photovoltaic Systems," *Energy Procedia*, vol. 111, pp. 914-923, 2017. doi: 10.1016/j.egypro.2017.03.254

[5] Ayang A., Wamkeue R., Ouhrouche M., and Saad M., "Faults Diagnosis and Monitoring of a Single Diode Photovoltaic Module Based on Estimated Parameters," *in Proceedings of the IEEE Electrical Power and Energy Conference*, Toronto, pp. 1-6, 2018. doi:10.1109/EPEC.2018.8598308

[6] Badr M., Hamad M., Abdel-Khalik A., Hamdy R., and Ahmed S., "Fault Identification of Photovoltaic Array Based on Machine Learning

Classifiers," *IEEE Access*, vol. 9, pp. 159113-159132, 2021. doi: 10.1109/ACCESS.2021.3130889

[7] Baradieh K. and Al-Hamouz Z., "Modelling and Simulation of Line Start Permanent Magnet Synchronous Motors with Broken Bars," *Journal of Electrical and Electronic Systems*, vol. 07, no. 2, pp. 1-7, 2018. doi: 10.4172/2332-0796.1000259

[8] Baradieh K. and Hamouz Z., "ANN Based Broken Rotor Bar Fault Detection in LSPMS Motors," *Journal of Electrical and Electronic Systems*, vol. 7, no. 4, 2018. doi: 10.4172/2332-0796.1000273

[9] Baradieh K., Zainuri M., Kamari N., Yusof Y., Abdullah H., and Zaman M., "Fault Detection and Classification in the Photovoltaic Arrays Using Machine Learning," *in Proceedings of the IEEE Industrial Electronics and Applications Conference*, Penang, pp. 177-182, 2023. doi:10.1109/IEACon57683.2023.10370647

[10] Bergstra J. and Bengio Y., "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281-305, 2012.

[11] Breiman L., "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001. DOI: 10.1023/A:1010950718922

[12] Brooks S., "A Discussion of Random Methods for Seeking Maxima," *Operation Research*, vol. 6, no. 2, pp. 165-302, 1958. https://doi.org/10.1287/opre.6.2.244

[13] Chen S., Yang G., Gao W., and Guo M., "Photovoltaic Fault Diagnosis via Semisupervised Ladder Network with String Voltage and Current Measures," *IEEE Journal of Photovoltaics*, vol. 11, no. 1, pp. 219-231, 2021. doi:10.1109/JPHOTOV.2020.3038335

[14] Chen Z., Han F., Wu L., Yu J., and Cheng S., "Random Forest Based Intelligent Fault Diagnosis for PV Arrays Using Array Voltage and String Currents," *Energy Conversion and Management*, vol. 178, pp. 250-264, 2018. doi:10.1016/j.enconman.2018.10.040

[15] Chen Z., Wu L., Cheng S., Lin P., Wu Y., and Lin W., "Intelligent Fault Diagnosis of Photovoltaic Arrays Based on Optimized Kernel Extreme Learning Machine and I-V Characteristics," *Applied Energy*, vol. 204, pp. 912-931, 2017. doi:10.1016/j.apenergy.2017.05.034

[16] Dhimish M., *Fault Detection and Performance Analysis of Photovoltaic Installations*, University of Huddersfield, 2018. http://eprints.hud.ac.uk/id/eprint/34576/

[17] Diantoro M., Suprayogi T., Hidayat A., Taufiq A., and Fuad A., "Shockley's Equation Fit Analyses for Solar Cell parameters from I-V Curves," *International Journal of Photoenergy*, 2018. doi:10.1155/2018/9214820

[18] Eskandari A., Milimonfared J., and Aghaei M., "Fault Detection and Classification for Photovoltaic Systems Based on Hierarchical Classification and Machine Learning Technique," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 12, pp. 12750-12759, 2021. doi:10.1109/TIE.2020.3047066

