AIMS Energy, 12(4): 727–750.
DOI: 10.3934/energy.2024034
Received: 26 February 2024
Revised: 31 May 2024
Accepted: 06 June 2024
Published: 02 July 2024

https://www.aimspress.com/journal/energy

Research article

Diagnosing faults in a photovoltaic system using the Extra Trees ensemble algorithm

Guy M. Toche Tchio[1,*], Joseph Kenfack[3], Joseph Voufo[3], Yves Abessolo Mindzie[3], Blaise Fouedjou Njoya[3] and Sanoussi S. Ouro-Djobo[1,2,*]

[1] Regional Center of Excellence for Electricity Management (CERME), University of Lomé, Togo
[2] Solar Energy Laboratory, Department of Physics, Faculty of Sciences, University of Lomé, Togo
[3] Laboratory on Small hydropower and hybrid systems, National Advanced School of Engineering of Yaoundé (NASEY), University of Yaoundé 1, Cameroon.

* **Correspondence:** Email: mtoche6@gmail.com, sourodjobo@univ-lome.tg; Tel: (+228) 93946807.

Abstract: The application of machine learning techniques for monitoring and diagnosing faults in photovoltaic (PV) systems has been shown to enhance the reliability of PV power generation. This research introduced a novel machine learning classifier for fault diagnosis in PV systems, utilizing an ensemble algorithm known as extra trees (ETC). The study initially proposed a system with two PV modules and developed a low-cost Arduino-based data logger to gather data from the PV system in free-fault and faulty conditions. Subsequently, the study evaluated six other advanced classifiers for fault diagnosis in PV systems, namely logistic regression (LR), k-nearest neighbor (kNN), support vector machine (SVM), decision tree (DT), AdaBoost, and random forest (RF) models using the collected data from the proposed PV system. The assessment of the various models' performance indicated that the extra trees model exhibits superior classification capabilities for partial shading (PS), open circuit (OCF), partial shading with bypass diode disconnected (PSBD), and combined partial shading with bypass diode disconnected plus open circuit (PSBDOC) faults. The results demonstrated that the new ETC classifier achieves an accuracy of 92%, surpassing the 91%, 87%, 7%, and 59% accuracy of the RF, DT, kNN, and LR classifiers, respectively. This highlights the effectiveness of the extra trees model in enhancing fault detection and classification by distinguishing between open circuits and twin faults. Consequently, these results can be utilized to develop advanced diagnostic tools for photovoltaic systems, thereby improving the reliability of solar technology and accelerating the rate of installation.

Keywords: photovoltaics; diagnosis; extra trees; fault; data acquisition system

Abbreviations: PV: photovoltaics; LR: logistics regression; KNN: k-nearest neighbor; RF: random forest; DT: decision tree; PS: partial shading; NF: no-fault; PSBD: partial shading with bypass diode disconnected; PSBDOC: partial shading with bypass diodes disconnected plus open circuit; OCF: open circuit fault; ETC: extra trees classifier

1. Introduction

1.1. Motivation

In recent decades, photovoltaic installations have become a reliable and popular way to produce renewable electricity. The International Energy Agency (IEA, 2023) predicts that the installed capacity of photovoltaic systems could reach 167.3 GW in 2024 [1]. This increased development of photovoltaic installations is driven by factors such as energy independence, government support policies, reduced production costs, reliability, and environmental friendliness of photovoltaic installations [2]. However, due to their exposure to harsh environmental conditions, photovoltaic systems are prone to anomalies and malfunctions that can reduce productivity or even lead to complete system shutdown if not detected in time [3]. For example, a study conducted by Firth et al. demonstrated that faults lead to an annual decline of 18.9% in the power generated by PV systems [4]. In light of these findings, it is evident that effectively addressing the growing energy demands of the modern world, while sustaining the rate of photovoltaic installations and their dependability, is a crucial concern for the photovoltaic community [5]. As a result, current research is focused on detecting and eliminating faults to maintain productivity, reduce maintenance costs, and extend the lifespan of photovoltaic (PV) systems [6]. A number of approaches to detecting and diagnosing faults in PV systems have become increasingly popular in recent decades, as shown in the literature [7,8].

1.2. literature review

Considering the extant literature, several methods for detecting and classifying faults in photovoltaic (PV) installations have been proposed. These methods vary in terms of their speed of execution, computational complexity, and effectiveness in identifying different types of faults. These fault diagnosis methods are typically classified into two main categories [7]. The first approach involves using diagnostic tools such as infrared cameras and transducers to identify defective panels. However, this technique requires additional investment in detection tools, which can reduce the profitability of the PV system [9]. The second technique utilizes operational data from solar panels and is categorized into three groups: methods based on the reference model [10] and methods for the statistical analysis of data from solar panels [11], which necessitate an understanding of the distribution properties of the analyzed objects, which can often be challenging to obtain [12]. The third category is based on artificial intelligence techniques. This method addresses the complex processes of previous methods through its straightforward implementation based on training models derived from a collection of data [11,13]. An analysis of these different diagnostic methods reveals that intelligent methods have been increasingly employed in recent decades due to their flexibility and reliability in

detecting and classifying faults. For instance, Pula et al. employed a novel methodology based on artificial neural networks (ANNs) to identify and categorize eight distinct fault types in a photovoltaic (PV) field with remarkable precision [14]. Similarly, the ANN algorithm is employed to identify short-circuit and bypass diode, open-circuit faults within a PV system. The experimental results of the ANN model developed yielded an average accuracy of 96.4% and a sensitivity of 92.6% [15]. In addition, Kumar et al. proposed a novel application of neural networks for the prediction of the power produced in a photovoltaic field with an installed capacity of approximately 2.7 kW in mountainous areas. The results obtained using the two parameters of temperature and irradiation demonstrate that this approach provides a more accurate prediction [16]. Madeti et al. [17] proposed the k-nearest neighbors (kNN) algorithm to detect open circuit, line-to-line, and partial shading faults in a photovoltaic system. The results yielded an accuracy of 98.7%. Furthermore, an SVM classifier optimized using the grid search method and k-fold cross-validation was employed to identify short-circuit, line-to-line, open-circuit, and irradiation faults in the PV system, with an accuracy of 97% [18,19]. Harrou et al. [20] proposed an unsupervised monitoring procedure to detect open circuit, short circuit, and partial shading faults using one-class SVM. Furthermore, some authors, such as Harrou et al. and Mellit et al., have proposed an intelligent approach based on ensemble learning algorithms to detect and classify faults in a PV system [21,22]. Benkercha et al. [23] developed a C4.5 decision tree model for diagnosing open circuit and line-to-line faults in a grid-connected photovoltaic (PV) array. The proposed model demonstrated 99.8% accuracy in detecting and classifying the simulated PV array. These results were obtained using a 10-fold cross-validation approach, with a 99% classification accuracy. Additionally, Gong et al. proposed random forest (RF) models for detecting open circuit, line-to-line, and partial shading faults in a PV system. A comparative study between the kNN, SVM, DT, and RF models showed that the RF model exhibited superior performance [24]. Another ensemble learning method based on the AdaBoost model was able to detect and classify short-circuit, open-circuit, and degradation faults in a PV array with an accuracy of 97.84%. The same AdaBoost model proposed by Ghoneim et al. [25] was able to detect string, string-to-string, and string-to-ground faults in a grid-connected PV array with an accuracy of 95%. In various research studies, it is clear that machine learning is being used in diagnosing PV systems and in other engineering fields. For example, Kumar et al. proposed the ANFIS, SVM, and Gaussian algorithms to evaluate the flexural strength of concrete containing marbles [26]. They also used neural networks, M5P, and random forest to predict water quality aeration. The results obtained showed the effectiveness of the ANN with a correlation coefficient of 0.9823, an MAE of 0.0098, and an RMSE of 0.0123 [27]. This paper introduces a novel learning ensemble classifier, extra trees, for the fault diagnosis of photovoltaic systems. To the best of the authors' knowledge, this model has not yet been adopted for photovoltaic system diagnostics. Additionally, Toche et al. demonstrated that the proposed model has been successfully applied in various fields such as medicine [28], economics [29], civil engineering [30], and telecommunications [31], outperforming models like kNN, DT, SVM, RF, and AdaBoost [32]. The main reasons for choosing the extra trees classifier are (1) the widespread use of artificial intelligence-based classifiers in recent decades, especially in machine learning and deep learning applications, making them faster and more efficient than traditional methods, and (2) the ability of ensemble algorithms like extra trees to handle noisy data and variance.

1.3. Contribution

The aim of this paper is to contribute to the advancement of fault detection and classification methodologies to enhance the reliability and safety of photovoltaic (PV) installations.

This study makes a significant contribution to the field of fault diagnosis in photovoltaic (PV) arrays.

- It introduces a new classifier called extra trees or extremely randomized trees (ETC), which has not been used in this context before.
- The experiment involves a simultaneous or accumulated partial shading fault with a disconnected bypass diode and a PV module open circuit (PSBDOC).
- The model's performance is evaluated using a number of metrics, including accuracy, precision, recall, F1 score, and calculation time.
- This evaluation is conducted in order to facilitate a comparison with the other models selected.

