Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 03 (2026)

Machine Learning-Based Intelligent Fault Diagnosis and Signal Forecasting for Photovoltaic Systems Using Performance Indicators

Hatira Bacem[1], Dhaoui Mehdi[2] and Sbita Lassaad[3]

[1,2,3]Electrical Department of Engineering, University of Gabes, National Engineering School, Gabes 6029, Tunisia.

L/R: Processes, Energy, Environment and Electrical Systems (PEESE).

[*]Correspondence: hatira.bacem@gmail.com; Tel: +216 29 51 25 43

Abstract: - Reliable fault diagnosis and predictive monitoring are essential for improving the performance, reliability, and operational safety of photovoltaic (PV) systems. This paper presents an intelligent framework for fault diagnosis and signal prediction based on performance indicators and machine learning techniques. First, a set of representative electrical and performance indicators is extracted from the PV system under normal and faulty operating conditions. These indicators are then used to identify different fault scenarios and to predict the evolution of key electrical signals, enabling early detection of abnormal behavior and supporting predictive maintenance. Several machine learning algorithms are investigated and evaluated according to their diagnostic accuracy, prediction capability, and computational efficiency. The proposed approach demonstrates high fault classification accuracy and reliable signal prediction under varying environmental conditions, while reducing false alarms and improving the robustness of the monitoring process. The results confirm that the integration of performance indicators with machine learning provides an effective solution for intelligent condition monitoring, fault diagnosis, and predictive maintenance of photovoltaic systems.

Keywords: Photovoltaic systems; Fault diagnosis; Signal prediction; Machine learning; Performance indicators; Predictive maintenance; Intelligent monitoring; Condition monitoring.

1. Introduction

Photovoltaic (PV) systems are widely deployed in modern power generation due to their ability to provide clean, renewable, and sustainable energy. However, PV installations are vulnerable to a variety of faults that can significantly degrade system performance and reliability. These faults can be classified into three main categories. Environmental faults include dust accumulation (soiling), partial shading, and temperature variations that affect solar irradiance and module efficiency. Electrical faults include open-circuit conditions, short-circuit faults, and degradation of module components such as bypass diodes and interconnections. Thermal faults arise from excessive heating, hotspot formation, and poor ventilation, which may lead to permanent damage of photovoltaic cells.

Conventional maintenance strategies—namely reactive maintenance, which intervenes only after a failure occurs, and preventive (scheduled) maintenance, which performs maintenance at predetermined intervals—are increasingly recognized as inefficient for PV systems. Reactive maintenance often leads to significant power losses and operational downtime, while preventive maintenance may result in unnecessary inspection and replacement of components, increasing operational costs. These limitations highlight the need for intelligent predictive maintenance systems capable of detecting, classifying, and identifying faults at early stages before significant performance degradation occurs.

The present research addresses this challenge by developing an intelligent predictive fault diagnosis framework tailored for grid-connected photovoltaic systems. The proposed approach relies on monitoring electrical and environmental parameters and calculating performance indicators to identify abnormal operating conditions. The specific objectives of this study are summarized as follows:

• To develop accurate mathematical models representing both healthy and faulty operating conditions of photovoltaic generators, including electrical and environmental fault mechanisms such as soiling, shading, degradation, open circuit, short-circuit, and hotspot conditions.

• To simulate realistic operating scenarios using MATLAB and Simulink, generating reliable datasets under both normal and faulty conditions for training and validation purposes.

• To compute key diagnostic performance indicators such as Performance Ratio (PR) and Fill Factor (FF), which provide quantitative information about system efficiency and fault conditions.

• To design and evaluate advanced artificial intelligence techniques, including Artificial Neural Networks (ANN) and Random Forest classifiers, for accurate fault detection and classification in photovoltaic systems.

• To analyze the performance of the proposed diagnostic framework using simulated data under various fault conditions, ensuring robustness and applicability to real-world photovoltaic installations.

The proposed methodology integrates physical modeling, simulation-based data generation, and intelligent diagnostic algorithms into a unified framework. Initially, a detailed photovoltaic system model is developed to represent both healthy operation and different fault conditions. Subsequently, simulation tools are used to generate representative datasets that capture variations in voltage, current, irradiance, and temperature signals. These datasets are then processed to compute performance indicators such as generated power, Performance Ratio, and Fill Factor. Finally, machine learning algorithms are trained to classify system conditions and detect faults at an early stage, enabling predictive maintenance and improved system reliability.

The remainder of this paper is organized as follows. Section 1 presents an overview of photovoltaic system operation and common fault mechanisms. Section 2 describes the mathematical modeling of the photovoltaic system and the simulation framework used to generate datasets. Section 3 presents the calculation of performance indicators and the fault detection methodology. Section 4 describes the machine learning-based classification techniques used for predictive diagnosis. Section 5 presents simulation results and discussion. Finally, Section 6 concludes the paper and outlines future research directions.

2. Diagnostic Methods and Common Faults in Photovoltaic Systems

2.1 Common Faults in Photovoltaic Systems

The diagnosis of photovoltaic (PV) systems is essential to ensure their reliability, efficiency, and long-term operation. Modern PV installations are exposed to various environmental, electrical, and thermal stresses that may lead to performance degradation or permanent damage. Therefore, systematic monitoring and diagnostic methods are required to identify faults at an early stage and minimize power losses.

Photovoltaic systems are subject to several types of faults that can significantly affect their performance. These faults are generally classified into three main categories according to their origin: environmental faults, electrical faults, and thermal faults.

Environmental faults are among the most common issues in PV systems. These faults include dust accumulation (soiling), partial shading caused by nearby objects, and variations in solar irradiance. Dust accumulation on the surface of PV modules reduces the amount of sunlight reaching the solar cells, resulting in decreased current generation and reduced output power. Partial shading creates non-uniform illumination across the module surface, leading to mismatch losses and reduced system efficiency.

Electrical faults are another major concern in photovoltaic installations. These include open-circuit faults, short-circuit faults, degradation of interconnections, and bypass diode failures. Open-circuit faults occur when electrical connections are interrupted, resulting in zero current flow and loss of power production. Short-circuit faults occur when unintended conductive paths are created between terminals, causing excessive current flow and potential damage to system components. Degradation of electrical connections and bypass diode failures can lead to localized power loss and abnormal voltage behavior.

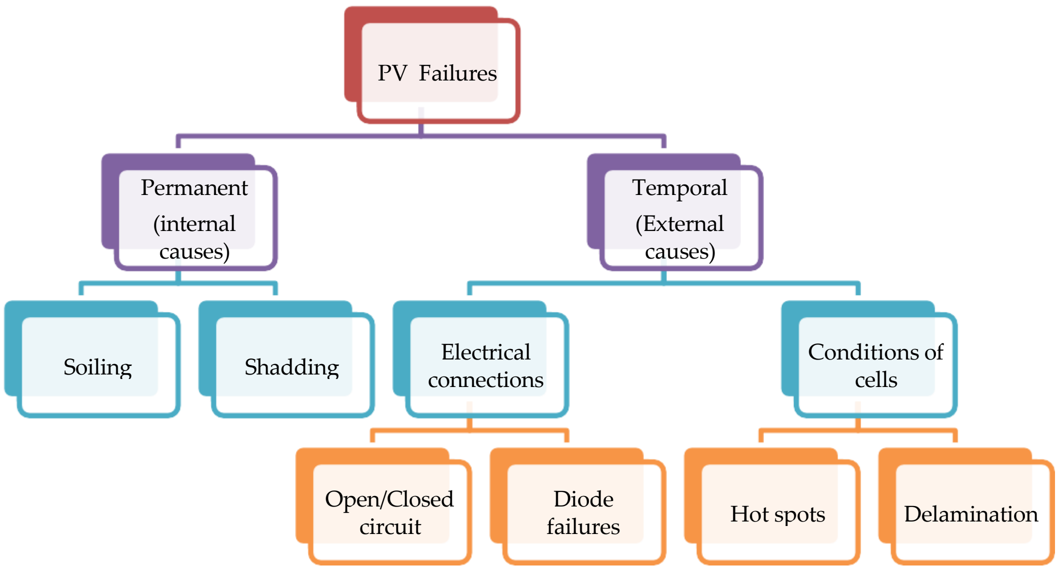


Figure 1. Classification of photovoltaic system faults according to their origin

Thermal faults are primarily associated with temperature-related stresses in photovoltaic modules. Hotspot formation is one of the most critical thermal faults and occurs when a portion of the module operates at significantly higher temperature than the surrounding cells. This condition may be caused by shading, cell damage, or poor electrical connections. Continuous hotspot formation may lead to irreversible module damage, reduced lifespan, and safety risks.

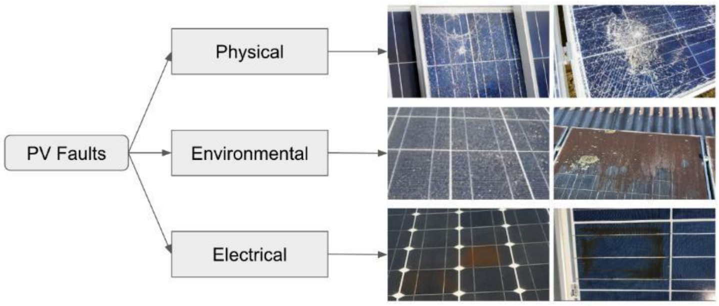


Figure 2. Examples of physical, environmental and electrical faults in PV modules.

Early detection of PV system faults is essential to prevent severe efficiency losses and equipment damage. Typical warning signs of PV faults include reduced output power, abnormal voltage-current behavior, excessive temperature rise, and irregular performance indicators. Continuous monitoring of electrical and environmental parameters allows operators to detect abnormal conditions and perform corrective actions before faults escalate.

Figure 3. Typical hotspot formation in photovoltaic modules

2.2 Diagnostic Methods for Photovoltaic Systems

Several diagnostic techniques are used to monitor photovoltaic systems and detect abnormal operating conditions. These methods rely on the measurement of electrical, environmental, and thermal parameters and their analysis using signal processing or intelligent algorithms.

2.2.1 Model-Free Diagnostic Methods

Model-free diagnostic methods are widely used due to their simplicity and effectiveness. These techniques rely on direct measurement and analysis of system signals without requiring detailed mathematical models.

- **Voltage and Current Monitoring**

Voltage and current measurements are fundamental indicators of PV system performance. Deviations from expected voltage and current values may indicate electrical faults such as open circuits, short circuits, or degradation of modules.

- **Power Analysis**

Power output monitoring is one of the most important diagnostic methods in photovoltaic systems. A decrease in output power under normal irradiance conditions may indicate the presence of faults such as soiling, shading, or module degradation.

- **Temperature Monitoring**

Temperature sensors are used to detect abnormal temperature rises associated with hotspot formation or overheating conditions. Excessive temperature levels significantly affect module efficiency and may lead to permanent damage.

- **Irradiance Monitoring**

Solar irradiance measurement is essential for evaluating the expected output power of PV modules. Comparing measured power with expected power based on irradiance allows detection of performance losses caused by environmental faults. Although these methods provide valuable information individually, combining multiple measurements improves fault detection accuracy and reliability.

2.2.2 Diagnosis Using Performance Indicators

Performance indicators are widely used for PV fault diagnosis because they provide quantitative measures of system efficiency. One of the most used indicators is the Performance Ratio (PR), which represents the ratio between the measured output power and the expected power under given environmental conditions. A significant decrease in PR indicates performance degradation or fault occurrence. Another important indicator is the Fill Factor (FF), which represents the quality of the PV module and reflects the relationship between maximum power, open-circuit voltage, and short-circuit current. A decrease in FF often indicates module degradation or electrical defects. Additional performance indicators include:

- Soiling Index (SI) for dust detection
- Heat Index (HI) for temperature-related faults
- Power Loss Ratio for performance evaluation

Performance indicator-based diagnosis is highly effective because it allows continuous monitoring and early detection of abnormal behavior without requiring complex instrumentation.

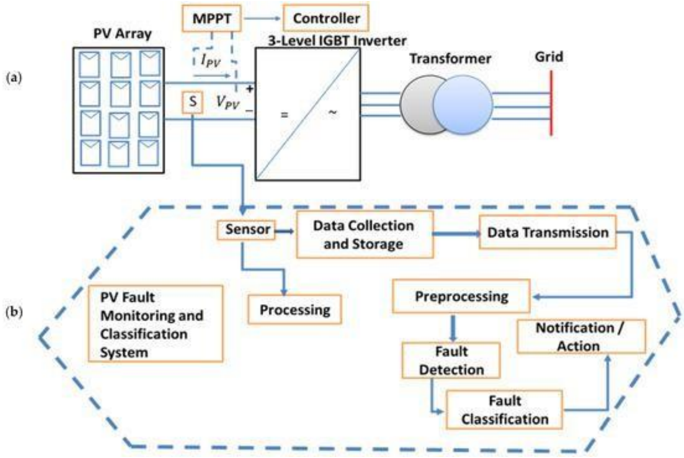


Figure 4. Block diagram of performance indicator-based fault detection system

2.2.3 Diagnosis Using Mathematical Modeling

Mathematical modeling plays an important role in predicting photovoltaic system behavior under different operating conditions. By developing accurate models of PV modules, it is possible to simulate both healthy and faulty conditions. Simulation tools such as MATLAB and Simulink allow the creation of realistic PV system models. These models incorporate electrical parameters, temperature effects, and environmental conditions to generate representative datasets.

Fault simulation techniques include:

- Reducing irradiance to simulate soiling
- Increasing series resistance to simulate degradation
- Introducing breaker elements to simulate open circuits
- Adding low-resistance paths to simulate short circuits
- Increasing temperature to simulate hotspot conditions

Simulation-based diagnosis provides a safe and cost-effective way to study PV faults without damaging physical equipment.

