Accepted manuscript (author version)

To appear in: **Majlesi Journal of Electrical Engineering (MJEE)**

Online ISSN: 2345-377X **Print ISSN:** 2345-3796

This PDF file is not the final version of the record. This version will undergo further copyediting, typesetting, and production review before being published in its definitive form. We are sharing this version to provide early access to the article. Please be aware that errors that could impact the content may be identified during the production process, and all legal disclaimers applicable to the journal remain valid.

Received: 18/04/2025	**Revised:** 18/06/2025
Accepted: 10/07/2025	
DOI: 10.57647/j.mjee.2025.17054	

Original Research

Photovoltaic Array Fault Detection Using Ensemble Learning-Based Technique

Authors:

Anshul Shekhar[1,2,*], M. Senthil Kumar[1]

[1] Electrical Engineering, National Institute of Technology, Patna, India.

[2] Electrical Engineering, Government Engineering College, Bhojpur, India.

** Corresponding Author:* shekhar.anshul@gmail.com 0000-0003-4662-775X

© Author(s) 2025

Abstract

Fault detection in the photovoltaic array aims to ensure a stable and continuous power supply. Detecting faults in photovoltaic arrays is challenging because normal and faulty conditions can sometimes exhibit similar characteristics. This paper presents an approach for photovoltaic array fault detection using an ensemble learning-based technique. A 3.2 kW MATLAB-Simulink photovoltaic array model is developed, and the fault characteristics of short circuit faults, line-ground faults, and hot spot faults are analyzed to identify the most suitable measurements for effective fault detection and classification. It is observed that photovoltaic array measurements such as voltage, current, power, rate of change of voltage and current over time (dV/dt and dI/dt), and change in power to voltage and current (dP/dV and dP/dI) exhibit distinct fault characteristics, making them valuable for accurate fault discrimination. The proposed model is trained using the selected photovoltaic array measurements, and its effectiveness is validated through a testing dataset, with performance indices derived from the confusion matrix. The proposed technique achieves a promising fault detection accuracy of 99.72%. Additionally, the performance of the proposed technique is evaluated and compared with other artificial intelligence-based techniques. The results demonstrate that the proposed method outperforms these alternatives.

Accepted Manuscript: Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

Keywords: Ensemble learning, Bagging, Fault detection, Photovoltaic fault characteristics, Photovoltaic system

1. Introduction

Power generation through solar photovoltaic (PV) arrays is the most favorable source of renewable energy. Recently, the rapid growth of PV plants has been driven by the availability of PV systems with diverse capacities suited for domestic, commercial, and industrial applications. Over the past decade, global installations of solar PV plants have increased by 7.66 [1]. This growing demand has contributed to the production of more environmentally friendly and efficient electricity. Although PV power generation is intermittent, advancements in technology continuously aim to ensure a reliable power supply [2]. However, operational challenges, such as shading, temperature fluctuations, dust accumulation on the PV array, and faults within the PV array, can disrupt consistent PV power generation [3]. Among these, faults in the PV array are a key factor affecting the reliability of PV power generation. The most common faults in PV arrays include open circuit faults (OCF), line-ground faults (LGF), hot spot faults (HSF), short circuit faults (SCF), and PV array degradation [4]. These faults interrupt power production and hinder the reliable supply of electricity. Therefore, it is essential to implement robust protection mechanisms to safeguard PV plants from malfunctions and enable rapid restoration of power generation.

Conventional methods such as circuit interrupter fuses (CIF), overcurrent protective devices (OCPD), and residual current devices (RCD) have been proposed for PV array fault detection [5]. While CIF and OCPD are cost-effective and simple to integrate into the system, they respond slowly to low fault currents and have limited sensitivity to high-impedance faults. Similarly, RCD has the drawback of being susceptible to false operation due to external noise and is only reliable when used in conjunction with CIF [6]. Consequently, protecting a single PV system requires multiple devices, and even then, their dependability is not guaranteed; therefore, alternate protection methods are adopted. These methods use the measured values of PV parameters to diagnose the faults. The author in [7] introduces a fault detection method by measuring voltage, current, and temperature at the PV module level and employs PLC PLC-based monitoring system to track individual module operations and detect faults. It compares measured and simulated DC power values to identify faults. However, there is a scope for reducing the cost and complexities by implementing a lower sensor count, smart microcontrollers, and simple detection algorithms. In [8], the author proposed a Cross-Domain Adaptive Generative Adversarial Network (CDAGAN) for fault diagnosis of PV arrays with varying degradation levels. The CDAGAN can maintain good performance despite the absence of fault data and domain shifts in the target domain. The study assumes alignment of fault types between source and target domains; however, new unseen faults may arise in real-world scenarios, necessitating further research. In [9], the author investigates the reliability of various solar PV array configurations under multiple electrical faults using MATLAB and prototype experiments, and develops a low-cost monitoring system for fault detection and notification. The series-parallel configuration has a higher average tolerance to electrical faults compared to honeycomb, bridge-linked, and total-cross-tied configurations. The proposed system effectively monitors system performance and detects faults in PV arrays. However, the current protection standards are limited in their applicability to the DC side of PV systems, and the DC side lacks a variety of protection systems, and available systems are costly. Also, standard PCA and PLS techniques are limited by their assumption of singular scale data. In [10], the author proposed a new fault detection and diagnosis method that combines an improved slime mould algorithm (ISMA) with a CatBoost model based on a polyloss function (PolyCatBoost). The study compares the performance of the ISMA-PolyCatBoost model with other machine learning algorithms such as random forest, K-nearest neighbors, decision tree, and CatBoost. However, there is scope to

Accepted Manuscript: Author Version

 This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

improve the accuracy, minimize the standard deviation, and reduce losses. Also, the model may be optimized further to cope with the complexities, such as missing data labels and extreme weather conditions.

Advanced technologies aided by artificial intelligence (AI) have opened new avenues for PV array fault identification and classification. AI techniques, including decision trees, random forest (RF), support vector machine (SVM), k-nearest neighbors (kNN), and artificial neural networks (ANN), have been implemented in various ways for PV array fault diagnosis in the literature [11]-[15]. In [11], a fault detection and classification model based on decision trees was explored to identify and classify line-to-line faults (LLF), shading faults, and open-circuit faults. The model was trained using real-world data from a PV system, achieving fault detection and classification accuracies varying under different conditions from 93.56% to 99.98% and 85.43% to 99.8%, respectively, on test data. However, the authors noted that variation in accuracy and the requirement for a large dataset limited the application of the implemented technique. In [12], a novel fault diagnosis method based on random forests was proposed to detect LLF, OCF, partial shading, and degradation faults in PV arrays. The method utilizes PV-array voltage and individual PV-string currents as fault features, achieving a fault detection accuracy of 99.24% and a classification accuracy of 99.13%. However, sampling and feature selection may affect the result. In [13], the author discussed the SVM-based fault classification technique, employing wavelet transform to decompose the collected voltage and current signals, with the resulting wavelet coefficients used to train the SVM classifier. The implemented method is economical due to the presence of low-cost sensors and achieves an accuracy of 91.40% under 25% mismatch conditions and 94.74% under 50% mismatch conditions. However, it is limited to identifying only line-to-line faults (LLF), and there is scope for improvement in accuracy. The kNN-based method presented in [14] achieves a fault detection accuracy of 98.7%, but it requires substantial data acquisition and additional signal processing tools, hence increasing the cost and complexity. The ANN-based fault classification technique discussed in [15] classifies HSF, SCF, and LGF in a 3.2 kW PV array using four distinct parameters: dI/dt, dV/dt, dP/dI, and dP/dV, achieving a fault classification accuracy of 99.7%. However, it will be interesting to check the performance of the proposed method with a large number of datasets. The fault detection technique discussed in [16] uses a one-dimensional convolutional neural network (1D-CNN) technique with Internet of Things (IoT) to detect LLF, OCF, and shading faults in a 15 kWp grid-connected PV system and achieve 98.15% accuracy during fault. The proposed method entails a high initial cost, may need further calibration, and could require more advanced models for large PV systems.

