

Fault Diagnosis in Photovoltaic Systems Using Machine Learning Algorithms

Patience Tifuh Taah¹, Derek Ajesam Asoh^{2,3*} , Jerome Ndam Mungwe²,
Jean-de-Dieu Nguimfack⁴, Daniel Agoons⁵

¹Department of Electrical Engineering, Pan African University Institute for Basic Sciences, Technology, and Innovation (PAUSTI), Nairobi, Kenya

²Department of Renewable Energy, Higher Technical Teachers Training College, The University of Bamenda, Bambili, Cameroon

³Electrical Engineering, Mechatronics, and Signal Processing Laboratory, ENSPY, UY-I, Yaounde, Cameroon

⁴Department of Electrical Power Engineering, Higher Technical Teachers Training College, The University of Bamenda, Bambili, Cameroon

⁵Agoons M&E Consultants, Yaounde, Cameroon

Email: *derekasoh@gmail.com

How to cite this paper: Taah, P.T., Asoh, D.A., Mungwe, J.N., Nguimfack, J.-D.-D., and Agoons, D. (2026) Fault Diagnosis in Photovoltaic Systems Using Machine Learning Algorithms. *Smart Grid and Renewable Energy*, 17, 131-156.

<https://doi.org/10.4236/sgre.2026.176007>

Received: March 10, 2026

Accepted: June 12, 2026

Published: June 15, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

The global energy landscape is undergoing a significant shift from fossil fuels to renewable energy sources, driven by environmental concerns. Solar energy, particularly through photovoltaic (PV) technology, has emerged as a prominent renewable energy source. However, PV systems face challenges due to the occurrence of faults, which negatively impact their efficiency and power output. Machine learning algorithms have revolutionized various sectors, including PV systems, offering potential solutions to fault diagnosis. This study focuses on utilizing machine learning for fault diagnosis in PV systems. Through a MATLAB simulation of a 7.5 kW PV system, three fault scenarios were implemented to generate the dataset. Machine learning algorithms were trained and tested using the classification application, with cross-validation and hold-out validation techniques employed for model validation. Results revealed that the wide neural network algorithm achieved the highest accuracy, reaching 98.88% and 98.23% with the cross-validation and hold-out respectively. The ensemble algorithms also demonstrated promising results, achieving an accuracy of 95.02%. These findings underscore the effectiveness of machine learning algorithms in accurately diagnosing faults in PV systems, offering valuable insights for enhancing system efficiency, reliability, and availability.

Keywords

Machine Learning, Fault Diagnosis, PV Systems, Neural Networks, PV Faults

1. Introduction

1.1. Background

In an era of a relentless quest for sustainable development solutions, energy stands at the forefront, presenting both formidable challenges and unprecedented opportunities. Global energy demand continues to surge, fueled by population growth, urbanization, and technological advancements [1]. Over the decades, the primary source of electrical energy has been derived from fossil fuels [2] [3]. However, fossil fuels are responsible for releasing greenhouse gases such as carbon dioxide (CO₂) and methane (CH₄). In the year 2021 electricity generation contributed 46% of global CO₂ emissions [4]. The effects of climate change have prompted the global community to come up with adaptation and mitigation strategies such as the transition to renewable energy sources such as wind, solar, and hydroelectric systems [5] [6].

The use of PV systems has a significant impact on decarbonizing the global energy mix. It is estimated that about 720 million tons of CO₂ have been saved as a result of PV system usage. For example in 2019, global CO₂ emissions were reduced by 2.2% for energy-related emissions and 5.3% for electricity-related emissions [7]. These renewable energy sources leave no carbon footprint and are indispensable in achieving Net Zero emissions.

In 2020, the total installed capacity of photovoltaic systems exceeded 760.04 GW [8]. The increasing installation of PV systems has raised concerns about the issue of maintenance. Typically, PV systems require minimal maintenance because no movable parts are involved [9]. However, in large PV systems, issues of efficiency, reliability, and availability (continuous power supply) are crucial. For example, failure of a module affects about 18.9% of the output power of the PV system [10].

In PV systems, routine maintenance is carried out to ensure the normal operation of the system. However, such maintenance efforts have proven to be inefficient in keeping the system running at optimum levels given the random occurrence of multiple faults in these systems [11]. PV systems may experience different types of faults which are often unpredictable; and given their exposure to a wide range of environmental conditions, the occurrence of faults is often inevitable [12]. In PV systems, where the currents and voltages are of high magnitude, the effect of these faults is often profound, sometimes resulting in a fire outbreak or a shutdown of the system [9] [11]. A recent example is the case of a 75 MW PV system in South Africa, which was destroyed due to module degradation [13].

The effective utilization of PV systems as an energy source requires constant monitoring for the diagnosis, identification, and elimination of faults. To ameliorate the situation, statistical methods and Machine Learning (ML) methods for fault diagnosis have been used. These methods make use of the large data sets acquired from PV systems.

ML algorithms have been applied to several areas of research, such as cybersecurity, traffic prediction, healthcare, energy optimization, and agriculture [14]

[15]. There has been significant advancement in the field of ML. ML is a unique tool for PV fault diagnosis because of its large data-handling capacity, pattern recognition, and low memory usage [16]. The objective of this study is to apply ML algorithms in the diagnosis of faults in PV systems using comprehensive machine learning algorithms. A fast fault detection approach to enable timely interventions to restore PV System functionality, thereby minimizing system downtime while maintaining the required energy output is required.

A PV system is used to convert sunlight into electrical energy. The fundamental component of the PV system is the solar module (panel) which is an interconnection of solar cells to ensure the required generated voltage. A solar cell is a p-n junction that converts incoming sunlight into direct current through the photovoltaic effect. Solar PV cells have been modelled as p-n junctions with non-linear characteristics using either the single diode and double diode models [17]. Solar PV cells are connected in series and parallel to form a PV module of desired power. Solar modules connected in series or parallel form an array. A PV array alongside DC-DC/DC-AC converters, battery banks, and inverters makes up a typical PV system [18].

For electrical power generation PV systems are implemented in various configurations namely; grid-connected, off-grid, and hybrid systems. An off-grid (standalone) system shown in **Figure 1** illustrates the energy flow from the PV array to the consumer's load. A grid connected system is a PV system which is linked to the Grid and incorporates a grid connected inverter and a net metering system; while a hybrid system consists of a PV system alongside one or more energy sources [19].

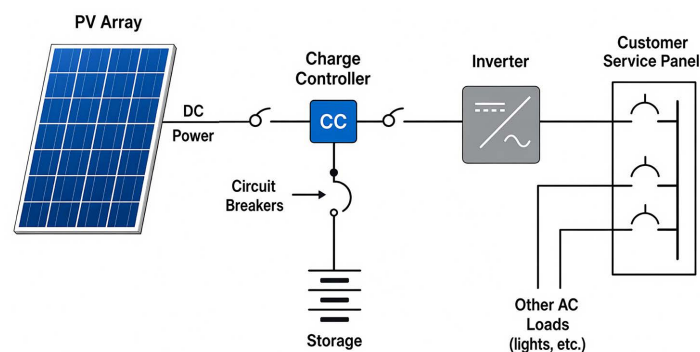


Figure 1. Off grid PV system [19].

1.2. Faults in PV System

A fault is any event that causes a solar module's output to deviate from the output defined by the manufacturer and ultimately the defined PV system [10]. Faults are classified based on several criteria including, current type, effects of fault, time characteristic, degree of damage and the cause of the fault. Based on the current type, there are DC and AC faults [16] [20] [21]; based on the effects of the fault, there are temporal and permanent faults [9]; based on time characteristics, there are intermittent, abrupt and incipient faults [22]; based on the degree of damage,

there are acute and chronic faults [23]; and based on the cause, there are physical, electrical, and environmental faults [10] [24] [25].

Faults are also broadly classified as open-circuit or short-circuits. Open-circuits faults are caused by a disconnection in the PV array, thus leading to an instant drop in power [12] [26]. Short-circuit faults are unintended connections between two points in the PV system at different potentials, leading to a high current. Short-circuit faults are typically caused by insulation damage, module damage and maintenance operations [27]. Short circuit faults are sub-classified as line-to-line and line-to-ground faults. Other faults types are termed arc faults; which stem from high current forced through air gaps [23]. Degradation, module damage, and bypass failure affect the frame of the module over time leading to a gradual decrease in power production [25] [28]. Some faults depend on environmental conditions such as permanent and temporal shading. Severe cases of shading result in hotspot faults evidenced in cracking and cell failure [10] [22] [29].

In order to mitigate the effects of faults in PV systems, appropriate maintenance strategies (routine, corrective and predictive maintenance) must be put in place. Routine maintenance is regular, time-based interval activities to ensure the smooth functioning of a system. Corrective maintenance refers to all activities carried out to restore a malfunctioning or faulty system [30]. Predictive maintenance evaluates the likelihood of a fault occurring based on collected data.

Fault diagnosis is a prerequisite for corrective maintenance; and is implemented using one of three approaches: Real Time Difference, Model Based Detection, and ML algorithms [23] [27]. ML algorithms have produced better results in identifying faults based on data inputs [31] [32].

1.3. Machine Learning and Machine Learning Algorithms

Machine Learning (ML) refers to the ability of machines to learn from data and make better predictions without using a conventional set of instructions [7]. ML is a subset of artificial intelligence, which is used to create models and algorithms that can recognise patterns and insights in data without explicit programming [33]. Algorithms used in ML are considered advanced data handling techniques. ML may be classified as either supervised or unsupervised learning as shown in **Figure 2** [34]. Supervised learning uses a labelled dataset to make predictions (regression) and identify distinct categories in a data sample (classification). Unsupervised learning involves the identification of patterns (clustering) in an unlabelled dataset [7].

Recent advances in research have led to a third perspective of ML called reinforcement learning using agents; with an agent being a virtual or real entity, responsible for decision making [35]. In reinforcement learning, an agent learns by adapting to its environment [7].