The proposed model is robust to noise and significantly reduces bias and variance errors compared with the other models selected [33]. Section 2 of this paper presents the proposed PV system, including an overview of its various components and the faults studied. Section 3 outlines the methodology for diagnosing the PV system using the proposed model, with the results and discussions reserved for Section 4. Finally, the conclusion is presented in Section 5.

2. Description of the PV system and its defects

2.1. Description of the PV system

The proposed photovoltaic system (Figure 1) comprises two parallel-mounted photovoltaic panels, a PWM charge controller, and a 12 V/9 Ah battery. The system operates at 12 V and includes two 12 V/50 W PV modules, the technical specifications of which are provided in Table 1. The PV modules convert sunlight into electricity using the photoelectric effect. The controller regulates the charge and discharge level of the battery and powers the load. Excess energy is stored in the battery for later use when there is no sunlight.

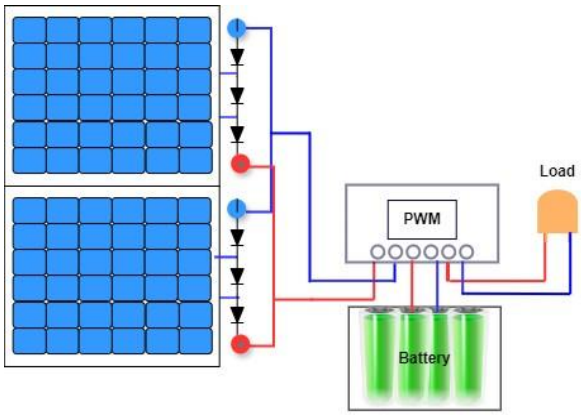


Figure 1. PV system wiring diagram.

The proposed PV module is a BLD SOLAR brand monocrystalline silicon module whose technical specifications are given in Table 1.

Table 1. Electrical parameters of a BLD SOLAR panel in STC conditions.

Electrical parameters	Values
Nominal power (P_{mp})	50 W
Open circuit voltage (V_{OC})	22.1 V
Short circuit current (I_{SC})	2.81 A
Voltage at maximum power (V_{mp})	18.1 V
Current at maximum power (I_{mp})	2.76 A
Number of cells	36

The fault detection and classification of the PV system described in this work was based on the electrical and meteorological parameters collected using an Arduino acquisition system. Four types of faults were created and physically tested, as described in the following section.

2.2. Description of defects

During operation, photovoltaic installations may experience various types of faults, including environmental factors such as shading, physical degradation, and electrical issues like open circuits. Multiple faults can occur simultaneously on a PV field. This study focuses on four specific types of faults that were intentionally created on one of two photovoltaic modules.

- Partial shading defects (PS)

Shading is the primary cause of PV system failure. It can be either permanent or temporary, depending on the source of obstruction, such as adjacent buildings, passing clouds, trees, or other objects [34]. A shading defect is considered partial when only part of the cells or modules are exposed to solar radiation, resulting in a significant decrease in current output. It is important to note that shading should be avoided as much as possible to ensure optimal performance of the PV system [35]. Furthermore, shading is considered uniform or total when all cells or modules receive uniform but low-intensity radiation. This type of shading leads to a constant reduction in the output current and voltage of individual cells in a string [36]. To create a partial shading defect in the proposed PV system, a panel was obscured for 3 h, resulting in half of the PV panel being affected. Figure 2 illustrates the various occultation processes of the module and cell that occurred throughout the day on October 20, 2023, from 8:30 a.m. to 5:30 p.m. The collected data was stored on a SD card.

(a) (b)

Figure 2. (a) Partial shading of a PV module; (b) Total shading of a PV module.

- Bypass diode fault with partial shading (PSBD)

The bypass diode is a protective device that limits reverse voltages in partially shaded cells. Photovoltaic (PV) modules typically integrate two to four bypass diodes in the junction box, with each connected in parallel to a group of cells in a solar panel [37]. The two main faults that can damage the bypass diode are short circuits and open circuits [38]. The short circuit fault of the bypass diode causes a significant decrease in power output as the voltage chain is absent. On the other hand, in the case of an open circuit, the diode breaks down completely and prevents power from passing through [37]. The origin of the fault may be linked to the non-functioning of the diode, reversed diodes during assembly, poor connection of diodes and/or disconnection, or corrosion of the junction boxes. Such a fault can cause hot spots, electric arcing, and the risk of fire if the diode is in an open circuit [39]. An open circuit bypass diode fault is only noticeable in the event of at least 5% shading [38]. This fault was obtained by disconnecting one of the two bypass diodes F_1 from the shaded module; the data was collected on October 22, 2023, from 8:52 a.m. to 4:50 p.m. on a storage medium. Figure 3 is an illustration of the partial shading fault with the bypass diode disconnected.

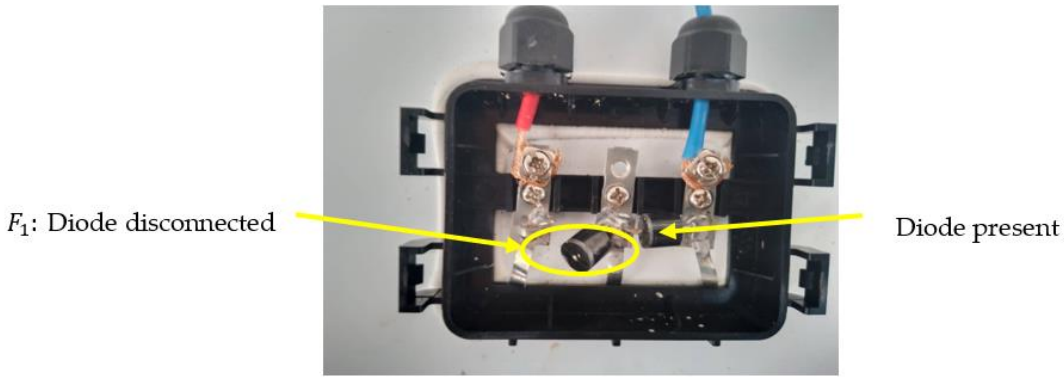


Figure 3. Bypass diode disconnected.

- Open-circuit fault (OCF)

The open-circuit fault (F_2) is typically caused by accidental breakage of connection wires between cells or strings of PV modules, faulty diodes, and deterioration of connection cables [40]. Additionally, an open-circuit fault results in a drastic drop in short-circuit current [41]. However, this type of fault can cause more damage than a short-circuit fault due to the increased current flow. A consequence of this fault is the line-to-line fault [42]. The open-circuit fault in this work was created by disconnecting the connecting wire between the positive terminal of module 1 and the positive terminal of module 2. Data from various sensors were automatically recorded and stored on the SD card on October 21, 2023, from 8:06 a.m. to 4:34 p.m. Figure 4 illustrates the open-circuit fault (F_2) on a PV system module.

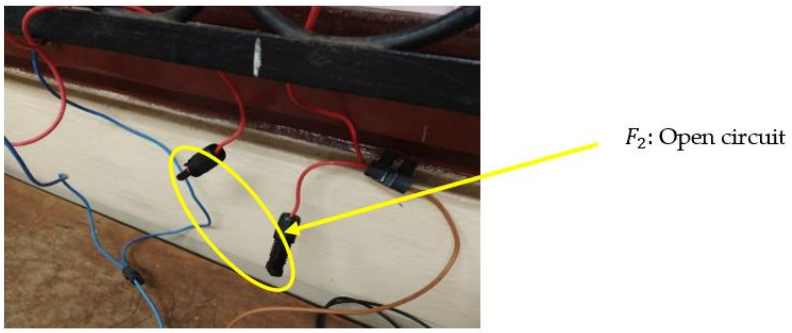


Figure 4. Experimental test of open-circuit fault of PV system.

- Partial shading with bypass diode disconnected and open circuit (PSBDOC)

A twin fault is a combination of two or more faults occurring simultaneously in a circuit. This type of fault results in a significant drop in the current and heating of the module if not addressed promptly. The impact of this fault is similar to that of an open-circuit fault in the case of two panels in parallel over a relatively short period. To avoid damaging the panel, the bypass diode on the partially shaded module was disconnected, and the open-circuit fault was introduced. The data was automatically collected on October 23, 2023, from 8:18 a.m. to 12:00 p.m. Figure 5 illustrates the twin fault F_3.

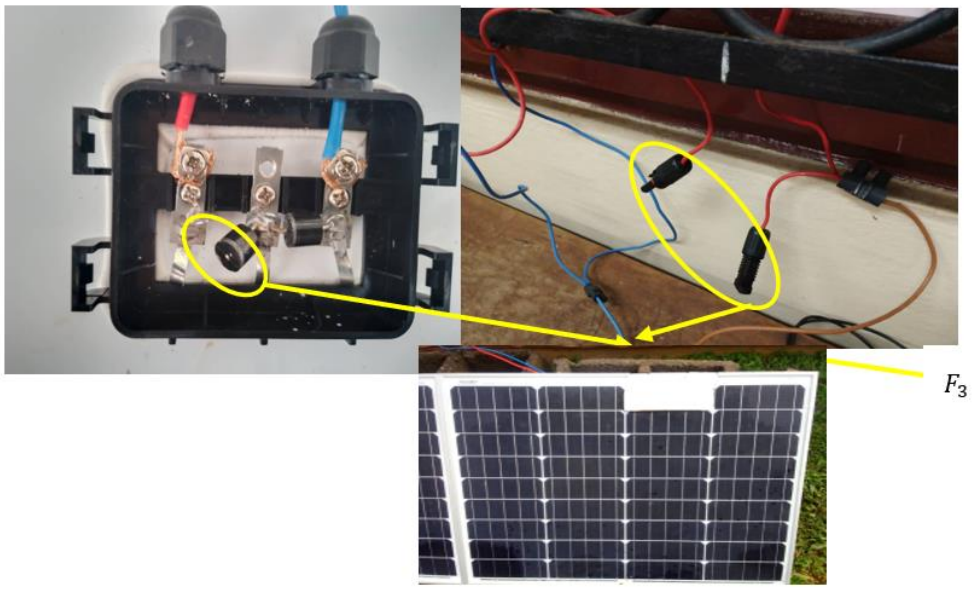


Figure 5. Experimental testing of partial shading faults with bypass diode disconnected and open circuit.

