



People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Echahid Hamma Lakhdar University – El Oued



Faculty of Technology
Department of Electrical Engineering

Thesis

Submitted in partial fulfillment of the requirements for the degree of

Academic Master

Specialty: Electrical Networks

Presented by:

1- Mekkaoui Oualid

2- Bouazza Ammar

Entitled:

**Intelligent Fault Detection and Diagnosis in Solar Power
Plants Using Statistical Analysis and Machine Learning**

Defended on : 13/06 /2026

Before the jury composed of :

Dr : Miloudi khaed

President

Dr : kaid imad eddine

Examiner

Dr : Toumi Riadh

Supervisor

Academic Year : 2025/2026

Dedication

With all love and respect, we dedicate this humble work, the fruit of two years of learning, discovery, and innovation, to the eminent professor, Dr. Ilyes Chaouki.

You were more than a teacher; you were a guide, a motivator, and a scientific father. We did not deserve all these gifts from you.

Our honorable Doctor,

These words cannot do you justice, but they carry within them a part of our infinite gratitude.

May Allah reward you on our behalf as a teacher is rewarded for his students.

I also dedicate this work to:

*“To the warm hearts that carried me in their arms,
To the safe hands that supported me when my steps became weak.”*

To my family,

My mother, who illuminated my path with her patience; my father, who taught me that success requires effort and determination; and my brothers, who were my support and companions.

And to my friends,

Companions who turned fatigue into beautiful memories, failure into lessons, and success into joy that we share together.

This achievement is the fruit of your love and support. I am deeply grateful and thankful to you.

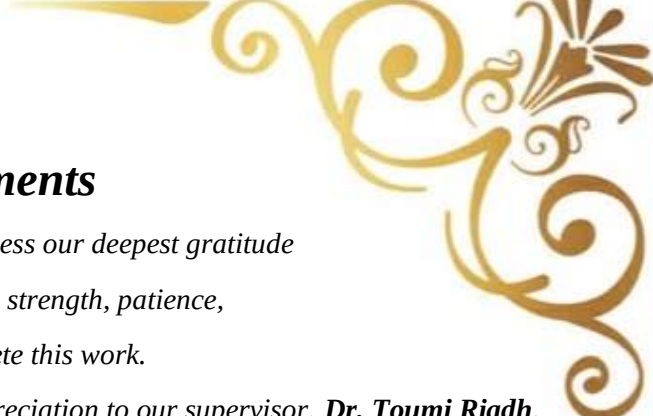
Thank you for being part of my story.

*“And if you are patient and fear Allah, indeed, that is of the matters worthy of determination.”
[Al-Imran: 186]*

Written by your grateful students:

***** Mekkaoui Oualid *****

***** Bouazza Ammar *****



Acknowledgements

*First and foremost, we would like to express our deepest gratitude
to Almighty Allah for giving us the strength, patience,
and determination to complete this work.*

*We would like to extend our sincere thanks and profound appreciation to our supervisor, **Dr. Toumi Riadh**,
for his valuable guidance, continuous support, encouragement, and constructive advice throughout the
preparation of this thesis. His knowledge, patience, and scientific supervision greatly contributed to the
completion of this research work.*

*We also wish to thank all the professors of the Department of Electrical Engineering for the knowledge and
skills they provided us during our academic journey.*

*Our special thanks go to the members of the jury for accepting to evaluate this work and for their valuable
remarks and suggestions.*

*Finally, we would like to express our heartfelt gratitude to our families and friends for their unconditional
love, support, encouragement, and understanding throughout our studies.*

*To all those who contributed directly or indirectly to the completion of this thesis, we sincerely say: **Thank
you.***



Abstract

This work presents an intelligent fault detection and diagnosis approach for solar power plants using statistical analysis and machine learning. The study focuses on photovoltaic systems, their operating characteristics, common faults, and the limitations of traditional diagnostic methods. The proposed framework combines data preprocessing, correlation analysis, Principal Component Analysis, Squared Prediction Error, Hotelling's T^2 statistic, contribution analysis, and machine learning classifiers to detect abnormal behavior, isolate faulty variables, and support maintenance decisions.

The proposed method was applied to the Oued El Kebrit photovoltaic power plant in Algeria. The results showed a strong physical correlation between solar irradiance and generated power, while the injected temperature bias and drift fault was successfully detected and isolated. The contribution analysis identified temperature as the dominant faulty variable, and the machine learning comparison showed that the SVM-RBF classifier achieved the best overall performance. These results confirm that combining statistical monitoring with machine learning can improve the reliability, interpretability, and efficiency of fault diagnosis in photovoltaic power plants.

Keywords: Photovoltaic system, Solar power plant, Fault detection, Fault diagnosis, Statistical analysis, PCA, SPE, Hotelling's T^2 , Machine learning, SVM, Random Forest.

الملخص:

يقدم هذا العمل منهجية ذكية لاكتشاف وتشخيص الأعطال في محطات الطاقة الشمسية الكهروضوئية باستعمال التحليل الإحصائي وتقنيات التعلم الآلي. ركزت الدراسة على أنظمة الطاقة الكهروضوئية، خصائص تشغيلها، أهم الأعطال التي قد تؤثر عليها، وحدود طرق التشخيص التقليدية. تعتمد المنهجية المقترحة على معالجة البيانات، تحليل الارتباط، تحليل المركبات الرئيسية، مؤشر خطأ التنبؤ التريبيعي، إحصائية Hotelling's T^2 ، تحليل المساهمات، وخوارزميات التعلم الآلي من أجل كشف السلوك غير الطبيعي، تحديد المتغيرات المسؤولة عن العطل، ودعم قرارات الصيانة.

تم تطبيق المنهجية المقترحة على محطة واد الكبريت للطاقة الشمسية الكهروضوئية في الجزائر. أظهرت النتائج وجود علاقة فيزيائية قوية بين الإشعاع الشمسي والقدرة المنتجة، كما تم كشف عطل الانحياز والانجراف المحقون في إشارة درجة الحرارة وعزله بنجاح. بين تحليل المساهمات أن درجة الحرارة هي المتغير الأكثر مسؤولية عن العطل، كما أوضحت مقارنة خوارزميات التعلم الآلي أن مصنف SVM-RBF حقق أفضل أداء عام. تؤكد هذه النتائج أن الجمع بين المراقبة الإحصائية والتعلم الآلي يساهم في تحسين موثوقية ووضوح وفعالية تشخيص الأعطال في محطات الطاقة الشمسية.

الكلمات المفتاحية: النظام الكهروضوئي، محطة الطاقة الشمسية، كشف الأعطال، تشخيص الأعطال، التحليل الإحصائي، تحليل المركبات الرئيسية، مؤشر SPE، إحصائية Hotelling's T^2 ، التعلم الآلي، SVM، Random Forest.