Supervised ML techniques have found applications in the PV industry for modelling, PV plant sizing, and fault diagnosis [36]. ML algorithms are evaluated based on performance measures such as precision, accuracy and confusion matrices [31]. Common ML Classification Techniques include the following:

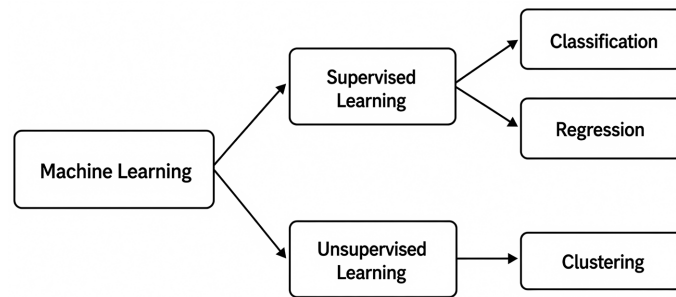


Figure 2. Machine learning taxonomy [34].

Decision Trees (DT): According to Gaboitaolelwe *et al.* [37] a Decision Tree is a type of ML algorithm that is used to classify the value of a target variable based on the values of other input variables. DTs are tree structures based on if/else statements and are constructed based on predefined conditions by splitting the data points using a splitting criterion. Typical splitting criteria include information gain, gain ratio and Gini index [38].

Random Forest (RF): RF is an ensemble ML algorithm obtained by combining a number of decision trees with the goal of improving model accuracy [37]. The RF algorithm was developed to overcome the major limitation of DT [7].

Linear Discriminant Analysis (LDA): LDA is a classification and dimensionality technique for binary and multi class classification problems. LDA creates a distinction between classes by projecting data from a higher-dimensional feature space to a lower-dimensional space thereby reducing the variability within classes. LDA makes use of statistical properties such as variance, covariance, and mean values of each class. In terms of dimensionality reduction, LDA uses two approaches; class dependent and class independent transformations depending on the type of data involved.

Support Vector Machines (SVM): SVMs map data into a higher dimensional space using hyperplane plots. These plots are such that two classes are separated optimally and the margin between the hyper plane and the observation is maximum. SVMs use kernel functions which are symmetrical functions that identify similarities between observations based on labels [37].

K-Nearest Neighbours (KNN): This algorithm classifies data by comparing the Euclidean distance between the data points in the training set. This algorithm identifies the unknown data element based on the labels of the nearest neighbours, hence the name KNN [39].

Ensemble: Ensembles are combinations of weaker algorithms to generate a powerful algorithm with less computational time. Ensembles are achieved through bagging, boosting and stacking [7]. Ensembles are either heterogeneous or homogenous. Heterogeneous ensembles use distinct learning algorithms while homogenous ensembles use single learning algorithm [37].

Neural Network (NN): A neural network (NN) is a simplified mathematical model which emulates the human brain. NN makes use of parallel-distributed signal processor composed of processing units called neurons which are capable of

learning using an algorithm and reproducing the information in the future when required. Structurally, NN consists of an input layer, a hidden layer, and an output layer [34]. The layers and neuron configurations can be varied to create specific types of NNs such as wide, medium, and Trilayered NNs.

1.4. Machine Learning Applications in Photovoltaic Systems

Fault diagnosis using ML algorithms has advanced over time to become an efficient technique in detecting irregularities in PV systems. Prominent fault diagnosis methods are based on specified data attributes such as non-electrical and electrical data [40]. Non electrical data consist of image and meteorological data typically used in assessing such faults as discoloration, cracks and soiling. For example, visual and thermography data have been used to detect faults [41] while meteorological data have been used for diagnosis in a 1MW PV system [42].

Electrical data consist of current-voltage (I-V) curves [43] [44] and current and voltage parameters; usually obtained either through real time data collection and/or simulations [45] [46]. A number of tools are used for the simulation of PV systems including MATLAB and PSCAD, which provide system modelling close to real world scenarios. These tools allow for extreme fault modelling which may result in hazards if implemented on real existing PV systems, making data simulation an essential data source for the application of ML in PV systems [47].

Recently there has been increased interest in research on the application of ML to fault diagnosis in PV systems; of which the following can be noted:

Badr *et al.* [48] obtained simulated data from MATLAB using fuzzy logic controlled MPPT and real time data from an experimental setup to detect faults in PV systems. The authors employed three ML algorithms, DT, KNN and SVM. Faults considered were arc fault, line to line fault, open circuit fault, MPPT unit failure, and partial shading fault. Ensemble learning models and semi-supervised methods were introduced to label the dataset. Results obtained show that SVM in combination with ensemble self-training algorithms had the best accuracy of 90.57% and 87.60% using simulated and experimental data respectively. Similarly, Chen *et al.* [49] used MATLAB generated data on a grid connected PV system and an experimental setup to diagnose faults. The random forest algorithm was used to identify faults such as line to line, shading, degradation, and open circuit. Fault diagnosis was achieved in two steps; detection and classification alongside a random validation and 10-fold cross validation. Classification using 10 fold cross validation, obtained an accuracy of 99.95% and 99.14% respectively for the simulated and experimental data. Attouri *et al.* [45] simulated an MPPT based grid connected PV system using bypass diode faults, string faults and line to ground faults. These faults were introduced using variable resistors. Principal component analysis (PCA) was employed for feature extraction followed by classification using SVM algorithm. A combination of all fault types produced an accuracy of 99.96%. Basnet *et al.* [46] proposed an intelligent detection model based on data collected under different environmental conditions during winter and summer. Data was

generated based on a simulation and an experimental setup used for validation. A probabilistic NN was used in the training and testing datasets, yielding a classification accuracy of 100%. Unlike previous authors, Basnet *et al.* [46] used solely an experimental setup to induce faults and collect data for fault diagnosis. The faults induced were partial shading (using translucent material) and short-circuit faults (using short circuited bypass diodes); and data collection was based on a remote cloud-based data management system.

In [50], the authors used an ensemble ML method alongside various optimisation parameters. These authors found that the best ensemble model was made up of quadratic discriminant analysis, extra trees, and DT with an accuracy of 97.46% and 97.67% respectively before and after optimization. Eskandari *et al.* [43] used MATLAB generated data and experimental data to diagnose faults using hierarchical classification (HC) and ML. Variations of line to line and line to ground faults were used to build the dataset using I-V curves. For ML, three algorithms were used: SVM, naive bayes, and logistic regression. The SVM emerged as the best classifier obtaining 96.66% and 91.66% classification accuracy respectively for line to line and line to ground faults.

1.5. Research Gap and Contributions

Despite the widespread use of machine learning methods in fault diagnosis of photovoltaic systems, some methodological limitations have been identified. Most of the existing literature focuses on specific types of faults, specific types of classifiers, or specific validation methods. Thus, the performance of the classification models is usually investigated in isolation, considering the impact of validation methodologies or the impact of fault severity on the performance of the classification models. In addition, most of the existing literature focuses on specific aspects of machine learning model optimization, which may not provide adequate insights into the impact of validation methodologies on classification model performance. Thus, the performance of classification models in the context of photovoltaic systems has not been adequately investigated in terms of the classes of ML models that provide stable performance under controlled electrical conditions. Unlike in [49], the authors evaluated a single algorithm family (Random Forest) using simulated MATLAB data from a grid-connected system, the present study performs a systematic cross-family comparison of five machine learning families comprising sixteen sub-models under three independent validation strategies. Furthermore, [48] combined simulated and experimental data within a semi-supervised framework. The present study adopts a fully supervised, simulation-only design on a standalone 7.5 kW system and explicitly includes a fault-free class within the multi-class decision framework, enabling simultaneous fault detection and classification in a single model. In view of the above, the following contributions are made in this study:

- 1) A new structured multi-fault photovoltaic dataset is developed based on a controlled 7.5 kW photovoltaic system simulator. The dataset includes string

faults, string-to-string faults, partial shading faults, with specific severity levels. Thus, the dataset allows for the discrimination of fault types under controlled conditions.

2) The performance of five classes of machine learning models, each with sixteen different models, is investigated under three independent validation methodologies: 5-fold cross-validation, 10-fold cross-validation, and hold-out validation.

3) A comparative evaluation of five machine learning families comprising sixteen sub-models is carried out using three independent validation strategies, which include 5-fold cross-validation, 10-fold cross-validation, and hold-out validation methods.

4) An integrated formulation of the multi-class classification problem is employed, wherein the identification of the presence and types of faults is addressed within a single decision framework through the explicit modelling of a fault-free class.

5) The results obtained from the study prove the superiority and consistency of wide neural network architectures in classification performance using a minimal set of electrical features, such as current, voltage, and power, making them suitable choices for computationally efficient PV fault diagnosis systems.

2. Methods and Material

2.1. Data Acquisition

A 7.5 kW PV array was designed on MATLAB using the SunPower SPR-415E-WHT-D module. The MATLAB Simulink was used on a Windows 10 HP AMD 2.10 GHz Ryzen processor computer with a 16 GB RAM. The electrical characteristics of the modules used to construct the 7.5 kW array are provided in **Table 1**.

The MPPT controller was implemented to facilitate convergence to the MPP at each simulated operating condition, although the extraction of the relevant characteristics for the machine learning process was done after stabilization at a steady state, and the transient MPPT process was not considered for the classification process.

Table 1. Characteristics of the solar module.

Feature	Value
Maximum power	414.8 W
Open circuit voltage	85.30 V
Short circuit current	6.09 A
Voltage at maximum power point	72.90 V
Current at maximum power point	5.69 A
Temperature coefficient of V_{oc}	-0.23
Temperature coefficient of I_{sc}	0.031

A standalone (off-grid) PV plant configuration under standard conditions of Irradiance (I_r) (1000 W/m^2) and temperature (T) (25°C) was used as presented in **Figure 3**. This PV configuration used for the simulation includes a DC-DC converter (boost converter) and MPPT inverter.