2.2.4 Diagnosis Using Artificial Intelligence Techniques

Artificial Intelligence (AI) techniques have recently become powerful tools for photovoltaic fault diagnosis. These methods enable automatic detection and classification of faults based on measured data. Machine learning algorithms can analyze large datasets generated from PV systems and identify patterns associated with different fault conditions. Common AI techniques used in PV diagnostics include:

- **Artificial Neural Networks (ANN)**

ANN models can learn nonlinear relationships between input signals and fault conditions. They are widely used for classification and prediction tasks.

- **Random Forest Classification**

Random Forest is an ensemble learning technique that uses multiple decision trees to improve classification accuracy. It is particularly effective for handling complex datasets and reducing classification errors.

- **Support Vector Machines (SVM)**

SVM models are used for classification problems where clear separation between classes is required.
The integration of AI techniques with real-time monitoring systems significantly improves fault detection accuracy, reduces maintenance costs, and enhances system reliability.

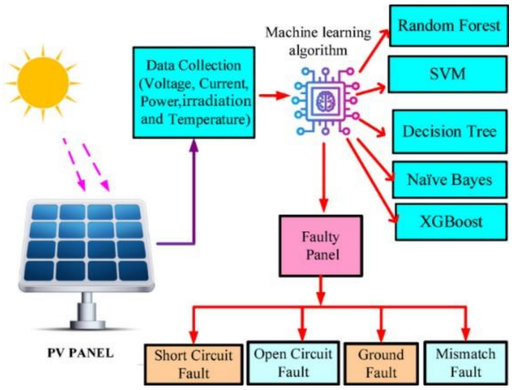


Figure 5. Intelligent photovoltaic fault detection framework using machine learning techniques

3. PV System and Fault Modeling

3.1 Photovoltaic System Configuration

The photovoltaic system considered in this study is based on a grid-connected architecture designed to generate electrical power under varying environmental conditions. The system consists of photovoltaic modules connected in series and parallel configurations to achieve the required output power level. In this work, a total of 180 photovoltaic modules are used, arranged in 20 modules connected in series per string and 9 strings connected in parallel, forming the complete photovoltaic generator.

Each photovoltaic module used in this study corresponds to the Trina Vertex 555 W photovoltaic module, selected due to its high efficiency and reliable performance under standard test conditions (STC). This configuration allows the system to provide sufficient voltage and current levels suitable for grid-connected operation while maintaining high energy conversion efficiency.

The electrical characteristics of the selected module are summarized in Table 1.

Table 1. Electrical Parameters of the Photovoltaic Module

Parameter	Value
Maximum Power P_{max}	555 W
Voltage at Maximum Power V_{mp}	31.8 V
Current at Maximum Power I_{mp}	17.45 A
Open-Circuit Voltage V_{oc}	38.1 V
Short-Circuit Current I_{sc}	18.56 A
Module Efficiency	21.2 %
Temperature Coefficient k_v	−0.33 %/°C
Temperature Coefficient k_i	0.047 %/°C
Power Coefficient γ	−0.377 %/°C
Nominal Operating Cell Temperature (NOCT)	43 ± 2 °C

3.2 Mathematical Modeling of the PV Module

The electrical behavior of a photovoltaic module can be represented using a single-diode equivalent circuit model (Figure 6). This model includes a current source representing the photocurrent, a diode modeling the p–n junction behavior, and series and shunt resistances representing internal losses.

The output current of the photovoltaic module is expressed as:

$$I_{pv} = I_{ph} - I_o\left[\exp\left(\frac{V_{pv} + R_s I_{pv}}{n.V_t}\right) - 1\right] - \frac{V_{pv} + R_s I_{pv}}{R_{sh}} \quad (1)$$

Where: I_{ph}: Photocurrent, I_o: Diode saturation current, R_s: Series resistance, R_{sh}: Shunt resistance, V_t: Thermal voltage and n: Ideality factor

The thermal voltage of the photovoltaic module is expressed as:

$$V_t = \frac{N_{cell}.\,k_B.\,T}{q} \quad (2)$$

With: T is the absolute temperature of the photovoltaic module, N_{cell} is the number of cells connected in series within the PV module and k_B is the Boltzmann constant.
This model enables accurate simulation of the PV module under different irradiance and temperature conditions.

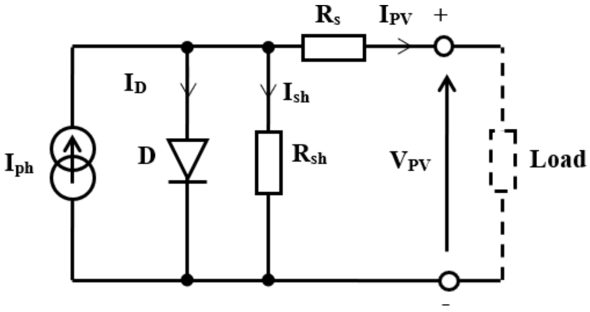


Figure 6. Equivalent circuit model of the photovoltaic module

3.3 Temperature and Irradiance Effects

Environmental conditions such as temperature and solar irradiance significantly influence the electrical output of photovoltaic modules. Therefore, temperature-dependent corrections are applied to the open-circuit voltage and short-circuit current.
The temperature-corrected open-circuit voltage is given by:

$$V_{oc}(T) = V_{oc,ref}[1 + k_v(T - T_{ref})] \quad (3)$$

The temperature and irradiation-corrected short-circuit current is given by:

$$I_{sc}(T) = I_{sc,ref} \times \frac{G}{G_{ref}} \times [1 + k_i(T - T_{ref})] \quad (4)$$

Similarly, the expected output power is calculated using:

$$P_{expected} = P_{ref} \times \frac{G}{G_{ref}} \times [1 + \gamma.(T - T_{ref})] \quad (5)$$

Where: P_{ref} : reference installed power (W), G: Solar irradiance (W/m²), T: Cell temperature (°C), G_{ref}, T_{ref} : reference irradiation and temperature at standard operating conditions (1000 W/m² and 25°C), k_v, k_i and γ : Voltage, current and power temperature coefficient (%/°C)
These equations are essential for calculating diagnostic performance indicators such as Performance Ratio (PR) and Fill Factor (FF).
Figure 7 illustrates the current–voltage (I–V) and power–voltage (P–V) characteristics of a photovoltaic module under various irradiance levels at a constant temperature of 25°C.

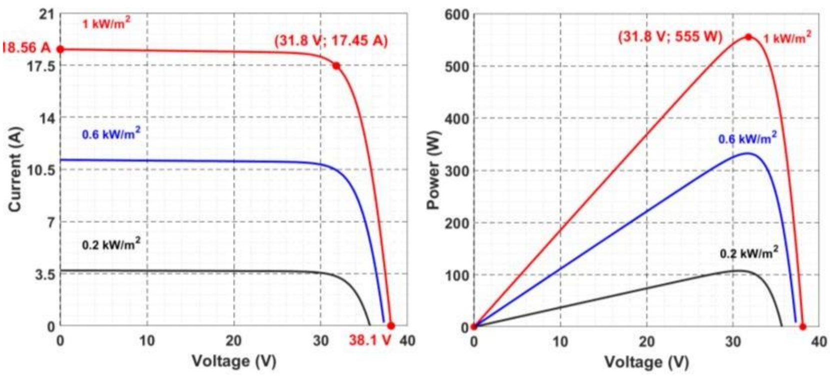


Figure 7. I–V and P–V characteristics of the PV module under different irradiance levels

Figure 8 illustrates the current–voltage (I–V) and power–voltage (P–V) characteristics of the PV module for different temperature values under standard irradiance conditions (1000 W/m²). The results clearly show that the output power of the PV array decreases when the temperature exceeds 25°C and increases when the temperature falls below 25°C.

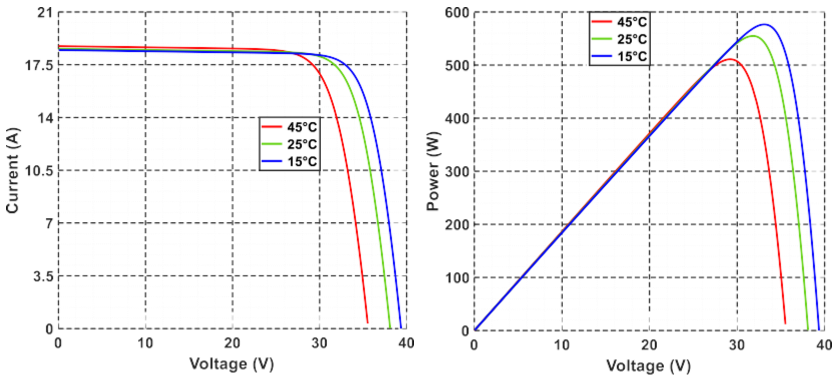


Figure 8. I–V and P–V characteristics of the PV module under different temperature levels

3.4 Key Performance Calculations

Based on the current values obtained from measurements and signals generated by the Simulink photovoltaic system model, several performance parameters are calculated in real time. These indicators are used to evaluate the dynamic behavior of the PV generator under simulated conditions, considering variations in irradiance, temperature, and electrical quantities. The objective is to analyze the system efficiency as well as the different sources of losses to validate the performance of the model.

3.4.1 Fill Factor (FF)

The Fill Factor represents the quality of the solar cell's I–V characteristic:

$$FF(\%) = \frac{V_{mp} \times I_{mp}}{V_{oc} \times I_{sc}} \times 100 \quad (6)$$

where: V_{mp}, I_{mp}: voltage and current at maximum power point, V_{oc} : open-circuit voltage and I_{sc} : short-circuit current

3.4.2 Performance Ratio (PR)

The Performance Ratio evaluates the actual system efficiency compared to its theoretical output:

$$PR(\%) = \frac{P_{actual} \times G_{ref}}{P_{ref} \times G} \times 100 \quad (7)$$

Where: P_{actual}: actual power output (W)

3.4.3 Soiling Index (SI)

The Soiling Index quantifies losses due to dust and dirt on the panel surface:

$$SI(\%) = (1 - \frac{P_{actual}}{P_{expected}}) \times 100 \quad (8)$$

3.4.4 Thermal Index (HI)

The Thermal Index estimates performance loss due to temperature increase:

$$HI(\%) = \frac{\mathrm{T} - (T_{ref} + 0.03 \times G)}{T_{ref}} \times 100 \quad (9)$$

Where: T : operating temperature and T_{ref} : reference temperature (usually 25°C)

3.5 Fault Modeling in Photovoltaic Systems

Fault modeling is a critical step in developing an intelligent diagnostic system. In this study, several common PV faults are simulated to analyze their impact on system performance. The following fault conditions are considered as priority:

3.5.1 Short-Circuit Fault

Short-circuit faults occur when unintended conductive paths form between terminals. This condition is simulated by introducing a very low resistance path between the output terminals:

$$\begin{cases} V \le 0.1 \times V_{oc} \\ and \\ I > 0.8 \times I_{sc} \end{cases} \tag{10}$$

3.5.2 Open Circuit or Connector Fault

Open-circuit faults occur when electrical connections are interrupted, preventing current flow. This condition is simulated by inserting an open breaker within the PV circuit, resulting in:

$$\begin{cases} I < 0.2 \times I_{sc} \\ and \\ P_{actual} < 0.2 \times P_{ref} \\ and \\ V > 0.8 \times V_{oc} \\ and \\ G > 300\ W/m^2 \end{cases} \tag{11}$$

3.5.3 Hotspot Fault

Hotspot formation occurs when a localized area of the module operates at elevated temperature levels. This condition is simulated by increasing the module temperature beyond normal operating limits:

$$\begin{cases} T > 65\ °C \\ and \\ G > 300\ W/m^2 \end{cases} \tag{12}$$

Hotspot faults lead to reduced output voltage and power.

3.5.4 Degradation or Bypass Fault

Degradation faults are caused by aging of photovoltaic cells and deterioration of internal components. This condition leads to increased internal resistance and reduced efficiency.

$$\begin{cases} PR < 80\ \% \\ and \\ SI > 40\ \% \\ and \\ FF < 60\ \% \end{cases} \tag{13}$$

The degradation condition is simulated by increasing the series resistance R_s of the PV module.