Table of Contents

Title	Page
Dedication	
Acknowledgements	
Abstract	
Arabic Abstract	
Table of Contents	
List of Figures	
List of Tables	
List of Abbreviations	
General Introduction	
Chapter I: Theoretical Foundations of Data-Driven Fault Detection and Machine Learning Diagnosis in PV Systems	
I.1. Introduction	4
I.2. Historical Background	4
I.3. General Overview of Solar Energy	5
I.4. Photovoltaic System	6
I.4.1. Definition and Main Components	6
I.4.2. Types of Photovoltaic Systems	6
I.5. Photovoltaic Generator	8
I.5.1. Definition and General Description	8
I.5.2. Electrical Characteristics of a PV Cell	8
I.5.3. Connection Configurations of PV Cells	9
I.5.4. Influence of Environmental Conditions	9
I.6. Electrical Characteristics of a PV Module	9
I.7. Performance Indicators for PV Systems	10
I.7.1. System Efficiency (η)	10
I.7.2. Performance Ratio (PR)	10
I.7.3. System Losses	10
I.8. Protection of Photovoltaic Modules	11
I.8.1. Hot Spot Formation	11
I.8.2. Bypass Diodes	11
I.8.3. Blocking (Anti-Reverse) Diodes	11
I.9. Advantages and Disadvantages of PV Systems	12
I.10. The Photovoltaic Cell	12
I.10.1. Types of Photovoltaic Cells	12
I.11. PV Panel Connection Configurations	13
I.11.1. Series Connection	14
I.11.2. Parallel Connection	14
I.11.3. Series-Parallel (Hybrid) Configuration	14
I.12. Conclusion	15
Chapter II: Fault Detection and Diagnosis in Photovoltaic Systems	
II.1. Introduction	17
II.2. Reliability and Failure Mechanisms in PV Systems	17
II.3. Classification of Faults in Photovoltaic Systems	18
II.4. Common Fault Topologies in PV Systems	20
II.4.1. Environment-Induced Anomalies	21
II.4.1.1. Partial Shading Real-Time Effects	21
II.4.1.2. Soiling and Dust Accumulation	21
II.4.2. Operational and Electrical Component Faults	22
II.4.2.1. Temperature Sensor Anomalies (Bias and Drift)	22
II.4.2.2. Open-Circuit and Short-Circuit Faults	22

II.4.2.3. Hot Spots and Bypass Diode Failures	22
II.5. Impact of Faults on PV Performance Indicators	23
II.6. Fault Detection and Diagnosis in PV Systems	23
II.6.1. Conventional Detection Methods	23
II.6.2. Fault Diagnosis Approaches	24
II.6.3. Data-Driven Detection Methods	26
II.7. Monitoring and Data Acquisition Systems in PV Plants	26
II.7.1. Sensors and Data Acquisition	27
II.7.2. Real-Time Monitoring Systems	28
II.7.3. SCADA and IoT-Based Monitoring	28
II.7.4. Data Quality and Big Data Challenges in PV Systems	29
II.8. Performance Indicators for Fault Detection	29
II.8.1. Performance Ratio (PR)	29
II.8.2. Efficiency Metrics	30
II.8.3. Energy Yield Analysis	30
II.8.4. Statistical Indicators and Threshold-Based Detection	30
II.9. Comparative Analysis of Fault Detection Methods	30
II.10. Challenges in Fault Detection and Diagnosis	31
II.11. Transition Toward Intelligent Fault Detection and Diagnosis	32
II.12. Conclusion	34
 Chapter III: Data-Driven and Machine Learning Approaches for Fault Detection in Photovoltaic Systems	
III.1. Introduction	36
III.2. Data-Driven Approaches for Fault Detection	36
III.2.1. Definition of Data-Driven Methods	36
III.2.2. Objectives of Data-Driven Fault Detection	37
III.2.3. Data-Driven Framework in PV Systems	37
III.3. Data Acquisition and Preprocessing	39
III.3.1. PV Monitoring Data	39
III.3.2. Data Cleaning	39
III.3.3. Normalization and Standardization	40
III.3.4. Feature Extraction and Selection	40
III.4. Statistical Methods for Fault Detection	41
III.4.1. Statistical Indicators	41
III.4.2. Correlation Analysis	42
III.4.3. Residual-Based Detection	42
III.4.4. Threshold-Based Detection	42
III.5. Principal Component Analysis (PCA)	43
III.5.1. Definition and Principle of PCA	43
III.5.2. Objective of PCA in PV Fault Detection	44
III.5.3. Mathematical Formulation of PCA	44
III.5.4. Selection of Principal Components	45
III.5.5. PCA Reconstruction and Residuals	45
III.6. PCA-Based Fault Detection Indices	46
III.6.1. Hotelling's T^2 Statistic	46
III.6.2. SPE/Q Statistic	46
III.6.3. Fault Detection Decision Rules	47
III.6.4. Fault Localization Using Contribution Plots	48
III.7. Advantages and Limitations of PCA	50
III.7.1. Advantages	50
III.7.2. Limitations	50
III.8. Machine Learning for Fault Diagnosis	51
III.8.1. Supervised Learning	51
III.8.2. Unsupervised Learning	51
III.9. Machine Learning Algorithms for PV Fault Diagnosis	51
III.9.1. Support Vector Machine (SVM)	52
III.9.2. K-Nearest Neighbors (KNN)	52
III.9.3. Random Forest (RF)	53
III.9.4. Model Training and Validation	53

III.10. Evaluation Metrics	54
III.10.1. Confusion Matrix	54
III.10.2. Performance Metrics	54
III.11. Comparative Analysis and Justification	55
III.12. Proposed Methodology Adopted in This Thesis	56
III.12.1. Workflow Description	57
III.12.2. Key Stages	57
III.12.3. Methodological Flowchart	59
III.13. Conclusion	60
Chapter IV: Implementation, Results and Discussion	
IV.1. Introduction	62
IV.2. Case Study: Oued El Kebrit Photovoltaic Power Plant	62
IV.3. Development Environment and Tools	63
IV.4. General Methodology of the Proposed Fault Detection Framework	63
IV.5. Correlation Analysis of PV Plant Variables	64
IV.6. SPE Control Chart for Anomaly Detection	65
IV.7. Hotelling's T^2 Control Chart	66
IV.8. Detection of a Hybrid Temperature Bias and Drift Fault	67
IV.9. Fault Isolation Using Variable Contribution Analysis	69
IV.10. Machine Learning Classification Results	70
IV.10.1. Confusion Matrix Analysis of the Best Classifier	72
IV.10.2. Feature Importance Analysis	73
IV.11. Summary of Important Quantitative Results	74
IV.12. Discussion	75
IV.13. Conclusion	76
General Conclusion	78
References	

List of Figures

	Title	Page
Figure I.1	Types of solar radiation received at the Earth's surface.	6
Figure I.2	Classification of photovoltaic systems.	8
Figure I.3	PV cell (a), PV module (b), and PV panel (c).	9
Figure I.4	Visual appearance of different PV cell types.	13
Figure I.5	Characteristic of a generator composed of $(n_p \times n_s)$ identical cells. [40]	14
Figure II.1	General classification of faults in PV systems.	19
Figure II.2	Circuit diagram of the proposed PV monitoring system [69].	28
Figure III.1	Global flowchart of the proposed PCA-contribution-plot-ML methodology.	59
Figure IV.1	Correlation matrix of monitored PV plant variables.	64
Figure IV.2	SPE (Q-statistic) control chart for anomaly detection in the PV plant.	65
Figure IV.3	Hotelling's T^2 control chart for multivariate fault detection.	66
Figure IV.4	High-FDR statistical detection report for the injected temperature bias and drift fault.	68
Figure IV.5	Fault isolation report based on variable contribution percentages.	70
Figure IV.6	Comparative performance of machine learning classifiers in the high-FDR context.	71
Figure IV.7	Confusion matrix of the SVM classifier.	72
Figure IV.8	Random Forest feature importance ranking.	73