It has been observed that the AI techniques mentioned here have certain limitations and perform well only for specific types of faults under particular conditions. The ensemble learning-based technique is another machine learning approach applied for fault detection in various domains [17]. In [18], the author proposed an ensemble learning algorithm and a probabilistic strategy for detecting and classifying faults at the DC side of PV systems. Although the proposed technique achieves a high classification accuracy of 99.5%, it is applied to detect only LL faults. The author in [19] uses an optimized ensemble learning technique to classify shading faults and SCF in a PV array, achieving an accuracy of 97.67%. Another study [20] discusses a voting-based ensemble learning algorithm with linear regression, decision tree, and support vector machine for PV fault detection and diagnosis, achieving 100% accuracy while classifying open circuit and line-line faults in a PV array. Studies have demonstrated that ensemble learning techniques excel in classification task applications [21]. Therefore, it motivates us to implement this technique for PV array fault detection and classification with new features. Table 1 presents a comparative overview of various PV array fault detection techniques, highlighting their benefits and limitations.

Table 1: Overview of various PV array fault detection techniques

Reference	Technique	Achievements	Limitations

Accepted Manuscript: Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

[5]	Conventional (CIF, OCPD)	• CIF and OCPD are cost-effective and simple to integrate into the system	• Slow response to low fault currents • Limited sensitivity to high-impedance faults
[6]	Conventional (RCD)	• Simple and cost-effective	• May exhibit false operation during noisy conditions
[7]	Electrical Characteristics Monitoring	• Cost-Effective Monitoring • Reduced Computational Requirements	• Requirement of a large number of sensors for monitoring individual PV modules • Increased cost and complexity
[8]	Electrical Characteristics Monitoring	• Effective Cross-Domain Adaptation • High Diagnostic Performance with improved data quality	• The study assumes aligned fault types between source and target domains, but new unseen faults may occur in real-world scenarios • Increases computational Complexity
[9]	Electrical Characteristics Monitoring	• Efficient Fault Detection and Monitoring • Includes a variety of possible faults in the PV system	• Increased System Complexity • Protection devices can fail due to distorted characteristic curves and other factors.
[10]	Improved slime mould algorithm (ISMA)	• Efficient Handling of Multi-Coupled Faults • Enhanced Diagnostic Accuracy	• The effectiveness of the model heavily relies on the quality of training data • Scope to improve the accuracy, minimize the standard deviation, and reduce losses
[11]	Decision Tree	• Effective for detecting LLF, OCF, and shading faults	• Large variation in accuracy • Need extensive datasets
[12]	Random Forest	• Effective for detecting LLF, OCF, partial shading, and degradation faults in PV arrays	• Sampling and feature selection may affect the result
[13]	SVM with Wavelet Transform	• Economical as it requires few sensors for parameter measurements	• Can not detect other faults except LLF • The fault detection accuracy may be improved
[14]	kNN	• Efficient for fault classification	• Requires a high number of tools for data acquisition and signal processing
[15]	ANN	• Provides options for multiple parameters for identifying PV array faults • Can detect important faults such as SCF, HSF, and LGF	• May be tested with a larger number of datasets
[16]	1D- CNN	• Efficient in detecting faults in the PV system	• High upfront cost • May require further calibration and the use of more advanced models for large PV systems
[18]	Ensemble Technique	• Achieve high accuracy	• Limited to detecting only LLF
[19]	Ensemble Technique	• Improve fault detection efficiency and accuracy	• Limited to detecting shading faults and SCF
[20]	Voting-based ensemble learning	• Includes linear regression, decision tree, and support vector machine	• Limited to detecting only shading faults

Accepted Manuscript: Author version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

		• Very high accuracy	

The literature review suggests that each technique has its own advantages and limitations. Existing techniques often struggle with low accuracy under varying environmental conditions, limited real-time applicability, and high computational costs. Additionally, many techniques lack robustness in distinguishing between different fault types, highlighting the need for more reliable, efficient, and adaptive fault detection methods. Hence, to address the limitations of various fault detection techniques, an ensemble-based PV array fault detection method is presented in this paper.

The objectives of the proposed work are as follows-

- Develop a PV array model using the MATLAB-Simulink platform, incorporating real-time parameter variations.
- Simulate the SCF, LGF, and HSF faults, considering different levels of irradiance, temperature, and fault resistance.
- Determine the most appropriate measurements for detecting the PV array faults.
- Develop an ensemble learning-based PV array fault detection technique with promising accuracy.
- Evaluate the effectiveness of the PV array fault detection technique and compare it with recently proposed artificial intelligence-based techniques.

The remaining part of the paper discusses PV array faults and their protection techniques as follows: Section 2 focuses on modeling the PV array, examining its fault characteristics, and identifying key parameters for fault detection. Section 3 introduces the proposed fault detection method and outlines the creation of a dataset for the ensemble learning-based approach. Section 4 presents the performance analysis of the presented method and a comparative analysis with other methods. Finally, Section 5 includes the findings, limitations, and future scope to conclude the presented work.

2. PV Array Modeling and Fault Analysis

2.1. PV Array Modeling

A PV model is developed using the MATLAB-Simulink platform to analyze the behavior of a PV system, as illustrated in Fig. 1. The model is separated into two parts: the PV array model and the boost converter, depicted in Fig. 1(a) and Fig. 1(b), respectively. The PV array includes sixteen PV modules (PVMs), grouped into four parallel strings, each of which has four series-connected PVMs. Each PV module (KC200GT) is rated at 200W, resulting in a maximum power output of 3.2 kW from the array. The specifications of KC200GT PV module are as follows: P_{max} = 200 W, V at P_{max}= 26.3 V, I at P_{max} = 7.61 A, short-circuit current (I_{SC}) = 8.21 A, Open-circuit voltage (V_{OC}) = 32.9 V, ideality factor = 1.3, temperature coefficient of I_{SC} = 0.0032 A/°C and number of PV cells connected in series in a PVM is 54 [11]. The PV array output (voltage and current) is measured at terminals (1 and 2) of the designed model, as shown in Fig. 1(a). The PV array's output is connected to the boost converter (DC-DC) circuit shown in Fig. 1(b). The boost converter (DC-DC) is controlled using an MPPT algorithm to ensure maximum power delivery [22]. A resistive load is connected to the output terminal of the boost converter (DC-DC).

2.2. Faults in PV Array

The PV array has many PVMs, which are organized in series and parallel combinations using cables and connector junction boxes in the field [23]. These cables are exposed to different weather conditions, allowing the presence of moisture in the junction box [24]. It creates chances for the occurrence of short circuits between cables, connecting PVMs in a string, called Intra-string SCF (ISCF). Similarly, a short circuit may develop between PVMs interconnected in distinct strings, called Cross-string SCF (CSCF) [25]-[26]. When shading occurs on a few PVMs, the current

Accepted Manuscript_Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

delivered from the shaded portion of the PVMs is decreased from the nominal value. The generated power of the unshaded cell is dispersed in the shaded cell, leading to the development of hot spot faults (HSF) within the PVM [27]. The development of a hot spot within the PV cell may be prevented by linking a bypass diode in parallel to each PVM in reverse polarity [28]. It facilitates the flow of current through the PVM under normal conditions, as in the case of forward bias, and allows the current to flow through the bypass diode during shading conditions. A PV array has multiple components carrying currents along with some non-current-carrying components. The non-current carrying components also have significant potential to create a fault, known as a line to ground fault (LGF), due to failure of insulation, physical damage, etc. [29].

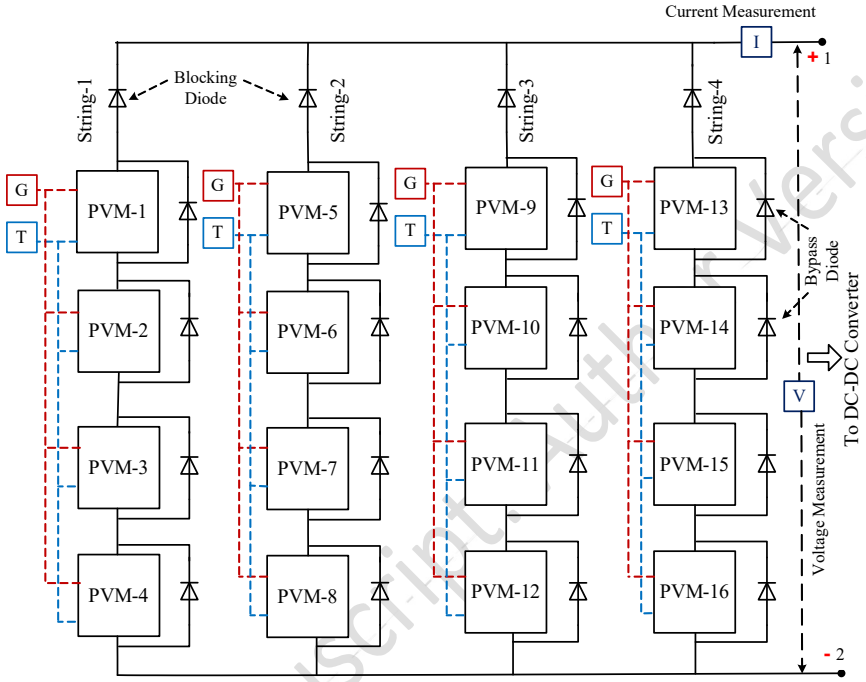


Fig. 1(a). PV array Simulink model.