3.5.5 Soiling or Shading fault

Soiling or partial Shading faults are caused by dust accumulation on the surface of photovoltaic modules or by a partial shading when a portion of the PV module receives less irradiance than the rest of the system. This condition results in mismatch losses and decreased power output. The soiling or shading fault is simulated by reducing irradiance on selected strings within the PV array. This condition reduces the effective irradiance received by the module, resulting in reduced current generation. A decrease in output current and power is observed under this condition.

$$\begin{cases} G > 400\ W/m^2 \\ and \\ PR < 80\ \% \\ and \\ SI > 20\ \% \\ and \\ FF > 50\ \% \end{cases} \tag{14}$$

4 Fault Detection System Model

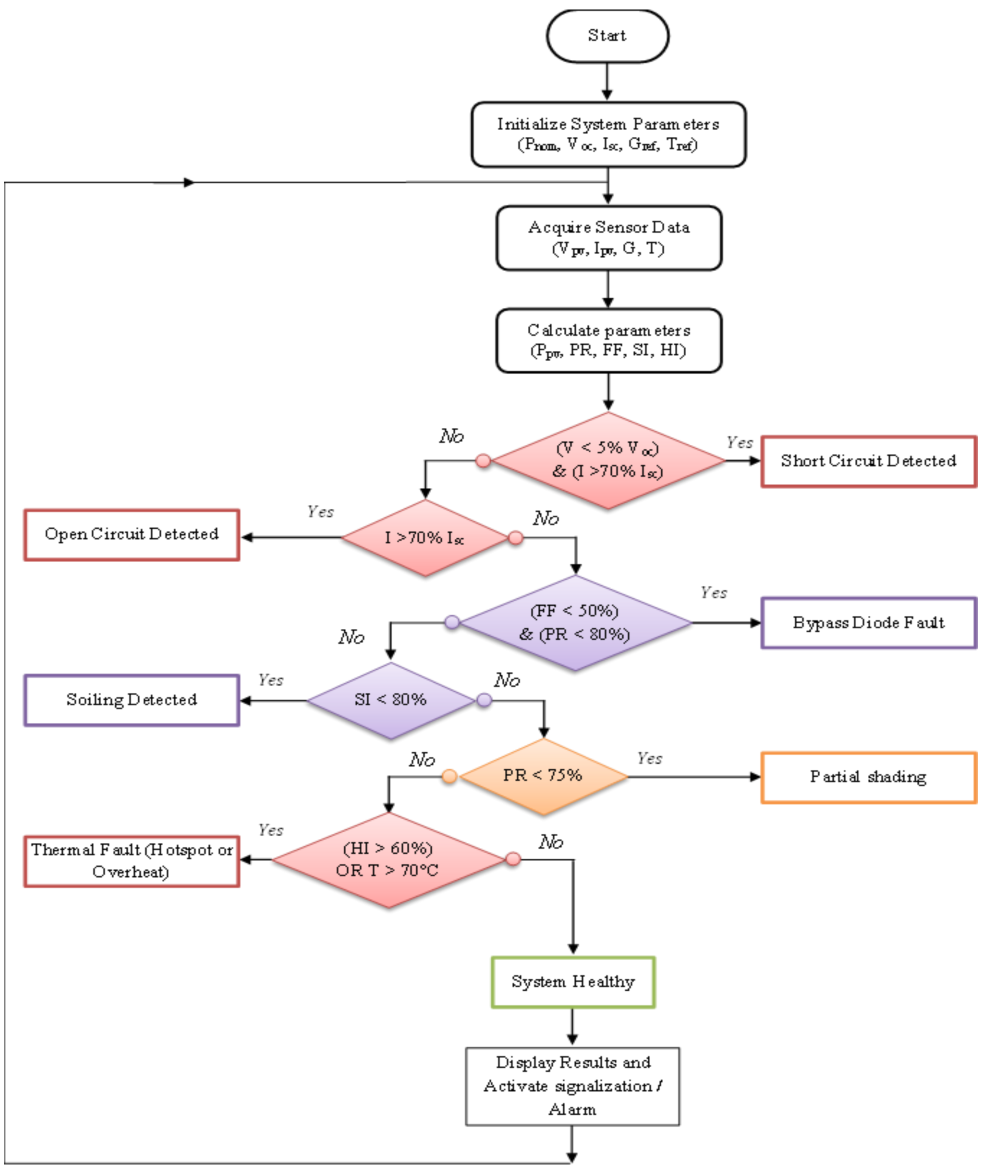
This section presents the proposed fault detection system model for photovoltaic (PV) systems, which is based on two complementary approaches: a logical method and an intelligent method. The main objective is to detect and classify faults occurring in a grid-connected PV generator using electrical and environmental measurements. The logical approach relies on rule-based performance monitoring using voltage, current, irradiance, and temperature signals to identify abnormal operating conditions through predefined thresholds and performance indicators. In parallel, the intelligent approach employs machine learning techniques to improve fault classification accuracy by learning patterns from historical and real-time data. This section describes the system architecture, data processing procedure, and the methodology used in both detection approaches.

4.1 Logical Algorithm Diagram

This mechanism ensures simple and efficient monitoring of the photovoltaic installation. Figure 9 illustrates the logical flow of the photovoltaic (PV) fault diagnosis algorithm. The process begins with the initialization of system parameters, followed by the acquisition of sensor data, including voltage (V), current (I), irradiance (G), and temperature (T). Based on these measurements, the system computes several performance indicators such as generated power (P), Performance Ratio (PR), Fill Factor (FF), Soiling Index (SI), and Heat Index (HI).

The algorithm then performs a series of rule-based conditional checks to detect possible faults, such as short circuit, open circuit, module degradation, bypass diode failure, soiling, partial shading, hotspot formation, and overheating. Once a fault condition is identified, the corresponding diagnostic message is displayed on the monitoring interface, and appropriate visual or alarm signals are activated to indicate the detected anomaly.

Figure 9. PV Fault logical Diagnosis Algorithm Flowchart



4.2 Intelligent fault detection Architecture and Methodology

The intelligent fault detection method is based on a data-driven approach that utilizes machine learning techniques, specifically a Feedforward Neural Network (FFNN), to forecast system measurements and automatically detect and classify faults in the photovoltaic (PV) system. Unlike the logical method, which relies on predefined rules and threshold values, the intelligent approach learns the nonlinear relationship between system measurements and fault conditions using historical and simulated datasets. In addition to fault classification, the FFNN model is employed to predict future values of key PV measurements, enabling early detection of abnormal system behavior. The structured methodology adopted for developing both the forecasting and fault classification models is illustrated in the flowchart shown in Figure 10.

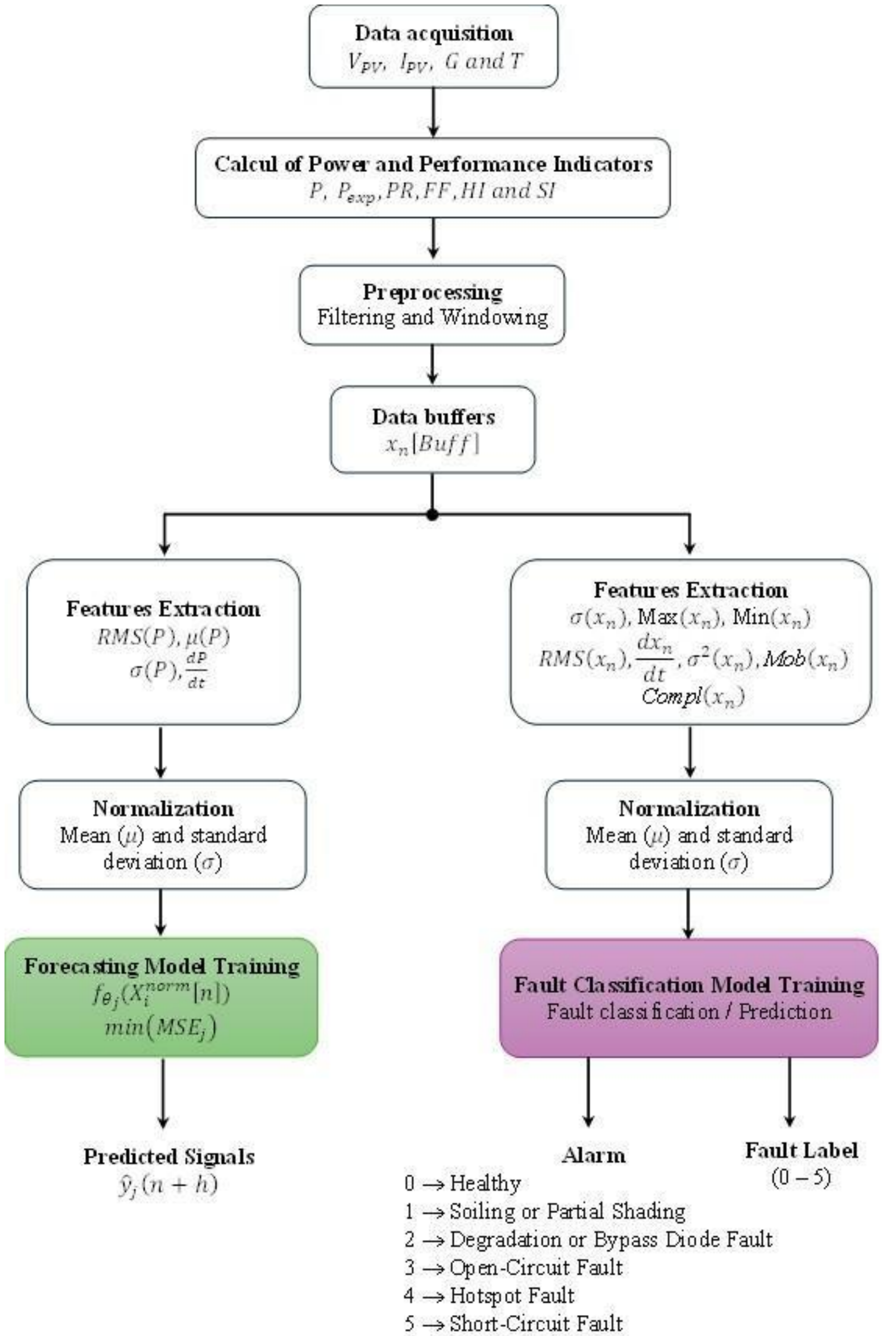


Figure 10. Intelligent System Flow Chart

The photovoltaic fault detection and diagnosis process consists of several sequential stages. First, electrical and environmental data are acquired from the PV system, including voltage, current, irradiance, and temperature measurements. These signals represent the operating conditions of the photovoltaic generator under varying environmental conditions. Following data acquisition, the measured signals undergo preprocessing operations to enhance data quality and reliability. These preprocessing steps include filtering to remove noise, normalization to scale the data into a consistent range, and validation to eliminate abnormal or missing values. After preprocessing, relevant features are extracted directly from the measured signals and used as input variables for the FFNN forecasting model. The trained FFNN predicts future values of the PV measurements, allowing the system to monitor expected system behavior. Deviations between predicted and actual measurements may indicate abnormal operating conditions. Finally, the measured and forecasted data are processed using FFNN-based classification to identify and categorize different fault types, such as soiling, partial shading, degradation, open-circuit faults, short-circuit faults, and hotspot formation. This intelligent methodology improves early fault detection capability and enhances the reliability and accuracy of photovoltaic system monitoring.

4.3 Forecast Model of PV Measurements

For forecasting electrical signals measurements and performance indicators (FF and PR) in PV systems, machine learning (ML) methods are widely applied due to their ability to model nonlinear system behavior. These methods mainly include (Le et al., 2023):

- **Classical ML methods** (Artificial Neural Networks (ANN), Support Vector Regression (SVR), Random Forest): which operate on pre-extracted features derived from measured signals.
- **Deep Learning methods** (LSTM, Bi-LSTM, GRU): which can capture temporal dependencies in raw or processed time-series data.
- **Hybrid models**: which combine feature extraction techniques with sequence learning models to achieve improved prediction accuracy.

In this work, forecasting of photovoltaic system measurements is performed using a supervised learning approach based on Feedforward Neural Networks for regression (FFNN) (Diab et al., 2022). This method consists of training a feedforward neural network to predict continuous output variables, such as voltage, current, irradiance, and temperature, from previously measured data. The FFNN model approximates the nonlinear relationship between past measurement values and their future behavior, enabling accurate prediction of PV system operating conditions.
The regression neural network is designed with multiple fully connected layers. Each neuron performs a weighted linear combination of its input signals, followed by a nonlinear activation function, allowing the network to capture complex interactions among the input variables. The final output layer produces the predicted values of the PV measurements. The implementation of the proposed FFNN forecasting model using MATLAB and Simulink is described in Section 5.2.

4.3.1 Preprocessing steps

- *FilteringThe filtered and sampled signal $x_f(n)$ is obtained by applying a filter $H(\cdot)$ to the raw of signal $x(t)$:*

$$x_f(n) = H(x(n)) \tag{15}$$

Where: $x(n) = [V(n) \quad I(n) \quad P(n) \quad FF(n) \quad PR(n)]$, the signal vector and $H(\cdot)$ can represent a low-pass, high-pass, or band-pass filter depending on the application.

- *Windowing*

Windowing ensures the FFNN sees local temporal patterns while keeping computation manageable and allows smooth, accurate segmentation of motor operating states. Let:
$x(n)$: the sampled signal ($n = \mathrm{T}_{\text{final}} / T_s$)
T_{final} : total running time of the simulation
W : window length (number of samples)
T_s : Sampling time of the simulation
H : hop size (step between consecutive windows)
The windowed segments x_k are defined as:

$$\left\{ \begin{array}{c} x_k[n] = x[n + k_H] \\ n = 0, 1, \dots, W-1, \qquad k = 0, 1, \dots, K-1 \end{array} \right. \tag{16}$$

Where K is the total number of windows:

$$K = \frac{T - W}{H} + 1 \tag{17}$$

With: $x_k[n]$: samples in the k-th window n *:* index within the window and k_H *:* start of the k-th window in the original signal

- $Overlap = W - H$ samples

4.3.2 Feature Extraction for Forecasting

For forecasting the PV signals, relevant time-domain features are extracted directly from the measured signals, including voltage, current, power, and performance indicator. Since PV signals are mainly slow-varying DC quantities influenced by environmental changes, statistical and differential features are more appropriate than frequency-based indicators. The selected features for PV signal forecasting are chosen to capture both the steady-state behavior and the dynamic variations of the PV system. The root mean square of power (RMS(P)) is used to represent the overall energy level of the generated power over a short time window, while the standard deviation of power (Std(P)) reflects the variability and fluctuations in power output caused by environmental changes or abnormal operating conditions. In addition, the power derivative (dP/dt) is included to describe the rate of change of the generated power, allowing the model to detect rapid transitions in system behavior. The moving averages of the Fill Factor (Mean (FF)) and Performance Ratio (Mean(PR)) are also incorporated to provide information about the efficiency and performance trend of the PV system. These combined features provide a compact and informative representation of the PV system dynamics, improving the accuracy and robustness of the forecasting model.
Thus, the feature vector at sample n is defined as:

$$features[n] = [RMS(P_n) \quad \sigma(P_n) \quad \frac{dP_n}{dt} \quad \mu(FF_n) \quad \mu(PR_n)] \tag{18}$$

a) Root Mean Square of Power

This feature represents the energy level of the power signal over the selected time window.

$$RMS(P_n) = \sqrt{\frac{1}{\omega_n}\sum_{k=n-\omega_n+1}^{n} P[k]^2} \tag{19}$$

Where: $P[k]$= power signal at sample k and $\omega_n = \min(n, W)$= effective window size

b) Standard Deviation of Power

The Std(P) feature measures the fluctuation and variability of the generated power.

$$\sigma(P_n) = \sqrt{\frac{1}{\omega_n}\sum_{k=n-\omega_n+1}^{n} (P[k] - \mu(P_n))^2} \tag{20}$$

Where the mean power is:

$$\mu(P_n) = \frac{1}{\omega_n}\sum_{k=n-\omega_n+1}^{n} P[k] \tag{21}$$

c) Power Derivative

This feature describes the rate of change of power, useful for detecting sudden variations.

$$\frac{dP_n}{dt} = \frac{P[n] - P[n-1]}{T_s} \tag{22}$$

d) Mean of Fill Factor and Performace Ratio

This feature reflects the average conversion efficiency of the PV module.

$$\begin{aligned} \mu(FF_n) &= \frac{1}{\omega_n}\sum_{k=n-\omega_n+1}^{n} FF[k] \\ \mu(PR_n) &= \frac{1}{\omega_n}\sum_{k=n-\omega_n+1}^{n} PR[k] \end{aligned} \tag{23}$$

Where: $FF[k]$ *and* $PR[k]$ are the Fill Factor and Performance Ratio at sample k

The input vector for the forecasting model is therefore defined as:

$$X[n] = [V[n], I[n], \mathrm{P}[n], FF[n], PR[n], \text{features}] \tag{24}$$

The output vector (predicted values) corresponds to the future values of the PV signals:

$$Y[n+1] = [V[n+1], I[n+1], \mathrm{P}[n+1], FF[n+1], PR[n+1]] \tag{25}$$

This formulation transforms the forecasting problem into a supervised learning task, where the Feedforward Neural Network (FFNN) learns the nonlinear mapping between past measurements and future PV system behavior.