List of Tables

	Title	Page
I.1	Comparison of Photovoltaic System Types	7
I.2	Main Losses in Photovoltaic Systems	11
I.3	Comparison of Different Types of Photovoltaic Cells	12
I.4	Comparison Between Series and Parallel Connections	13
II.1	Classification of Faults in Photovoltaic Systems	20
II.2	Technical and Meteorological Signatures of Common PV System Faults	21
II.3	Comparative Overview of Fault Detection Techniques in PV Systems	23
II.4	Comparative Overview of Fault Diagnosis Approaches	25
II.5	Functional Components of PV Monitoring and Data Acquisition Systems	27
II.6	Challenges in Fault Detection and Diagnosis	31
III.1	Data-Driven Framework in PV Systems	37
III.2	Typical PV Monitoring Variables and Diagnostic Relevance	38
III.3	Main Preprocessing Operations for PV Data-Driven FDD	41
III.4	Statistical Methods Commonly Used in PV Fault Detection	43
III.5	Selection of Principal Components	44
III.6	Fault Detection Decision Rules	47
III.7	PCA-Based Indices for PV Fault Detection	48
III.8	Diagnostic Interpretation of Contribution Plots in PV Systems	49
III.9	Advantages of PCA	50
III.10	Limitations of PCA	50
III.11	Comparison of ML Algorithms for PV Fault Diagnosis	52
III.12	Evaluation Metrics for PV Fault Detection and Diagnosis	54
III.13	Comparative Analysis of Statistical and ML Methods	55
III.14	Reasons for Selecting the Practical ML Algorithms	56
III.15	Workflow Description	57
IV.1	Python Libraries	63
IV.2	Summary of Statistical Fault Detection Results and Interpretation	74
IV.3	Machine Learning Performance Comparison	74

List of Abbreviations

AI: Artificial Intelligence

ANN: Artificial Neural Network

AC: Alternating Current

DC: Direct Current

FDD: Fault Detection and Diagnosis

FDR: Fault Detection Rate

FPGA: Field-Programmable Gate Array

GPU: Graphics Processing Unit

IoT: Internet of Things

KNN: K-Nearest Neighbors

ML: Machine Learning

MPPT: Maximum Power Point Tracking

P&O: Perturb and Observe

PCA: Principal Component Analysis

PR: Performance Ratio

PV: Photovoltaic

RF: Random Forest

RBF: Radial Basis Function

SCADA: Supervisory Control and Data Acquisition

SPE: Squared Prediction Error

SVM: Support Vector Machine

T²: Hotelling's T-Squared Statistic

TPU: Tensor Processing Unit



The growing global demand for clean and sustainable energy has accelerated the deployment of renewable energy technologies. Among these, photovoltaic (PV) solar energy has emerged as one of the most promising solutions, particularly for large-scale power plants integrated into modern electrical networks.

Despite their advantages, PV systems are exposed to various faults and degradation mechanisms throughout their operational lifetime. Electrical faults, partial shading, soiling, hot spots, inverter malfunctions, and component aging can significantly reduce energy production, compromise system reliability, and increase maintenance costs. In large solar power plants, undetected faults may lead to substantial economic losses and, in some cases, safety hazards. Therefore, ensuring continuous monitoring and rapid fault diagnosis has become a critical challenge.

Traditional fault detection methods—such as visual inspections, thermal imaging, and basic electrical measurements—are often time-consuming, costly, and difficult to implement continuously in large-scale installations. With the increasing availability of monitoring data from Supervisory Control and Data Acquisition (SCADA) systems, data-driven approaches have gained considerable attention as effective alternatives for intelligent fault management. These approaches exploit historical and real-time operational data to identify abnormal system behavior without requiring detailed physical models.

In recent years, statistical analysis and machine learning techniques have demonstrated significant potential for enhancing fault detection and diagnosis in PV systems. Statistical methods, particularly Principal Component Analysis (PCA), enable the detection of abnormal operating conditions by modeling normal relationships among system variables and identifying deviations from expected behavior. Contribution analysis then provides valuable insights into which variables are responsible for detected anomalies, facilitating fault localization.

Machine learning algorithms add an intelligent layer by enabling automatic fault classification. Techniques such as Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) are capable of learning complex patterns from labeled data and distinguishing between different fault types with high accuracy. The integration of statistical monitoring and machine learning therefore provides a powerful framework for developing reliable, accurate, and explainable fault diagnosis systems.

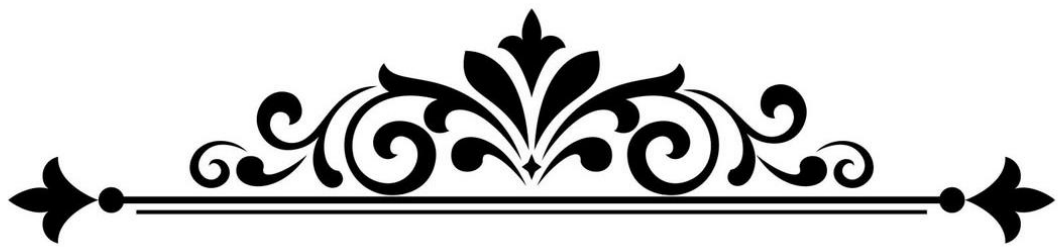
The main objective of this thesis is to develop an intelligent framework for fault detection, localization, and diagnosis in solar power plants using statistical analysis and machine learning techniques. The proposed methodology combines:

- PCA-based fault detection using SPE (Q-statistic) and Hotelling's T^2 indices
- Contribution-plot-based fault localization to identify the most influential variables
- Machine-learning-based fault classification using SVM, Random Forest, and KNN

By leveraging operational data collected from a real PV plant (Oued El Kebrit, Algeria), the study aims to improve the reliability, interpretability, and efficiency of photovoltaic system monitoring, and to contribute to the development of advanced maintenance strategies.

This dissertation is organized into four chapters:

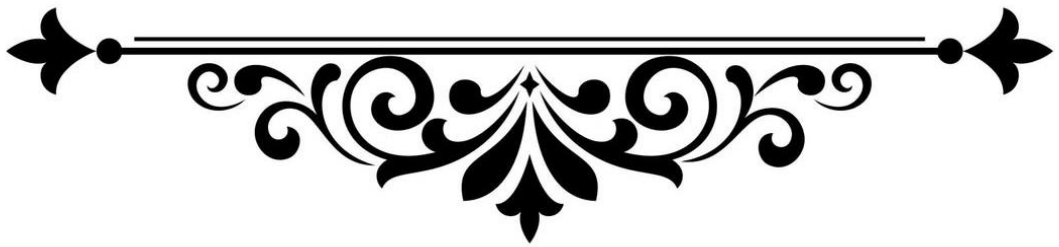
- Chapter I introduces photovoltaic systems and their operating principles.
- Chapter II reviews the different types of faults affecting solar power plants and existing fault diagnosis techniques.
- Chapter III presents the theoretical foundations of data-driven methods, PCA-based monitoring (SPE, T^2), and machine learning algorithms (SVM, RF, KNN) used in this work.
- Chapter IV describes the experimental framework, presents the obtained results, discusses their significance, and highlights future research perspectives.



Chapter I

Theoretical Foundations of Data-Driven Fault Detection and Machine

Learning Diagnosis in PV Systems



I.1. Introduction

Photovoltaic (PV) solar energy has emerged as one of the most promising renewable energy sources due to its abundance, scalability, and decreasing cost of conversion technologies [1], [2]. Unlike fossil fuels, which are limited and non-renewable, solar energy is available on a human timescale and can be harvested in most geographical regions [3].

The direct conversion of sunlight into electricity is achieved through the photovoltaic effect, primarily using cells made of crystalline silicon. Silicon is non-toxic, readily available, and represents the most mature technology from both industrial and technological perspectives [4].