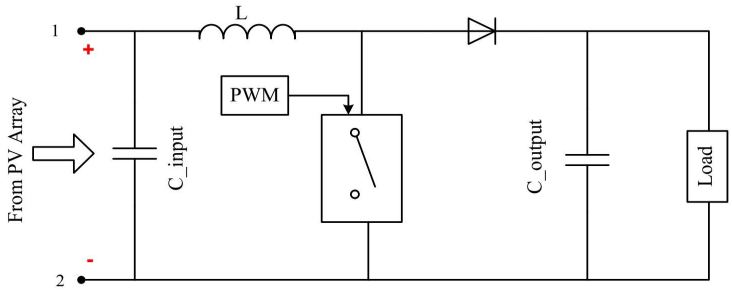


Fig. 1(b). Boost converter (DC-DC) circuit.

This work aims to detect SCF, LGF, and HSF in the PV array. The possible fault cases considered to train and test the proposed method and the typical case of each fault are listed in Table 2 and represented in Fig. 2, where the ISCF occurs between two terminals of the PVM-2 with a fault resistance (R) in the first string. The CSCF fault occurs between the terminal of the PVM-12 in the third string and the terminal of the PVM-15 in the fourth string, with a fault resistance. The HSF fault involves the PVMs 3-4-6-7-8-10-11 with a lower shading level compared to the other PVMs.

Table 2. Faults in a PV array.

Fault Category	Fault Representation	Faults description
SCF	ISCF	Intra-string SCF across a PVM with fault resistance

Accepted manuscript Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

	CSCF	Cross-string SCF with fault resistance
LGF	LGF	Line to Ground Fault with fault resistance
HSF	HSF	Hot spot fault

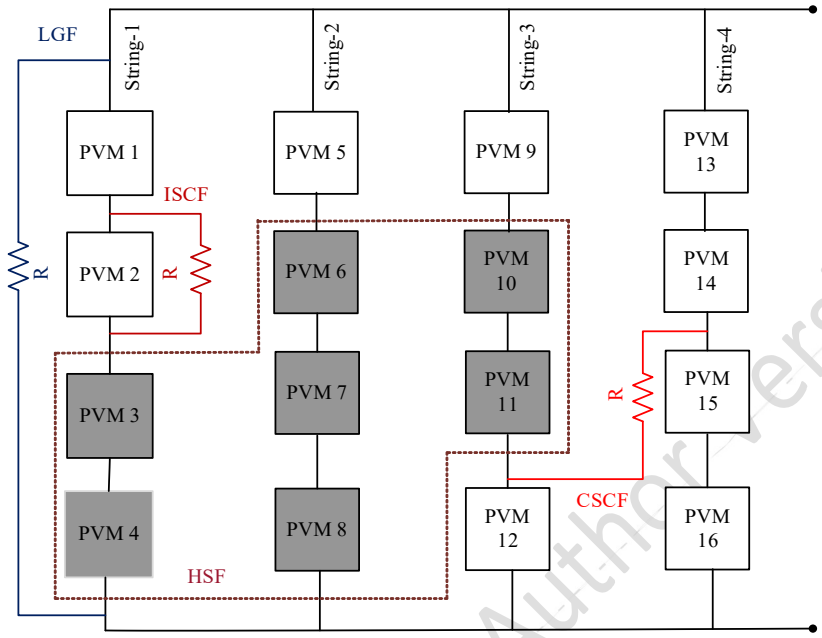


Fig. 2. Modeling of PV Array faults.

2.3. Faults characteristic of PV array

The I-V and P-V characteristics of the developed PV array system under various SCF, LGF, HSF, and normal operating conditions (NOC) at the standard testing conditions are shown in Fig. 3(a) and 3(b), respectively. It is observed that the voltage under ISCF and CSCF with fault resistance is higher than that of the same fault with zero fault resistance (represented as ISCF-R0 and CSCF-R0 in Fig. 3(a) and 3(b)). It is worth noting that the LGF with zero fault resistance (denoted as LGF-R0 in Fig. 3(a) and 3(b)) produces the lowest short-circuit current compared to other faults. The HSF exhibits the lowest maximum power and a significant reduction in short-circuit current across the voltage range. It also produces multiple peaks on P-V characteristics [28],[30]. However, the fault detection method uses real-time measurements of PV parameters, which provide distinct characteristics for each fault.

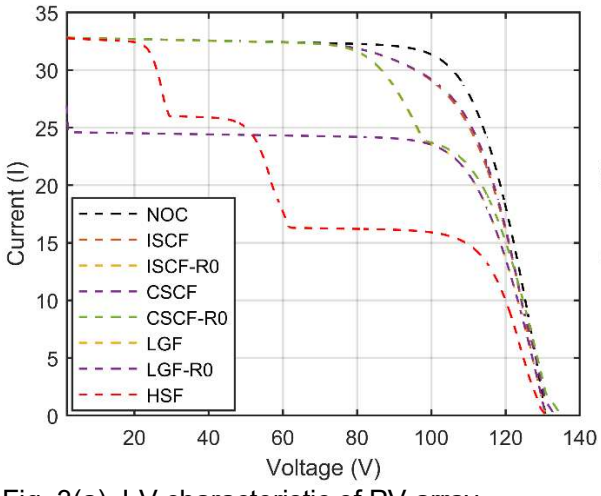


Fig. 3(a). I-V characteristic of PV array

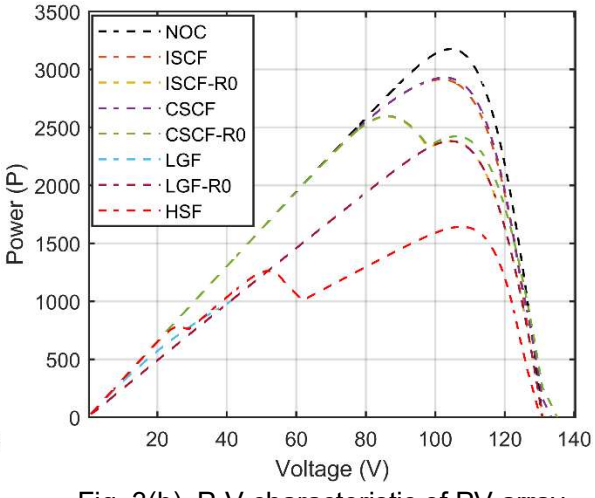


Fig. 3(b). P-V characteristic of PV array

Accepted Manuscript: Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

The effectiveness of fault detection is strongly influenced by the measurements used for fault detection. Therefore, it is beneficial to include real-time measurements that exhibit unique characteristics for each fault type considered in this study, presented in Table 3. The fault was applied in the MATLAB-Simulink model for the time interval of 0.5 to 1.0 seconds, with a resistance value of 5 ohms for faults involving resistance. It has been observed that real-time measurements of I, V, and P exhibit similar signatures across different fault cases and are not sufficient for detecting and classifying faults. Therefore, additional parameters, such as the rate of change of current (dI/dt), rate of change of voltage (dV/dt), the rate of change of power to current (dP/dI), and the rate of change of power to voltage (dP/dV) have been identified that show distinct patterns amid distinct fault scenarios. Fig. 4(a) illustrates the variation in these parameters for ISCF and ISCF-R0, while Fig. 4(b) presents the same for CSCF and CSCF-R0. The corresponding behavior for LGF and LGF-R0 is shown in Fig. 5, and for HSF, it is depicted in Fig. 6.

Table 3. PV Array's Fault Representation.