4.3.3 Z-Score normalization

Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 03 (2026)

Z-Score normalization (also called standardization) transforms data so that each variable has zero mean (μ=0) and unit variance (σ=1). Normalized signals are computed as:

$$\begin{cases} y_j^{norm}[n] = \dfrac{y_j[n] - \mu_y}{\sigma_y} \\ y_j^{norm}[n] = \dfrac{y_j[n] - \mu_y}{\sigma_y} \end{cases} \tag{26}$$

Where: μ_x, σ_x are mean and standard deviation of feature i, and μ_y, σ_y are mean and standard deviation of feature j,

$$\begin{cases} \mu_x = \frac{1}{L}\sum_{i=1}^{L} x_i[n] \\ \sigma_x = \sqrt{\frac{1}{L}\sum_{i=1}^{L} (x_i[n] - \mu_x)^2} \end{cases} \tag{27}$$

$$\begin{cases} \mu_y = \frac{1}{L}\sum_{i=1}^{L} y_j[n] \\ \sigma_y = \sqrt{\frac{1}{L}\sum_{j=1}^{L} (y_j[n] - \mu_y)^2} \end{cases} \tag{28}$$

Where: $L = length(x) - 1$.

4.3.4 Forecast Model Training

For each target variable ($V(t)$, $G(t)$ and $T(t)$), a regression neural network ($fitrnet$) is trained. Each model approximates the function:

$$\hat{y}_j^{norm}(n + h) = f_{\theta_j}(x_i^{norm}[n]) \tag{29}$$

Where: $\hat{y}_j^{norm}(n + h)$ is the normalized value predicted h steps ahead, $x_i^{norm}[n]$ are historical normalized measurements and f_{θ_j} is the Neural Network function parameterized by weights and biases θ_j.

The training minimizes mean squared error (MSE):

$$min(MSE_j) = \frac{1}{L}\sum_{n=1}^{L} (y_j^{norm}[n] - f_{\theta_j}(x_i^{norm}[n]))^2 \tag{30}$$

The model predicts a normalized output $y_j^{norm}[n]$, so we must return to the physical scale. We apply the inverse of normalization (Denormalization):

$$\hat{y}_j = \hat{y}_j^{norm}.\sigma_y + \mu_y \tag{31}$$

4.3.5 Calculation of Forecasting Errors

To validate the PV system voltage, current, irradiation and temperature prediction operation, the performance is evaluated using absolute errors, relative errors and mean square errors. The mathematical relationships of the different errors are given by:

a) Absolute Error

The absolute error directly gives an amplitude of the difference between the actual behavior of the induction motor and that predicted by the forecasting model, expressed in the units of the quantities concerned.

$$\Delta y_j = |y_j[n] - \hat{y}_j[n + h]| \tag{32}$$

Where: $y_j[n]$: actual value and $\hat{y}_j[n + h]$: predicted value h steps ahead, with $j = 1, 2, \dots n$: (number of observations)

b) Relative Error

The relative error allows us to judge the severity of the error in relation to the actual signal from the motor.

$$\frac{\Delta y_j}{y_j}(\%) = \frac{\Delta y_j[n]}{|y_j[n]|} \times 100 \tag{33}$$

c) Mean Square Error

The mean squared error (MSE) is the average squared difference between the measured and predicted values. For time-series forecasting, this can be interpreted as the average power of the prediction error signal.

$$MSE(y_j) = \frac{1}{L}\sum_{n=1}^{L}(\Delta y_j[n])^2 \quad (34)$$

d) Root Mean Square Error

The Root Mean Squared Error (RMSE) is the square root of the average squared difference between the measured and predicted values. In time-series forecasting, it can be interpreted as the root of the average power of the prediction error signal.

$$RMSE(y_j) = \sqrt{\frac{1}{L}\sum_{n=1}^{L}(\Delta y_j[n])^2} \quad (35)$$

. Random Forest for Fault Classification ModelRandom Forest-based forecasting model is given by Figure 11.

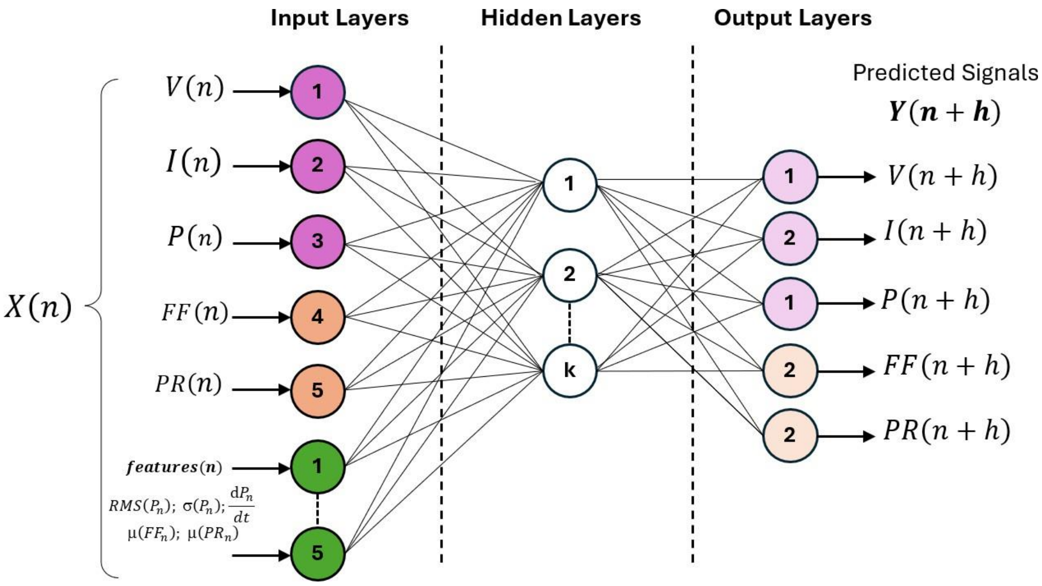


Figure 11. Random Forest for Forecasting Model

4.4 Faults Classification/Prediction Model

To diagnose and detect anomalies (bearings) in the PV system, features are extracted from the dataset produced by the simulated healthy and faulty PV system signals to train a Random Forest model. Then feature selection process step is added to select the 30 best features that produces that most accurate results before training the ML algorithm. This section describes the steps used to build the fault prediction model.

4.4.1 Signal Buffering

Each predicted signal $\hat{y}_j$ is buffered at the frequency:

$$f_s = \frac{1}{T_s} \quad (36)$$

4.4.2 Features Extraction for Fault Prediction Model

To characterize the PV system behavior and detect abnormal operating conditions, statistical, dynamic, and nonlinear features are extracted from signal vector measurement. Statistical features (mean, standard deviation, min, max, RMS) describe the overall behavior and stability of the system. Dynamic features, based on first-order differences, capture rapid variations and transient changes. Nonlinear features, derived from Hjorth parameters (activity, mobility, and complexity), quantify signal intensity, rate of change, and irregularity. Together, these features provide a compact and comprehensive representation of PV system conditions for effective monitoring and fault detection.

Thus, the feature vector at sample n is defined as:

$$features = [\mu(x_n)\ \ \sigma(x_n)\ \mathrm{Max}(x_n)\ \mathrm{Min}(x_n)\ RMS(x_n)\ \ \sigma^2(x_n)\ \ \mathrm{Mob}(\mathrm{x_n})\ \ \mathrm{Compl}(\mathrm{x_n})] \tag{37}$$

The signal vector at sample n is defined as:
$x(n) = [G(n), T(n), V(n), I(n), P(n), FF(n), PR(n), SI(n), HI(n)]$.
Mathematical equations of a different feature extraction are defined as follows:

a) Basic statistical features

- *Mean*

$$\mu(x_n) = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{38}$$

- *Standard deviation*

$$\sigma(x_n) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu(x_n))^2} \tag{39}$$

- *Maximum / Minimum*

$$\begin{cases} \mathrm{Max}(x_n) = \max(x_1, x_2, \ldots, x_N) \\ \mathrm{Min}(x_n) = \min(x_1, x_2, \ldots, x_N) \end{cases} \tag{40}$$

- *Root Mean Square (RMS)*

$$RMS(x_n) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n)^2} \tag{41}$$

- *First derivative (discrete difference)*

$$\frac{dx_n}{dt} = \frac{x_n - x_{n-1}}{T_s} \tag{42}$$

b) Hjorth Parameters

- *Activity (variance)*

$$Var(x_n) = \sigma^2(x_n) = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2 \tag{43}$$

- *Mobility*

$$\mathrm{Mob}(\mathrm{x_n}) = \sqrt{\frac{\mathrm{var}\left(\frac{\mathrm{dx_n}}{\mathrm{dt}}\right)}{\mathrm{Var(x)}}} \tag{44}$$

- *Complexity*

$$\mathrm{Compl}(\mathrm{x_n}) = \frac{\sqrt{\frac{\mathrm{var}\left(\frac{\mathrm{d^2x_n}}{dt^2}\right)}{\mathrm{var}\left(\frac{\mathrm{dx_n}}{\mathrm{dt}}\right)}}}{Mobility} \tag{45}$$

4.4.3 Fault Classification Model Training

In this step, a supervised classification model is trained using the selected diagnostic features extracted from electrical and environmental measurements to predict the operating condition of the photovoltaic (PV) system. The input features include voltage, current, irradiance, temperature, generated power, Performance Ratio (PR), Fill Factor (FF), and statistical indicators derived from the power signal. The classification model is designed to identify multiple PV system operating states corresponding to healthy and faulty conditions. The fault classes considered in this study are defined as follows:

- Class 0 : Healthy condition
- Class 1 : Soiling or partial shading fault
- Class 2 : Degradation or bypass diode fault
- Class 3 : Open-circuit fault
- Class 4 : Hotspot fault
- Class 5 : Short-circuit fault

The training process involves feeding labeled datasets generated from simulation scenarios into the machine learning model. Each dataset sample contains the extracted features and the corresponding fault label representing the system condition. The model learns the relationship between feature patterns and fault categories to enable accurate classification of system states. After training, the classification model produces an output label corresponding to the detected fault type. In addition, an alarm signal is generated whenever a fault condition is detected, allowing real-time monitoring and early fault identification in the PV system. . Random Forest for Fault Classification ModelRandom Forest-based fault classification model for photovoltaic system diagnosis.

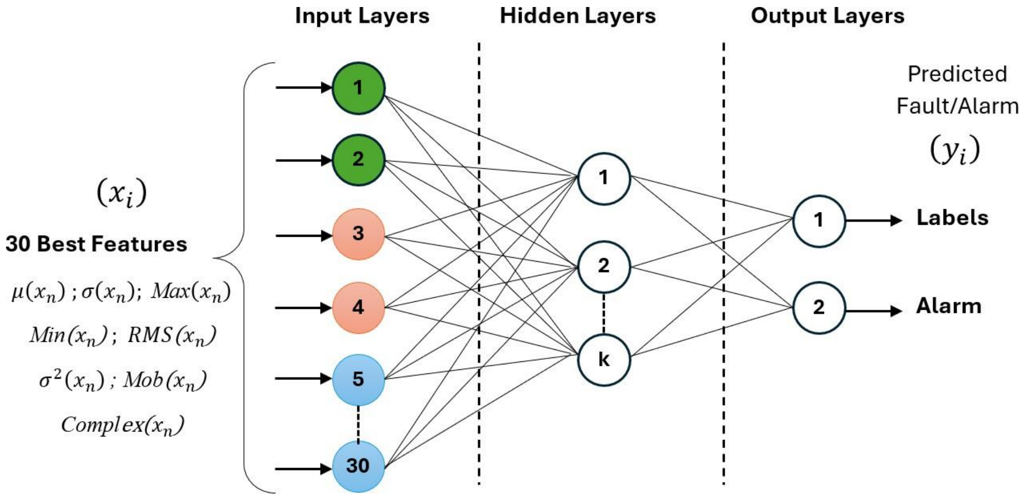


Figure 12. Random Forest for Fault Classification Model

Training is performed using Random Forest with 5-fold cross-validation. This algorithm is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and robustness (Jiang, 2023). Each tree is trained on a bootstrap sample of the dataset, drawn with replacement, and at each node split only a random subset of features is considered to reduce correlation between trees. The Random Forest is built as a two-level ensemble model:

a) *Input Layer (Data + Features)*

The dataset $D = \{(x_i, y_i)\}_{i=1}^{L}$, where: $x_i \in \mathbb{R}^{72}$: feature vector. y_i : class label (Healthy, DF1, DF2, …, DF5).

b) *Ensemble of Decision Trees (Forest Layer)*

The Random Forest consists of D decision trees; each trained on a different subset of data and features:

- *Bootstrap Sampling (Bagging):*

Each decision tree is trained using a randomly selected subset of data, with replacement.