[19] González M., Raison B., Bacha S., and Bun L., "Fault Diagnosis in A Grid-Connected Photovoltaic System by Applying a Signal Approach," *in Proceedings of the 37th Annual Conference of the IEEE Industrial Electronics Society*, Melbourne, pp. 1354-1359, 2011. doi:10.1109/IECON.2011.6119505

[20] Haque A., Bharath K., Khan M., Khan I., and Jaffery Z., "Fault Diagnosis of Photovoltaic Modules," *Energy Science and Engineering*, vol. 7, no. 3, pp. 622-644, 2019. doi:10.1002/ese3.255

[21] Harrou F., Saidi A., Sun Y., and Khadraoui S., "Monitoring of Photovoltaic Systems Using Improved Kernel-Based Learning Schemes," *IEEE Journal of Photovoltaics*, vol. 11, no. 3, pp. 806-818, 2021. doi:10.1109/JPHOTOV.2021.3057169

[22] Hashunao S. and Mehta R., "Fault Analysis of Solar Photovoltaic System," *in Proceedings of the 5th International Conference on Renewable Energies for Developing Countries*, Marrakech, pp. 1-6, 2020. doi: 10.1109/REDEC49234.2020.9163847

[23] Huang J., Wai R., and Gao W., "Newly-Designed Fault Diagnostic Method for Solar Photovoltaic Generation System Based on IV-Curve Measurement," *IEEE Access*, vol. 7, pp. 70919-70932, 2019. doi:10.1109/ACCESS.2019.2919337.

[24] IEA, *Renewables 2019-Analysis and Forecast to 2024*, International Energy Agency, 2019. https://doi.org/10.1787/b3911209-en

[25] IEA, *Renewables 2021*, International Energy Agency, 2021. https://doi.org/10.1787/6dcd2e15-en

[26] Jain P., Xu J., Panda S., Poon J., and Spanos C., "Fault Diagnosis via PV Panel-Integrated Power Electronics," *in Proceedings of the IEEE 17th Work, Control and Modeling for Power Electronics*, Trondheim, pp. 1-6, 2016. doi: 10.1109/COMPEL.2016.7556716.

[27] Kumar B., Ilango G., Reddy M., and Chilakapati N., "Online Fault Detection and Diagnosis in Photovoltaic Systems Using Wavelet Packets," *IEEE Journal of Photovoltaics*, vol. 8, no. 1, pp. 257-265, 2018. doi:10.1109/JPHOTOV.2017.2770159

[28] Lu X., Lin P., Cheng S., Lin Y., and Chen Z., "Fault Diagnosis for Photovoltaic Array Based on Convolutional Neural Network and Electrical Time Series Graph," *Energy Conversion and Management*, vol. 196, pp. 950-965, 2019.

https://doi.org/10.1016/j.enconman.2019.06.062

[29] Maree M., Eleyat M., and Mesqali E., "Optimizing Machine Learning-based Sentiment Analysis Accuracy in Bilingual Sentences via Preprocessing Techniques," *The International Arab Journal of Information Technology*, vol. 21, no. 02, pp. 257-270, 2024. doi: 10.34028/iajit/21/2/8

[30] Markoulidakis I., Kopsiaftis G., Rallis I., and Georgoulas I., "Multi-Class Confusion Matrix Reduction Method and its Application on Net Promoter Score Classification Problem," *in Proceedings of the 14th Pervasive Technologies Related to Assistive Environments Conference*, New York, pp. 412-419, 2021. doi:10.1145/3453892.3461323

[31] Minh N., Mai D., and Nguyen H., "PV Array Fault Classification based on Machine Learning," *in Proceedings of the 11th International Conference on Control, Automation and Information Sciences*, Hanoi, pp. 322-326, 2022. doi:10.1109/ICCAIS56082.2022.9990272

[32] Nazarudin N., Ariffin N., and Maskat R., "Leveraging on Synthetic Data Generation Techniques to Train Machine Learning Models for Tenaga Nasional Berhad Stock Price Movement Prediction," *The International Arab Journal of Information Technology*, vol. 21, no. 3, pp. 483-494, 2024. doi:10.34028/iajit/21/3/11