Fault Representation	*ISCF*	*ISCF-R0*	*CSCF*	*CSCF-R0*	*LGF*	*LGF-R0*	*HSF*
Fault Resistance (Ω)	5	0	5	0	5	0	0

It is observed that the measurements dI/dt, dV/dt, and dP/dV demonstrate decreasing patterns, whereas the measurements dP/dI show a significant increase for ISCF and CSCF. The same measurements display more pronounced characteristics in a similar trend for ISCF-R0 and CSCF-R0. The measurements of dI/dt and dP/dV are increasing, while dV/dt and dP/dI are decreasing for both LGF and LGF-R0. The measurements of dV/dt and dP/dI exhibit downward trends, while dI/dt and dP/dV are on the rise for HSF. Additionally, it is noted that the magnitude of the measurements dI/dt, dV/dt, dP/dI, and dP/dV varies during each fault case, thus aiding in distinguishing the nature of the fault. Consequently, the additional measurements of dI/dt, dV/dt, dP/dI, and dP/dV are chosen along with I, V, and P as inputs for the fault detection method, improving the proposed method's accuracy. This specific combination of parameters has not been previously reported in the literature for the identification of SCF, HSF, and LGF.

Accepted Manuscript- Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

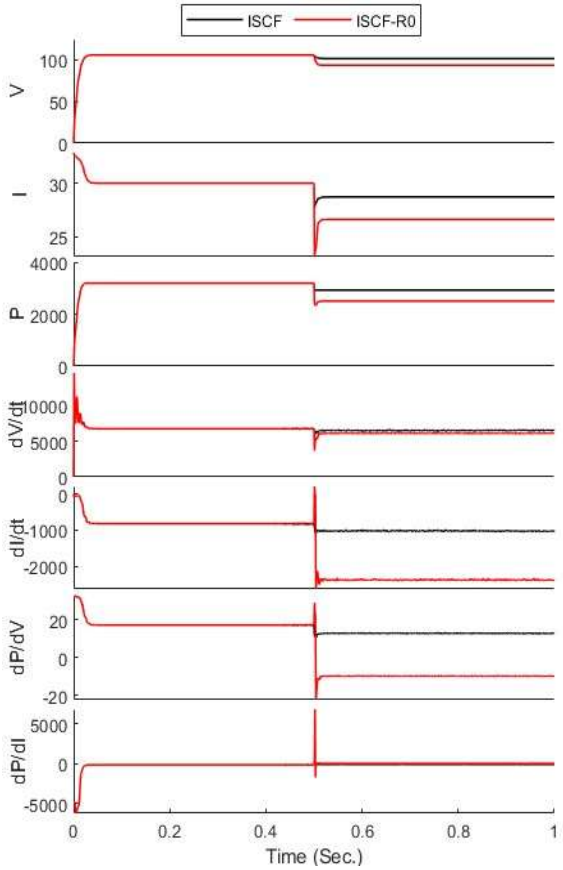


Fig. 4(a). Real-time measurement of various parameters for ISCF and ISCF-R0 faults

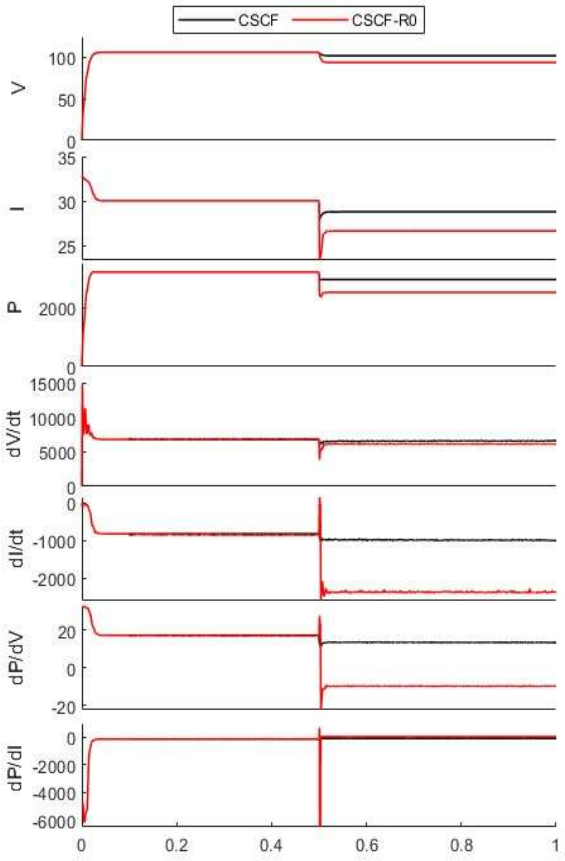


Fig. 4(b). Real-time measurement of various parameters for CSCF and CSCF-R0 faults

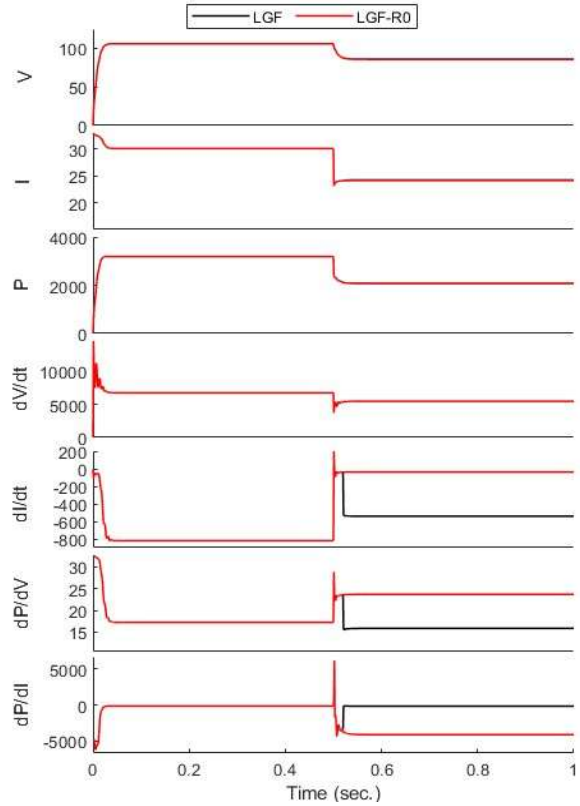


Fig. 5. Real-time measurement of various parameters for LGF and LGF-R0 faults

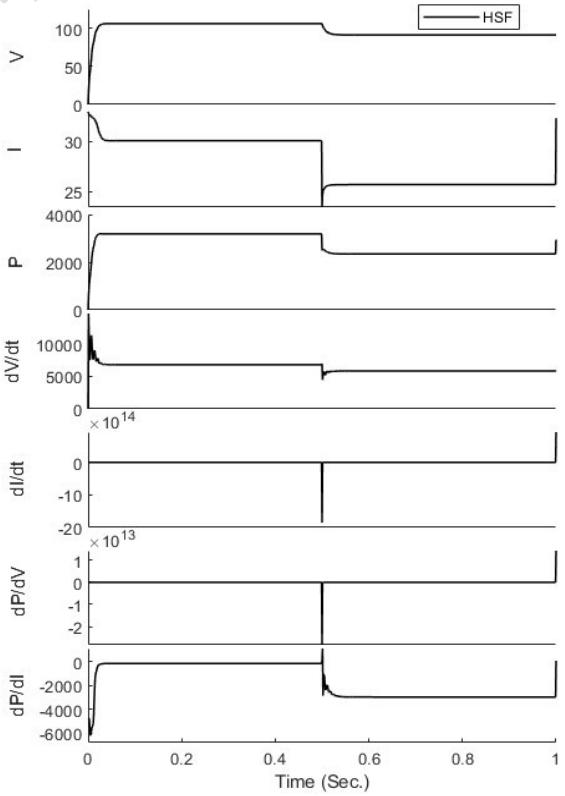


Fig. 6. Real-time measurement of various parameters for HSF

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

3. Proposed Fault Detection Method

The fault detection method should be able to detect and classify the faults quickly using the real-time measurements available from the field. This work proposes an innovative methodology for detecting PV array faults utilizing an ensemble-learning-based technique.

3.1 Ensemble-learning-based technique

This study introduces a novel approach for detecting and classifying PV array faults using an ensemble learning technique based on measurements of PV arrays. Ensemble learning is an effective method that merges predictions from multiple models and offers more robust and reliable forecasts. Ensemble learning improves the overall performance of the learning system by considering multiple perspectives and utilizing the strengths of different models [21]. This approach enhances accuracy and also provides resilience against uncertainties in the data. Ensemble learning offers multiple techniques, such as bagging, boosting, and stacking [31].