The term "photovoltaic" itself is derived from the Greek word "*photo*" (light) and the name of the Italian physicist Alessandro Volta, reflecting the direct conversion of light into electricity [4].

To develop effective fault detection and diagnosis (FDD) strategies for PV systems, a solid understanding of their fundamental operating principles is essential. Therefore, this chapter provides a structured overview of the photovoltaic effect, the electrical characteristics of PV cells, and the architecture of the PV generator.

I.2. Historical Background

Although the theoretical foundations of the photovoltaic effect were established earlier, the first practical PV cell with reasonable efficiency was developed at Bell Telephone Laboratories in 1954 [4]. This breakthrough marked the beginning of modern PV technology. During the 1970s, significant progress was made in silicon purification processes, and laboratory efficiencies for single-crystal gallium arsenide (GaAs) cells reached 20%, although these levels were difficult to maintain at an industrial scale [4], [5].

From the 1970s onward, silicon-based cells were commercialized and gradually came to dominate the market. Subsequently, polycrystalline silicon cells offered a more cost-effective solution, while amorphous silicon cells found applications in low-power devices such as calculators and watches [4].

The first practical applications of PV electricity were in the space sector. Later, industrial production for terrestrial domestic applications began, enabling the electrification of isolated households. Research continues today, focused on developing more efficient and cost-effective PV products [4], [5].

I.3. General Overview of Solar Energy

The Sun delivers an enormous amount of energy to the Earth's surface, with an instantaneous power of approximately 1 kWp/m² outside the Earth's atmosphere. This power is distributed across the electromagnetic spectrum, from ultraviolet to infrared [6], [7]. Desert regions receive, in just six hours, more solar energy than humanity consumes in an entire year.

Solar energy can be harnessed using two main technologies:

- **Solar Thermal Energy:** This method generates heat using solar collectors. The heat can be used directly or converted into electricity via steam turbines.
- **Photovoltaic Solar Energy:** This technology directly converts sunlight into electricity using PV panels. It is the primary focus of this thesis.

❖ Solar Radiation Components

Once solar radiation enters the Earth's atmosphere, it is modified by atmospheric gases, clouds, and aerosols. For PV system modeling and fault detection, it is useful to distinguish between three components of solar radiation [8], [9]:

- **Direct Radiation:** Radiation that reaches the Earth's surface in straight, parallel rays directly from the solar disk. This is the main contributor to PV power generation under clear-sky conditions and is highly dependent on the sun's position and atmospheric clarity.
- **Diffuse Radiation:** Radiation that reaches the surface after being scattered by air molecules, water vapor, dust, and clouds. It arrives from multiple directions and contributes significantly to total irradiance on cloudy days.
- **Reflected (Albedo) Radiation:** Radiation reflected by the Earth's surface or surrounding objects (e.g., snow, water, sand). Its magnitude is quantified by the *albedo* coefficient and is particularly important for inclined PV panels [10].

Understanding these components is essential for accurately modeling the incident irradiance on PV modules, which is a critical input for any data-driven fault detection method

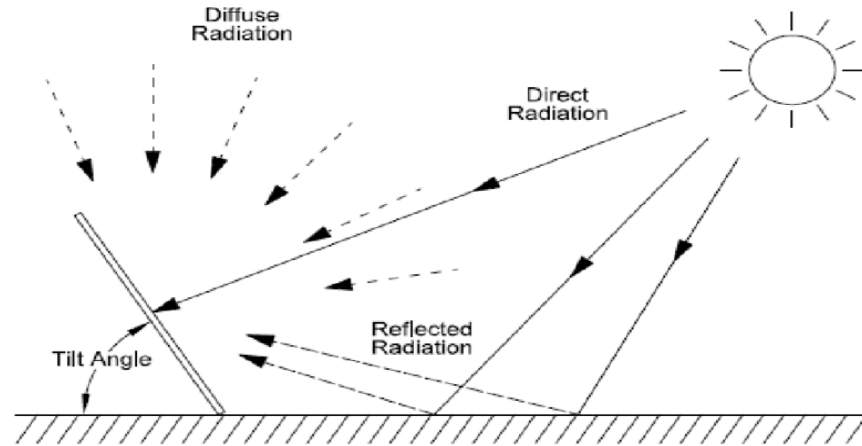


Figure I.1 Types of solar radiation received at the Earth's surface.

I.4. Photovoltaic System

I.4.1. Definition and Main Components

A photovoltaic (PV) system converts solar energy into usable electrical energy. It mainly consists of three core components [12], [35]:

1. The PV Generator (Array): The energy source, composed of interconnected PV modules that convert sunlight directly into direct current (DC).
2. The Power Conditioning Unit: Typically a DC-DC converter and an inverter (DC-AC), equipped with a control system. Its primary function is impedance adaptation and maximum power extraction. Due to the varying nature of irradiance and temperature, Maximum Power Point Tracking (MPPT) algorithms are essential to ensure the generator operates at its optimal point [13], [16].
3. The Electrical Load: The consumer of the generated energy.

I.4.2. Types of Photovoltaic Systems

PV systems are generally classified into three main categories based on their grid connection and energy storage strategy. **Table I.1** provides a comparative summary.

Table I.1 Comparison of Photovoltaic System Types [35]

System Type	Grid Connection	Storage Required	Typical Applications
Off-grid	No	Yes (or water reservoir)	Remote areas, rural electrification, water pumping
Grid-connected	Yes	Optional	Urban buildings, large-scale solar power plants
Hybrid	Partial	Yes	Critical systems (telecom, hospitals), isolated communities

- **Autonomous (Off-grid) Systems:** Operate independently from the public grid. They are essential for remote locations. Two configurations exist: without battery (e.g., water pumping, where the water acts as storage) and with battery (e.g., residential lighting), which ensures continuous power supply at the cost of higher system complexity [14], [15], [36], [37].
- **Grid-connected Systems:** Directly connected to the electrical grid via an inverter. The grid itself acts as an infinite energy sink/storage. The main operational strategy is **total production injection**, where all generated electricity is fed into the grid, often compensated via feed-in tariffs. This is the standard for large-scale solar power plants [16], [39], [40].
- **Hybrid Systems:** Integrate multiple energy sources (e.g., PV + diesel generator). They combine the renewable source (reducing fuel consumption) with a conventional backup (guaranteeing power continuity), making them suitable for high-reliability applications [5], [6], [38].