[33] Nie S., Chen Y., Pei X., Wang H., and Kang Y., "Fault Diagnosis of a Single-Phase Inverter Using the Magnetic Field Waveform Near the Output Inductor," *in Proceedings of the 26th Annual IEEE Applied Power Electronics Conference and Exposition*, Texas, pp. 1648-1655, 2011. doi:10.1109/APEC.2011.5744816

[34] Paing M. and Choomchuay S., "Improved Random Forest Classifier for Imbalanced Classification of Lung Nodules," *in Proceedings of the 4th International Conference on Engineering, Applied Sciences, and Technology*, Tokyo, pp. 1-4, 2018. doi:10.1109/ICEAST.2018.8434402

[35] Qi C. and Ming Z., "Photovoltaic Module Simulink Model for a Stand-Alone PV System," *Physics Procedia*, vol. 24, pp. 94-100, 2012. doi:10.1016/j.phpro.2012.02.015

[36] Quaschning V. and Hanitsch R., "Numerical Simulation of Current-Voltage Characteristics of Photovoltaic Systems with Shaded Solar Cells," *Solar Energy*, vol. 56, no. 6, pp. 513-520, 1996. doi:10.1016/0038-092X(96)00006-0

[37] Rajak P., Bharadwaj S., and Gawre S., "PV Module Fault Detection and Diagnosis," *International Research Journal of Engineering and Technology*, vol. 35, no. 05, pp. 3809-3813, 2018. https://doi.org/10.1007/s00521-023-09041-7

[38] Rakesh N., Banerjee S., Subramaniam S., and Babu N., "A Simplified Method for Fault Detection and Identification of Mismatch Modules and Strings in a Grid-Tied Solar Photovoltaic System," *International Journal of Emerging Electric Power Systems*, vol. 21, no. 4, pp. 1, 2020. doi:10.1515/ijeeps-2020-0001

[39] Román E., Alonso R., Ibañez P., Elorduizapatarietxe S., and Goitia D., "Intelligent PV Module for Grid-Connected PV Systems," *IEEE Transactions on Industrial Electronics*, vol. 53, no. 4, pp. 1066-1073, 2006. doi:10.1109/TIE.2006.878327

[40] Saravanan C. and Srinivasan K., "Optimal Extraction of Photovoltaic Model Parameters Using Gravitational Search Algorithm Approach," *Circuits and Systems*, vol. 7, no. 11, pp. 3849-3861, 2016. doi: 10.4236/cs.2016.711321

[41] Li W., "Retracted: Optimization and Application of Random Forest Algorithm for Applied Mathematics Specialty," *Security and Communication Networks*, vol. 2023. pp. 1-1, 2023. doi: 10.1155/2023/9818912

[42] Shahriari B., Swersky K., Wang Z., Adams R., and Freitas N., "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *in Proceedings of the IEEE*, vol. 104, no. 1, pp. 148-175, 2016. doi:10.1109/JPROC.2015.2494218

[43] Taghezouit B., Harrou F., Sun Y., Arab A., and Larbes C., "Multivariate Statistical Monitoring of Photovoltaic Plant Operation," *Energy Conversion and Management*, vol. 205, pp. 112317, 2020. doi:10.1016/j.enconman.2019.112317

[44] Taghezouit B., Harrou F., Sun Y., Arab A., and Larbes C., "A Simple and Effective Detection Strategy Using Double Exponential Scheme for Photovoltaic Systems Monitoring," *Solar Energy*, vol. 214, pp. 337-354, 2021. doi:10.1016/j.solener.2020.10.086

[45] Talbi M., Mensia N., and Ezzaouia H., "Modeling of a PV Panel and Application of Maximum Power Point Tracking Command Based," *The International Arab Journal of Information Technology*, vol. 18, no. 4, pp. 76-85, 2021. doi:10.34028/18/4/9