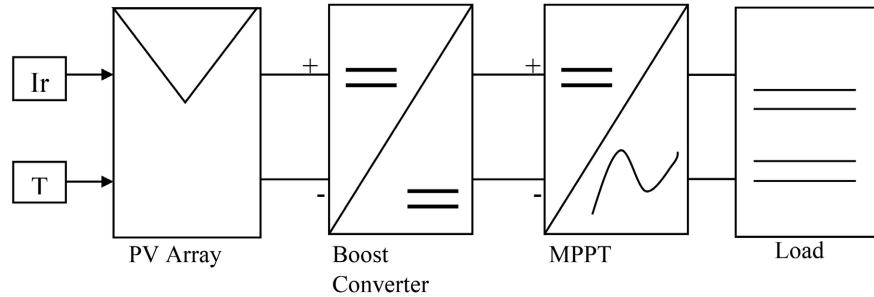
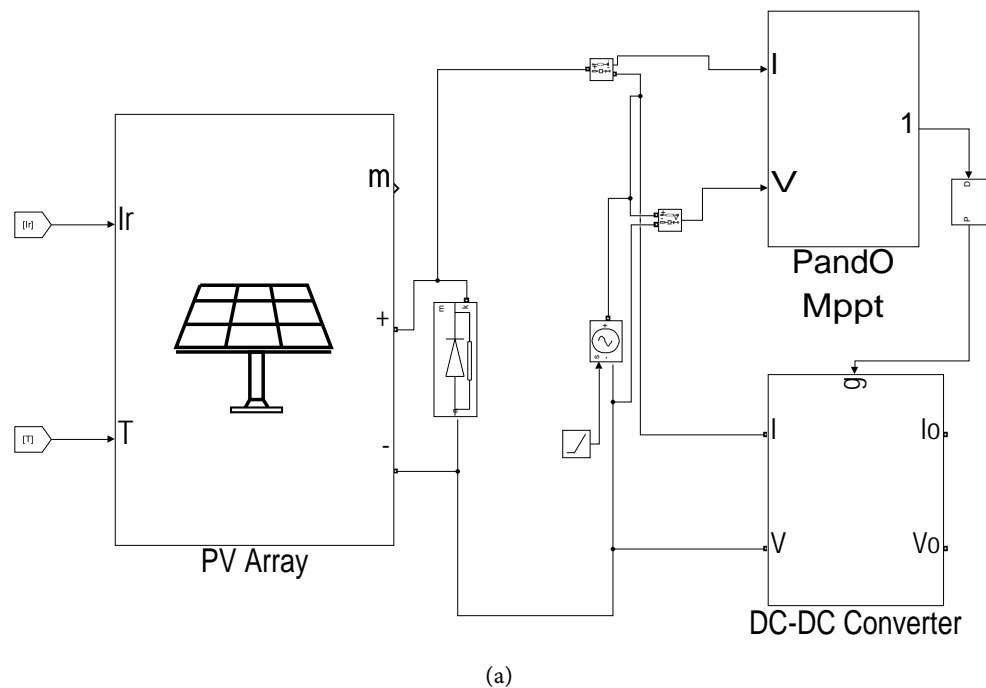


Figure 3. Configuration of a 7.5 kW PV system.

The boost converter optimises the DC voltage generated by the array. In PV systems, the boost converter efficiently increases the low voltage output from the PV modules to a higher level suitable for charging the battery bank or powering the load. This vital function enables the system to maximize power transfer from the PV array, especially during periods of low sunlight or varying weather conditions. The MPPT inverter operates using the Perturb and Observe (P&O) algorithm; and ensures that maximum power is processed by continually adjusting the input impedance to match the solar module operating point thus maximizing energy conversion efficiency. The PV system simulation setup is shown in **Figure 4(a)**; while an expanded view of the PV array is presented in **Figure 4(b)**, consisting of a 6×3 array. This setup was used to generate the data set used in the study.



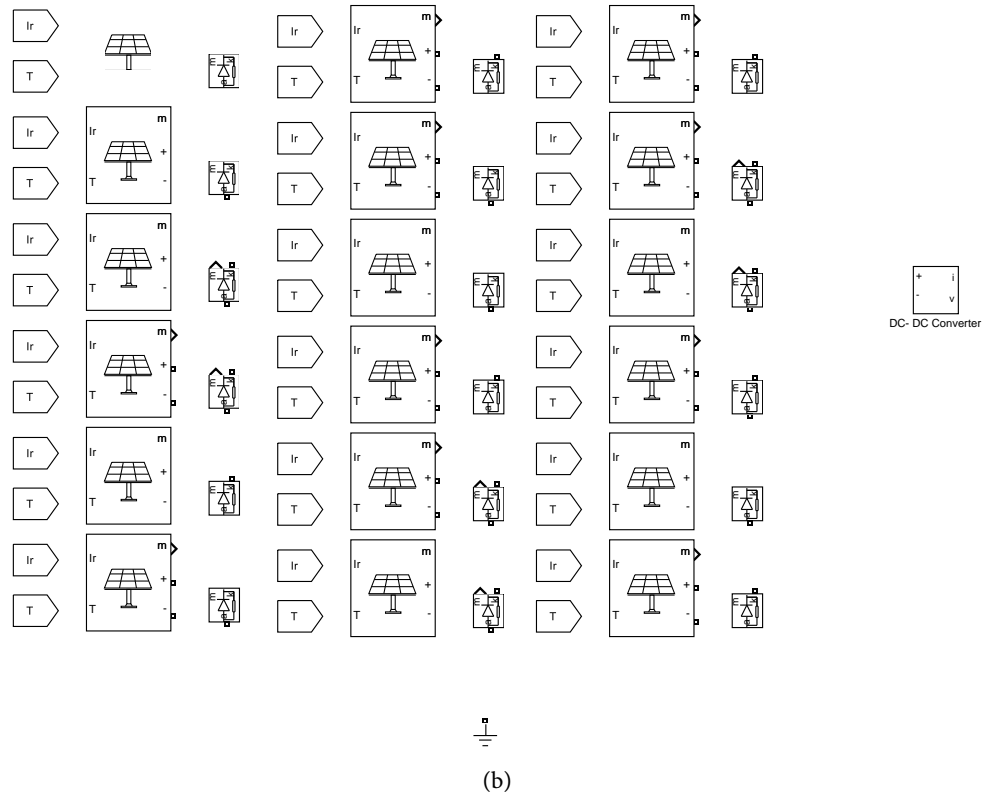


Figure 4. (a) PV Simulation Setup and (b) PV System Simulation of 6×3 array.

The classification of faults based on the locations on the PV array enabled the simulation of three faults: string fault, string-to-string fault, and partial shading fault. String faults, which are inclusive of all faults that can arise within a string in an array, are common and typically degrade the performance of the PV system. The choice of these fault types, aligns with the modular design of the PV system, where optimising individual strings ensures better performance of the PV system. A string fault, shown in **Figure 5** (F1), creates an unintended path for current flow between two points in the array. This fault was simulated using an RLC branch with resistance values of 5Ω , 10Ω , and 20Ω . A string-to-string fault, shown in **Figure 5** (F2), was simulated by creating a low resistance path between two strings in the array.

The partial shading fault was implemented by varying the irradiance levels across the strings, with a shading intensity of 60% per string. Teo *et al.* [51] examined the effects of partial shading at 20%, 60%, and 80%, while Badr *et al.* [9] focused on 20% and 80% for fault diagnosis using machine learning. Given this trend, 60% shading was selected for this study. To enhance the analysis, two scenarios were considered: Case 1 involved 60% shading on a single string, while Case 2 applied the same shading level to two strings.

In the context of practical photovoltaic supervisory control systems, it is worth noting that a high level of computational efficiency represents a critical design requirement, particularly in the context of large-scale photovoltaic arrays that ne-

cessitate continuous monitoring. In order to minimise unnecessary levels of computational complexity, advanced diagnostic algorithms are usually only activated in the event that abnormal electrical characteristics are detected via lightweight monitoring strategies that focus on voltage, current, and power level thresholds. In the context of this research, the proposed classification model includes a fault-free class (F0), which allows the model to simultaneously verify fault presence and discriminate fault types. As a result, it can be stated that the proposed model will not require unnecessary post-processing in the event that the predicted class corresponds to the fault-free class (F0), and it will directly classify fault types in the event that the predicted class corresponds to fault classes F1, F2, and/or F3.

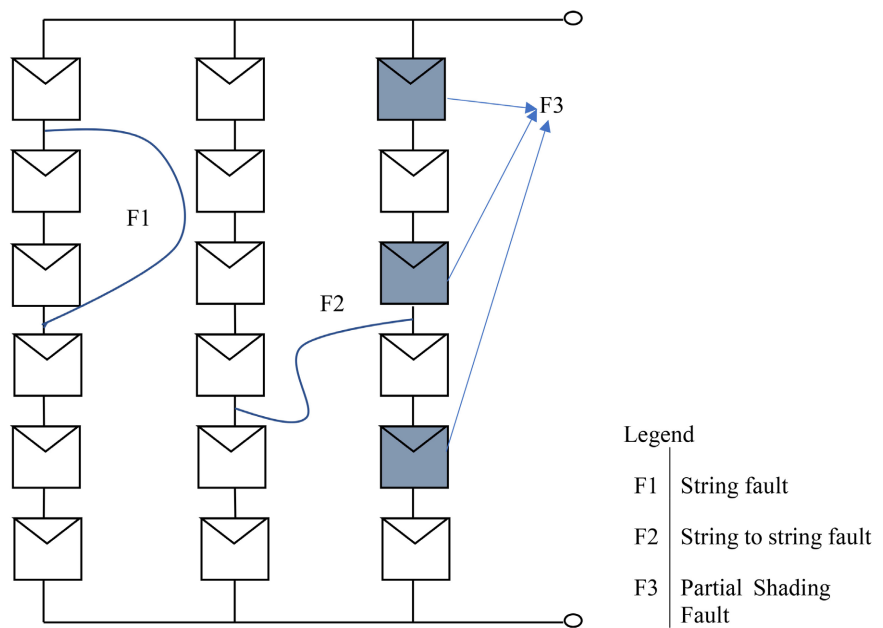


Figure 5. Fault simulation.