After reviewing the proposed PV system and identifying various faults, we used a low-cost acquisition system based on the Arduino IDE to collect data from the sensors detecting these faults. The collected data was used to train the selected models. Figure 6 shows the data collection process

for training the models. The next section will provide a detailed description of the fault diagnosis procedure for the proposed PV system.

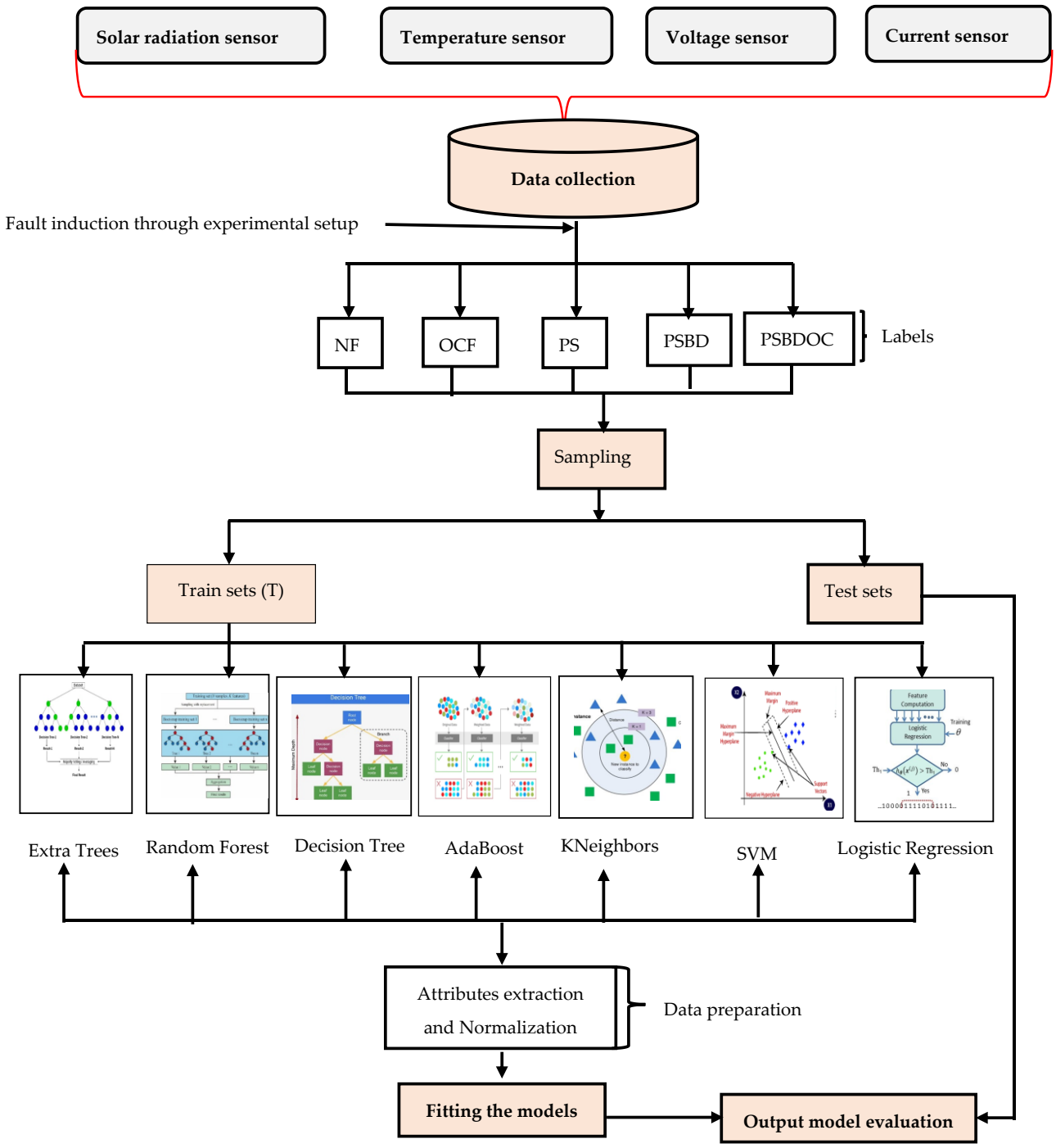


Figure 6. Framework of applied model for PV fault diagnosis.

3. PV system diagnostic methodology

The process of diagnosing faults in a photovoltaic system using a digital approach involves two important phases: acquiring and processing raw data and using a model for fault prediction [43]. The fault diagnosis process is given by the diagram in Figure 7.

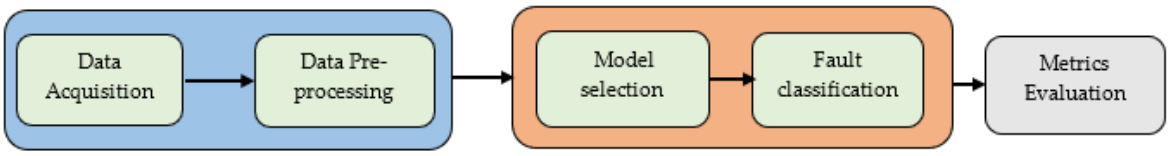


Figure 7. Diagram of the proposed fault diagnosis program.

3.1. Acquisition of data

A test bench for data acquisition using dedicated sensors is implemented in the data acquisition phase. To accommodate budgetary constraints, a low-cost acquisition system based on Arduino was developed to collect data from both healthy and faulty panels using sensors. Current, voltage, irradiation, and temperature data are used for PV fault diagnosis. Figure 8 shows the simplified diagram of the developed acquisition device.

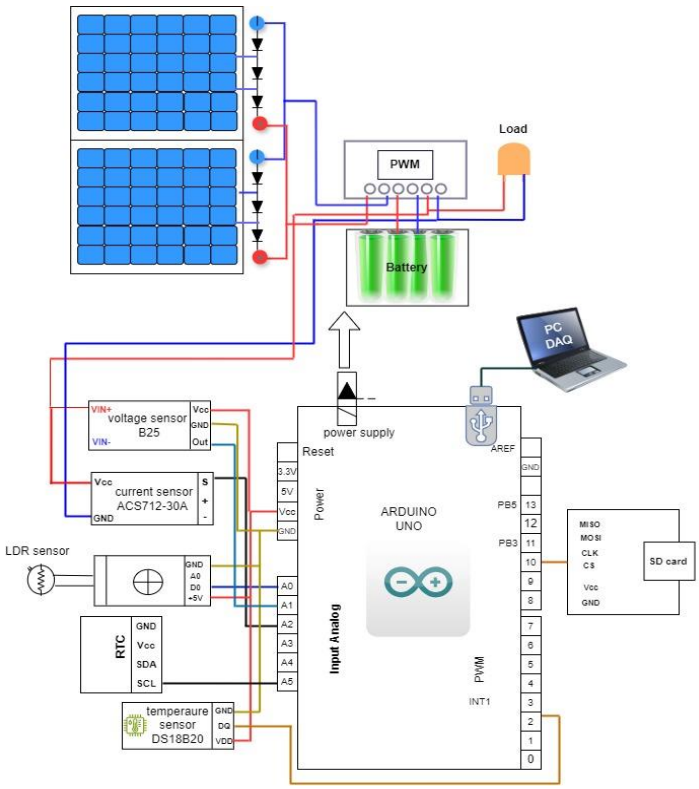


Figure 8. Structure of the proposed acquisition system.

The data was collected over a period of five days at one-minute intervals, corresponding successively to no fault (NF), partial shading fault (PS), open circuit fault (OCF), partial shading with bypass diode disconnected (PSBD), and bypass diode disconnected fault with partial shading and open circuit (PSBDOC). The data was saved as a text file on the micro-SD card and then converted into an Excel file for pre-processing purposes. The data collected at the output was obtained through a program that converts the analog outputs of the sensors into their digital equivalents using the Arduino Uno microcontroller, following the flowchart in Figure 9.