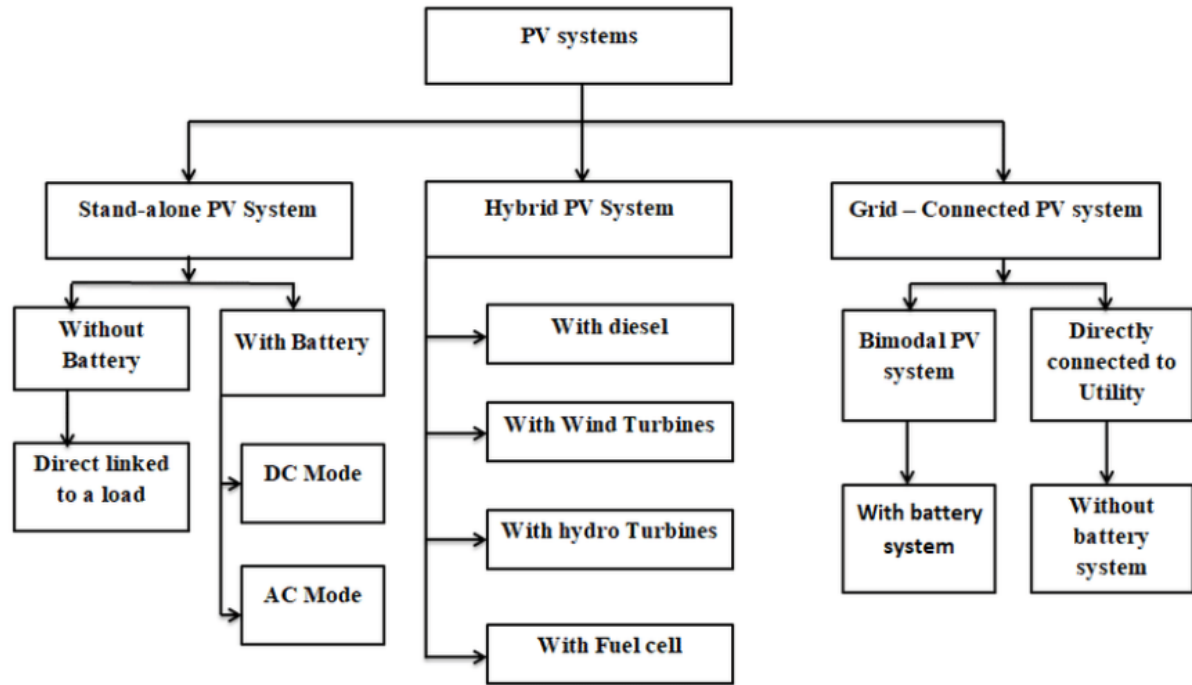


Figure I.2 Classification of photovoltaic system

I.5. Photovoltaic Generator

I.5.1. Definition and General Description

A PV generator is the fundamental energy conversion unit. It is formed by interconnecting PV cells in series and/or parallel to create modules and arrays capable of producing usable electrical power [17], [18].

I.5.2. Electrical Characteristics of a PV Cell

The performance of a PV cell is defined by several key parameters [16], [37], [39]:

- **Open-Circuit Voltage (V_{oc}):** The maximum voltage when the cell is not connected to a load. It is highly sensitive to temperature (decreases as temperature increases).
- **Short-Circuit Current (I_{sc}):** The maximum current when the cell's terminals are shorted. It is directly proportional to solar irradiance.
- **I-V and P-V Characteristics:** The electrical behavior is described by the current-voltage (I-V) curve. The **Maximum Power Point (MPP)** is the operating point where

the product of voltage and current ($P = V \times I$) is maximized. **MPPT algorithms** are used to track this point under varying environmental conditions.

I.5.3. Connection Configurations of PV Cells

To achieve desired voltage and current levels, cells/modules are interconnected as follows:

- **Series Connection:** Increases voltage; current remains constant. Sensitive to partial shading.
- **Parallel Connection:** Increases current; voltage remains constant. More robust under shading.
- **Series-Parallel (Hybrid) Configuration:** Combines both to optimize voltage, current, and reliability [10], [19].

I.5.4. Influence of Environmental Conditions

PV generator performance is highly dependent on environmental factors. Variations in **solar irradiance**, **ambient temperature**, **partial shading**, **dust accumulation (soiling)**, and **module aging** directly modify the I-V and P-V curves, leading to fluctuations in power generation [10], [40]. This variability is the primary reason why **continuous monitoring and intelligent fault detection are crucial** for maintaining efficient operation.

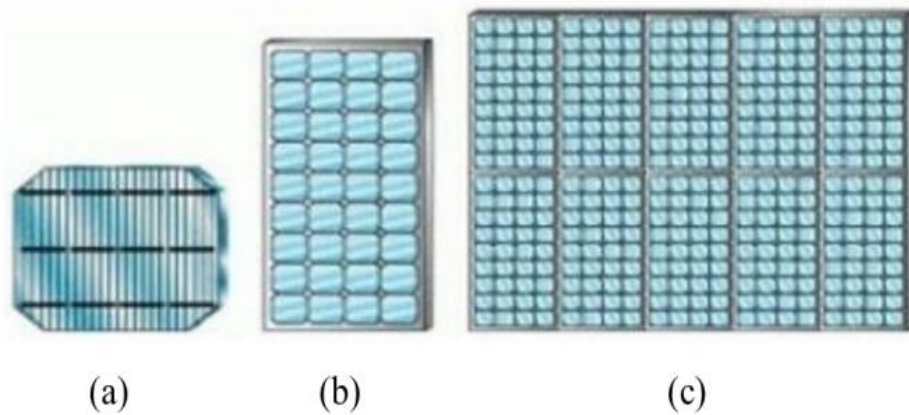


Figure I.3 PV cell (a), PV module (b), and PV panel (c) [23].

I.6. Electrical Characteristics of a PV Module

A PV module's electrical behavior is characterized under Standard Test Conditions (STC): 25°C and 1000 W/m² irradiance [20]. The key parameters are:

Peak Power (P_{max}): The maximum power delivered under STC.

I-V Curve: The relationship between current and voltage.

Fill Factor (FF): A measure of cell quality, defined as:

$$FF = \frac{P_{max}}{V_{oc} \times I_{sc}}$$

A higher FF indicates better module performance and lower internal losses [21].

I.7. Performance Indicators for PV Systems

To evaluate PV system performance under real conditions, several key indicators are used. These indicators are also essential for fault detection, as deviations from expected values often signal underlying issues [41], [42].

I.7.1. System Efficiency (η):

$$\eta_{system} = \frac{E_{out}}{E_{in}}$$

Where E_{out} is the electrical energy produced and E_{in} is the incident solar energy. System efficiency is lower than module efficiency due to thermal, inverter, wiring, and soiling losses [43].

I.7.2. Performance Ratio (PR):

The PR is the most important indicator for PV system monitoring. It represents the ratio between the actual energy output and the theoretical energy output under ideal conditions [10].

$$PR = \frac{\text{Actual Energy Output}}{\text{Theoretical Energy Output}}$$

A well-designed system typically has a PR between 0.75 and 0.90. Sudden reductions in PR are a strong indicator of potential faults such as shading, soiling, degradation, or component malfunction [42].

I.7.3. System Losses:

The main losses affecting PV performance are summarized in Table I.2. Monitoring these losses is essential for detecting performance degradation and identifying possible system faults [16], [22], [37].

Table I.2 Main Losses in Photovoltaic Systems

Type of Loss	Cause	Impact on Performance
Thermal Loss	High temperature	Reduced voltage and efficiency
Inverter Loss	Conversion inefficiency	Energy loss (2–5%)
Shading & Mismatch	Partial obstruction	Significant power drop, hot spot risk
Soiling Loss	Dust accumulation	Reduced irradiance absorption
Wiring Loss	Resistance in cables	Power dissipation

I.8. Protection of Photovoltaic Modules

I.8.1. Hot Spot Formation

Partial shading is a critical issue. When a cell in a series string is shaded, its current decreases. To maintain the same current, the shaded cell becomes reverse-biased and dissipates power instead of generating it, leading to localized overheating known as a hot spot. This can permanently damage the cell [22], [23], [44].

I.8.2. Bypass Diodes

To prevent hot spots, bypass diodes are connected in parallel with groups of cells. Under normal conditions, they are reverse-biased (off). When a cell is shaded, the diode becomes forward-biased (on), providing an alternative path for the current and bypassing the shaded group. This protects the module at the cost of a slight voltage drop [45].