2.2. Application of Machine Learning to the Dataset

The procedure used for machine learning application to fault diagnosis is shown in Figure 6.

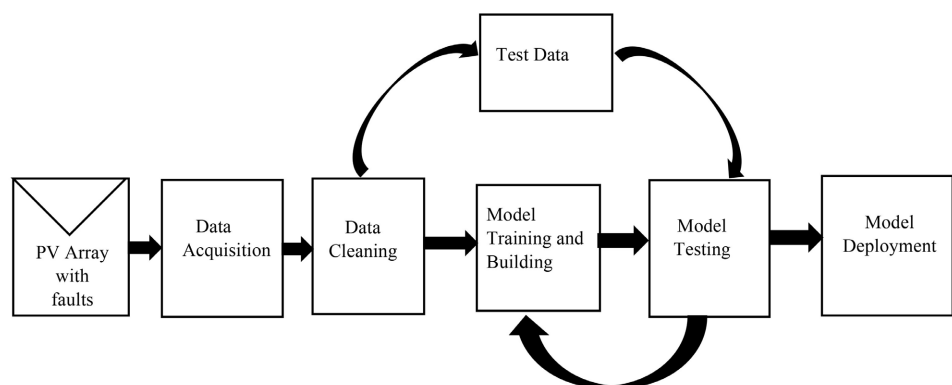


Figure 6. Procedure for fault diagnosis using machine learning.

The dataset was obtained from the PV simulation under four scenarios: 1) a fault-free (normal working) condition (F_0) and 2) three fault conditions (F_1 , F_2 , F_3) illustrated in **Figure 5**. The designation of the fault, fault condition (type) and label (class) is depicted in **Table 2**.

Table 2. Fault scenarios.

Fault	Fault Condition (Type)	Label (Class)
F_0	Normal (fault-free)	0
F_1	String fault	1
F_2	String-to-string fault	2
F_3	Partial shading fault	3

The faults were characterized/analysed using values of three parameters: current (I) and voltage (V) (measured), and power (P) (computed as the product of current and voltage). For practical purposes, the power captures the operating point of the PV system, which is the most diagnostically relevant quantity for fault discrimination.

Each scenario was given a label (class), k , and the dataset for the analyses was represented by the matrix depicted in Equation (1). In order to generate a high-resolution dataset for fault diagnosis, each scenario was simulated N times and all three parameters (I , V , P) obtained in each case.

$$F_{k,i} = \begin{bmatrix} k & I_{k,i} & V_{k,i} & P_{k,i} \end{bmatrix} \quad (1)$$

where;

$k = 0, 1, \dots, M$ ($M = 3$, the number of fault conditions, $k = 0$, the fault-free condition).

$i = 1, 2, \dots, N$ ($N = 1000$ is number of times each parameter was measured).

The final dataset was obtained by merging the four datasets from Equation (1).

Each of the four scenarios (F_0 - F_3) was simulated under fixed standard test conditions of irradiance (1000 W/m^2) and temperature (25°C). The simulated data often contains missing data points or transient points which affect the accuracy of ML algorithms. The data generated was cleaned by removing outliers. The dataset contained 4000 data points before cleaning and 3116 data points after cleaning. The cleaning was achieved by deleting outliers and redundant values. The main criteria were removal of values that arise from the numerical transient behaviour of the MATLAB Simulink solver during initialisation and MPPT convergence, before the system reaches steady-state; they do not represent any real operating condition of the PV array.

To enhance model stability and ensure balanced feature contribution during training, data normalization was applied as part of the preprocessing stage. The dataset comprises electrical parameters with different numerical scales, namely current, voltage, and power. Normalization reduces sensitivity to changes in the magnitude of these values, which can disproportionately influence distance-based

and gradient-based learning algorithms, potentially biasing model training and degrading generalization performance. The min-max normalization was applied to each feature independently, scaling values to the interval [0, 1] according to (2)

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (2)$$

where:

x is the original feature value.

x_{\max} , x_{\min} are the minimum and maximum values computed from the training dataset.

The 3116 data points were split into two using the 80:20 ratio for training and testing respectively [46]. The training and testing dataset therefore consisted of 2493 and 623 samples respectively. It is important to clarify the scope of the normalization procedure with respect to data splitting. The min-max normalization described in Equation (2) was applied after the 80:20 train-test split. The minimum and maximum values (x_{\min} and x_{\max}) were computed exclusively from the training set and subsequently applied to scale the test set. This ensures that no information from the test set influenced the normalization parameters, thereby preventing data leakage. Within the cross-validation folds, normalization was applied independently within each fold; the scaling parameters were derived from the training fold only and applied to the corresponding validation fold. This procedure is consistent with best practice for supervised learning pipelines and ensures that all reported validation accuracy figures are uncontaminated by test-set information. The 80:20 train-test split was applied at the sample level (row-wise) across the merged four-class dataset.

In this study, 5 ML algorithms and 16 sub-algorithms (Table 3) were trained using the classification learner application in the statistics and machine learning toolbox of MATLAB. The data was imported into the classification learner application followed by selecting a validation criterion. Three validation criteria were used ($k = 5$ cross validation, $k = 10$ cross validation, and hold-out validation) to evaluate the performance of the algorithms.

The k - cross validation involved partitioning the data set into k subsets which were used to train and validate the model. In the case of hold-out validation, the data was divided into two groups, for training and validation.

The testing phase is crucial in refining the model. Testing these ML algorithm is achieved through running the already trained model on a new dataset. These ML algorithms were fed with the 20% data for testing and subsequently classified based on previous learning experience.

To ensure consistency and avoid manual bias, the default model configurations were adopted for each algorithm, followed by performance-based selection using cross-validation and hold-out validation techniques.

For neural network, the hyperparameters adjusted were the number of hidden layers and the number of neurons per layer, corresponding to MATLAB's predefined narrow, medium, wide, bilayered, and trilayered neural network architec-

tures. The wide neural network employed a single hidden layer with 100 neurons, while the trilayered neural network consisted of three hidden layers with progressively increasing neuron counts.

Table 3. Machine learning algorithms [52].

Algorithm	Sub-Algorithm	Description
Tree	Fine Tree	Consist of numerous small leaves and the maximum number of splits are 100
	Medium Tree	Smaller number of leaves are required and the maximum number of splits are 20
SVM	Linear SVM	Low classification accuracy though less complex
	Quadratic SVM	Provides medium flexibility and are complex to interpret.
	Fine Gaussian SVM	High model flexibility and allows rapid variations in the response function.
	Medium Gaussian SVM	It gives a less flexible response function.
KNN	Fine KNN	Provides a finely detailed distinctions between classes and number of neighbours is set to 1.
	Weighted KNN	Medium distinctions between classes, using a distance weight and the number of neighbours is set to 10.
Ensemble	Boosted Tree	Comprises of shallow trees which uses relatively lower memory and time
	Bagged Tree	Bootstrap aggregation using deep trees and are complex to build.
	RUSBoosted Tree	Random Under sampling Boosting, suited for binary and multiclass classification.
Neural Network	Narrow Neural Network	Increases with the first layer size setting with focus on the depth
	Medium Neural Network	Creates a balance between the depth and the breath.
	Wide Neural Network	Increases the first layer size setting by expanding the breath of each layer
	Bilayered Neural Network	Increases the first layer size and Second layer size settings
	Trilayered Neural Network	Increases with the first layer size, second layer size, and third layer size settings

For SVM classifiers, kernel functions (linear, quadratic, and Gaussian) were selected based on predefined configurations, with kernel scale set to automatic. Tree-based models used predefined maximum split limits, while ensemble models employed bagging and boosting strategies. Model selection was performed strictly based on validation accuracy and generalization performance, rather than hyperparameter, to maintain model robustness and reproducibility.

In the deployment phase, the best performing algorithms were exported as

models which were used for fault diagnosis on the PV system; and can be subsequently applied to new datasets for fault diagnosis in a predictive maintenance system.

3. Results and Discussions

3.1. Machine Learning Algorithm for Fault Diagnosis in PV Systems

Table 4 presents the classification accuracy results obtained for the five ML algorithms and 16 sub algorithms for the cases of $k = 5$ and $k = 10$ cross validation and holdout validation.

The neural network algorithm had the best accuracy compared to the other algorithms. Accuracy in training and testing was done to select the best algorithm amongst the neural network algorithms. During testing, neural network obtained best accuracy of 98.88% and 98.23% using k fold validation of 5 and holdout respectively. While reported accuracies in the literature vary depending on dataset composition, PV configuration, and evaluation protocol, the present results demonstrate consistent performance under a controlled multi-fault dataset such as those obtained in Kurukuru *et al.* [41] where ANN achieved 93.40% accuracy during training.

From **Table 4**, the trees, ensembles and neural networks obtained were the promising models due to their accuracy score. The best performing algorithms are represented in **Table 5**, with respect to the training and testing accuracy.

Table 4. Classification accuracy.