- *Random Feature Selection:*

At every split within the tree, a random portion of the available features is chosen (30 best features in this work). This approach lowers correlation among trees and prevents overfitting.

- *Decision Tree Structure:*

Root Node → all data from the bootstrap sample. Internal Nodes → split conditions chosen to minimize impurity.

- *Gini Index*

$$G(p) = 1 - \sum_{k=1}^{K} (p_k)^2 \tag{46}$$

- *Entropy*

$$H(p) = -\sum_{k=1}^{K} p_k . log_2 . (p_k) \tag{47}$$

Leaf Nodes → predicted class (classification) or numeric value (regression).

c) *Output Layer (Aggregation)*

The predictions of all trees are combined classification and regression. For classification, the final prediction is obtained by majority voting:

$$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_D(x)\} \tag{48}$$

While for regression, the output is the average of all tree predictions:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{49}$$

where $h_t(x)$ is the prediction of the t-th tree and T is the total number of trees. Random Forest reduces variance through bagging, controls overfitting via random feature selection, and provides robust, generalizable models with built-in estimates of error using out-of-bag (OOB) samples.
Model performance is measured by classification loss and accuracy, providing robust estimation while reducing overfitting. The system jointly optimizes forecasting and classification:

$$\mathcal{L}_{total} = \alpha \cdot MSE_{forecast} + \beta \cdot CE_{classification} \tag{50}$$

Where:
$MSE_{forecast}$ is the mean squared error for prediction,
$CE_{classification}$ is the cross-entropy loss for classification.
α and β the weight the relative importance of the two tasks.

4.4.4 Fault Classification Performance Metrics

To evaluate the performance of the proposed photovoltaic fault classification model, several evaluation metrics are used, including **Accuracy**, **Precision**, **Recall**, and **F1-score**. These metrics provide quantitative measures of the classification model's ability to correctly identify different PV fault conditions. This section describes the calculation and interpretation of these performance indicators.

a) Accuracy

Accuracy represents the ratio of correctly classified samples to the total number of samples. It provides a general indication of the overall performance of the classification model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{51}$$

Where:
TP: True Positives
TN: True Negatives
FP: False Positives
FN: False Negatives
Accuracy provides an overall measure of model performance. However, in multi-class fault classification problems, accuracy alone may not fully reflect the reliability of the system, especially when class distributions are unbalanced. Therefore, additional metrics such as precision, recall, and F1-score are also considered.

b) Precision

Precision measures the ability of the classification model to correctly identify fault conditions without generating false alarms.

$$Precision = \frac{TP}{TP + FP} \tag{52}$$

A high precision value indicates that the system produces fewer false alarms, which is essential for reliable fault diagnosis in photovoltaic systems.

c) Recall

Recall represents the ability of the classification model to detect actual fault occurrences.

$$Recall = \frac{TP}{TP + FN} \tag{53}$$

A high recall value indicates that the model successfully detects most of the fault events and minimizes missed fault detections.

d) F1-Score

The F1-score is the harmonic mean of precision and recall. It provides a balanced evaluation between false alarms and missed fault detections.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{54}$$

A high F1-score indicates that the classification model achieves both high precision and high recall, making it suitable for reliable photovoltaic fault detection. In this study, these metrics are computed for each fault class and averaged using macro-averaging to evaluate the overall performance of the multi-class photovoltaic fault classification model.

5 MATLAB and Simulink Implementation

In this section, MATLAB R2024b and Simulink are used to implement the photovoltaic (PV) system model and the proposed fault detection method in Sections 3 and 4. A detailed Simulink model of the PV generator is developed, including both normal operating conditions and various fault scenarios such as open circuit, short-circuit, partial shading, soiling, and hotspot faults. This simulation model is used to generate data under different environmental and operating conditions, which is then used to train both the forecasting model and the fault classification model. The trained Feedforward Neural Network (FFNN) models are also integrated into Simulink to evaluate the performance of the proposed fault detection system in an online simulation environment. The results obtained are presented in the next section.

5.1 Modeling of the PV System and Fault Scenarios in Simulink

The developed Simulink model represents a grid-connected photovoltaic (PV) system operating under both healthy and faulty conditions. Figure 13 illustrates the overall structure of the PV system model implemented in Simulink. The model is composed of a photovoltaic array subjected to variable solar irradiance and temperature profiles to simulate real environmental conditions. The PV generator is connected to a load or power conversion stage, where electrical output variables such as voltage and current are continuously measured. To simulate fault conditions, different fault modules are introduced into the PV system, including open-circuit faults, short-circuit faults, partial shading, soiling effects, and hotspot conditions. These faults are applied at different time intervals to generate representative datasets for both healthy and faulty operation. The model records key electrical and environmental signals, including PV voltage, current, irradiance, and temperature. These signals are used for feature extraction, forecasting, and fault classification using the developed machine learning models. The simulation is executed with a fixed sampling time suitable for accurate dynamic analysis, ensuring high-resolution data for condition monitoring and fault diagnosis studies.

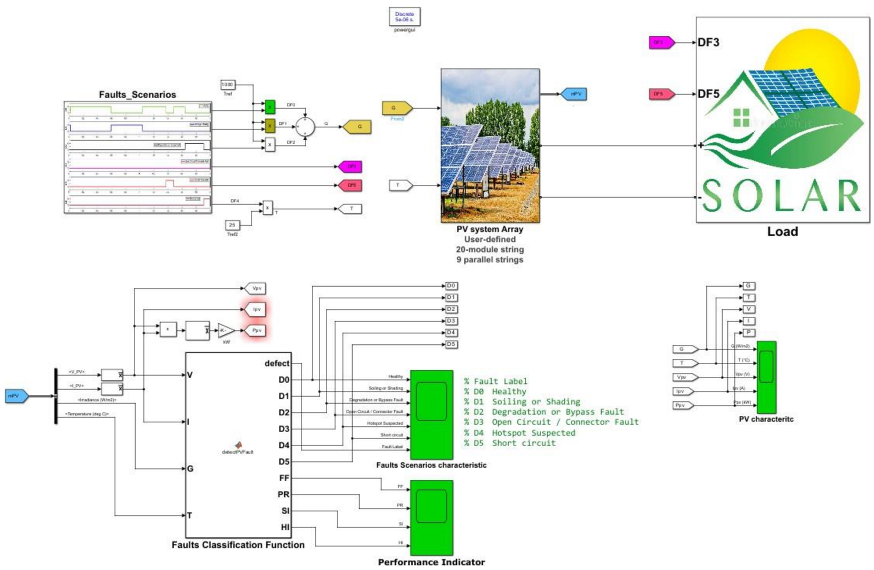


Figure 13. Simulink Model of the PV System with logical Faults diagnosis algorithm

5.2 Forecasting Model of PV Electrical Signals

Figure 14 presents the forecasting model designed to predict the future behavior of key photovoltaic system variables, including voltage, current, output power, and performance indicators (PR, FF). These parameters are critical for monitoring the operational state of the PV system and for early fault detection. The forecasting structure consists of a signal buffer and two MATLAB Function blocks responsible for feature extraction and forecasting operations. The buffer stores recent historical data over a selected time window, while the feature extraction module computes representative statistical characteristics of the signals. The forecasting module then generates predicted values, which are compared with real-time measurements to identify deviations that may indicate abnormal system behavior or developing faults. The simulation is conducted over a total duration of 2 hours, corresponding to multiple healthy and faulty operating intervals, enabling the generation of comprehensive datasets suitable for training and validating forecasting-based fault diagnosis models.

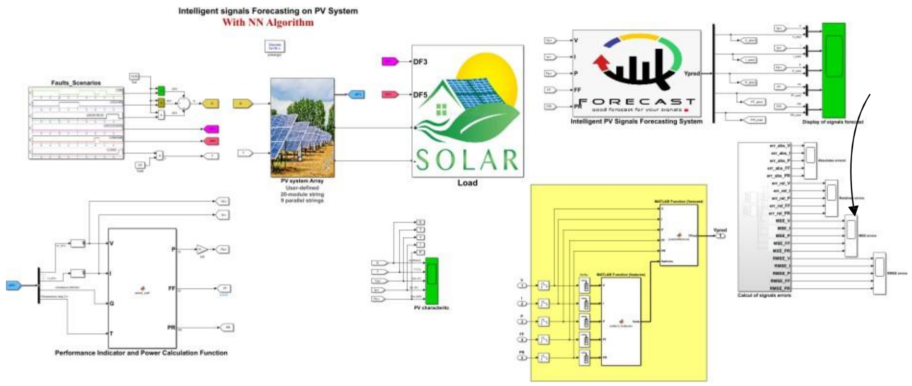


Figure 14. Online Forecasting Model of PV Signals

5.3 PV Fault Prediction Model

Figure 15 presents the online PV fault prediction and forecasting model. The system combines signal forecasting with a fault diagnosis block that processes predicted PV variables (V_{pv}, I_{pv}, P_{pv}, FF, PR). Feature extraction and classification functions are used to identify multiple fault conditions, including soiling, degradation, open circuit, short circuit, and hotspot. The model outputs both a fault label and an alarm signal, enabling real-time detection with high accuracy.

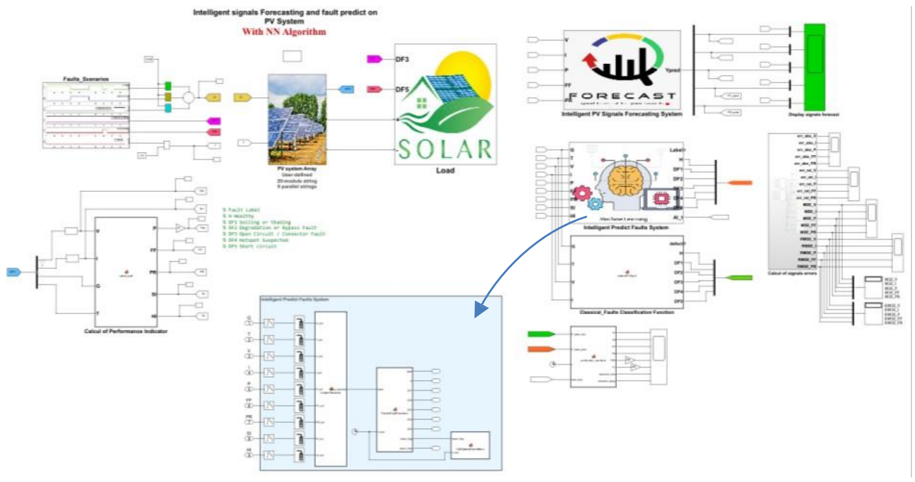


Figure 15. Online PV fault forecast and predict model

6. Simulation Results and Discussion

In this section, the results will be shown from the model developed in section 3. This section is organized as follows. subsection 6.1 contains the results of running the PV system model. This is to validate the results of the healthy PV system, in addition to the different faults. Subsection 6.2 presents the results of the forecasting model. The model will be evaluated with relation to the error calculations. Subsection 6.3 shows the results of the fault prediction system for the PV system. The fault classification model will be evaluated by itself, then the system results will be demonstrated from the simulation using the forecasted signals and fault prediction blocks.

6.1 PV System Model Results

To extract simulation results in the MATLAB/Simulink environment, a PV system model was developed based on the electrical characteristics of high-power PV modules. The system operates under varying environmental conditions, including changes in solar irradiance and temperature, to simulate realistic operating scenarios. In this study, MATLAB/Simulink simulations are used to analyze the behavior of the photovoltaic system under both healthy and faulty operating conditions. Two main modes of operation are considered:

- Healthy mode, where the photovoltaic system operates under normal environmental conditions without any faults, providing a reference for standard performance behavior.
- Fault mode, where different fault conditions are introduced into the PV array to evaluate their impact on system performance. The considered faults include soiling or shading, degradation or bypass diode fault, open-circuit fault, hotspot faults and short-circuit fault.

The simulation procedure generally follows several steps:

- Development of the photovoltaic system model in Simulink, including the mathematical representation of PV modules and power generation characteristics.
- Definition of environmental operating conditions, such as solar irradiance and module temperature, to simulate realistic weather variations.
- Implementation of fault models, introducing disturbances that represent the physical effects of faults, such as reduced irradiance (soiling or shading), electrical disconnections (open-circuit faults), internal degradation (bypass diode faults), short-circuit conditions, and localized overheating (hotspot faults).
- Execution of simulations for both healthy and faulty operating conditions.
- Comparative analysis of simulation results, focusing on key variables such as output voltage, current, generated power and Performance Indicators (FF, PR, SI and HI).

The objective of this analysis is to highlight the differences between normal and faulty operation, providing insight into the influence of photovoltaic faults on system performance and establishing a reliable foundation for the development of intelligent fault detection and diagnostic strategies. The Numerical Fault Codes are given in table 2.