I.8.3. Blocking (Anti-Reverse) Diodes

In battery-based systems, a blocking diode is placed in series with the PV module. It prevents reverse current from flowing from the battery back into the module at night or during complete shading, which could damage the cells [21]

I.9. Advantages and Disadvantages of PV Systems

PV systems offer significant advantages: they utilize a free and inexhaustible energy source, produce no direct emissions, have no moving parts (high reliability), low maintenance costs, and operate silently. They are also modular and scalable [11], [40].

However, the main disadvantages include a high initial investment cost, relatively low conversion efficiency compared to other technologies, power reduction under cloudy conditions, and the need for energy storage in off-grid applications, which adds cost and complexity [15], [37]

I.10. The Photovoltaic Cell

A photovoltaic (PV) cell is the smallest functional unit of a PV system. It is a semiconductor device that directly converts solar radiation into direct current (DC) electricity. Under standard test conditions, a single silicon cell typically produces an output voltage of approximately 0.5–0.6 V [16], [17].

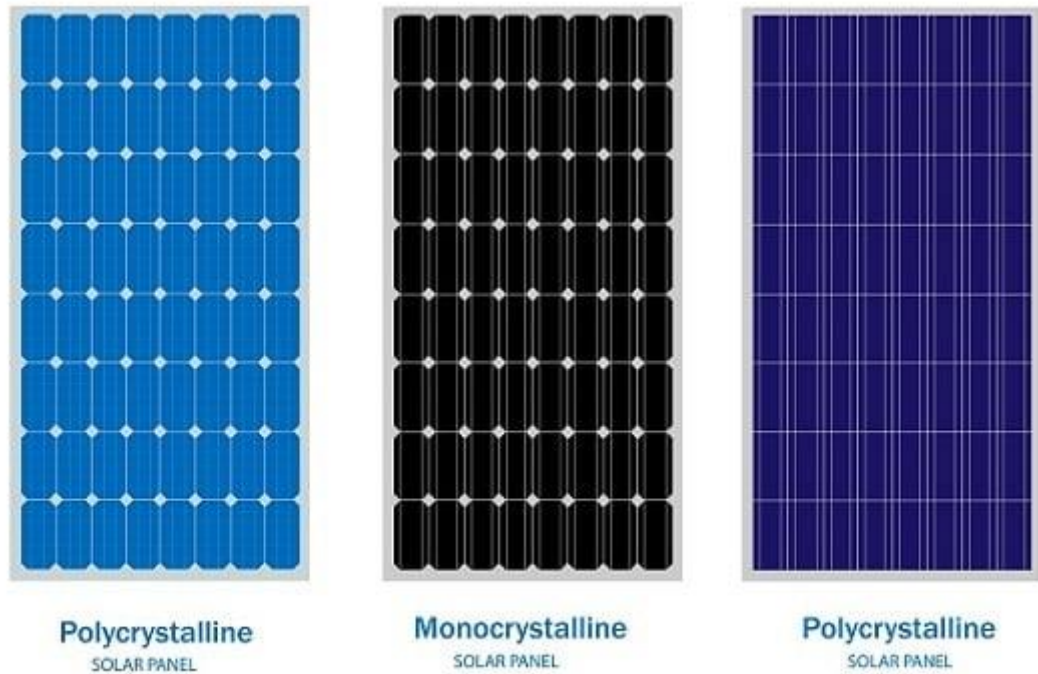
Most commercial PV cells are manufactured from crystalline silicon, which offers a good balance between efficiency, cost, and availability [38] , [40].

I.10.1. Types of Photovoltaic Cells

PV cells are classified into three main technologies based on their manufacturing process and crystalline structure. Table I.3 summarizes their key electrical and economic characteristics.

Table I.3 Comparison of Different Types of Photovoltaic Cells [15], [20], [22]

Type of Cell	Efficiency (%)	Cost Level	Key Characteristics
Monocrystalline	15 – 24	High	Highest efficiency; requires high-purity silicon
Polycrystalline	11 – 19	Medium	Lower cost; slightly lower efficiency than mono
Amorphous (Thin-film)	5 – 13	Low	Flexible; works under low light; lowest efficiency



I.11. PV Panel Connection Configurations

To achieve the desired voltage and current levels for a PV array, individual modules are interconnected using three main configurations, compared in Table I.4.

Table I.4 Comparison Between Series and Parallel Connections [10], [14], [15]

Parameter	Series Connection	Parallel Connection
Voltage	Adds (increases)	Remains constant
Current	Remains constant	Adds (increases)
Sensitivity to Shading	High (string current limited by shaded panel)	Low (only the shaded panel is affected)
Typical Application	High-voltage systems	High-current systems

I.11.1. Series Connection

Modules are connected end-to-end (positive to negative). The total voltage is the sum of individual voltages, while the current is limited by the lowest-current module in the string. This makes series strings highly sensitive to partial shading; a single shaded panel can drastically reduce the power output of the entire string [14], [35], [37].

I.11.2. Parallel Connection

Modules are connected positive-to-positive and negative-to-negative. The total current is the sum of individual currents, while the voltage remains constant. This configuration is more robust under shading but requires that all modules have the same nominal voltage [15], [16], [37].

I.11.3. Series-Parallel (Hybrid) Configuration

In large-scale PV installations, a hybrid configuration is used (series strings connected in parallel). This approach allows designers to optimize both voltage and current levels while improving system reliability and adaptability to partial shading conditions [10], [40], Figure I.5 shows the resulting I-V characteristic of such an array.

I.12. Conclusion

This chapter has presented the fundamental concepts of photovoltaic systems, including the structure of the PV generator, the electrical characteristics of PV cells, and the different configurations of PV panels. The influence of environmental conditions (irradiance, temperature, shading) on system performance was also discussed.

Understanding these basic operating principles is essential before addressing the main topic of this thesis: **intelligent fault detection and diagnosis in solar power plants**. The next chapter will provide a comprehensive review of fault types, their impact on PV systems, and existing detection techniques.

Chapter II

**Fault Detection and Diagnosis in
Photovoltaic Systems**

II.1 Introduction

The rapid expansion of photovoltaic (PV) installations worldwide has made system reliability and operational efficiency critical concerns. PV systems operate under harsh and variable environmental conditions, making them inherently susceptible to a wide range of faults and degradation mechanisms. These anomalies can significantly reduce energy yield, increase maintenance costs, accelerate component aging, and, in severe cases, create safety hazards [46], [47].

Consequently, **fault detection and diagnosis (FDD)** has become an essential component in the management of modern solar power plants. Early and accurate identification of faults allows operators to minimize energy losses, plan targeted maintenance interventions, and improve the overall economic viability of PV installations [48].

This chapter provides a structured overview of FDD in PV systems. It begins by examining reliability and failure mechanisms, followed by a classification of fault types. The impact of faults on system performance is then discussed, and both conventional and advanced detection techniques are reviewed. Finally, the chapter presents diagnostic approaches, monitoring infrastructures, performance indicators, and a comparative analysis of existing methods, establishing a solid foundation for the data-driven and machine-learning models developed in subsequent chapters.

II.2. Reliability and Failure Mechanisms in PV Systems

Reliability is a fundamental requirement for PV systems, which are designed to operate efficiently over 20 to 30 years. However, continuous exposure to environmental and operational stresses leads to gradual degradation or sudden failures [46].