Algorithm	Sub Algorithm	$K = 5$ Accuracy (%)	$K = 10$ Accuracy (%)	Holdout Validation Accuracy (%)
Tree	Fine tree	85.71	84.11	85.71
	Medium tree	59.23	61.48	59.23
SVM	Linear SVM	56.98	53.13	52.97
	Quadratic SVM	56.98	53.93	58.43
	Fine Gaussian SVM	57.30	53.93	57.30
	Medium Gaussian SVM	48.96	48.31	48.96
KNN	Fine KNN	60.35	56.66	60.35
	Weighted KNN	61.16	57.95	61.16
Ensemble	Boosted tree	69.98	68.06	69.98
	Bagged tree	94.54	93.90	95.02
	RUSBoosted Tree	66.93	67.09	66.45
Neural Network	Narrow Neural Network	95.99	94.38	98.23
	Medium Neural Network	97.91	97.27	98.23
	Wide Neural Network	98.88	98.07	97.43
	Bilayered Neural Network	91.81	96.31	97.27
	Trilayered Neural Network	68.38	97.27	94.38

Accuracy in Training and Testing: The accuracy in training and testing of the best performing sub neural network algorithm (Trilayered, narrow, medium, and wide neural network) were compared with respect to the validation criteria (cross and holdout validations) as presented in **Table 5**. The neural network algorithm demonstrated significant performance in fault diagnosis due to its multilayer architecture, which enhances both learning capacity and predictive accuracy. The wide neural network configuration, with a layer size of 100, yielded the best accuracy among the models tested. The increased width allowed the network to capture complex patterns within the input features, making it particularly effective for fault diagnosis.

Table 5. Accuracy in training and testing models.

Validation Criteria	Parameter	ML Algorithm	Accuracy (%)	
			Training	Testing
Cross	$k = 5$	Wide NN	97.79	98.88
		Medium NN	97.31	97.91
	$k = 10$	Wide NN	98.03	98.07
		Medium NN	97.11	97.27
Hold out validation	80:20	Medium/Narrow NN	97.43	98.23
		Wide NN	97.27	97.43

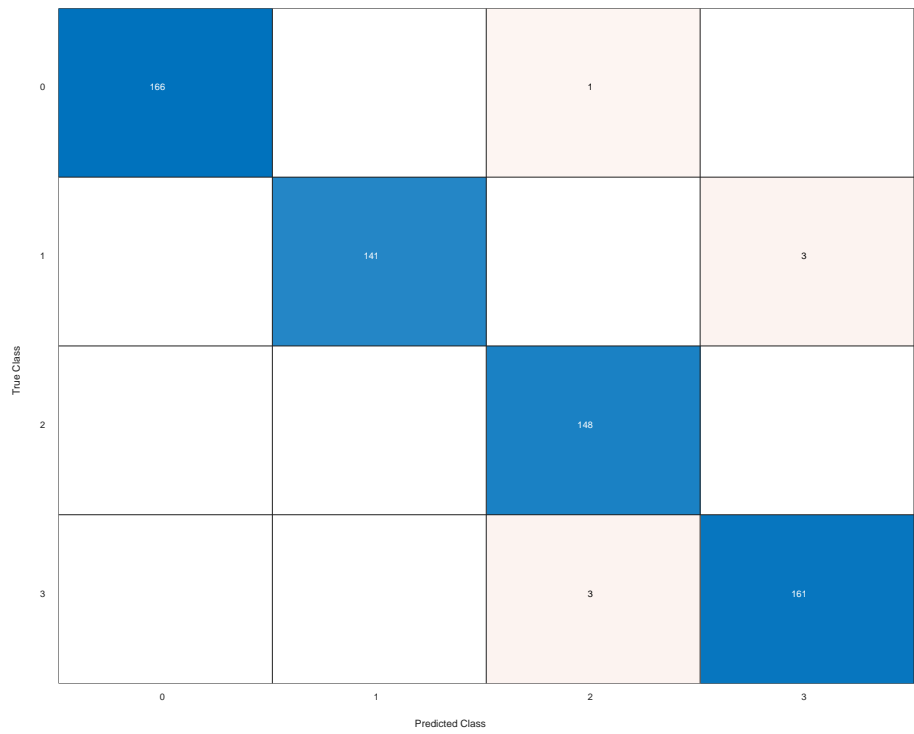
Table 6 shows the hyperparameter settings for the neural network presets. All models have a single fully connected hidden layer with ReLU activation and an iteration limit of 1000, guaranteeing convergence for all models. The only difference between models is the number of neurons in the first layer, with 100 neurons for the Wide NN, while the Medium and Narrow NN have only 25 and 10 neurons, respectively. The improved performance of the Wide NN is due to its greater representational power, allowing for better separation of overlapping fault classes within the three-feature electrical space.

Table 6. Hyper parameter tuning for the Wide NN.

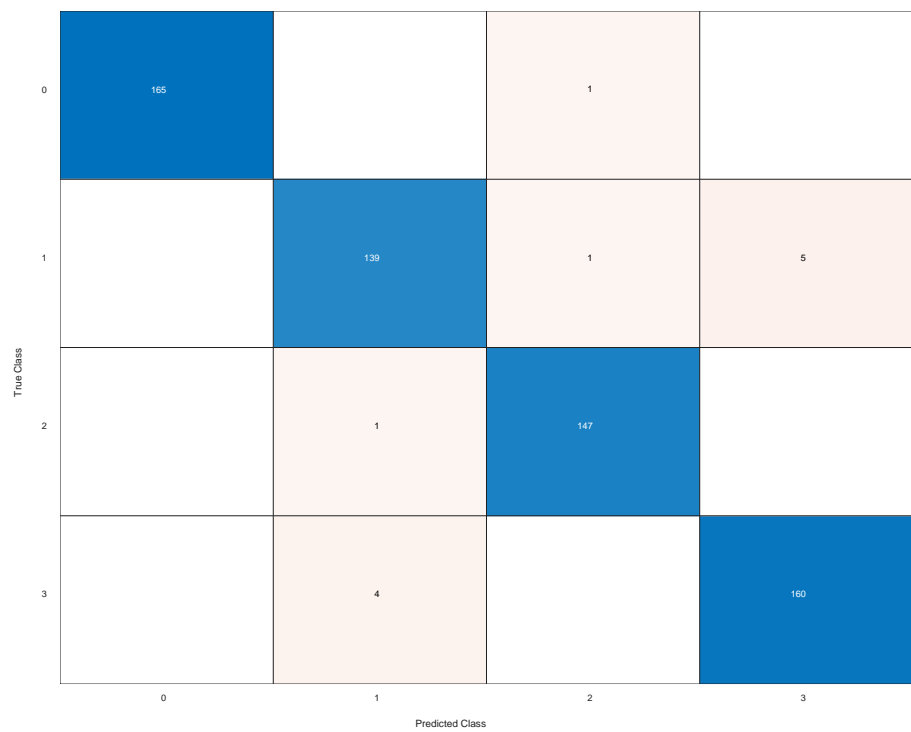
Preset	Wide NN	Medium NN	Narrow NN
Number of fully connected layers	1	1	1
First layer size	100	25	10
Activation	ReLU	ReLU	ReLU
Iteration limit	1000	1000	1000

Confusion Matrix: The performances of the above three ML algorithms were further examined using the confusion matrix measure. A confusion matrix uses two parameters, true class and predicted class to represent the misclassification

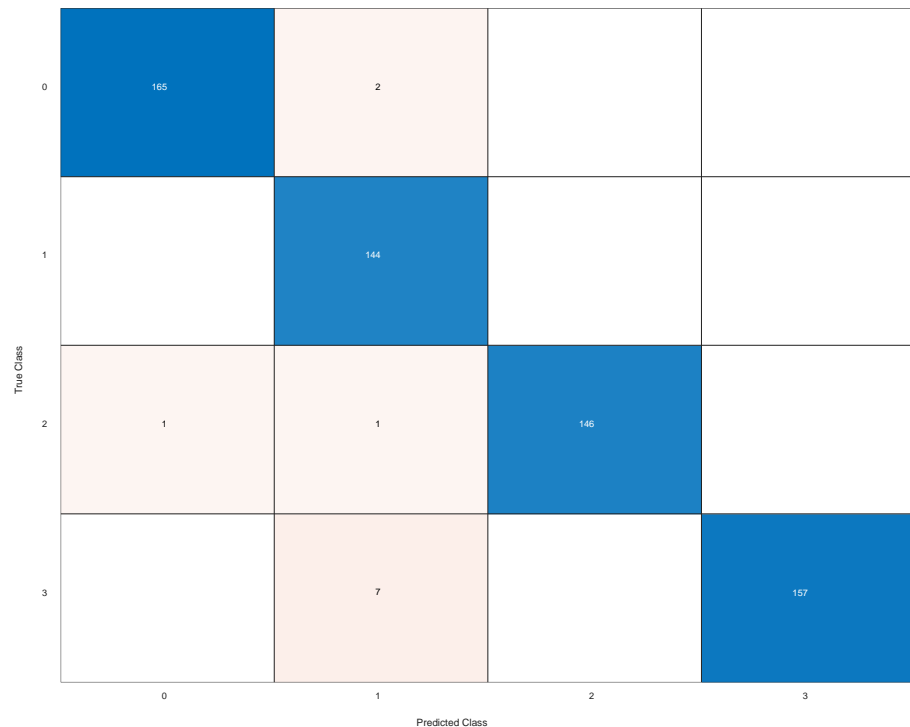
cost. **Figures 7(a)-(c)** show the confusion matrices of the best algorithm. The confusion matrices indicate the classification accuracy with respect to a defined fault category.



(a)



(b)



(c)

Figure 7. Confusion matrices of; (a) Wide Neural Networks ($k=5$), (b) Wide Neural Networks ($k=10$), and (c) Medium Neural Network (hold out).

Based on training and testing, the Wide Neural Network ($k=5$) emerged as best performing algorithm. The model was subsequently exported for use in the prediction of faults.