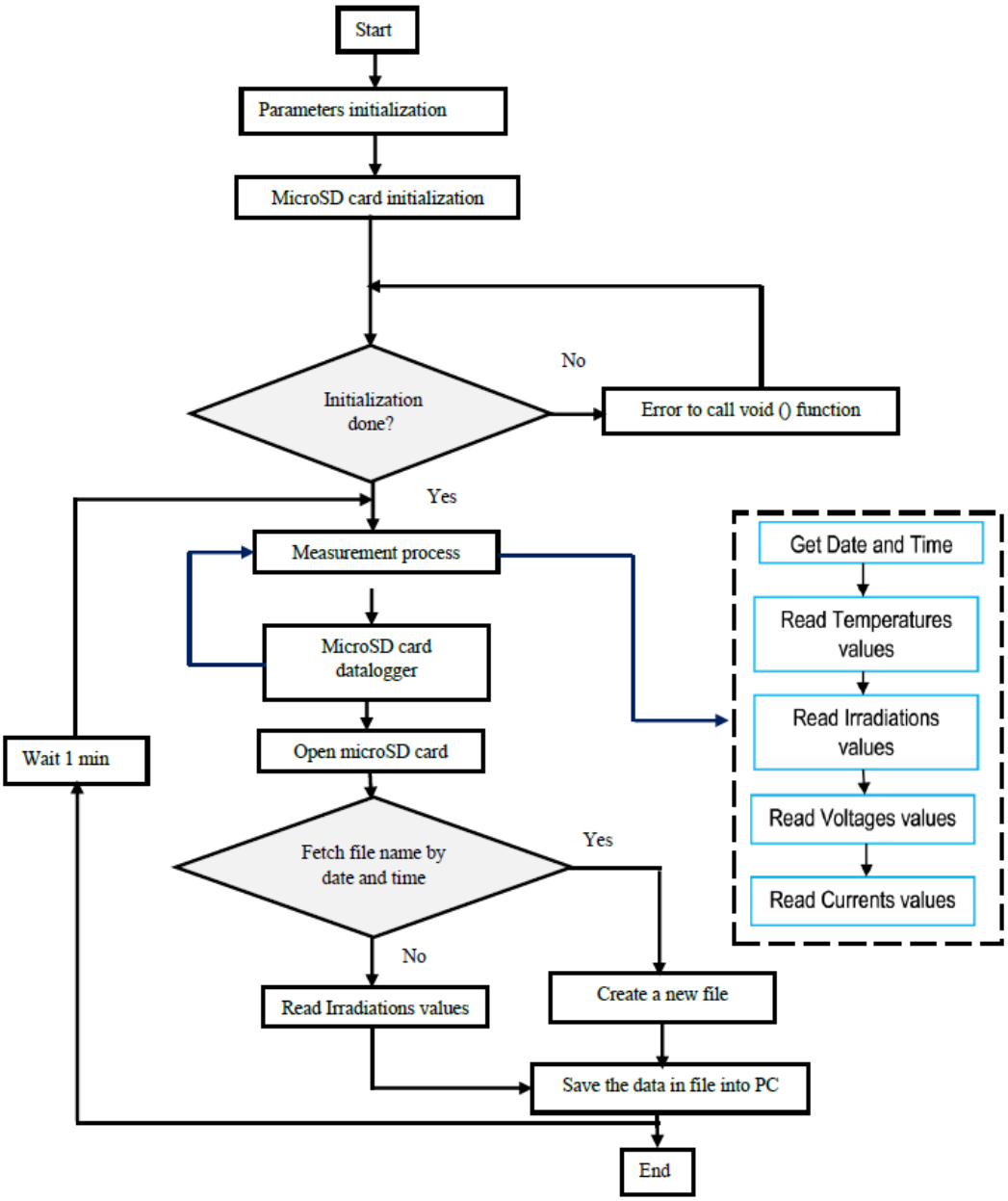


Figure 9. Flowchart Arduino ADC conversion of PV system data acquisition.

3.2. Development and application of the extra trees model

3.2.1. Model development

The extra trees algorithm, also known as extra trees randomized, is an ensemble learning algorithm based on a decision tree forest. It was first proposed and implemented in 2006 by Geurst et al. and can be used for both classification and regression [33]. Similar to random forest (RF), the extra trees (ET) model relies on a large number of decision trees to build a group of unpruned decision trees to reduce the risk of overfitting [44]. However, the ET algorithm differs from the RF algorithm in that it uses the entire training sample to train each tree, rather than a Bootstrap sample. Additionally, the ET algorithm randomly selects cutoff points, whereas the RF model uses an optimal distribution. The execution algorithm for the ET model is based on the training dataset (T) and three main hyperparameters:

- The number of trees (M) to train based on the number of training samples T.
- The number of attributes (K) to be randomly selected and used in each node for each trained ensemble tree.
- The minimum number of samples/instances (n_{min}) needed to split a node of each trained ensemble tree.

After training the trees, the final prediction of the algorithm is made based on test data by majority voting for classification or by calculating the arithmetic mean for regression [29]. The procedure for implementing the model proposed in this work is represented in Figure 10.

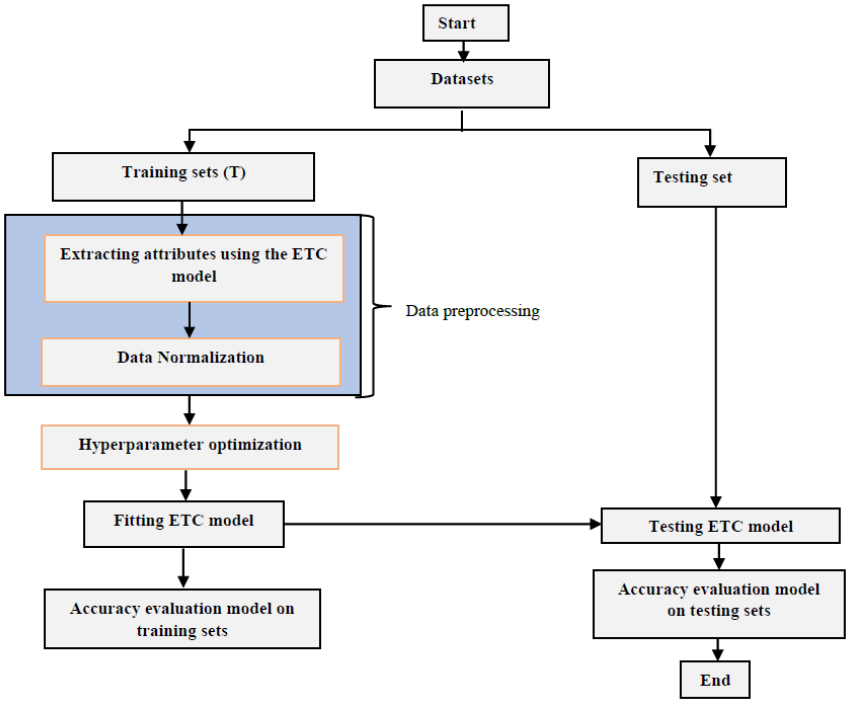


Figure 10. ETC model construction procedure for diagnosis.

The extra trees algorithm proposed in this work is an ensemble learning method that generates a decision function from a collection of decision trees [31].

3.2.2. *Extra trees (ETC) ensemble model application*

- Data preparation

Python was used for data pre-processing and implementing the model using the JupyterLab toolkit. After correcting the values, data were ready for analysis. During pre-processing, missing data was observed due to various causes such as sensor defects, incorrect storage of input data, and others. After correcting the outliers, we harmonized the data size for each defect case. A total of 2380 data points and a column of five defect classes were recorded, with an average of 476 data points per class. Figure 11 shows the distribution of data for each defect class.

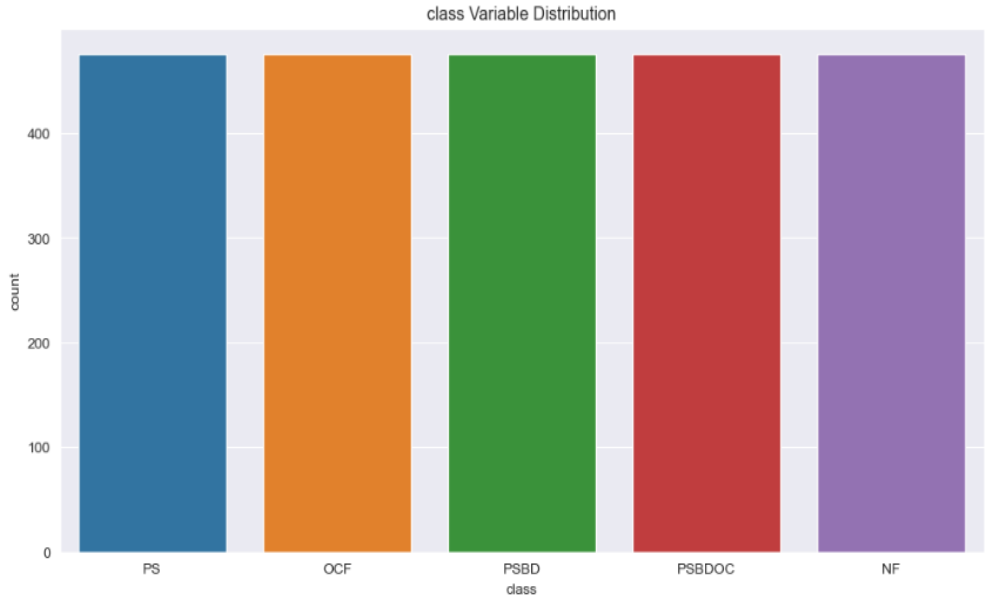


Figure 11. Data distribution for each class.