Failure mechanisms are generally divided into two categories [47]:

- **Intrinsic mechanisms:** Related to material properties and manufacturing defects, such as microcracks, encapsulant discoloration, and solder bond degradation.
- **Extrinsic mechanisms:** Arising from environmental conditions, including thermal cycling, humidity ingress, ultraviolet (UV) radiation, and mechanical stress from wind or snow loads.

The most significant degradation processes include:

- **Thermal cycling:** Causes repeated expansion and contraction, leading to mechanical fatigue in solder joints and interconnections.
- **Humidity ingress:** Penetrates module layers, resulting in corrosion and delamination.
- **UV radiation:** Degrades polymeric materials, reducing optical transmittance and energy conversion efficiency.
- **Potential-Induced Degradation (PID):** Occurs due to voltage differences between PV cells and the grounded frame, causing leakage currents and substantial power losses.
- **Light-Induced Degradation (LID):** Affects crystalline silicon modules during initial sunlight exposure, causing early-stage performance reduction.

From a system reliability perspective, these degradation mechanisms manifest as measurable changes in electrical parameters such as current, voltage, and power output. Therefore, their identification is not only a physical inspection task but also a **data analysis problem**. Statistical indicators and time-series analysis can be employed to detect abnormal trends, while machine learning techniques enable the modeling of complex nonlinear degradation patterns [46].

Understanding these failure mechanisms is essential for designing robust FDD systems, as it allows the development of predictive maintenance strategies and the application of advanced data-driven techniques.

II.3. Classification of Faults in Photovoltaic Systems

Faults in PV systems can be classified based on their origin, impact, and physical characteristics. A systematic classification is essential for accurate detection, diagnosis, and mitigation. In general, PV system faults are grouped into four main categories: electrical faults, environmental and external faults, degradation and aging faults, and system-level faults [46].

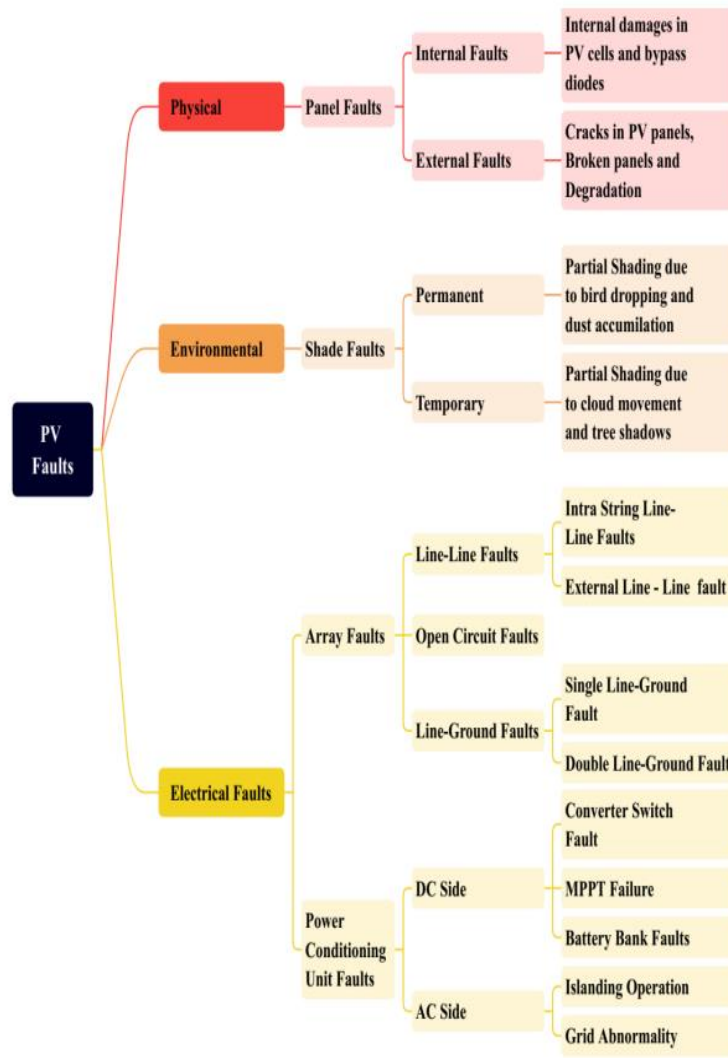


Figure II.1 illustrates a general classification of faults in PV systems.

This classification plays a crucial role in structuring fault detection methodologies, as each category exhibits distinct signatures in electrical, thermal, and performance data. For instance, electrical faults often result in abrupt changes in current or voltage, whereas degradation faults appear as gradual performance decline over time [46].

Moreover, this categorization facilitates the application of tailored diagnostic approaches. In particular, statistical analysis techniques rely on identifying deviations from normal operating conditions, while machine learning models depend on recognizing patterns associated with specific fault types. Therefore, a well-defined classification framework enhances the effectiveness of intelligent fault detection and diagnosis systems.

Table II.1 Classification of Faults in Photovoltaic Systems

Fault Category	Origin / Cause	Key Characteristics	Typical Examples
Electrical Faults	Internal electrical components (modules, cables, inverters)	Abrupt changes in current/voltage; safety hazards	Open-circuit, short-circuit, ground fault, arc fault
Environmental & External Faults	External conditions (temporary)	Reduce output without permanent internal damage	Partial shading, soiling, temperature variations, wind, hail
Degradation & Aging Faults	Gradual material deterioration	Slow progression; subtle performance decline	Encapsulant browning, delamination, microcracks, PID, LID
System-Level Faults	Control, communication, integration	Affect entire system; often related to inverters or SCADA	Inverter malfunction, sensor failure, grid disconnection, mismatch losses

II.4. Common Fault Topologies in PV Systems

Photovoltaic systems are dynamically exposed to harsh environmental factors and operational stresses that directly influence their efficiency, safety, and reliability. From a data-driven diagnostic perspective, any systemic anomaly or physical degradation manifests as a distinct signature across the core monitored parameters of the plant: Power (P), Solar Irradiance (G), Temperature (T), Wind Speed (W), Relative Humidity (H), and Atmospheric Pressure (PR) [51].

In this thesis, the diagnostic framework is built around the variables available from the PV plant monitoring system: generated power (P), solar irradiance (G), temperature (T), wind speed (W), relative humidity (H), and atmospheric pressure (PR). These variables were selected because the studied faults disturb their physical correlations: shading and soiling mainly affect the relationship between irradiance and power, temperature sensor bias and drift affect the thermal channel and its correlation with humidity and power, while open-circuit and short-circuit events modify the expected power response under normal weather conditions.

To establish an effective baseline for fault diagnosis, it is essential to categorize these failures based on their physical nature, electrical behavior, and impact on multi-variable correlations.

Table II.2 Technical and Meteorological Signatures of Common PV System Faults.

Fault/Anomaly Type	Primary Cause	Physical	Main Impacts	Variable	Operational Effect
Partial Shading	Clouds, obstacles, trees.	building	↓ Power, approximately normal irradiance.		Mismatch losses and localized cell stress.
Soiling & Dust	Arid accumulation, storms.	sand dust	Gradual ↓ Power, correlated with W and H.		Continuous attenuation of light absorption.
Sensor Anomalies	Calibration loss, thermal aging.		Sudden step (Bias) or linear rise/fall (Drift) in T.		Erroneous data and feedback reporting errors.
Open-Circuit	Cable cut, faulty connectors, blown fuse.		Power drops instantly to zero, normal weather metrics.		Total loss of generation in the affected string.
Short-Circuit	Insulation failure, water ingress.		Severe ↓ Power, localized ↑↑ Module Temp.		High safety risk and protective fuse activation.