3.2. Simulated PV System and Fault Diagnosis

3.2.1. P-V Curves of the PV System

The P-V curves portray the variation of power against the voltage of the PV array. The curves in **Figures 8(a)-(d)** were obtained at standard conditions of 1000 W/m^2 and 25°C . These curves indicate the variation of power to the voltage for the fault-free and three fault scenarios. **Figure 8(a)** is the curve for a fault-free PV system. **Figure 8(b)** presents the scenario with a string fault with different resistances used: case 1, 5 Ohms; case 2, 10 Ohms; and case 3, 20 Ohms. **Figure 8(c)** presents the scenario with a string-to-string fault. **Figure 8(d)** refer to instances of different shading uses: case 1, one string is partially shaded by 60%; and case 2, two strings are partially shaded by the same percentage.

Figure 8(a) shows the P-V curve for a fault-free PV system, with maximum power of 7.5 kW which serves as a standard to compare the power output whenever a fault occurs. Based on previous works, string faults have been modelled using 10Ω or 20Ω independently [45] [48]. In **Figure 8(b)** a string fault modelled at three instances, representing resistances of 5Ω , 10Ω , and 20Ω respectively. The highest power loss was experienced in case 1 (5Ω) with the power dropping

down by 4 kW.

The analysis of the curves indicates the influence of increasing resistance on the system's maximum power and voltage at the maximum power point (MPP). A key observation is the progressive steepening of the negative slope in the descending region of the P-V curve as resistance increases. In case 3, the negative slope is the steepest, indicating a rapid and abrupt power drop-off after the MPP. As a result, while the MPP is higher compared to lower resistance cases, the system fails rapidly beyond this point, leading to a sharper transition from functional operation to power loss.

In **Figure 8(c)**, a low resistance path was created between two strings (string-to-string) in the array, leading to a power loss of 1.5 kW.

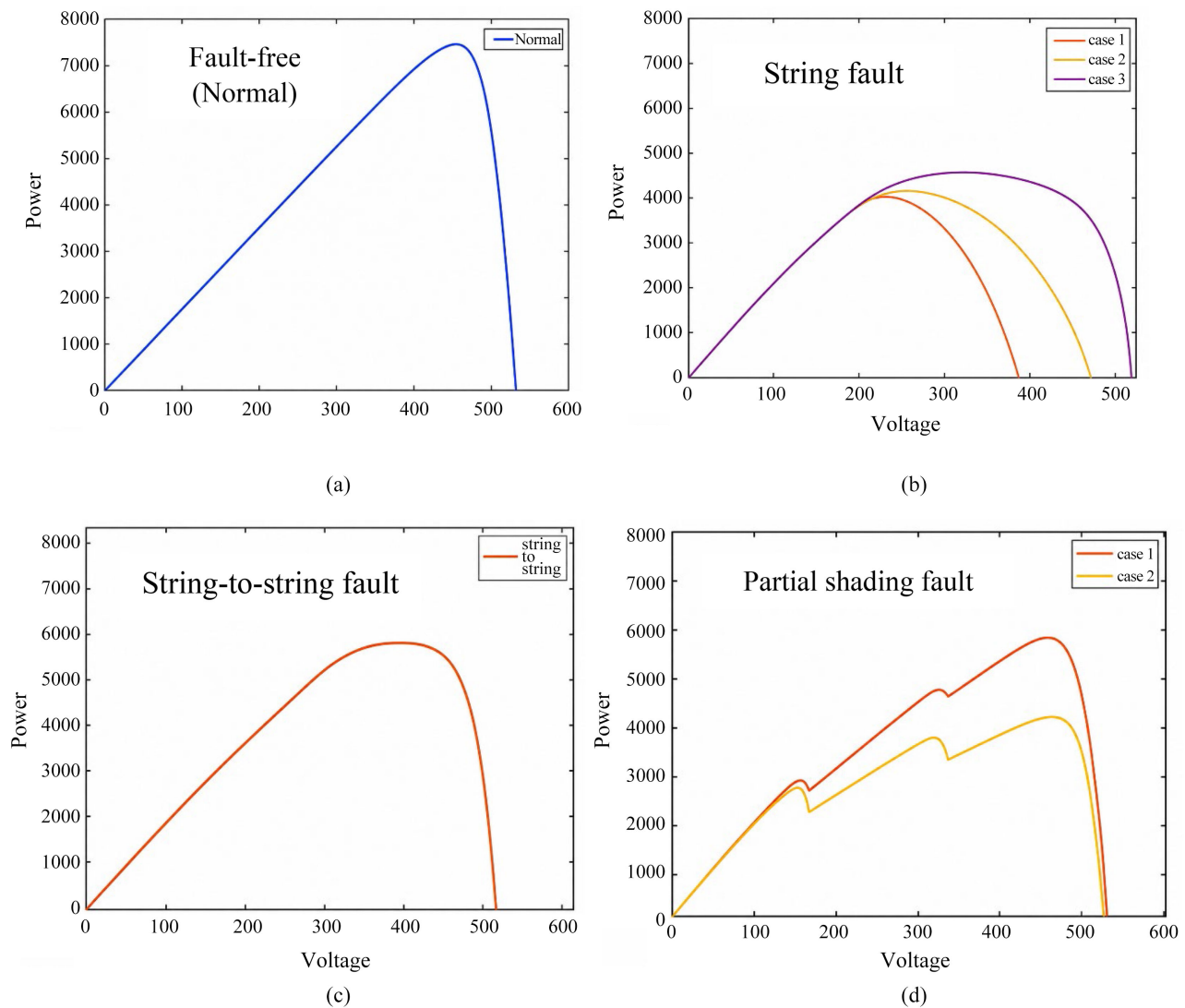


Figure 8. P-V Curves: (a) fault-free, (b) string fault, (c) string-to-string fault, (d) partial shading fault.

In the partial shading fault in **Figure 8(d)**, two cases of partial shading were simulated. In the first one, one string is partially shaded by 60%, and in the second

case, two strings are partially shaded by the same percentage. The curves illustrate the fluctuating power output under partial shading conditions. The nature of the curves shows that the output does not go above normal and creates multiple MPPs. A varying power output hurts the supply side of the systems. These results align with Badr *et al.* [9] here fault cases were compared to normal case and results show non-uniform P-V curve under shading conditions.

3.2.2. I-V Curves of the PV System

Figure 9(a) is the curve for a fault-free PV system. **Figure 9(b)** presents the scenario with a string fault with different resistances used: case 1, 5 Ohms; case 2, 9 Ohms; and case 3, 20 Ohms. **Figure 9(c)** presents the scenario with a string-to-string fault. **Figure 9(d)** refer to instances of different shading uses: case 1, one string is partially shaded by 60%; and case 2, two strings are partially shaded by the same percentage.

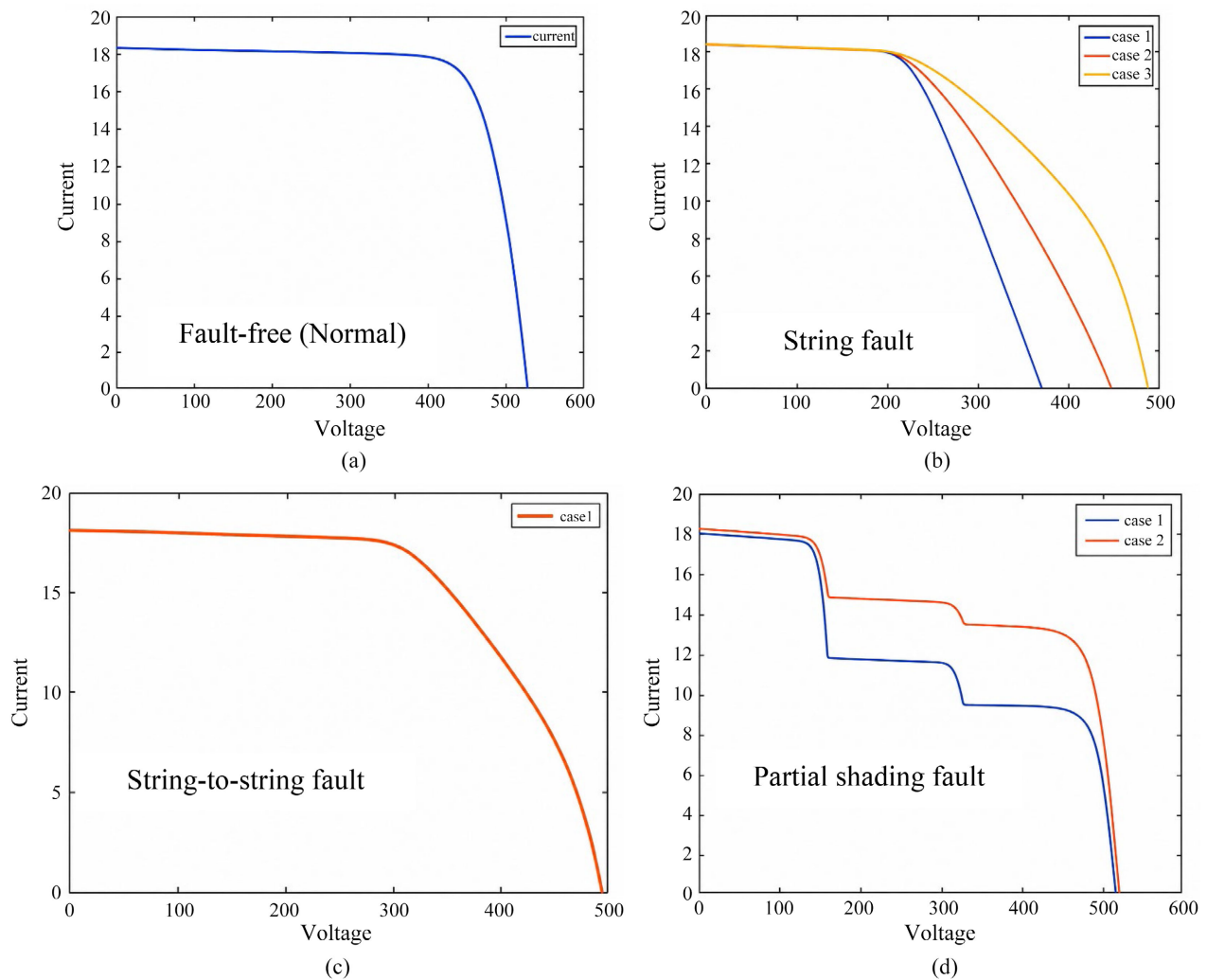


Figure 9. I-V Curves: (a) fault-free, (b) string fault, (c) string-to-string fault, (d) partial shading fault.