3.1.1 Bagging

Bagging or bootstrap aggregation is a popular ensemble learning approach that outperforms classical classification methods, especially for imbalanced datasets [32]. Bagging generates diverse classifiers by manipulating training data and makes final predictions by combining the predictions of multiple models to improve the overall performance through majority voting. This technique has proven effective in improving classification accuracy for various artificial and real-world datasets.

3.1.2 Boosting

Boosting is a powerful ensemble learning technique used in machine learning classification to improve model accuracy by combining multiple weak learners into a strong learner. It works by training models sequentially, where each new model corrects errors made by the previous ones. Boosting assigns higher weights to misclassified instances, ensuring subsequent models focus on difficult cases. Popular boosting algorithms include AdaBoost, Gradient Boosting, and XGBoost. Boosting reduces bias and variance, making it effective for complex classification tasks [32]. However, it can be prone to overfitting and may require careful tuning. Overall, boosting enhances predictive performance.

3.1.3 Stacking

Stacking is an advanced ensemble learning technique that improves classification performance by combining multiple base models using a meta-learner. Unlike bagging and boosting, stacking trains diverse models in parallel and then uses a higher-level model to aggregate their predictions. The base models generate outputs, which serve as inputs for the meta-learner, optimizing final predictions. This approach leverages the strengths of different algorithms, reducing bias and variance [33]. Commonly, decision trees, support vector machines, and neural networks are used as base learners. While stacking enhances accuracy, it requires careful selection of models and tuning to prevent overfitting, making it computationally intensive.

3.2 Methodology

The proposed ensemble-learning-based bagging methodology is divided into four steps, as presented in Fig. 7 [31]. The mathematical formulation of the bagging process started with the original dataset (D) represented as follows:

$$D = (x_1 y_1), (x_2 y_2), (x_3 y_3) \dots \dots \dots (x_n y_n) \quad (1)$$

where x_i are the features and y_i are the labels [31]. It involves creating multiple subsets, called bootstrap samples (D_i), from the original dataset (D) through random sampling. A bootstrap sample D_i is represented as

$$D_i = (x'_1 y'_1), (x'_2 y'_2), (x'_3 y'_3) \dots \dots \dots (x'_n y'_n) \quad (2)$$

Accepted Manuscript- Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

where i = 1, 2, ... n; n is the number of base models. The second step involves the creation of multiple base models, also called weak learners. The number of base models is the same as the number of bootstrap samples [34].

Fig. 7. A typical bagging process

The third step involved the training of a base model $f_i(x)$ for each bootstrap sample D_i. An individual base model is trained on each bootstrap sample, with each model learning independently from its specific subset of data. After the completion of training, each model generates individual predictions. These predictions are then combined using a voting mechanism, typically selecting the majority class as the final prediction for the classification task [35], [36]. The aggregated result serves as the output of the proposed model. If $f_i(x)$ is the prediction of the i^{th} model, the ensemble prediction $\hat{y}$ is given as follows [33], [37]:

$$\hat{y} = \text{majority vote } [f_1(x), f_2(x), \ldots, f_n(x)] \quad (3)$$

The key benefit of bagging is the reduction in variance, as multiple models reduce the overall model's sensitivity due to variations in training data. This ensemble method helps to reduce overfitting and improve predictive accuracy by averaging biases and variances across multiple models [38].

3.4 Architecture of the proposed model

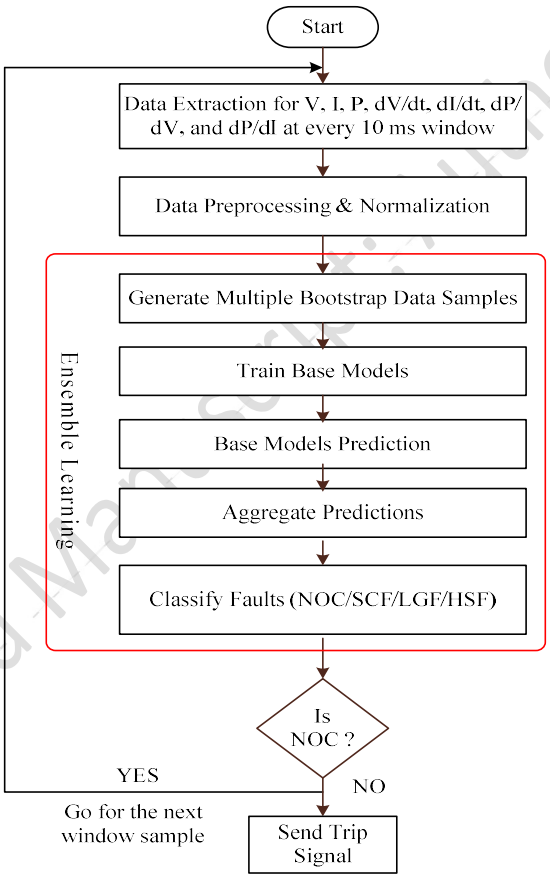
A flowchart illustrating the proposed method to detect and classify faults is shown in Fig. 8. The proposed methodology begins with the initialization of the data extraction process for selected parameters by simulating the MATLAB-Simulink model (Fig. 1) under different fault conditions and a wide range of irradiance, temperature, and fault resistance as shown in Table 3. Simulating the model under these conditions produces 600 samples for each case, NOC, SCF, HSF, and LGF; thus, altogether, 2400 samples are generated.

Accepted Manuscript - Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

Fig. 8. Flowchart of proposed ensemble learning-based bagging technique

The proposed method is implemented with a sampling frequency of 1 kHz and a 10-msec window size. Hence, 10 samples per parameter are obtained for detecting the fault using the proposed technique. It creates a 2400x70 dataset to train the model. Data samples collected at specified frequencies give less computational burden as well as better promising accuracy. Once the input dataset is received, it is pre-processed to reduce the effects of asymmetry in size and dimension and expedite training convergence [39]. A transformation or scale between 0 and 1 is applied to the data during preprocessing to ensure that every feature contributes equally. The normalized pre-processed dataset is given to the proposed model. The proposed technique begins with the original dataset and creates multiple bootstrap samples from the training dataset. These samples are the same size as the original dataset but may include duplicate entries. An individual base model is trained on each bootstrap sample, with each model learning independently from its specific subset of data. Once trained, the base models generate predictions. These predictions are then combined and used in a voting mechanism, typically selecting the majority class as the final prediction for one of the selected classes: NOC/SCF/HSF/LGF. The aggregated result serves as



the output of the proposed model. In the case of NOC, the process is repeated for the next window sample dataset.

4. Results and Discussion

The ensemble-based model is trained and tested using a dataset generated from a developed PV array simulation. This model was subjected to a wide range of operating conditions, including various fault types, temperature levels, irradiance intensities, and fault resistance values, to generate a comprehensive dataset for training and testing. The dataset incorporates both extreme and noisy conditions to ensure robustness and generalization. This work collects the data through simulation of the proposed model, during which communication delay is considered negligible. A

Accepted Manuscript - Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

detailed description of the dataset generation methodology and the performance assessment of the proposed model is presented in the following section.

4.1 Dataset generation for training

The data extraction process is initiated for selected measurements of I, V, P, dV/dt, dI/dt, dP/dV, and dP/dI by simulating the MATLAB-Simulink model (Fig. 1) under different fault conditions and a wide range of temperature, irradiance, and fault resistance as illustrated in Table 4. Generating a training dataset under a wide range of temperature and irradiance conditions increases data diversity, allowing the model to learn general patterns rather than memorizing specific cases. This variability improves the model's robustness to unseen scenarios and reduces the chances of overfitting by preventing bias toward limited operating conditions. Simulating the model under these conditions produces 600 datasets for each case, NOC, SCF, LGF, and HSF; thus, altogether, 2400 datasets are generated. The selected measurements are sampled at an optimal frequency of 1 kHz and a window size of 10 msec to minimize the computation burden, such that it produces 10 samples of each measurement per window (for a size of 10 msec). Therefore, it generates 70 samples corresponding to the selected seven measurements in a window, resulting in a dataset size of 70x2400 for training the model. It creates a 70x2400 dataset to train the model. The sequential information of the measurements I, V, P, dI/dt, dV/dt, dP/dI, and dP/dV are supplied to the model as input to detect the faults. The input dataset is always pre-processed using a scale transformation between 0 and 1 to reduce the effects of asymmetry scale and expedite the training convergence.