[46] Tharwat A., "Classification Assessment Methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168-192, 2018. doi:10.1016/j.aci.2018.08.003

[47] Tyagi S., Dhingra B., and Tomar A., "Condition Monitoring and Fault Detection in Photovoltaic Modules Using Machine Learning," *in Proceedings of the* 1st *International Conference on Sustainable Technology for Power and Energy Systems*, Srinagar, pp. 1-6, 2023. doi:10.1109/stpes54845.2022.10006619

[48] Wang G., Youn C., and Stankovic A., "DC-Side

High Impedance Ground Fault Detection for Transformerless Single-Phase PV Systems," *in Proceedings of the North American Power Symposium*, Charlotte, pp. 1-6, 2015. doi:10.1109/NAPS.2015.7335209

[49] Watt J., and Borhani R., *Machine Learning Refined: Foundations, Algorithms, and Applications*, Cambridge University Press, 2020. https://doi.org/10.1017/9781108690935

[50] Yang B., Di X., and Han T., "Random Forests Classifier for Machine Fault Diagnosis," *Journal of Mechanical Science and Technology*, vol. 22, no. 9, pp. 1716-1725, 2008. doi:10.1007/s12206-008-0603-6

[51] Yi Z. and Etemadi A., "Fault Detection for Photovoltaic Systems Based on Multi-Resolution Signal Decomposition and Fuzzy Inference Systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 3, pp. 1274-1283, 2017. doi:10.1109/TSG.2016.2587244

[52] Yi Z. and Etemadi A., "Line-To-Line Fault Detection for Photovoltaic Arrays Based on Multi-Resolution Signal Decomposition and Two-Stage Support Vector Machine," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 11, pp. 8546-8556, 2017. doi: 10.1109/TIE.2017.2703681

[53] Zabinsky Z., *Random Search Algorithms*, Wiley Encyclopedia of Operations Research and Management Science, 2011. doi:10.1002/9780470400531.eorms0704

[54] Zainuri M., Radzi M., Soh A., and Rahim N., "Development of Adaptive Perturb and Observe-Fuzzy Control Maximum Power Point Tracking for Photovoltaic Boost DC-DC Converter," *IET Renewable Power Generation*, vol. 8, no. 2, pp. 183-194, 2014. doi:10.1049/iet-rpg.2012.0362

[55] Zainuri M., Radzi M., Soh A., Mariun N., and Rahim N., "DC-link Capacitor Voltage Control for Single-Phase Shunt Active Power Filter with Step Size Error Cancellation in Self-Charging Algorithm," *IET Power Electronics*, vol. 9, no. 2, pp. 323-335, 2016. doi: 10.1049/iet-pel.2015.0188

[56] Zbib B. and Al-Sheikh H., "Fault Detection and Diagnosis of Photovoltaic Systems through I-V Curve Analysis," *in Proceedings of the 2nd International Conference on Electrical, Communication, and Computer Engineering*, Istanbul, pp. 12-13, 2020. doi:10.1109/ICECCE49384.2020.9179390

[57] Zhang M., Li H., Pan S., Lyu J., and Ling S., "Convolutional Neural Networks-Based Lung Nodule Classification: A Surrogate-Assisted Evolutionary Algorithm for Hyperparameter Optimization," *IEEE Transactions on Evolutionary Computation*, vol. 25, no. 5, pp. 869-882, 2021. doi:10.1109/TEVC.2021.3060833

[58] Zhao Y., Ball R., Mosesian J., Palma J., and Lehman B., "Graph-Based Semi-Supervised Learning for Fault Detection and Classification in Solar Photovoltaic Arrays," *IEEE Transactions on Power Electronics*, vol. 30, no. 5, pp. 2848-2858, 2015. doi:10.1109/TPEL.2014.2364203

[59] Zhao Y., *Fault Analysis in Solar Photovoltaic Arrays*, Northeastern University, 2010.