In **Figure 9(a)** the curve begins at a constant current, indicating that the overall

irradiation and system conditions remain constant. In **Figure 9(b)** case 1, the steep slope illustrates a sharpest decline in current as voltage increases. This illustrates that at low resistance, a significant amount of current is diverted into the unintended fault path, leading to higher power dissipation and losses. The reduced slope steepness in case 2 suggests that the array retains more of its current output, making the power degradation less severe. Case 3 exhibits the least steep decline in current beyond the maximum power point, meaning that higher resistance limits fault current significantly. In **Figure 9(d)**, a drop in irradiance results in an instant drop in the current generated

4. Conclusions

The occurrence of faults in PV systems undermines their usefulness; making it an imperative to restore functionalities whenever faults occur. In this study, 16 machine learning algorithms, based on five main algorithms were trained, tested and validated using cross-validation and holdout validations. During training, the Wide Neural Network obtained the best accuracy of 98.03%. In the testing phase, Wide Neural Network sub-algorithms had the best accuracy of 98.88%. These results were obtained from the simulation of a 7.5 kW PV system, made up of three strings containing six modules each. Three faults were simulated; string fault, string-to-string fault, and partial shading fault. The effects of the faults were illustrated by P-V and I-V curves for each case. The string fault was modelled at three instances, representing resistances of 5 Ω , 10 Ω , and 20 Ω respectively. The highest power loss was experienced in 5 Ω with the power dropping down by 4 kW. A low resistance path was created between two strings (string-to-string) in the array leading to a power loss of 1.5 kW. Two cases of partial shading demonstrated varying power peaks with an overall decline in energy production.

The dataset used in this study was obtained through controlled modelling of a 7.5 kW PV system using MATLAB. Although simulation models cannot replicate the stochastic variability of real-world PV systems, they offer vital advantages in controlled fault analysis. Simulation models allow for controlled electrical fault injection, which can include low-resistance string faults and string-to-string faults. These types of electrical faults can pose safety hazards in real-world scenarios. In addition, simulation models enable controlled fault severity variation, ensuring that the dataset used in supervised learning is balanced and well-classified. However, it is also acknowledged that real-world PV systems can experience measurement noise, irradiance changes, sensor inaccuracies, and component aging. These factors can impact the performance of the classifiers. The performance of the classifiers in this study was therefore under controlled electrical conditions.

Future work will therefore focus on validating the proposed framework using experimentally acquired datasets, expanding the fault taxonomy to include degradation and arc-related conditions, and investigating the integration of dynamic features for improved robustness under real-world variability.

Acknowledgements

The authors are grateful to the Responsible Artificial Intelligence Network for Climate Action in Africa (RAINCA) Consortium, made up of the Regional Universities Forum for Capacity Building in Agriculture (RUFORUM), the West African Science Centre on Climate Change and Adapted Land Use (WASCAL), and AKADEMIYA2063 for providing funding for this research with the support of IDRC (Grant #: 109705-001/002). The Authors acknowledge the institutional support of The University of Bamenda, Cameroon.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Zohuri, B. (2023) Navigating the Global Energy Landscape Balancing Growth, Demand, and Sustainability. *Journal of Material Science & Manufacturing Research*, **2**, 1-7.
- [2] Chandel, R., Chandel, S.S., Prasad, D. and Dwivedi, R.P. (2024) Sustainable Passive Solar and Photovoltaic Integrated Technology Interventions for Climate Responsive Net Zero Energy Buildings in Western Himalayan Mountainous Terrain of India. *Next Sustainability*, **3**, Article 100039. <https://doi.org/10.1016/j.nxsust.2024.100039>
- [3] Tchao, E.T., Asakipaam, S.A., Fiagbe, Y.A.K., Yeboah-Akowitz, B., Ramde, E., Agbemenu, A.S., *et al.* (2022) An Implementation of an Optimized Dual-Axis Solar Tracking Algorithm for Concentrating Solar Power Plants Deployment. *Scientific African*, **16**, e01228. <https://doi.org/10.1016/j.sciaf.2022.e01228>
- [4] International Energy Agency (IEA) (2022) Africa Energy Outlook 2022. <https://www.iea.org/reports/africa-energy-outlook-2022>
- [5] Abdulai, D., Gyamfi, S., Diawuo, F.A. and Acheampong, P. (2023) Data Analytics for Prediction of Solar PV Power Generation and System Performance: A Real Case of Bui Solar Generating Station, Ghana. *Scientific African*, **21**, e01894. <https://doi.org/10.1016/j.sciaf.2023.e01894>
- [6] Mellit, A. and Kalogirou, S.A. (2008) Artificial Intelligence Techniques for Photovoltaic Applications: A Review. *Progress in Energy and Combustion Science*, **34**, 574-632. <https://doi.org/10.1016/j.pecs.2008.01.001>
- [7] Mellit, A. and Kalogirou, S. (2022) Handbook of Artificial Intelligence Techniques in Photovoltaic Systems Modelling, Control, Optimization, Forecasting and Fault Diagnosis. Academic Press.
- [8] International Energy Agency Photovoltaic Power Systems Programme (IEA-PVPS) (2021). Snapshot of Global PV Markets 2021. IEA-PVPS. <https://iea-pvps.org/snapshot-reports/snapshot-2021/>
- [9] Badr, M.M., Hamad, M.S., Abdel-Khalik, A.S., Hamdy, R.A., Ahmed, S. and Hamdan, E. (2021) Fault Identification of Photovoltaic Array Based on Machine Learning Classifiers. *IEEE Access*, **9**, 159113-159132. <https://doi.org/10.1109/access.2021.3130889>
- [10] Appiah, A.Y., Zhang, X., Ayawli, B.B.K. and Kyeremeh, F. (2019) Review and Performance Evaluation of Photovoltaic Array Fault Detection and Diagnosis Techniques. *International Journal of Photoenergy*, **2019**, 1-19. <https://doi.org/10.1155/2019/6953530>