The data were divided into training and test data following the 80/20 principle, with 1904 (80%) of the data allocated for training and 476 (20%) for testing. The distribution of all training data is shown in Table 2. The different faults are represented by class numbers (0, 1, 2, 3, 4) corresponding to cases without fault (NF), open circuit fault (OCF), partial shading (PS), partial shading with bypass diode disconnected (PSBD), and partial shading with bypass diode disconnected and open circuit (PSBDOC).

Table 2. Distribution of training data for each fault class.

Description	Data proportion	Percentage	Class
NF	380.8	20%	0
OCF	380.8	20%	1
P.S.	380.8	20%	2
PSBD	380.8	20%	3
PSBDOC	380.8	20%	4
Total	1904	100%	05

After the data splitting process, the MinMaxScaler function was used for data normalization, followed by hyperparameter optimization using the GridSearchCV function. To evaluate the performance of the proposed model, six other models capable of detecting various types of defects were selected: random forest (RF), decision tree (DT), logistic regression (LR), k-nearest neighbors (KNN), AdaBoost, and support vector machine (SVM). The performance of each model was improved using the GridSearchCV function. Each model was trained on the training data with a specific range of parameters. The optimal parameters were identified using the best parameters function and the best score function, which returned the highest corresponding score. The hyperparameters listed in Table 3 are those that achieved the best performance of each model.

Table 3. Hyperparameters used for machine learning models.

Models	Hyperparameters tuning
Extra trees	n_estimator = 100, max_depth = 20, min samples split = 2
Random forest	n_estimator = 100, max_depth = 15, min samples split = 3
Decision tree	Criterion = Gini, max_depth = 15, min samples split = 2
Adaboost	n_estimators = 200, learning rate = 0.1, algorithm = SAMME
KNeighbors	Default
SVM	C = 1,kernel = rbf, gamma = scale
Logistic regression	C = 0.1, max_iter = 150, intercept scaling = 1

The models were chosen based on their superior performance and implementation speed for diagnosing faults in photovoltaic systems [23,45]. The main reasons for choosing the extra trees classifier were (1) the widespread use of artificial intelligence–based classifiers in recent decades, especially in machine learning and deep learning applications, making them faster and more efficient than traditional methods, and (2) the ability of ensemble algorithms like extra trees to handle noisy data and variance. Table 4 provides a summary of the advantages of the extra trees model.

Table 4. Advantages of extra trees model.

Models	Advantages
ETC	Improve robustness
	Reduce bias
	Reduce variance
	Reduce overfitting
	Improve accuracy
	Faster and easier to implement

- Evaluation of performances

To evaluate and analyze any classification model, it is crucial to select appropriate metrics such as the confusion matrix, accuracy, precision, recall, and f1_score. The choice of relevant metrics depends on the distribution of available data. When data is balanced, the accuracy metric is appropriate for evaluating the performance of a classification model [43]. Furthermore, the evaluation of the performance of a model on the basis of accuracy is not feasible when there is data imbalance [43]. In

this paper, we use the classification report function to evaluate different metrics. These metrics are defined by the following equations:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$f_{1\_score} = 2 \times \frac{Recall*Precision}{Recall+Precision} \tag{4}$$

where $TP\ et\ TN$ represents true positives and true negatives while $FP\ et\ FN$ represents false positives and false negatives.

4. Results

Using the acquisition device implemented, a total of 2543 data was collected in five days, including one day where the two modules were in healthy operation and four days where four faults were successively created. Figure 12 is an overview of the testbed developed for data collection.

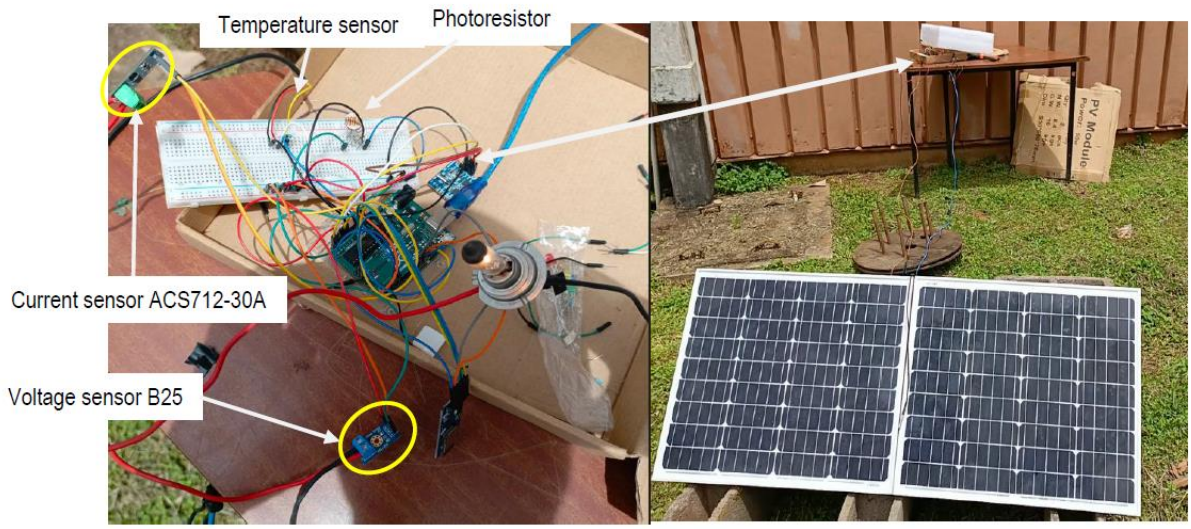


Figure 12. Experimental setup.

After filtering, 2380 data points were used to construct the proposed diagnostic model. The distribution of the different classes of defects according to the attributes was visualized using the heatmap function of the seaborn library of Python (see Figure 13).

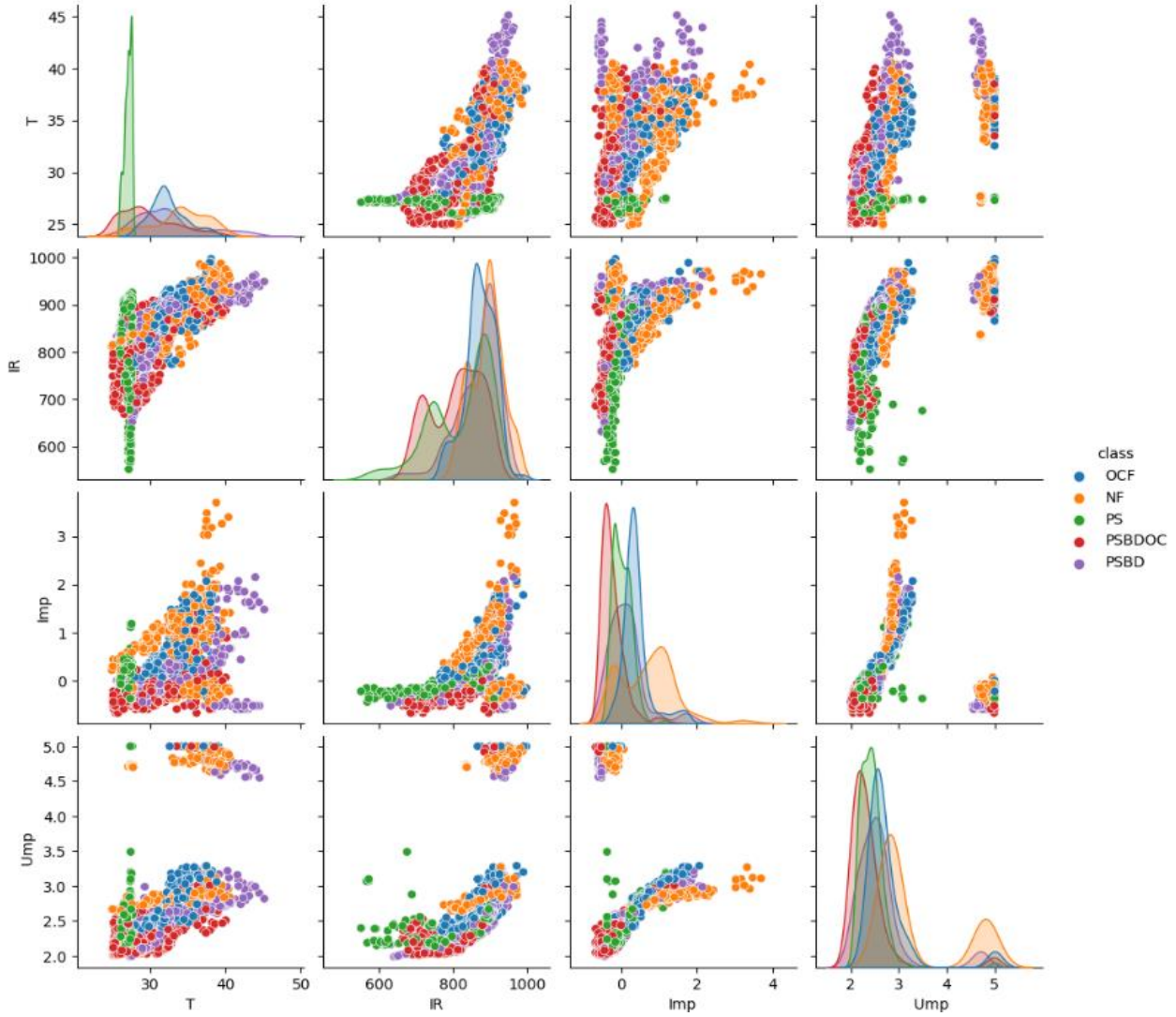
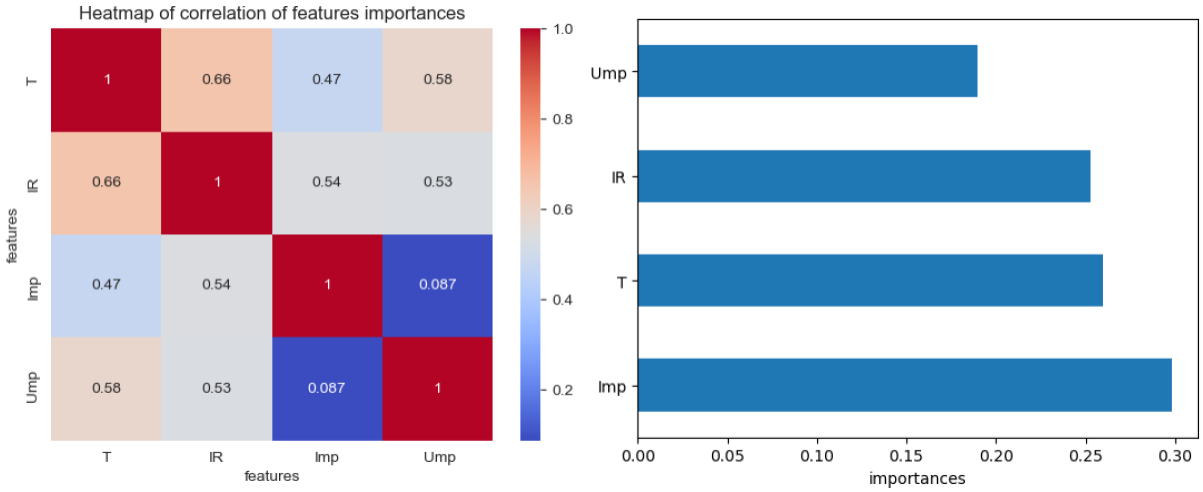