Table 4. Simulation cases for generating a training dataset.

Fault/ Class	Cases	Irradiance (w/m^2)	Temperature (˚C)	Fault Resistance (Ω)/ Shading	Dataset (Nos.)
NOC	NOC	[410:10:1000]	[05:05:50]	[0]	600
SCF	ISCF	[500:10:690]	[20:15:50]	[0, 1, 3, 5, 7]	600
	CSCF	[500:10:690]	[20:15:50]	[0, 1, 3, 5, 7]	
LGF	LGF	[500:10:730]	[10:10:50]	[0, 1, 3, 5, 7]	600
HSF	HSF	Normal- [550:50:1000] Partial- [300:50:500]	[20:10:50]	Shaded PVMs- [(4-8), (3-4-8), (4-7-8-11)]	600

4.2 Data pre-processing and normalization

Data pre-processing is a fundamental step in the machine learning pipeline, aimed at transforming raw, unstructured, or noisy data into a clean and analyzable format suitable for model development. It involves a series of techniques to enhance data quality, ensure consistency, and facilitate accurate learning. Key processes include data cleaning (handling missing values, removing duplicates, and correcting inconsistencies), feature scaling (normalization), and encoding categorical variables into numerical formats. Mathematically, normalization can be expressed as

$$X' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (4)$$

Additionally, dimensionality reduction and outlier detection may be applied to improve computational efficiency and model robustness. Effective data pre-processing is critical for optimizing model performance and ensuring reliable outcomes [40], [41].

The pre-processed and normalized dataset is evaluated using k-fold cross-validation, a

Accepted Manuscript- Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

technique that enhances the reliability of model performance estimation. In this approach, the training data is randomly divided into *k* equal-sized folds. The model is trained k times, each time using a different fold as the validation set and the remaining k−1 folds as the training set. In this study, 5-fold cross-validation (*k = 5*) with shuffled data is employed. As a result, the proposed algorithm is tested across five different training-validation splits, each yielding unique performance metrics. This method helps ensure robust and consistent results for PV fault detection by leveraging the entire dataset in multiple combinations. Consequently, it reduces the risk of overfitting and improves the generalization capability of the model [20].

4.3 Training performance of the proposed method

The ensemble learning-based categorization of 2400 training data samples into the four classes NOC, SCF, LGF, and HSF is shown in Table 5. Notably, all training samples are correctly classified, achieving a 100% accuracy rate. The training accuracy and the confusion matrix suggest that the model does not suffer from variance issues. Thus, the proposed model is expected to perform reliably when applied to the test dataset. During training, the model achieved a final size of approximately 415 kB, a prediction/inference speed of about 7,500 observations per second, and a total training time of 3.0447 seconds on the following system configuration: 11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40GHz. These characteristics confirm that the proposed model has a low computational burden, making it suitable for all kinds of PV systems.

Table 5. Confusion Matrix for the training of the bagging ensemble model.

Output Class / Target Class	NOC	SCF	LGF	HSF	
NOC	**600**	0	0	0	100%
SCF	0	**600**	0	0	100%
LGF	0	0	**600**	0	100%
HSF	0	0	0	**600**	100%
					100%

To interpret the impact of each measurement used for PV array fault detection, a detailed analysis has been conducted on the trained ensemble-based model using Shapley explanation values. Shapley provides a unified framework to interpret the impact of each measurement used for PV array fault detection by computing the contribution of each measurement to the output. The Shapley analysis reveals that the measurements dI/dt and voltage (V) have the highest average Shapley values, indicating their strong influence on the model's ability to distinguish between different fault types (Fig. 9). Conversely, measurements like dP/dI, dV/dt, dP/dV, moderately support the fault detection. Though the measurements current (I) and power (P) have comparatively less impact on detection, they support strengthening the transparency and interpretability. It enhances the transparency and interpretability of the ensemble-based model, providing further confidence in its predictive capabilities and the relevance of the selected features.

Accepted Manuscript: Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

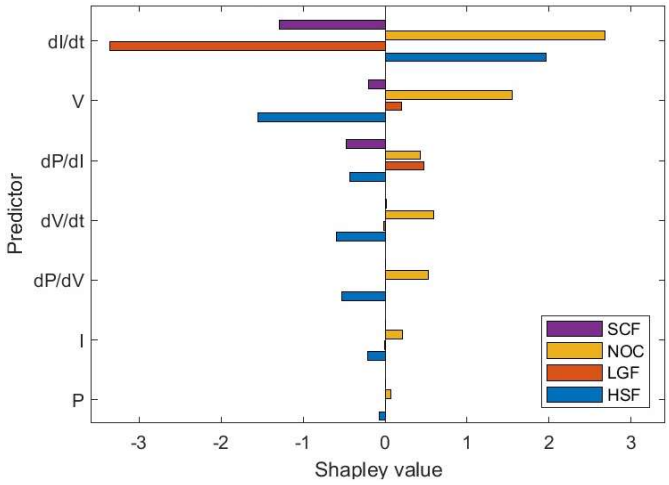


Fig. 9. Shapley analysis

4.4 Dataset generation for testing

The distinct testing datasets are generated with different operating conditions achieved using varying values of temperature, irradiance, and fault resistance, which are different from the training datasets and are shown in Table 6. Each case of NOC, SCF, LGF, and HSF has 90 testing datasets, resulting in 360 datasets, used to evaluate the proposed model.

Table 6. Simulation cases for generating a testing dataset.

Fault/ Class	Cases	Irradiance (w/m^2)	Temperature (°C)	Fault Resistance (Ω)/ Shading	Dataset (Nos.)
NOC	NOC	[100:50:500]	[05:05:50]	[0]	90
SCF	ISCF	[800:100:1000]	[25:10:45]	[0, 2, 4, 6, 8]	90
	CSCF	[800:100:1000]	[25:10:45]	[0, 2, 4, 6, 8]	
LGF	LGF	[750:50:1000]	[25:10:45]	[0, 2, 4, 6, 8]	90
HSF	HSF	Normal- [780:50:980] Partial- [200:40:280]	[25:10:45]	Shaded PVMs- [(3-8), (4-6-10)]	90

4.5 Testing performance of the proposed method

An additional dataset is used to assess the effectiveness of the PV array fault detection method based on the ensemble learning model. Table 6 presents the testing pattern generation under various operating conditions simulated across different parameter ranges (temperature, irradiance, and fault resistance), ensuring no overlap with the training data. Each fault category—NOC, SCF, LGF, and HSF—produces 90 testing datasets, resulting in a total of 360 datasets for evaluating the proposed model. The confusion matrix in Table 7 summarizes the testing results, demonstrating that all samples in the testing dataset are accurately classified into their respective categories except one sample belonging to SCF, which is misclassified as HSF. The ensemble learning-based PV fault detection method achieves a 99.72% accuracy rate, confirming that the predicted values closely align with the true values.

Table 7. Confusion Matrix for testing of the bagging ensemble model.

Output Class / Target Class	NOC	SCF	LGF	HSF	

Accepted Manuscript: Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

NOC	**90**	0	0	0	100%
SCF	0	**89**	1	0	98.88%
LGF	0	0	**90**	0	100%
HSF	0	0	0	**90**	100%
					99.72%

4.6 Performance analysis using confusion matrix analysis

Confusion matrix analysis aids in analyzing the effectiveness of the classification method. The labels displayed in Table 8 represent the multi-class confusion matrix for the NOC, SCF, LGF, and HSF [42]. The confusion matrix presented in Table 8 assesses the following performance indices: accuracy, error rate and sensitivity. The effectiveness of a truly classified dataset is described by its accuracy, whereas the overall representation of the misclassified datasets is called the error. However, sensitivity evaluates the effectiveness of the classifier in recognizing true instances of a particular class, relative to the total number of actual instances of that class [43], [44]. The dataset in this study is divided into multiple classes, and the effectiveness of the multi-class model is evaluated using the performance indices listed below from (5) to (7).

Table 8. Multi-class confusion matrix.