- [11] Alam, M.K., Khan, F., Johnson, J. and Flicker, J. (2015) A Comprehensive Review of Catastrophic Faults in PV Arrays: Types, Detection, and Mitigation Techniques. *IEEE Journal of Photovoltaics*, **5**, 982-997. <https://doi.org/10.1109/jphotov.2015.2397599>
- [12] AbdulMawjood, K., Refaat, S.S. and Morsi, W.G. (2018) Detection and Prediction of Faults in Photovoltaic Arrays: A Review. 2018 *IEEE 12th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG2018)*, Doha, 10-12 April 2018, 1-8. <https://doi.org/10.1109/cpe.2018.8372609>
- [13] Groenendaal, B. and Gifford, J. (2020) Reports of Backsheet Failures at 75 MW Mulilo Sonnedix Prieska Solar Farm in South Africa. *PV Magazine*. <https://www.pv-magazine.com/2020/02/11/reports-of-backsheet-failures-at-75-mw-mulilo-sonnedix-prieska-solar-farm-in-south-africa/#:~:text=The%20South%20African%20project%20is%20reportedly%20experiencing%20serious.Chinese%20manufacturer%20BYD%20and%20South%20African%20producer%20Artsolar>
- [14] Ezugwu, A.E., Oyelade, O.N., Ikotun, A.M., Agushaka, J.O. and Ho, Y. (2023) Machine Learning Research Trends in Africa: A 30 Years Overview with Bibliometric Analysis Review. *Archives of Computational Methods in Engineering*, **30**, 4177-4207. <https://doi.org/10.1007/s11831-023-09930-z>
- [15] Olawade, D.B., Wada, O.Z., David-Olawade, A.C., Fapohunda, O., Ige, A.O. and Ling, J. (2024) Artificial Intelligence Potential for Net Zero Sustainability: Current Evidence and Prospects. *Next Sustainability*, **4**, Article 100041. <https://doi.org/10.1016/j.nxsust.2024.100041>
- [16] Voutsinas, S., Karolidis, D., Voyiatzis, I. and Samarakou, M. (2023) Development of a Machine-Learning-Based Method for Early Fault Detection in Photovoltaic Systems. *Journal of Engineering and Applied Science*, **70**, Article No. 27. <https://doi.org/10.1186/s44147-023-00200-0>
- [17] Alrahim Shannan, N.M.A., Yahaya, N.Z. and Singh, B. (2013) Single-Diode Model and Two-Diode Model of PV Modules: A Comparison. 2013 *IEEE International Conference on Control System, Computing and Engineering*, Penang, 29 November-1 December 2013, 210-214. <https://doi.org/10.1109/iccsce.2013.6719960>
- [18] Tiwari, G.N., Tiwari, A. and Shyam, (2016) Solar Radiation. In: Tiwari, G.N., Tiwari, A. and Shyam, Eds., *Handbook of Solar Energy*, Springer, 1-49. https://doi.org/10.1007/978-981-10-0807-8_1
- [19] Luque, A. and Hegedus, S. (2011) *Handbook of Photovoltaic Science and Engineering*. 2nd Edition, Wiley.
- [20] Fazai, R., Abodayeh, K., Mansouri, M., Trabelsi, M., Nounou, H., Nounou, M., et al. (2019) Machine Learning-Based Statistical Testing Hypothesis for Fault Detection in Photovoltaic Systems. *Solar Energy*, **190**, 405-413. <https://doi.org/10.1016/j.solener.2019.08.032>
- [21] Naser Kabour, S., Adnan Almalki, R., A. Alghamdi, L., Alharthi, W. and B. Alshagi, N. (2023) Fault Classification and Detection for Photovoltaic Plants Using Machine Learning Algorithms. *Indonesian Journal of Electrical Engineering and Computer Science*, **32**, Article 353. <https://doi.org/10.11591/ijeecs.v32.i1.pp353-362>
- [22] Lipták, R. and Bodnár, I. (2021) Simulation of Fault Detection in Photovoltaic Arrays. *Analecta Technica Szegedinensia*, **15**, 31-40. <https://doi.org/10.14232/analecta.2021.2.31-40>
- [23] Khalil, I.U., Ul-Haq, A., Mahmoud, Y., Jalal, M., Aamir, M., Ahsan, M.U., 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>
- [24] Venkatakrishnan, G.R., Rengaraj, R., Tamilselvi, S., Harshini, J., Sahoo, A., Saleel, C.A., et al. (2023) Detection, Location, and Diagnosis of Different Faults in Large Solar PV System—A Review. *International Journal of Low-Carbon Technologies*, **18**, 659-674. <https://doi.org/10.1093/ijlct/ctad018>
- [25] Memon, S.A., Javed, Q., Kim, W., Mahmood, Z., Khan, U. and Shahzad, M. (2022) A Machine-Learning-Based Robust Classification Method for PV Panel Faults. *Sensors*, **22**, Article 8515. <https://doi.org/10.3390/s22218515>
- [26] Kase, R. and Nishikawa, S. (2016) Fault Detection of Bypass Circuit of PV Module—Detection Technology of Open Circuit Fault Location. *19th International Conference on Electrical Machines and Systems (ICEMS)*, Chiba, 13-16 November 2016, 1-4.
- [27] Pillai, D.S., Blaabjerg, F. and Rajasekar, N. (2019) A Comparative Evaluation of Advanced Fault Detection Approaches for PV Systems. *IEEE Journal of Photovoltaics*, **9**, 513-527. <https://doi.org/10.1109/jphotov.2019.2892189>
- [28] Piliouline, M., Guejia-Burbano, R.A., Petrone, G., Sánchez-Pacheco, F.J., Mora-López, L. and Sidrach-de-Cardona, M. (2021) Parameters Extraction of Single Diode Model for Degraded Photovoltaic Modules. *Renewable Energy*, **164**, 674-686. <https://doi.org/10.1016/j.renene.2020.09.035>
- [29] Sameera, Tariq, M. and Rihan, M. (2025) Impact of Shading Heaviness on Voltage, Current and Power of the Solar Photovoltaic String. *Next Sustainability*, **5**, Article 100062. <https://doi.org/10.1016/j.nxsust.2024.100062>
- [30] Osmani, K., Haddad, A., Lemenand, T., Castanier, B. and Ramadan, M. (2020) A Review on Maintenance Strategies for PV Systems. *Science of The Total Environment*, **746**, Article 141753. <https://doi.org/10.1016/j.scitotenv.2020.141753>
- [31] Hammoudi, Y., Idrissi, I., Boukabous, M., Zerguit, Y. and Bouali, H. (2022) Review on Maintenance of Photovoltaic Systems Based on Deep Learning and Internet of Things. *Indonesian Journal of Electrical Engineering and Computer Science*, **26**, Article 1060. <https://doi.org/10.11591/ijeecs.v26.i2.pp1060-1072>
- [32] Naimi, A., Deng, J., Doney, P., Sheikh-Akbari, A., Shimjith, S.R. and Arul, A.J. (2022) Machine Learning-Based Fault Diagnosis for a PWR Nuclear Power Plant. *IEEE Access*, **10**, 126001-126010. <https://doi.org/10.1109/access.2022.3225966>
- [33] El-Bouzaidi, Y.E.I. and Abdoun, O. (2023) Advances in Artificial Intelligence for Accurate and Timely Diagnosis of COVID-19: A Comprehensive Review of Medical Imaging Analysis. *Scientific African*, **22**, e01961. <https://doi.org/10.1016/j.sciaf.2023.e01961>
- [34] Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., et al. (2019) Machine Learning Methods for Wind Turbine Condition Monitoring: A Review. *Renewable Energy*, **133**, 620-635. <https://doi.org/10.1016/j.renene.2018.10.047>
- [35] Oudaa, T., Gharsellaoui, H. and Ben Ahmed, S. (2021) An Agent-Based Model for Resource Provisioning and Task Scheduling in Cloud Computing Using DRL. *Procedia Computer Science*, **192**, 3795-3804. <https://doi.org/10.1016/j.procs.2021.09.154>
- [36] Engel, E. and Engel, N. (2022) A Review on Machine Learning Applications for Solar Plants. *Sensors*, **22**, Article 9060. <https://doi.org/10.3390/s22239060>
- [37] Gaboitaolelwe, J., Zungeru, A.M., Yahya, A., Lebekwe, C.K., Vinod, D.N. and Salau, A.O. (2023) Machine Learning Based Solar Photovoltaic Power Forecasting: A Review and Comparison. *IEEE Access*, **11**, 40820-40845. <https://doi.org/10.1109/access.2023.3270041>
- [38] Jo, T. (2021) Machine Learning Foundations: Supervised, Unsupervised, and Ad-

vanced Learning. Springer International Publishing.

- [39] da Costa, C.H., Moritz, G.L., Lazzaretti, A.E., Mulinari, B.M., Ancelmo, H.C., Rodrigues, M.P., et al. (2019) A Comparison of Machine Learning-Based Methods for Fault Classification in Photovoltaic Systems. 2019 *IEEE PES Innovative Smart Grid Technologies Conference—Latin America (ISGT Latin America)*, Gramado, 15-18 September 2019, 1-6. <https://doi.org/10.1109/isgt-la.2019.8895279>
- [40] Ren, B., Chi, Y., Zhou, N., Wang, Q., Wang, T., Luo, Y., et al. (2024) Machine Learning Applications in Health Monitoring of Renewable Energy Systems. *Renewable and Sustainable Energy Reviews*, **189**, Article 114039. <https://doi.org/10.1016/j.rser.2023.114039>
- [41] Kurukuru, V.S.B., Haque, A., Khan, M.A. and Tripathy, A.K. (2019) Fault Classification for Photovoltaic Modules Using Thermography and Machine Learning Techniques. 2019 *International Conference on Computer and Information Sciences (IC-CIS)*, Sakaka, 3-4 April 2019, 1-6. <https://doi.org/10.1109/iccisci.2019.8716442>
- [42] Ncube, P.D.N., Meyer, E.L. and Shibeshi, Z.S. (2021) Towards the Development of a Photovoltaic Array Fault Detection and Diagnosis (PVAFFD) System. 2021 *International Conference on Electrical, Computer and Energy Technologies (ICECET)*, Cape Town, 9-10 December 2021, 1-6. <https://doi.org/10.1109/icecet52533.2021.9698581>
- [43] Eskandari, A., Milimonfared, J. and Aghaei, M. (2021) Fault Detection and Classification for Photovoltaic Systems Based on Hierarchical Classification and Machine Learning Technique. *IEEE Transactions on Industrial Electronics*, **68**, 12750-12759. <https://doi.org/10.1109/tie.2020.3047066>
- [44] Li, B., Delpha, C., Migan-Dubois, A. and Diallo, D. (2021) Fault Diagnosis of Photovoltaic Panels Using Full I-V Characteristics and Machine Learning Techniques. *Energy Conversion and Management*, **248**, Article 114785. <https://doi.org/10.1016/j.enconman.2021.114785>
- [45] Attouri, K., Hajji, M., Mansouri, M., Harkat, M., Kouadri, A., Nounou, H., et al. (2020) Fault Detection in Photovoltaic Systems Using Machine Learning Technique. 2020 *17th International Multi-Conference on Systems, Signals & Devices (SSD)*, Monastir, 20-23 July 2020, 79-84. <https://doi.org/10.1109/ssd49366.2020.9364094>
- [46] Basnet, B., Chun, H. and Bang, J. (2020) An Intelligent Fault Detection Model for Fault Detection in Photovoltaic Systems. *Journal of Sensors*, **2020**, 1-11. <https://doi.org/10.1155/2020/6960328>
- [47] Yilmaz, M. and Martinez-Morales, A.A. (2024) Fault Detection in Photovoltaic Panels Using Digital Twin Technology: A Comprehensive Study. *The Journal of Cognitive Systems*, **8**, 28-32. <https://doi.org/10.52876/jcs.1407133>
- [48] Badr, M.M., Abdel-Khalik, A.S., Hamad, M.S., Hamdy, R.A., Hamdan, E., Ahmed, S., et al. (2023) Intelligent Fault Identification Strategy of Photovoltaic Array Based on Ensemble Self-Training Learning. *Solar Energy*, **249**, 122-138. <https://doi.org/10.1016/j.solener.2022.11.017>
- [49] Chen, Z., Han, F., Wu, L., Yu, J., Cheng, S., Lin, P., et al. (2018) Random Forest Based Intelligent Fault Diagnosis for PV Arrays Using Array Voltage and String Currents. *Energy Conversion and Management*, **178**, 250-264. <https://doi.org/10.1016/j.enconman.2018.10.040>
- [50] Kapucu, C. and Cubukcu, M. (2021) A Supervised Ensemble Learning Method for Fault Diagnosis in Photovoltaic Strings. *Energy*, **227**, Article 120463. <https://doi.org/10.1016/j.energy.2021.120463>
- [51] Teo, J.C., Tan, R.H.G., Mok, V.H., Ramachandramurthy, V.K. and Tan, C. (2018)

Impact of Partial Shading on the P-V Characteristics and the Maximum Power of a Photovoltaic String. *Energies*, **11**, Article 1860. <https://doi.org/10.3390/en11071860>

- [52] MATLAB (2024) Statistics and Machine Learning Toolbox™ User's Guide. <https://www.mathworks.com/help/stats/index.html>