Figure 13. Distribution of each fault class based on attributes.

Upon analysis of Figure 14, it is evident that the current and voltage at the maximum power point exhibit a normal distribution. In addition, these dimensions allow a clear distinction between the fault-free class and the other classes, as indicated by the visible separation of the orange color. To further visualize the relationship between the attributes and the fault class, refer to Figure 14a, which employs the heatmap function.



(a) (b)

Figure 14. (a) Correlation matrix between the different attributes; (b) Distribution of observations in order of importance.

Examining Figure 14(a), it is evident that the defect is strongly correlated with the parameter that appears to have the least correlation with the other parameters in the dataset. Specifically, there is a strong correlation of approximately 54% between current and irradiation, and a correlation of 58% between voltage and temperature. Figure 14(b) confirms the strong correlation observed in Figure 14(a), indicating that the variables of current and temperature are the most significant. It is important to note that all statements made are objective and supported by the data presented. The correlation matrix confirms that irradiation has an influence on the current value, and temperature affects the voltage.

The performance of the model was obtained by evaluating the confusion matrices of the different selected models (see Figures 15 and 16), accuracy, precision, recall, and f1_score. For the confusion matrix, the values shown on the diagonal represent the correctly classified data.

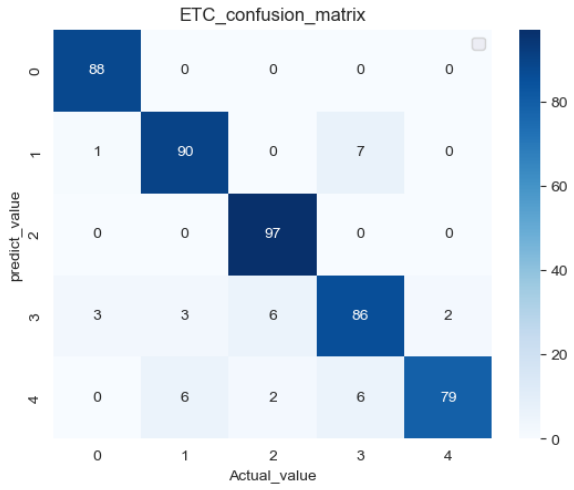


Figure 15. Confusion matrix of the developed extra trees model.

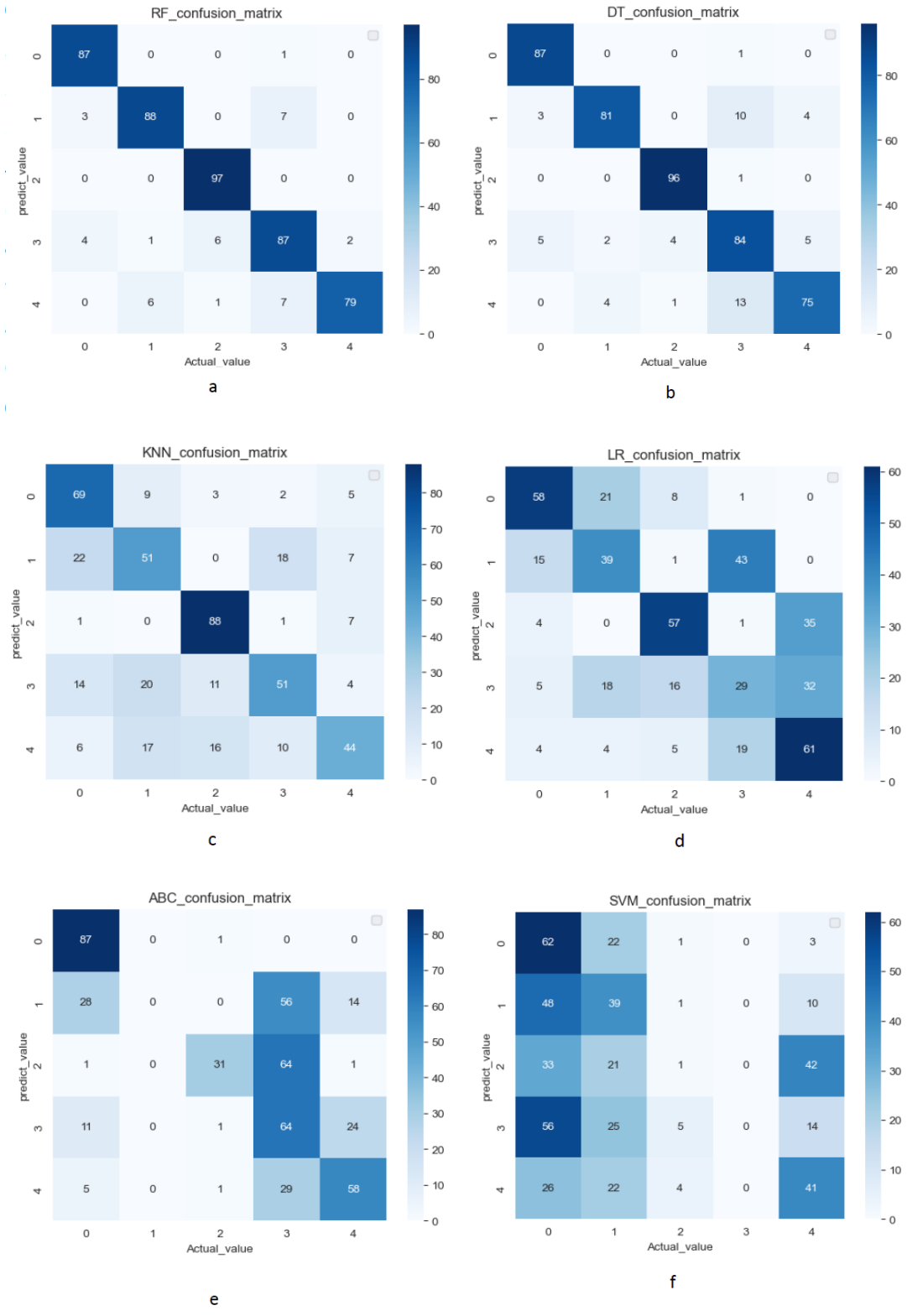


Figure 16. Confusion matrix. (a) Random forest; (b) decision tree; (c) k-nearest neighbors; (d) logistic regression; (e) AdaBoost; (f) SVM.

Figure 16 illustrates that the ETC model proposed in this work achieves a higher data classification rate than other models. The accuracy of each model is calculated by dividing the sum of

the diagonal elements by the total number of samples. Table 5 presents the performance of the selected models in training data.

Table 5. Performance summary of different algorithms on training data.

MODELS	Training time (S)	Accuracy (%)	Precision (%)	Recall (%)	f1 score (%)
ETC	0.3	1	1	1	1
RF	0.373	1	1	1	1
DT	0.015	1	1	1	1
KNN	0.019	75	75	75	75
LR	0.091	59	58	59	58
AdaBoost	0.557	55	52	55	48
SVM	0.22	32	21	32	24

The analysis of the table indicates that the ETC, RF, and DT classifiers achieved the highest scores, with 100% accuracy. It can be observed that the training time for these three algorithms is relatively low, with the DT model exhibiting the shortest training time and the ETC and RF models exhibiting similar training times. In summary, the time taken for the testing process is relatively brief for the majority of the models.