Predicted Class / Target Class	Class-1 NOC	Class-2 SCF	Class-3 LGF	Class-4 HSF
Class-1 NOC	**TRUE Class-1-1 NOC**	FALSE Class-2-1 SCF	FALSE Class-3-1 LGF	FALSE Class-4-1 HSF
Class-2 SCF	FALSE Class-1-2 NOC	**TRUE Class-2-2 SCF**	FALSE Class-3-2 LGF	FALSE Class-4-2 HSF
Class-3 LGF	FALSE Class-1-3 NOC	FALSE Class-2-3 SCF	**TRUE Class-3-3 LGF**	FALSE Class-4-3 HSF
Class-4 HSF	FALSE Class-1-4 NOC	FALSE Class-2-4 SCF	FALSE Class-3-4 LGF	**TRUE Class-4-4 HSF**

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} \text{True Class k}}{\sum_{k=1}^{K} \text{True Class k} + \sum_{k=1}^{K} \sum_{\substack{n=1 \\ n \neq k}}^{K} \text{False Class k} - \text{n}} \quad (5)$$

$$\text{Error Rate} = \frac{\sum_{k=1}^{K} \sum_{\substack{n=1 \\ n \neq k}}^{K} \text{False Class k} - \text{n}}{\sum_{k=1}^{K} \text{True Class k} + \sum_{k=1}^{K} \sum_{\substack{n=1 \\ n \neq k}}^{K} \text{False Class k} - \text{n}} \quad (6)$$

$$\text{Sensitivity} = \frac{\sum_{k=1}^{K} \sum_{\substack{n=1 \\ n \neq k}}^{K} \text{True Class k} - \text{n}}{\sum_{k=1}^{K} \sum_{\substack{n=1 \\ n \neq k}}^{K} \text{True Class k} - \text{n} + \sum_{k=1}^{K} \sum_{\substack{n=1 \\ n \neq k}}^{K} \text{False Class k} - \text{n}} \quad (7)$$

An analysis of performance indices shows that the proposed method offers the least percentage error with good accuracy and sensitivity, as shown in Table 9. The performance of the proposed technique is evaluated by comparing its results with other AI techniques discussed in the literature and is shown in Table 10. A comparison of performance metrics reveals that the proposed

Accepted Manuscript (author version)

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

technique achieves higher accuracy, sensitivity, and minimum error rate with the same dataset. Therefore, the proposed methodology, when compared with various techniques such as RF, SVM, kNN, and ANN, performs with better accuracy.

Table 9. Confusion matrix analysis for an Ensemble Learning based PV array fault detection method.

Class	NOC	SCF	LGF	HSF	Overall
Accuracy (%)	100	98.88	100	100	99.72
Error Rate (%)	0.00	1.12	0.00	0.00	0.28
Sensitivity (%)	100	98.88	100	100	99.72

Table 10. Comparative analysis of PV array fault detection techniques.

Reference	Technique	Parameters	Faults	Fault Classification Accuracy
[12]	RF	I, V	LLF, OCF, partial shading	99.24%
[13]	SVM	I, V	LLF	91.40% – 94.74%
[14]	kNN	I, V, P	LLF, OCF, HSF, Bypass Diode Faults	98.7%
[15]	ANN	dI/dt, dV/dt, dP/dI, dP/dV	SCF, HSF, LGF	98.3%
Proposed method	Ensemble-Technique	I, V, P, dI/dt, dV/dt, dP/dI, dP/dV	SCF, LGF, HSF	99.72%

5. Conclusion

An ensemble learning-based PV array fault detection methodology is suggested to identify and discriminate the faults. The proposed method focused on detecting the SCF, LGF, and HSF faults in the PV array effectively. The fault analysis identifies the seven measurements that have unique characteristics for each fault case: I, V, P, dI/dt, dV/dt, dP/dI, and dP/dV. The sampling of all seven measurements for every 10 ms window is collected to generate the dataset. The proposed model is trained and tested with different datasets that are generated with operating conditions without overlapping the parameter range used for simulation. The effectiveness of the proposed method is assessed using confusion matrix analysis and various performance indices. Further, the accuracy of the proposed method is also compared with the recent AI techniques presented in the literature, such as RF, SVM, kNN, and ANN. The comparison results show that the proposed PV array fault detection method outperforms the other methods and offers a promising accuracy of 99.72%, while adding a computational burden. Nevertheless, substantial extreme variations in the environmental conditions may challenge the performance of the proposed technique. The proposed method may be extended to detect other kinds of faults, such as open-circuit faults, bypass diode faults, and degradation faults.

References:

[1] H. H. Pourasl, R. V. Barenji, and V. M. Khojastehnezhad, "Solar energy status in the world: A comprehensive review," Energy Reports 10 (2023): 3474-3493.

[2] A. Prakash, A. Kumar, A. Ranjan, and S. Kumar, "Technological Assessment of a 130 KW

Accepted Manuscript: Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

Grid-Connected Rooftop Solar Power Plant: A Case Study at BAU Sabour, India," 2023 First International Conference on Cyber-Physical Systems, Power Electronics and Electric Vehicles (ICPEEV), Hyderabad, India, 2023, pp. 1-6, doi: 10.1109/ICPEEV58650.2023.10391924.

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

[4] D. S. Pillai and N. Rajasekar, "A comprehensive review on protection challenges and fault diagnosis in PV systems," Renew. Sustain. Energy Rev., vol. 91, pp. 18–40, Aug. 2018, doi: https://doi.org/10.1016/j.rser.2018.03.082.

[5] A. T. Lahiani, A. B. B. Abdelghani, I. S. Belkhodja, "Fault detection and monitoring systems for photovoltaic installations: A review", Renewable and Sustainable Energy Reviews, vol. 82, pp. 2680-2692, Sept. 2017, doi: https://doi.org/10.1016/j.rser.2017.09.101.

[6] Y. Zhao and R. Lyons, Jr., "Ground-fault analysis and protection in PV arrays," in Proc. Photovoltaic Protection, pp. 1–4, 2011.

[7] B. Aljafari, S. R. K. Madeti, P. R. Satpathy, S. B. Thanikanti, and B. V. Ayodele, Automatic monitoring system for online module-level fault detection in grid-tied photovoltaic plants. *Energies*, vol. *15, no.* 20, pp.7789, 2022.

[8] P. Lin, F. Guo, Y. Lin, S. Cheng, X. Lu, Z. Chen, and L. Wu, Fault diagnosis of photovoltaic arrays with different degradation levels based on cross-domain adaptive generative adversarial network. *Applied Energy*, vol. *386*, pp.125578, 2025.

[9] P. R. Satpathy, B. Aljafari, S. B. Thanikanti, and S. R. K. Madeti, Electrical fault tolerance of photovoltaic array configurations: Experimental investigation, performance analysis, monitoring, and detection. *Renewable Energy*, Vol. *206*, pp.960-981, 2023.

[10] H. Fu, H. Liu, S. Xie, S. Liu, H. Han, and J. Ma, Multi-coupling fault detection and diagnosis of photovoltaic arrays with improved slime mould algorithm and PolyCatBoost. *Process Safety and Environmental Protection*, vol. *194*, pp.523-541, 2025.

[11] Y. Zhao, L. Yang, B. Lehman, J.-F. de Palma, J. Mosesian, and R. Lyons, "Decision tree-based fault detection and classification in solar photovoltaic arrays," in 2012 Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC), pp. 93–99, Orlando, FL, February 2012, doi: 10.1109/APEC.2012.6165803.

[12] Z. Chen, F. Han, L. Wu, J. Yu, S. Cheng, P. Lin, and H. Chen, "Random Forest based intelligent fault diagnosis for PV arrays using array voltage and string currents" Energy conversion and management, vol. 178, pp.250-264, 2018, doi: https://doi.org/10.1016/j.enconman.2018.10.040.

[13] Z. Yi and A. H. Etemadi, ''Line-to-line fault detection for photovoltaic arrays based on multi-resolution signal decomposition and two-stage support vector machine,'' IEEE Trans. Ind. Electron., vol. 64, no. 11, pp. 8546–8556, Nov. 2017, doi: 10.1109/TIE.2017.2703681.

[14] S. R. Madeti and S. N. Singh, "Modeling of PV system based on experimental data for fault detection using kNN method," Solar Energy, vol. 173, pp. 139-151, Oct. 2018, doi: https://doi.org/10.1016/j.solener.2018.07.038.

[15] A. Shekhar, and M. Senthil Kumar, "Identification and Classification of PV Array Faults Using Artificial Neural Network" 2023, International Conference on Soft Computing: Theories and Applications, Singapore: Springer Nature Singapore, pp. 339-351, doi: https://doi.org/10.1007/978-981-97-2089-7_30.