[60] Zhao Y., *Fault Detection in Protection Photovoltaic Solar Arrays*, Northeastern University, 2015. https://repository.library.northeastern.edu/downloads/neu:m039kr12f?datastream_id=content

[61] Zhao Y., Yang L., Lehman B., Palma J., and Mosesian J., "Decision Tree-based Fault Detection and Classification in Solar Photovoltaic Arrays," *in Proceedings of the 27th Annual IEEE Applied Power Electronics Conference and Exposition*, Orlando, pp. 93-99, 2012. doi:10.1109/APEC.2012.6165803

Khaled Baradieh was born in Hebron, Palestine in 1990. He received the B.S. degree in electrical and electronics engineering from Palestine Polytechnic University, Hebron, Palestine, in 2013 and the M.S. degree in electrical engineering from King Fahad University of Petroleum and Minerals, Dammam, KSA, 2017. He is currently pursuing the Ph.D. degree in electrical engineering at Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia. From 2017 to 2018, he was a researcher in the Research institute with King Fahad University of Petroleum and Minerals, Dhahran, KSA. Since 2018 he has been an instructor in electrical engineering department with American College of the Middle East, Kuwait. His research interests include Faults Detection and Classification in Electrical and Electronic Devices, Renewable Energy, And Machine Learning.

Mohd Zainuri was born in Kuala Lumpur, Malaysia in 1988. He received the Bachelor degree in Electrical and Electronic Engineering in 2011, MSc (Electrical Power Engineering) in 2013 and Doctorate Ph.D (Electrical Power Engineering) in 2017 from Universiti Putra Malaysia. He is currently Senior Lecturer at Department of Electrical, Electronic and Systems Engiuneering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Selangor, Malaysia. His areas of research interests are Power Electronic, Power Quality, Microgrid, Renewable Energy System and Electrical Vehicle.

Mohamed Kamari received Bachelor in Electrical and Electronic Engineering from Meiji Univeristy, Japan, M.Sc. in Electrical and Electronic Engineering from Ehime University, Japan and Ph.D. in Electrical Engineering from Universiti Teknologi Mara, Malaysia in 2000, 2004 and 2016, respectively. He has been an academician for over 10 years at several universities before joining Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia in 2017. His areas of research interests are Power Engineering, Energy Management and Artifcial Intelligence.

Yushaizad Yusof received the B.Eng. degree in electrical and electronic engineering from Kagoshima University, Japan, in 1999, the M.Eng. Degree in electrical engineering from Universiti Teknology Malaysia, Johor, Malaysia in 2002, and Ph.D. degree in power electronics from Universiti Malaya, Kuala Lumpur, Malaysia in 2019. He is currently a Senior Lecturer with the Department of Electrical, Electronic and System Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Selangor, Malaysia. His research interests include Power Electronics, Motor Drives, and Control Engineering.

Huda Abdullah is a professor in Department of Electrical, Electronic and System Engineering, Faculty of Engineering of and Built Environment, Universiti Kebangsaan Malaysia. My research strength lies in the field of Nano Functional Material with a specialization in material for energy and sensor application. She has published more than 241 research papers in journals (ISI/SCOPUS), and has obtained more than 18 research grants as a project leader

Mohd Zaman is currently a senior lecturer in the Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia. He obtained a Bachelor of Engineering (Electrical–Electronic) from Universiti Teknologi Malaysia in 2001 and holds MSc and Ph.D. degrees from Universiti Kebangsaan Malaysia in 2012 and 2019, respectively. His research interests are Robotics, Control System and Artificial Intelligence.

Mohd Zulkifley is an associate professor at the Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia. He received his B. Eng. (Mechatronics) from the International Islamic University Malaysia (2008), Ph.D. (Electrical and Electronic Engineering) from The University of Melbourne (2012), and did his postdoctoral at the University of Oxford (Deep Learning Localization). His current research interests Are Deep Learning Applications in Image and Video Analysis.