Table 6. Performance summary of different algorithms on testing data.

MODELS	ACCURACY (%)	PRECISION (%)	RECALL (%)	F1 SCORE (%)
ETC	92	92	92	92.2
RF	91	92	91	91
DT	87	88	88	88
KNN	64	63	64	63
LR	56	55	56	56
AdaBoost	52	51	52	46
SVM	30	21	31	23

A comparison of Table 6 reveals that the ETC model outperforms the RF, DT, KNN, LR, AdaBoost, and SVM models with an average accuracy of 92%, 91%, 87%, 64%, 56%, 52%, and 30%, respectively. Additionally, the LR, KNN, and DT models have shorter training times compared with the RF and ETC models. In summary, the extra trees model has the best performance with a high classification rate compared to the other selected models. Table 7 summarizes the classification rate of each fault class provided by the proposed ETC model.

Table 7. Data point classification rate by fault class.

Class of fault	Well ranked points	Misclassified points	Total points	Prediction rate (%)
NF	88	0	88	100
OCF	90	8	98	91.83
PS	97	0	97	100
PSBD	86	14	100	86
PSBDOC	79	14	93	84.94

In order to assess and compare the effectiveness of the proposed ETC model in classifying various fault conditions, we conducted an evaluation using the receiver operating characteristic (ROC) curve. This curve shows the trade-off between correctly classified points and those that are incorrectly classified. Figure 17 shows the ROC curve, which demonstrates the probability of correctly classified points (TPR) in relation to incorrectly classified points (FPR).

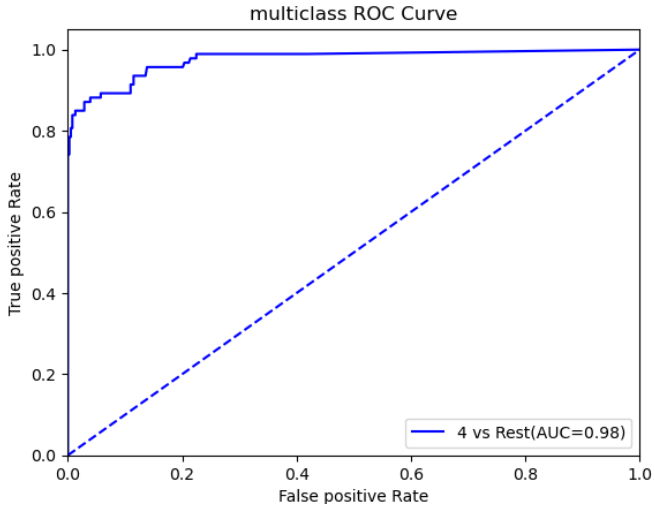


Figure 17. ROC function of the developed ETC model.

Upon examination of the ROC curve in Figure 17, it is evident that the proposed model has a high capacity for detecting and accurately classifying defects, with an AUC (area under curve) of 98% representing the model's overall performance.

5. Discussion

The effectiveness of the proposed new classifier for identifying faults in photovoltaic systems has been assessed by comparing it with several other approaches. These include basic supervised learning models like logistic regression (LR), k-nearest neighbor (kNN), and support vector machines (SVM), as well as ensemble algorithms such as decision trees (DT), random forests (RF), and AdaBoost. To ensure a fair and thorough comparison of the different models, the execution steps were carefully monitored. The internal hyperparameters of each model were adjusted using the GridSearchCV function to identify the best values, as indicated in Table 5. An evaluation of the different metrics based on the aforementioned hyperparameters revealed that the ETC, RF, and DT models demonstrated excellent learning performance on the training dataset, with an overall score of 100%. Conversely, the KNN, LR, AdaBoost, and SVM models demonstrated limited learning on the training dataset, with accuracy scores of 75%, 59%, 55%, and 32%, respectively, as shown in Table 5. In comparison, the SVM model exhibited the lowest accuracy, being below 50%, specifically 32%. A review of the confusion matrix for the SVM model reveals that the majority of points in each class were poorly classified. For instance, only one point in class 2 (PS) was classified correctly, and none of the points in class 3 (PSBD) were classified correctly, resulting in a success rate of 0%. In terms of calculation time, the DT classifier was the fastest, with a training time of 0.015 seconds. However, an analysis of the training times of the three ensemble algorithms, ETC, RF, and AdaBoost, reveals that

the ETC model has the lowest execution time, at 0.3 s, compared with 0.337 s and 0.557 s for the RF and AdaBoost models. The speed of the ETC model can be attributed to its decision tree-based structure, which does not consider the errors of previous trees. In contrast, the RF model allocates time to identifying the optimal node, whereas the AdaBoost model has a high computation time due to the complexity of processing data to adjust weights and correct errors. Following the training phase, the models were tested on the test set to assess their ability to detect and classify the various faults. The results obtained and shown in Table 6 reveal that during the test phase, the deployed ETC model outperformed all the other models. In fact, we recorded an accuracy of 92% for the ETC model, followed by an accuracy of 91% for RF. In comparison with the other metrics like precision, recall, and F1 score, the DT model's accuracy was 87%, which was lower by 1%. This could be due to the DT classifier being either too selective or too lenient, leading to a significant number of false negatives being accepted. In our evaluation, we prioritized accuracy as the main performance measure due to the balance between classes and tasks performed. As a result, the ETC model outperformed the other models we considered. The model's validity is supported by its high AUC score of 0.98, indicating a 98% probability of correctly distinguishing between well-classified and poorly classified points when identifying defects. However, in the event of a tie, the model will randomly assign points to the "well-classified" or "misclassified" categories. In such cases, the AUC will be 50%, represented by the first broken bisector.

6. Conclusions

The study introduces a new classifier called extra trees set (ETC) for identifying faults in a system comprising two photovoltaic panels. The defects were intentionally created, and data was collected using a low-cost Arduino-based recording system. The collected data was used to train seven classifiers: logistic regression (LR), k-nearest neighbor (KNN), decision tree (DT), random forest (RF), AdaBoost, and extra trees set (ETC). A comparative study was conducted to evaluate the performance of these classifiers, with the ETC model proving to be highly effective in accurately detecting and classifying various defects. The ETC model outperformed the other classifiers, achieving 92% accuracy compared to 91%, 87%, 64%, 56%, 52%, and 30% for RF, DT, KNN, LR, AdaBoost, and SVM, respectively.

These results highlight the importance of choosing an optimal model for fault detection in photovoltaic installations. The ETC model also shows promise for practical monitoring of the health of a photovoltaic system, suggesting its potential for developing advanced diagnostic tools for photovoltaic systems, thereby improving the reliability of solar technology and accelerating the installation rate. Although the ETC model produced satisfactory results, further analysis revealed certain limitations that underscore the need for additional research. One of the identified constraints is the failure to consider environmental variables such as humidity and wind speed, which can affect the performance of solar modules. Additionally, low data values due to the power of the proposed PV module and the limited number of usable parameters, such as current and voltage values, are specific to the type of photovoltaic system developed. Therefore, future research should explore the effectiveness of different classifiers on large-scale photovoltaic installations to enhance the reliability of the models. This could involve applying the model to fault diagnosis in a grid-connected photovoltaic system with numerous input parameters and conducting a similar study on a four-panel solar array, incorporating humidity and wind speed data. It would also be prudent to test the model on

degradation faults and accumulated faults on the direct current and alternating current sides in the diagnosis of photovoltaic systems.

Use of AI tools declaration

It is declared that, we didn't used AI Tools.

Authors contributions

GMTT worked research for the paper, while JK, JV, contributed to the writing and provided the necessary materials, YAM, BFN participated in proofreading the paper and SSO-D supervised the project. All authors have read and approved the final version of the manuscript.

Funding

This work has been supported by the World Bank through the Regional Center of Excellence for Electricity Management (CERME).

Acknowledgments

The authors express their gratitude to the World Bank for funding this work through the Regional Center of Excellence for Electricity Management (CERME). They also extend their appreciation to the small hydropower and hybrid systems laboratory of National Advanced School of Engineering of Yaoundé (NASEY) for their warm welcome and material support. The reviewers' insightful and positive comments and suggestions greatly assisted in improving this paper.

Conflict of interest

The authors declare no conflicts of interest.