Table 2. Numerical codes used for PV fault classification

Code	Fault Type
0	Healthy
1	Soiling or Shading
2	Degradation or Bypass Fault
3	Open-Circuit Fault
4	Hotspot Fault
5	Short-Circuit Fault

In this study, representative PV fault scenarios were selected based on their practical occurrence and impact on system performance. Several representative photovoltaic fault scenarios were simulated to evaluate the proposed diagnostic system. These scenarios (Table 3 and Figure 16) include healthy operation, soiling or shading, degradation or bypass diode fault, open-circuit fault, hotspot fault and short-circuit fault. Each fault was activated during specific time intervals to generate realistic operating conditions for training and validating the fault classification model.

Table 3. Fault scenarios used for classification

Fault Type	Time Interval (h)	Description
Healthy	0 – 0.6	Normal operation
Soiling or Shading	0.6 – 1.0	Reduced current and power due to dust or shading
Healthy	1.0 – 1.4	System returns to normal condition
Hotspot Fault	1.35 – 1.45	Localized temperature rise
Healthy	1.45 – 1.5	Recovery to normal state
Degradation or Bypass Fault	1.6 – 1.9	Reduction in Fill Factor
Short-Circuit Fault	1.9 – 2.0	Voltage drops toward zero
Open-Circuit Fault	Not activated	Reserved scenario

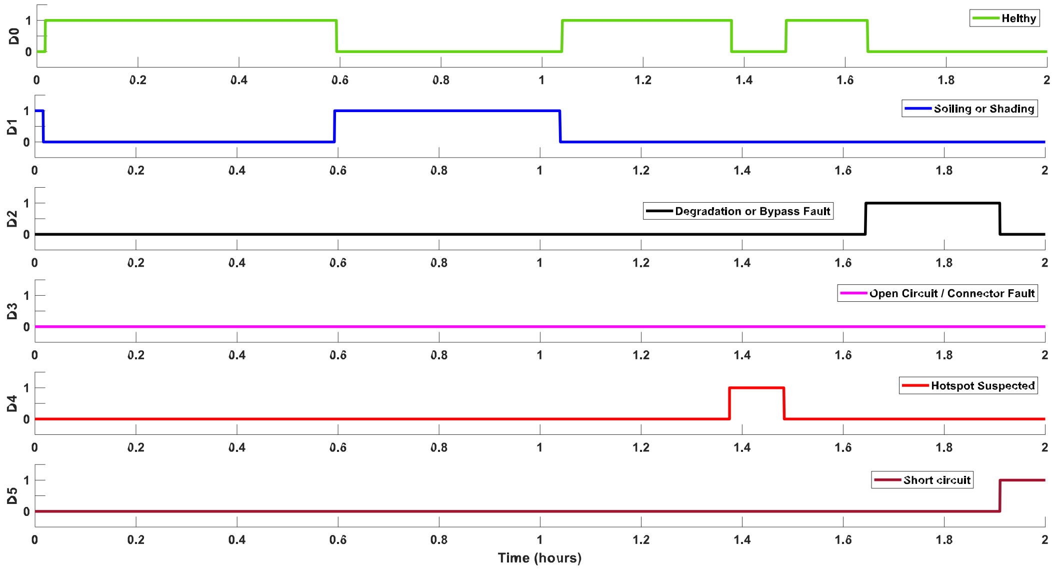


Figure 16. Fault scenarios characteristics

Figure 17 shows the variations of irradiance, temperature, voltage, current, and output power under different fault scenarios. Distinct reductions in voltage, current, and power are observed during soiling, hotspot, degradation, and short-circuit conditions, confirming the sensitivity of electrical parameters to photovoltaic faults.

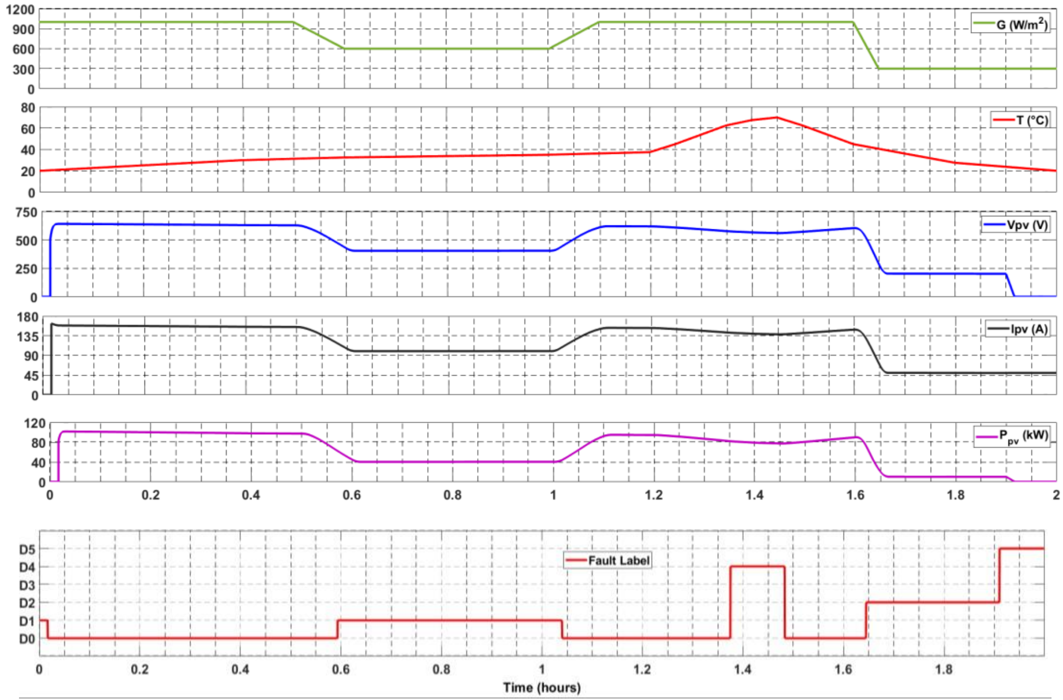


Figure 17. Simulation results of PV system under different fault scenarios

Figure 18 shows the evolution of the main photovoltaic performance indicators (FF, PR, SI, and HI) under different fault scenarios. Distinct variations are observed during soiling, hotspot, degradation, and short-circuit conditions, confirming the effectiveness of these indicators for photovoltaic fault detection and diagnosis. Each fault condition produces distinct changes in performance indicators. Healthy operation shows stable FF and PR values with low SI, while soiling increases SI and reduces PR. Hotspot faults increase HI, degradation reduces FF, and short-circuit faults cause sharp drops in FF and PR with high SI values.

Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. xx No. xx(2026)

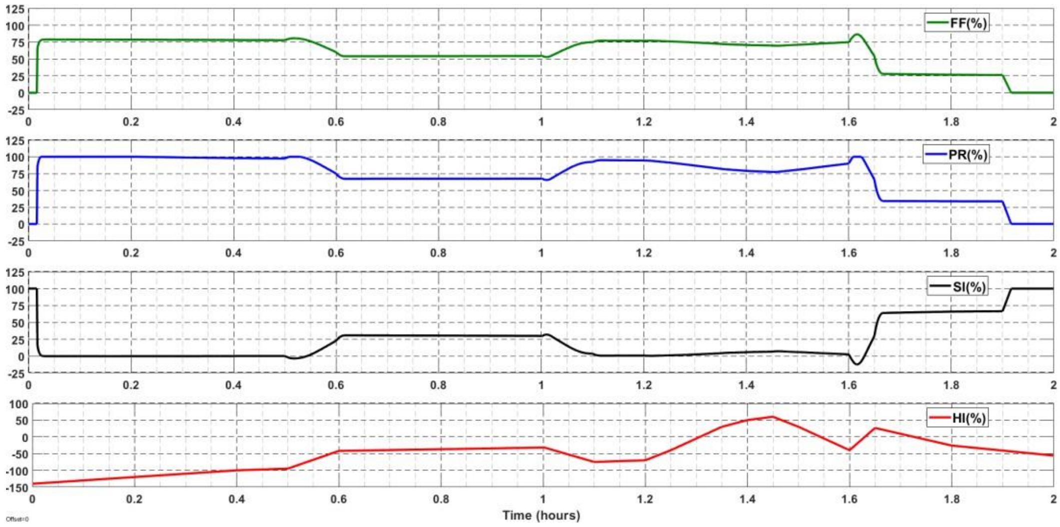


Figure 18. Performances Indicator characteristic under different simulated fault scenarios

6.2 Forecasting Model Results

The forecasting model presented in Section 5.2 was evaluated using an offline photovoltaic dataset generated from simulation under healthy and faulty conditions. The model was then applied to the photovoltaic system to predict future values of key signals such as PV voltage, current, power, and performance indicators. This section presents the forecasting results obtained in real-time photovoltaic simulation. To evaluate the forecasting model in real time under normal and faulty operating conditions, the offline prediction database (system history) is used. The online test forecasts, in real time, the future values of PV system signals. Figure 19 presents the results of the online forecasting model, illustrating the predicted and actual voltage, current, power, and performance indicator signals throughout the simulation period.

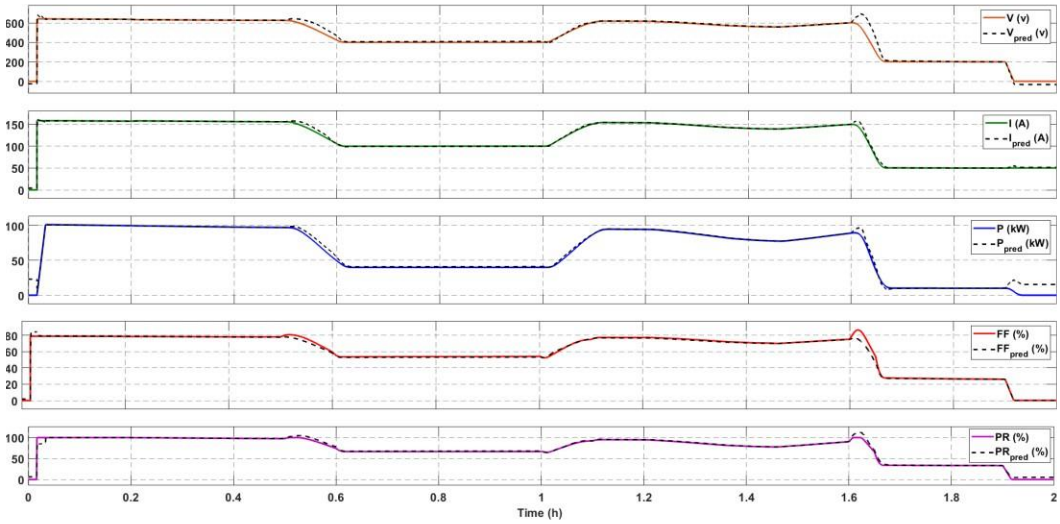


Figure 19. Online Forecasting for PV signals and performance indicators

The online forecasting results demonstrate a clear distinction between healthy and faulty operating conditions of the photovoltaic PV system. Under healthy conditions, the electrical signals, including voltage, current, and power, remain stable with smooth variations, while the performance indicators such as fill factor and performance ratio maintain high and consistent values. In this regime, the predicted signals closely match the measured ones, indicating that the forecasting model accurately captures the normal behavior of the PV system. In contrast, during faulty conditions, noticeable deviations appear in both the signals and performance indicators. Sudden drops or fluctuations in V, I, and P are observed, accompanied by a degradation in FF and PR. Moreover, the discrepancy between the predicted and measured values increases significantly, particularly during abrupt transitions, reflecting the model's sensitivity to abnormal system dynamics. These growing prediction errors serve as reliable indicators of faults, enabling effective detection of anomalies in real time. Overall, the forecasting approach proves to be both accurate under normal operation and responsive to fault-induced disturbances, making it a valuable tool for PV system monitoring and diagnosis. To better evaluate the performance of the forecasting model, the absolute and relative errors of the predicted PV signals are illustrated in Figures 20 and 21. The results show the evolution of the forecasting deviations for key electrical and performance indicators, including voltage (ΔV), current (ΔI), power (ΔP), fill factor (ΔFF), and performance ratio (ΔPR) over time. Overall, the model maintains low prediction errors during steady-state operation, indicating good agreement between predicted and measured PV behavior under normal conditions. However, noticeable error peaks appear during transient events, where rapid variations in the signals are observed. These peaks correspond to sudden changes in system operating conditions, where the forecasting model temporarily struggles to follow fast dynamics. Among all variables, some signals exhibit more pronounced sensitivity to these transitions, particularly during abrupt changes in irradiance or operating conditions, which directly affect voltage, current, and power behavior. Despite these fluctuations, the errors return to low levels after the transient phases, confirming the model's ability to recover and maintain stable long-term predictions.