[16] B. Aljafari, P. R. Satpathy, S. B. Thanikanti, and N. Nwulu, Supervised classification and fault detection in grid-connected PV systems using 1D-CNN: Simulation and real-time validation. *Energy Reports*, vol. *12*, pp.2156-2178, 2024.

[17] M. Wozniak, M. Grana, and E. Corchado, ''A survey of multiple classifier systems as hybrid systems,'' Inf. Fusion, vol. 16, pp. 3–17, Mar. 2014, doi: https://doi.org/10.1016/j.inffus.2013.04.006.

[18] A. Eskandari, J. Milimonfared, and M. Aghaei, "Line-line fault detection and classification for photovoltaic systems using ensemble learning model based on IV characteristics", *Solar Energy*, vol. 211, pp.354-365, Nov. 2020, doi: https://doi.org/10.1016/j.solener.2020.09.071.

[19] C. Kapucu, and M. Cubukcu, "A supervised ensemble learning method for fault diagnosis in photovoltaic strings", *Energy*, vol. 227, pp.120463, July 2021, doi: https://doi.org/10.1016/j.energy.2021.120463.

Accepted Version - Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

[20] N. C. Yang, and H. Ismail, “Voting-based ensemble learning algorithm for fault detection in photovoltaic systems under different weather conditions”, *Mathematics*, vol. 10, no. 2, p.285, Jan 2022, doi: https://doi.org/10.3390/math10020285.
[21] Z. Mian, X. Deng, X. Dong, Y. Tian, T. Cao, K. Chen, and T. Al Jaber, “A literature review of fault diagnosis based on ensemble learning”, *Engineering Applications of Artificial Intelligence*, vol. 127, p.107357, Jan 2024, doi: https://doi.org/10.1016/j.engappai.2023.107357.
[22] R. A. Muhammed, and D. Sulaiman, “Particle Swarm Optimization (PSO) Based MPPT controller Modeling and Design of Photovoltaic Module”, *Majlesi Journal of Electrical Engineering*, vol. 16, no. 4, pp.167-175, 2022.
[23] R. Nitheesh, B. P. Kumar, M. Chakkarapani, G. S. Ilango, and C. Nagamani, "Detection and quantification of degradation using time constant of PV voltage," 2017 National Power Electronics Conference (NPEC), Pune, India, 2017, pp. 90-95, doi: 10.1109/NPEC.2017.8310440.
[24] B. P. Kumar, R. Nitheesh, M. Chakkarapani, G. S. Ilango, and C. Nagamani, “Estimation of PV module degradation through the extraction of I–V curve at inverter pre-startup condition,” in IET Renewable Power Generation, vol. 14, no. 17, pp. 3479-3486, Dec. 2020, doi: https://doi.org/10.1049/iet-rpg.2020.0316.
[25] Y. Zhao, J. de Palma, J. Mosesian, R. Lyons and B. Lehman, "Line–Line Fault Analysis and Protection Challenges in Solar Photovoltaic Arrays," in IEEE Transactions on Industrial Electronics, vol. 60, no. 9, pp. 3784-3795, Sept. 2013, doi: 10.1109/TIE.2012.2205355.
[26] B. P. Kumar, D. S. Pillai, N. Rajasekar, M. Chakkarapani, and G. S. Ilango, "Identification and Localization of Array Faults with Optimized Placement of Voltage Sensors in a PV System," in IEEE Transactions on Industrial Electronics, vol. 68, no. 7, pp. 5921-5931, July 2021, doi: 10.1109/TIE.2020.2998750.
[27] J. P. Ram, N. Rajasekar, “A new global maximum power point tracking technique for solar photovoltaic (PV) system under partial shading conditions (PSC)”, Energy, vol. 118, pp. 512-525, Jan. 2017, doi: https://doi.org/10.1016/j.energy.2016.10.084.
[28] T.S.Babu N. Rajasekar, K. Sangeetha, “Modified particle swarm optimization technique based maximum power point tracking for uniform and under partial shading condition” in Applied Soft Computing, vol. 34, pp. 613-24, Sept. 2015, doi: https://doi.org/10.1016/j.asoc.2015.05.029.
[29] W. I. Bower and J. C. Wiles, "Analysis of grounded and ungrounded photovoltaic systems," Proceedings of 1994 IEEE 1st World Conference on Photovoltaic Energy Conversion - WCPEC (A Joint Conference of PVSC, PVSEC and PSEC), Waikoloa, HI, USA, 1994, pp. 809-812 vol.1, doi: 10.1109/WCPEC.1994.520083.
[30] V. L. Mishra, Y. K. Chauhan, and K. S. Verma, “Various modeling approaches of photovoltaic module: A comparative analysis”, *Majlesi Journal of Electrical Engineering*, vol. 17, no. 2, pp.117-131, 2023.
[31] G. Tuysuzoglu and D. Birant, ‘‘Enhanced bagging (eBagging): A novel approach for ensemble learning,’’ Int. Arab J. Inf. Technol., vol. 17, no. 4, pp. 515–528, Jul. 2020, doi: 10.34028/iajit/17/4/10.
[32] H. P. Koapaha, and N. Ananto, “Bagging based ensemble analysis in handling unbalanced data on classification modeling”, *Klabat Accounting Review*, vol. 2, no. 2, pp.165-178, Sept. 2021, doi: https://doi.org/10.60090/kar.v2i2.589.165-178.
[33] Zhou, Zhi-Hua. 2012. Ensemble Methods: Foundations and Algorithms. Hoboken: CRC Press.
[34] Breiman, L., 1996. Bagging predictors. Machine learning, 24, pp.123-140.
[35] T. Hothorn, and B. Lausen, “Double-bagging: combining classifiers by bootstrap aggregation”, *Pattern Recognition*, vol. 6, pp.1303-1309, June 2003.
[36] Q. Kadhim, A. Q. A. S. Al-Sudani, I. A. Almani, T. Alghazali, H. K. Dabis, A. T. Mohammed, S. G. Talib, R. A. Mahmood, Z. T. Sahi, and Y. Mezaal, “IOT-MDEDTL: IoT Malware Detection based on Ensemble Deep Transfer Learning”, *Majlesi Journal of Electrical Engineering*, vol. 16, no. 3, 2022.
[37] E. Bauer, and R. Kohavi, “An empirical comparison of voting classification algorithms: Bagging, boosting, and variants”, *Machine learning*, vol. 36, pp.105-139, July 1999, doi: https://doi.org/10.1023/A:1007515423169.

Accepted manuscript - Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/

[38] N. Altman, and M. Krzywinski, “Ensemble methods: bagging and random forests”, *Nature Methods*, vol. 14, no. 10, pp.933-935, Oct. 2017.
[39] S. Zhang, Y. Wang, M. Liu, and Z. Bao, ‘‘Data-based line trip fault prediction in power systems using LSTM networks and SVM,’’ IEEE Access, vol. 6, pp. 7675–7686, Dec. 2018, doi: 10.1109/ACCESS. 2017.2785763.
[40] M. Kang and J. Tian, Machine learning: Data pre-processing. Prognostics and health management of electronics: fundamentals, machine learning, and the internet of things, pp.111-130, 2018.
[41] V. Kumar, and R. K. Mandal, “Comparison of data filtering methods effects on smartgrid load forecasting”, *Majlesi Journal of Electrical Engineering*, Vol. 18, no. 3, pp.1-10, 2024.
[42] S. P. Simon, M. S. Kumar, K. Sundareswaran and C. C. Columbus, "Performance analysis of empirical Fourier transform based power transformer differential protection," 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE), Eilat, Israel, 2016, pp. 1-5, doi: 10.1109/ICSEE.2016.7806152.
[43] Marina Sokolova, Guy Lapalme, “A systematic analysis of performance measures for classification tasks,” Information Processing and Management, Vol. 45, pp. 427–437, July 2009, doi: https://doi.org/10.1016/j.ipm.2009.03.002.
[44] M. M. Mahdi, M. A. Mohammed, H. Al-Chalibi, B. S. Bashar, H. A. Sadeq, and T. M. J. Abbas, “An ensemble learning approach for glaucoma detection in retinal images”, *Majlesi Journal of Electrical Engineering*, vol. 16, issue. 4, 2022.

Accepted Manuscript: Author Version

This article has license CC BY 4.0 https://creativecommons.org/licenses/by/4.0/