References

1. International Energy Agency. Renewable energy market update—June 2023. Available from: www.iea.org/t&c/.
2. Hong YY, Pula RA (2023) Diagnosis of PV faults using digital twin and convolutional mixer with LoRa notification system. *Energy Rep* 9: 1963–1976. https://doi.org/10.1016/j.egyr.2023.01.011
3. Harrou F, Saidi A, Sun Y, et al. (2021) Monitoring; of photovoltaic systems using improved kernel-based learning schemes. *IEEE J Photovoltaics* 11: 806–818. https://doi.org/10.1109/JPHOTOV.2021.3057169
4. Firth SK, Lomas KJ, Rees SJ (2010) A simple model of PV system performance and its use in fault detection. *Sol Energy* 84: 624–635. https://doi.org/10.1016/j.solener.2009.08.004
5. Bendary AF, Abdelaziz AY, Ismail MM, et al. (2021) Proposed anfis based approach for fault tracking, detection, clearing and rearrangement for photovoltaic system. *Sensors* 21: 2269. https://doi.org/10.3390/s21072269

6. Hussain I, Khalil IU, Islam A, et al. (2022) Unified fuzzy logic based approach for detection and classification of PV faults using I-V trend line. *Energies* 15: 5106. https://doi.org/10.3390/en15145106
7. Mellit A, Tina GM, Kalogirou SA (2018) Fault detection and diagnosis methods for photovoltaic systems: A review. *Renewable Sustainable Energy Rev* 91: 1–17. https://doi.org/10.1016/j.rser.2018.03.062
8. Boubaker S, Kamel S, Ghazouani N, et al. (2023) Assessment of machine and deep learning approaches for fault diagnosis in photovoltaic systems using infrared thermography. *Remote Sens* 15: 1686. https://doi.org/10.3390/rs15061686
9. Osmani K, Haddad A, Lemenand T, et al. (2023) A critical review of PV systems' faults with the relevant detection methods. *Energy Nexus* 12: 100257. https://doi.org/10.1016/j.nexus.2023.100257
10. Taghezouit B, Harrou F, Sun Y, et al. (2024) Model-based fault detection in photovoltaic systems: A comprehensive review and avenues for enhancement. *Results Eng*, 21. https://doi.org/10.1016/j.rineng.2024.101835
11. Zhao J, Sun Q, Zhou N, et al. (2020) A photovoltaic array fault diagnosis method considering the photovoltaic output deviation characteristics. *Int J Photoenergy*. https://doi.org/10.1155/2020/2176971
12. Garoudja E, Harrou F, Sun Y, et al. (2017) Statistical fault detection in photovoltaic systems. *Sol Energy* 150: 485–499. https://doi.org/10.1016/j.solener.2017.04.043
13. Akram MN, Lotfifard S (2015) Modeling and health monitoring of DC side of photovoltaic array. *IEEE Trans Sustainable Energy* 6: 1245–1253. https://doi.org/10.1109/TSTE.2015.2425791
14. Chine W, Mellit A, Lughi V, et al. (2016) A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks: *Renewable Energy* 90: 501–512. https://doi.org/10.1016/j.renene.2016.01.036
15. Dhimish M, Tyrrell AM (2023) Photovoltaic bypass diode fault detection using artificial neural networks. *IEEE Trans Instrum Meas*, 72. https://doi.org/10.1109/TIM.2023.3244230
16. Kumar R, Sharma N, Chahat, et al. *(2024)* Prediction of jet impingement solar thermal air collector thermohydraulic performance using soft computing techniques. *Case Stud Therm Eng*, 55. https://doi.org/10.1016/j.csite.2024.104144
17. Madeti SR, Singh SN (2018) Modeling of PV system based on experimental data for fault detection using kNN method. *Sol Energy* 173: 139–151. https://doi.org/10.1016/j.solener.2018.07.038
18. Eskandari A, Milimonfared J, Aghaei M (2020) Optimization of SVM classifier using grid search method for line-line fault detection of photovoltaic systems. *IEEE Photovoltaic Specialists Conference*, 1134–1137. https://doi.org/10.1109/PVSC45281.2020.9300846
19. Wang J, Gao D, Zhu S, et al. (2019) Fault diagnosis method of photovoltaic array based on support vector machine. *Energy Sources* 45: 5380–5395. https://doi.org/10.1080/15567036.2019.1671557
20. Harrou F, Dairi A, Taghezouit B, et al. (2018) An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class Support Vector Machine. *Sol Energy* 179: 48–58. https://doi.org/10.1016/j.solener.2018.12.045
21. Mellit A, Zayane C, Boubaker S, et al. (2023) A sustainable fault diagnosis approach for photovoltaic systems based on stacking-based ensemble learning methods. *Mathematics* 11: 936. https://doi.org/10.3390/math11040936

22. Harrou F, Taghezouit B, Khadraoui S, et al. (2022) Ensemble learning techniques-based monitoring charts for fault detection in photovoltaic systems. *Energies* 15: 6716. https://doi.org/10.3390/en15186716
23. Benkercha R, Moulahoum S (2018) Fault detection and diagnosis based on C4.5 decision tree algorithm for grid connected PV system. *Sol Energy* 173: 610–634. https://doi.org/10.1016/j.solener.2018.07.089
24. Gong S, Wu X, Zhang Z (2020) Fault diagnosis method of photovoltaic array based on random forest algorithm. *2020 39th Chinese Control Conference (CCC)*, 4249–4425. https://doi.org/10.23919/CCC50068.2020.9189016
25. Ghoneim SSM, Rashed AE, Elkalashy NI (2021) Fault detection algorithms for achieving service continuity in photovoltaic farms. *Intell Autom Soft Comput* 30: 467–479. https://doi.org/10.32604/iasc.2021.016681
26. Sharma N, Thakur MS, Kumar R, et al. (2022) Assessing waste marble powder impact on concrete flexural strength using gaussian process, SVM, and ANFIS. *Processes* 10: 2745. https://doi.org/10.3390/pr10122745
27. Puri D, Kumar R, Kumar S, et al. (2024*)* Performance analysis and modelling of circular jets aeration in an open channel using soft computing techniques. *Sci Rep*, 14. https://doi.org/10.1038/s41598-024-53407-3
28. Mathew TE (2022) An optimized extremely randomized tree model for breast cancer classification. *J Theor Appl Inf Technol*. Available from: www.jatit.org.
29. Mahkya DAl, Notodiputro KA, Sartono B (2022) Extra trees method for stock price forecasting with rolling origin accuracy evaluation. *Media Stat* 15: 36–47. https://doi.org/10.14710/medstat.15.1.36-47
30. Almohammed F, Thakur MS, Lee D, et al. (2024) Flexural and split tensile strength of concrete with basalt fiber: An experimental and computational analysis. *Constr Build Mater*, 414. https://doi.org/10.1016/j.conbuildmat.2024.134936
31. Saeed U, Jan SU, Lee YD, et al. (2020) Fault diagnosis based on extremely randomized trees in wireless sensor networks. *Reliab Eng Syst Saf*, 205. https://doi.org/10.1016/j.ress.2020.107284
32. Toche Tchio GM, Kenfack J, Kassegne D, et al. (2024) A comprehensive review of supervised learning algorithms for the diagnosis of photovoltaic systems, proposing a new approach using an ensemble learning algorithm. *Appl Sci* 14: 2072. https://doi.org/10.3390/app14052072
33. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Mach Learn* 63: 3–42. https://doi.org/10.1007/s10994-006-6226-1
34. Maghami MR, Mutambara AGO (2022) Challenges associated with hybrid energy systems: An artificial intelligence solution. *Energy Rep* 9: 924–940. https://doi.org/10.1016/j.egyr.2022.11.195
35. Pei T, Hao X (2019) A fault detection method for photovoltaic systems based on voltage and current observation and evaluation. *Energies* 12: 1712. https://doi.org/10.3390/en12091712
36. Khalil IU, Ul-Haq A, Mahmoud Y, et al. (2020*)* Comparative analysis of photovoltaic faults and performance evaluation of its detection techniques. *IEEE Access* 8: 26676–26700 https://doi.org/10.1109/ACCESS.2020.2970531
37. Dhimish M, Tyrrell AM (2018) Photovoltaic bypass diode fault detection using artificial neural networks. *IEEE Trans Instrum Meas* 72: 1–10. https://doi.org/10.1109/TIM.2023.3244230

38. Dhakshinamoorthy M, Sundaram K, Murugesan P, et al. (2022) Bypass diode and photovoltaic module failure analysis of 1.5 kW solar PV array. *Energy Sources* 44*:* 4000–4015. https://doi.org/10.1080/15567036.2022.2072023
39. Platon R, Martel J, Woodruff N, et al. (2015) Online fault detection in PV systems. *IEEE Trans Sustainable Energy* 6: 1200–1207. https://doi.org/10.1109/TSTE.2015.2421447
40. Guerriero P, Piegari L, Rizzo R, et al. (2017) Mismatch based diagnosis of PV fields relying on monitored string currents. *Int J Photoenergy*. https://doi.org/10.1155/2017/2834685
41. Roger PY, Emilio CCJ, Rubén RH (2021) Fault diagnostic methodology for grid-connected photovoltaic systems. *J La Multiapp* 2: 10–30. https://doi.org/10.37899/journallamultiapp.v2i2.339
42. Berghout T, Benbouzid M, Bentrcia T, et al. (2021) Machine learning-based condition monitoring for PV systems: State of the art and future prospects. *Energies* 14: 6316. https://doi.org/10.3390/en14196316
43. Lodhi E, Wang FY, Xiong G, et al. (2023) A novel deep stack-based ensemble learning approach for fault detection and classification in photovoltaic arrays. *Remote Sens* 15: 1277. https://doi.org/10.3390/rs15051277
44. Goude Y (1996) Methodes d'ensemble et forets aléatoires. Available from: https://www.imo.universite-paris-saclay.fr/~yannig.goude/Materials/ProjetMLF/rf_web.html.
45. Camana Acosta MR, Ahmed S, Garcia CE, et al. (2020) Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks. *IEEE Access* 8: 19921–19933. https://doi.org/10.1109/ACCESS.2020.2968934

AIMS Press