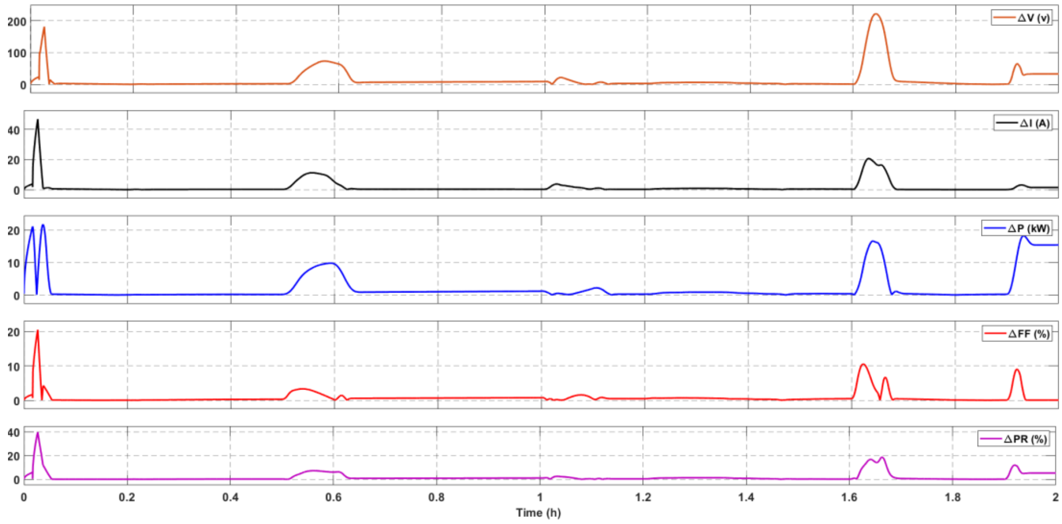


Figure 20. Absolute Error in Online Forecasting PV signals

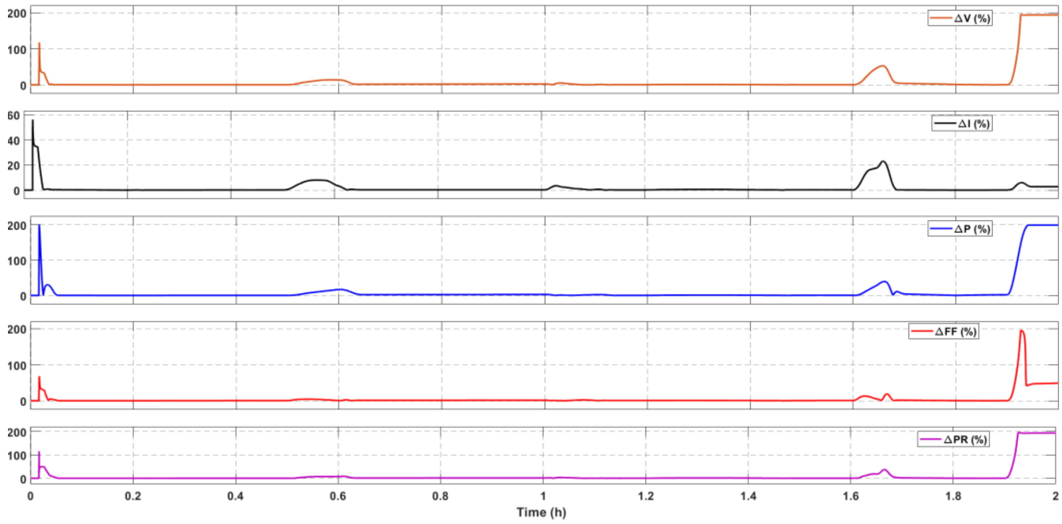


Figure 21. Relative Error in Online Forecasting PV signals

Figure 22 shows the Mean Squared Error (MSE) of the online forecasting PV signals during the simulation period. The MSE is computed over the entire time window, including the initial transient phase, which results in higher error values at the beginning of the signals. For the voltage and current signals, the MSE remains generally low, with only brief spikes occurring during the fault injection events. The power signal shows slightly higher error levels compared to V and I, reflecting its greater sensitivity to system variations, but it still stabilizes during steady-state operation.

The FF and PR signals exhibit comparatively higher MSE values, particularly during transient periods and fault injection instances. These increases highlight their stronger responsiveness to system disturbances, especially when

Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. xx No. xx(2026)

the short-circuit fault is introduced at 1.9 hours, where a noticeable rise in error is observed before returning toward steady-state behavior.

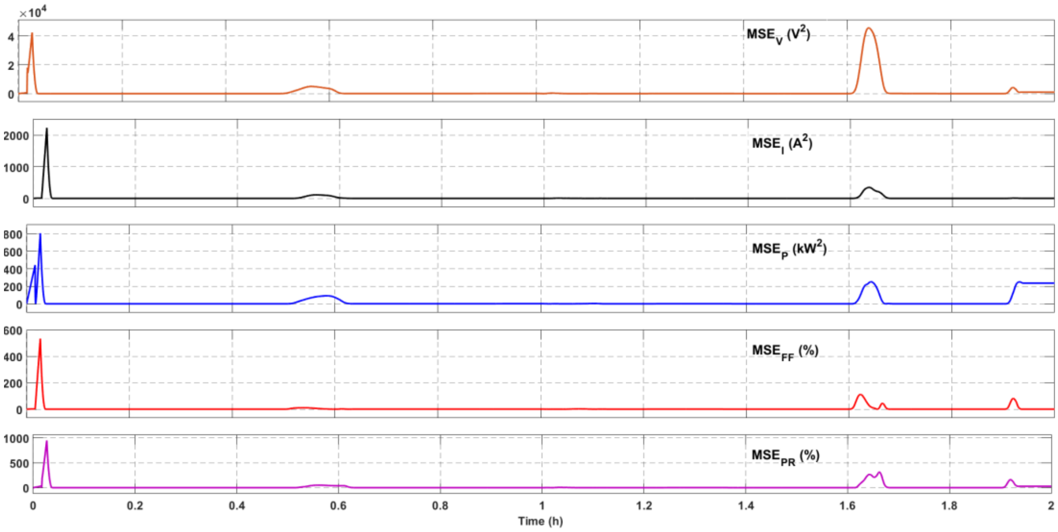


Figure 22. Mean square Errors (MSE) of Online Forecasting PV signals

When utilizing the RMSE values as shown in Figure . RMSE values for signals are all considered low. The voltage signal is not performing as well as the current signal, however, when compared to the performance ratio, the final RMSE value being 8 %, the error is relatively small.

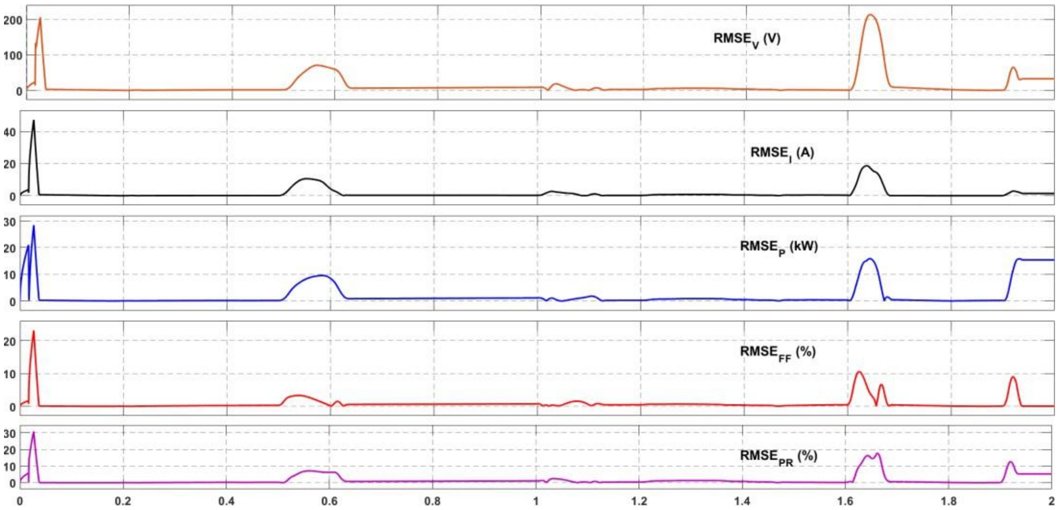


Figure 23. Root Mean square Errors (RMSE) of Online Forecasting PV signals

6.3 Fault Prediction Model Results

The Fault prediction model introduced in section 3 is the combination of the forecasting model feeding the PV signals to a fault classification model, and in this section the fault classification model is tested, proceeded by the testing of the fault prediction model. The objective is to evaluate the capability of the fault prediction model under both healthy and faulty conditions.

6.3.1 Fault Classification Model Results

To understand the capability of a fault classification model, a confusion matrix is presented in Figure to illustrate its performance in distinguishing between healthy (label 0) and faulty (labels 1, 2, 3, 4 and 5) conditions. Data of the PV system is taken for 2 hours with the faults injected at different times. The faults were detected immediately with no delays.

Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. xx No. xx(2026)

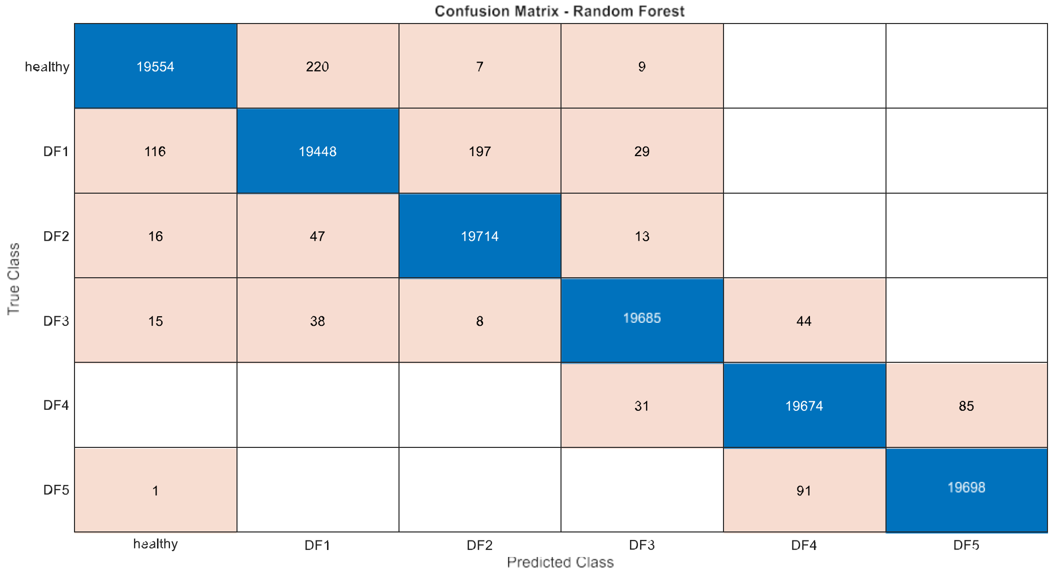


Figure 24. Confusion Matrix for Fault Classification Model

Table 4 presents the Random Forest performances.

Table 4. Random Forest performances

Classe	**Fault Type**	Recall (%)
0	Healthy	98.807
DF1	Soiling or Shading	98.272
DF2	Degradation or Bypass Fault	99.616
DF3	Open-Circuit Fault	99.469
DF4	Hotspot Fault	99.414
DF5	Short-Circuit Fault	99.535
CrossVal Loss:		0.81 %
Accuracy:		99.19%

The results show that the Random Forest model achieves very high performance in classifying the states of the photovoltaic (PV) system, effectively distinguishing between healthy and faulty conditions (DF1 to DF5). The overall accuracy of 99.19% indicates that almost all observations are correctly classified, while the low cross-validation loss (0.0081) suggests that the model generalizes well to unseen data, with no significant overfitting. Looking at the recall, which measures the model's ability to correctly identify each class:

- The healthy state has a recall of 0.988, meaning that nearly all truly healthy systems are correctly detected.
- The fault classes DF1 to DF5 also show very high recall values (between 0.982 and 0.996), indicating that the model successfully detects most fault conditions.
- DF2 to DF5 are extremely well identified (around 0.99+), suggesting that these fault types have clearly distinguishable patterns in the data.

The classification system is highly reliable, both for confirming normal operation and for accurately detecting and distinguishing between multiple fault types. This makes it a strong tool for automated monitoring and fault diagnosis in PV systems.

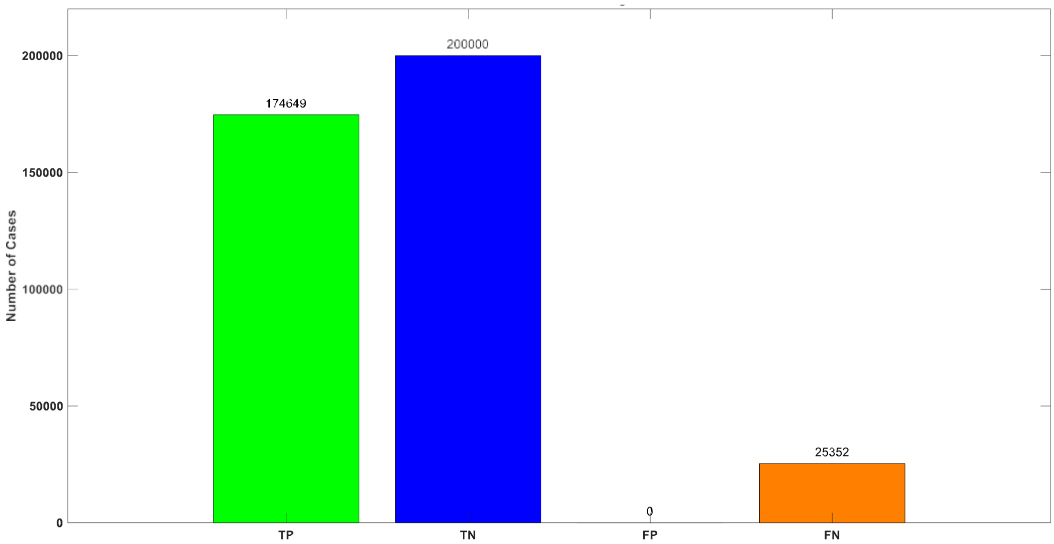


Figure 25. Confusion Matrix Breakdown for Fault Classification Model

6.3.2 Fault Prediction Model Results

In this section, the overall system described in section 3 is tested with the forecasted signals directly in the Simulink model. The system runs for 2 hours, with the fault injected in diffents times, then the system will be evaluated using TP, TN, FP, and FN. The results of the fault prediction system are shown in Figure . When the fault occurs after 1 second elapses, TP begins to increase rapidly, reaching approximately 3.5×10^5, indicating that the system immediately starts detecting the fault correctly. The False Positives (FP) remain at zero throughout the period, demonstrating that the model does not produce any false alarms and thus resulting in precision being 100%. TN increases steadily from the start of the simulation and stops at 4×10^5 before the fault, and as all the health period is correctly classified. The system still shows the same weaknesses as the fault classification test results with FN increase to about the same ratio. This explains the F1-score rising to 100% at the fault occurrence due to perfect precision and recall at that instant, and then stabilizes at 93% as more data accumulates, reflecting FN reducing recall over time. It is also important to note that the system utilizes the forecasted signal and is still able to detect the fault the instant it was injected.

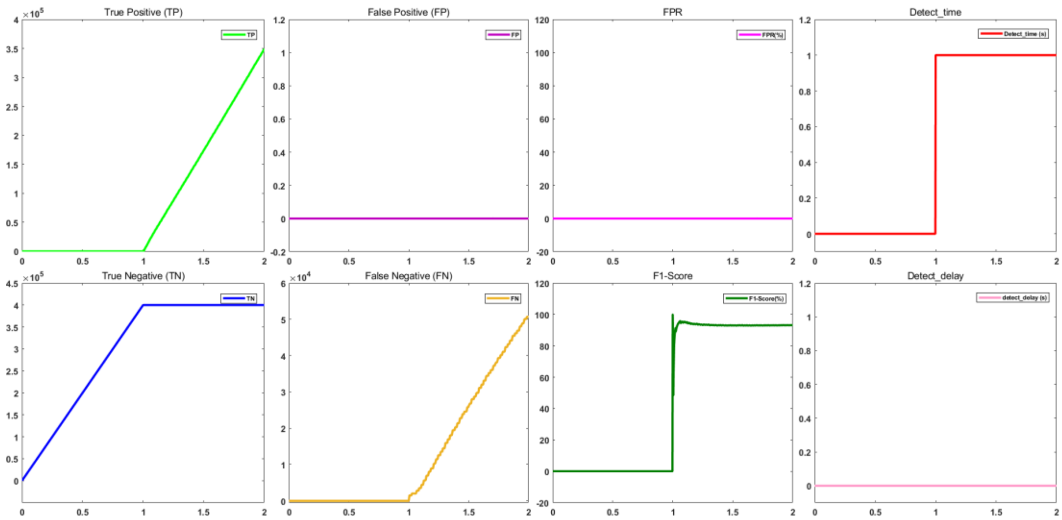


Figure 26: Real-Time Fault Prediction Performance Results

7. Conclusion

This work presented the development and evaluation of a machine learning-based forecasting and fault prediction system for photovoltaic (PV) systems, with the aim of providing a practical approach to improve operational reliability and reduce energy losses caused by faults and performance degradation in PV installations. A simulation framework was developed in MATLAB/Simulink to model both normal operating conditions and multiple fault scenarios in PV systems, enabling the generation of datasets for analysis, training, and validation.

A feed-forward neural network (FFNN) was used to forecast key electrical and performance variables, including voltage, current, power, and performance indicators. The forecasting outputs were then compared with measured signals to compute error indicators, including relative errors and RMSE-based features, which were subsequently used as inputs to a Random Forest classification model for fault detection and classification. The integration of forecasting and residual-based learning enabled the development of a robust fault prediction framework for PV systems under different operating and environmental conditions.

The forecasting model demonstrated accurate prediction of PV electrical variables, particularly under steady-state and quasi-steady irradiance conditions. The low error metrics, including RMSE and relative error values across the predicted signals, confirmed the ability of the FFNN model to capture the underlying dynamics of the PV system. Among the predicted signals and performance indicators showed higher sensitivity to operating changes, making them particularly informative for downstream fault analysis. The subsequent fault classification model achieved strong performance across multiple fault types, including soiling/shading, degradation or bypass faults, open-circuit faults, hotspot faults, and short-circuit faults, in addition to the healthy operating state. The use of error-based features significantly improved the separability between fault classes and enhanced classification robustness.

While the proposed framework shows promising results, several directions remain open for improvement and further research. Future work should consider extending the model to include real-time experimental PV data to validate its robustness under real-world conditions such as partial shading variability, sensor noise, and rapidly changing weather patterns. In addition, incorporating temporal learning approaches such as Long Short-Term Memory (LSTM) and Transformer-based architectures could further improve the ability of the system to capture time-dependent degradation patterns and transient fault behaviors.

Moreover, the integration of prognostic capabilities, such as remaining useful life (RUL) estimation and fault progression modeling, would enable the transition from fault detection to predictive maintenance decision support. Expanding the framework to handle multiple simultaneous faults and developing adaptive learning strategies would also enhance its applicability in large-scale PV plants with heterogeneous operating conditions.

From an engineering and energy management perspective, this approach supports more informed operational decision-making in PV plant maintenance. By combining machine learning-based forecasting with residual-driven fault detection, using only standard electrical and environmental measurements (irradiance, temperature, voltage, current, and power), the proposed system offers a cost-effective and scalable solution for photovoltaic monitoring. It avoids reliance on expensive or intrusive sensing technologies, making it suitable for widespread deployment in utility-scale and distributed PV systems.

Overall, the research demonstrates that combining data-driven forecasting with residual-based fault classification provides an effective and practical framework for intelligent PV system monitoring. The proposed methodology bridges the gap between predictive modeling and operational reliability, contributing to the development of more autonomous, efficient, and resilient photovoltaic energy systems.

Competing Interest declaration: Authors have no conflict of interest to declare.

Funding: The authors did not receive support from any organization for the submitted work.

Availability of data and material: Not applicable.

Ethics, Consent to Participate, and Consent to Publish declarations: Not applicable.

Clinical trial registration: This study is not a clinical trial.

Author contributions: Hatira Bacem and Dhaoui Mehdi: Original Idea, Conceptualization, Data Analysis, Paper Writing, Paper Review. Hatira Bacem: Data Collection, Conceptualization, Paper Writing. Dhaoui Mehdi: Conceptualization, Paper Review. Sbita Lassaad: Paper Review, Hatira Bacem: Paper writing.

References

[1] Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2018). Tree-based ensemble methods for PV power forecasting. Renewable Energy, 136, 1213–1225. https://doi.org/10.1016/j.renene.2018.09.015
[2] Pedro, H. T. C., & Coimbra, C. F. M. (2012). Assessment of forecasting techniques for solar power prediction. Solar Energy, 86(10), 2656–2664. https://doi.org/10.1016/j.solener.2012.06.006

[3] Inman, R. H., Pedro, H. T. C., & Coimbra, C. F. M. (2013). Solar forecasting methods for renewable energy integration. Progress in Energy and Combustion Science, 39(6), 535–576. https://doi.org/10.1016/j.pecs.2013.06.002
[4] Dhimish, M., Holmes, V., Mehrdadi, B., & Dales, M. (2017). PV fault detection using machine learning and performance monitoring. Renewable Energy, 119, 1–14. https://doi.org/10.1016/j.renene.2017.11.006
[5] Korkmaz, D., et al. (2020). Data-driven fault diagnosis of photovoltaic systems using electrical signals. IEEE Access, 8, 176377–176389. https://doi.org/10.1109/ACCESS.2020.3026567
[6] Kumari, P., & Toshniwal, D. (2021). Deep learning for solar PV fault detection: A review. Renewable and Sustainable Energy Reviews, 135, 110122. https://doi.org/10.1016/j.rser.2020.110122
[7] Mellit, A., & Kalogirou, S. A. (2014). Artificial intelligence techniques for PV systems: A review. Renewable and Sustainable Energy Reviews, 26, 1–15. https://doi.org/10.1016/j.rser.2013.05.015.
[8] Ahmed, H. O. A., Wong, M. L. D., & Nandi, A. K. (2018). Intelligent condition monitoring method for bearing faults from highly compressed measurements using sparse over-complete features. Mechanical Systems and Signal Processing, 99, 459–477. https://doi.org/10.1016/j.ymssp.2017.06.027
[9] Aires, F. L., Galeno, G. D., Belchior, F. N., Oliveira, A. M., & Hunt, J. D. (2025). Enhancing three-phase induction motor reliability with health index and artificial intelligence-driven predictive maintenance. Royal Society Open Science, 12(5), 241946. https://doi.org/10.1098/rsos.241946
[10] Bahgat, B. H., Elhay, E. A., Sutikno, T., & Elkholy, M. M. (2024). Revolutionizing motor maintenance: A comprehensive survey of state-of-the-art fault detection in three-phase induction motors. International Journal of Power Electronics and Drive Systems, 15(3), 1968.
[11] Barcelos, A. S., & Cardoso, A. J. M. (2021). Current-based bearing fault diagnosis using deep learning algorithms. Energies, 14(9), 2509. https://doi.org/10.3390/en14092509
[12] Batista, L., Badri, B., Sabourin, R., & Thomas, M. (2012). Detecting bearing defects under high noise levels: A classifier fusion approach. IECON 2012, 3880–3886. https://doi.org/10.1109/IECON.2012.6389272
[13] Cerrada, M., et al. (2018). A review on data-driven fault severity assessment in rolling bearings. Mechanical Systems and Signal Processing, 99, 169–196. https://doi.org/10.1016/j.ymssp.2017.06.012
[14] Cirrincione, G., et al. (2020). Shallow versus deep neural networks in gear fault diagnosis. IEEE Transactions on Energy Conversion, 35(3), 1338–1347. https://doi.org/10.1109/TEC.2020.2978155
[15] Diab, A. A. Z., et al. (2022). Artificial neural-based speed and flux estimators for induction machines. Mathematics, 10(8), 1348. https://doi.org/10.3390/math10081348
[16] Ece, D. G., & Başaran, M. (2011). Condition monitoring of speed-controlled induction motors using wavelet packets. Expert Systems with Applications, 38(7), 8079–8086. https://doi.org/10.1016/j.eswa.2010.12.149
[17] Eftekharnejad, B., et al. (2011). Spectral kurtosis for bearing fault detection. Mechanical Systems and Signal Processing, 25(1), 266–284. https://doi.org/10.1016/j.ymssp.2010.06.010
[18] El Bouchikhi, E. H., et al. (2015). Induction machine diagnosis using stator current analysis. International Journal on Energy Conversion, 3(3), 76–87.
[19] El Idrissi, A., et al. (2024). Bearing fault diagnosis using current envelope analysis and neural networks. Electronics, 13(16), 3195. https://doi.org/10.3390/electronics13163195
[20] Ghods, A., & Lee, H.-H. (2016). Wavelet-based bearing fault detection. Neurocomputing, 188, 206–216. https://doi.org/10.1016/j.neucom.2015.06.100
[21] Gundewar, S. K., & Kane, P. V. (2021). Condition monitoring of induction motors. Journal of Vibration Engineering & Technologies, 9(4), 643–674. https://doi.org/10.1007/s42417-020-00253-y
[22] Guo, X., Chen, L., & Shen, C. (2016). Deep CNN for bearing fault diagnosis. Measurement, 93, 490–502. https://doi.org/10.1016/j.measurement.2016.07.054
[23] Han, Q., et al. (2019). Stator current model for bearing fault detection. Mechanical Systems and Signal Processing, 131, 554–575. https://doi.org/10.1016/j.ymssp.2019.06.010
[24] Huang, L., et al. (2022). Electromagnetic–dynamic modeling for induction motor faults. Machines, 10(8), 682. https://doi.org/10.3390/machines10080682
[25] IEEE Std 1415-2006. (2007). IEEE guide for induction machinery maintenance testing and failure analysis.
[26] Janssens, O., et al. (2015). Thermal image-based fault diagnosis. Infrared Physics & Technology, 73, 78–87. https://doi.org/10.1016/j.infrared.2015.09.004
[27] Jiang, Y. (2023). Fault detection using random forest algorithm. ICMIII 2023, 311–314. https://doi.org/10.1109/ICMIII58949.2023.00065
[28] Kumar, R. R., et al. (2022). Review of fault diagnosis in induction motors. Energies, 15(23), 8938. https://doi.org/10.3390/en15238938
[29] Lau, E. C. C., & Ngan, H. W. (2010). Motor current analysis for bearing fault detection. IEEE TIM, 59(10), 2683–2690. https://doi.org/10.1109/TIM.2010.2045927
[30] Le, T.-T.-H., et al. (2023). Time series forecasting using Bi-LSTM. Sensors, 23(17), 7647. https://doi.org/10.3390/s23177647

[31] Liu, J., et al. (2012). Cost of downtime in manufacturing systems. Journal of Manufacturing Science and Engineering, 134(2), 021016. https://doi.org/10.1115/1.4005789
[32] Nandi, S., et al. (2005). Condition monitoring of electrical motors. IEEE Transactions on Energy Conversion, 20(4), 719–729. https://doi.org/10.1109/TEC.2005.847955
[33] Oo, A. T., et al. (2024). Machine learning-based motor fault detection. IJC, 51(1), 149–169.
[34] Pohakar, P., et al. (2025). Multiple fault analysis in induction motors. E-Prime, 12, 101007. https://doi.org/10.1016/j.prime.2025.101007
[35] Prieto, M. D., et al. (2013). Statistical features for fault detection. IEEE Transactions on Industrial Electronics, 60(8), 3398–3407. https://doi.org/10.1109/TIE.2012.2219838
[36] Ruzimov, S., et al. (2024). Fault detection in induction motors: Recent advances. PIERE 2024, 1420–1427. https://doi.org/10.1109/PIERE62470.2024.10804864
[37] Tavner, P. J. (2008). Condition monitoring of electrical machines. IET Electric Power Applications, 2(4), 215–247.
[38] Wang, C., et al. (2020). Model-based bearing fault detection. Journal of Physics: Conference Series, 1650, 032130. https://doi.org/10.1088/1742-6596/1650/3/032130
[39] Zhang, S., et al. (2020). Bearing fault quantification. IEEE Transactions on Industry Applications, 56(3), 2158–2170. https://doi.org/10.1109/TIA.2020.2979383.