II.4.1. Environment-Induced Anomalies

II.4.1.1. Partial Shading Real-Time Effects

Partial shading is a frequent and highly non-linear physical phenomenon where a localized section of a PV array receives less solar radiation than the surrounding modules [55].

- **Variable Interaction:** Shading fundamentally breaks the expected linear coupling between solar inputs and electricity generation. While the ambient meteorological Irradiance (\$G\$) sensor captures a clear-sky reading, the actual plant Power (P) output exhibits a sharp drop. This distinct mismatch serves as a clear indicator of localized obstruction.

II.4.1.2. Soiling and Dust Accumulation

In arid and semi-arid operational environments, dust and fine sand particles continuously deposit on the glass protective layers of PV modules [22].

- **Variable Interaction:** The accumulation rate and severity of soiling are governed by weather dynamics, specifically Wind Speed (W), which carries the particulates, and Relative Humidity (H). High humidity levels often trigger cementation, turning loose dust into a persistent crust. This causes a slow, progressive decline in daily Power (P) generation while raw solar Irradiance (G) remains high.

II.4.2. Operational and Electrical Component Faults

II.4.2.1. Temperature Sensor Anomalies (Bias and Drift)

Automated monitoring systems rely heavily on accurate thermal feedback to gauge panel health and compute expected performance metrics. Temperature sensors are highly susceptible to harsh field conditions, introducing two primary fault signatures [37]:

- **Constant Bias:** A sudden, instantaneous step-change offset in the data stream (e.g., an abrupt -2.5°C shift), caused by sudden calibration loss or electrical surges.
- **Accelerated Drift:** A slow, time-dependent linear degradation over consecutive observations (e.g., a cumulative -0.09°C per sample degradation), typically resulting from material aging or loose terminal connections.
- **Variable Interaction:** During these sensor faults, the actual plant Power (P) matches the real weather conditions, but appears mathematically mismatched with the faulty Temperature (T) data stream, creating a distinct multi-variable correlation anomaly.

II.4.2.2. Open-Circuit and Short-Circuit Faults

- **Open-Circuit Faults:** Triggered by physical breaks in the conductive paths, such as degraded solder joints, broken DC cables, or disconnected string fuses [53]. The immediate signature is a total collapse of **Power (P)** to zero, while environmental inputs (G , T , W) continue to report normally.

- **Short-Circuit Faults:** Occur due to insulation failure, mechanical cable damage, or moisture ingress inside junction boxes [54]. They disturb the DC-side voltage and current signatures at the string level and usually lead to a substantial decrease in the net output power (P) measured at the plant output. In addition, affected cells may dissipate energy locally, causing abnormal temperature (T) rises.

II.4.2.3. Hot Spots and Bypass Diode Failures

- **Hot Spots:** Localized thermal stresses where a damaged or shaded cell acts as a reverse-biased resistive load [10]. This anomaly causes a sharp local increase in Temperature (T) while overall system Power (P) degrades slightly.

- **Bypass Diode Failures:** Diodes can fail in a short-circuit mode due to overvoltage stress [56]. This permanently bypasses the protected cell group, resulting in a distinct, quantized step-down reduction in Power (P) while environmental variables remain stable.

The present study focuses on data-driven detection using environmental and power measurements rather than direct electrical signature analysis.

II.5. Impact of Faults on PV Performance Indicators

Any physical anomaly or electrical failure directly compromises the net energy output and alters the operational metrics of the system. Under normal, fault-free operating conditions, the relationship between actual energy yield and ambient climate parameters remains stable, maintaining a healthy Performance Ratio (PR) [57, 58].

When faults occur, they disturb this balance. Abrupt electrical faults (open/short circuits) cause immediate collapses in efficiency, while environmental anomalies (shading and soiling) introduce non-linear or long-term gradual decay into the plant's yield. Distinguishing whether a performance drop is caused by a true physical power loss or an artificial sensor reporting error is a critical requirement for modern intelligent diagnostic systems.

II.6. Fault Detection and Diagnosis in PV Systems

Fault detection and diagnosis (FDD) in PV systems involves identifying abnormal operating conditions and determining their root causes. These techniques can be broadly classified into conventional methods and advanced approaches [57], [62].

Table II.3 Comparative Overview of Fault Detection Techniques in PV Systems [58]

Technique	Core Principle	Main Strength	Main Limitation	Best Use Case
Visual inspection	Manual observation of visible defects	Simple and low cost	Subjective and incomplete	Surface damage and discoloration
Electrical signal-based methods	Analysis of voltage, current, and power	Fast and widely available	Limited fault specificity	Open-circuit and short-circuit faults
Thermal imaging	Detection of temperature anomalies	Non-invasive and effective for hot spots	Affected by ambient conditions	Hot spots and diode problems
I-V curve analysis	Comparison with reference electrical curves	Highly informative	Requires special equipment	Shading and mismatch faults
Power loss analysis	Deviation between expected and actual power	Scalable for large plants	Depends on model accuracy	Degradation and underperformance

Data-driven methods	Statistical and machine learning analysis	Adaptive and accurate	Requires quality data	Complex and mixed fault patterns
----------------------------	---	-----------------------	-----------------------	----------------------------------

II.6.1. Conventional Detection Methods

Visual Inspection is the simplest method, involving manual examination of PV modules, wiring, and structural components to identify visible defects such as cracks, discoloration, corrosion, or loose connections [59]. Although low-cost and easy to implement, it is highly dependent on human expertise and cannot detect hidden or early-stage faults.

Electrical Signal-Based Methods rely on the analysis of voltage, current, and power measurements to detect abnormalities. By comparing real-time measurements with expected values derived from physical or statistical models, deviations indicative of faults can be identified. Techniques such as residual analysis, control charts, and statistical thresholds are commonly applied [60].

Thermal Imaging (Infrared Thermography) is a powerful non-invasive technique used to detect temperature anomalies in PV modules, particularly effective for identifying hot spots, defective cells, and faulty interconnections [61]. Inspections can be performed manually or using drones, though accuracy may be influenced by environmental conditions such as wind speed and ambient temperature.

I-V Curve Analysis involves measuring the current-voltage characteristics of PV modules or strings and comparing them with reference curves. Different fault types produce distinct distortions in the I-V curve, making this method highly informative. However, it requires specialized equipment and is generally not suitable for continuous real-time monitoring in large-scale systems [62].

Power Loss Analysis focuses on evaluating the difference between expected and actual power output. Using models that incorporate irradiance, temperature, and system parameters, expected power generation can be estimated with reasonable accuracy. Significant discrepancies indicate potential faults, making this approach particularly useful for utility-scale PV systems where direct inspection is not feasible [62].

II.6.2. Fault Diagnosis Approaches

Fault diagnosis involves identifying, isolating, and classifying faults after their detection. Unlike fault detection, which focuses on recognizing abnormal behavior, diagnosis aims to determine the root cause and nature of the fault [62].

Fault diagnosis approaches in PV systems can be broadly categorized into four main types, as summarized in **Table II.4**.

Table II.4 Comparative Overview of Fault Diagnosis Approaches [63]

Diagnosis Approach	Knowledge Source	Interpretability	Data Dependence	Typical Output
Model-based diagnosis	Physical and mathematical models	High	Moderate	Residual-based fault indication
Signal processing-based diagnosis	Time and frequency signal features	Moderate	Moderate	Feature-based fault identification
Knowledge-based diagnosis	Expert rules and heuristics	Very high	Low to moderate	Rule-driven diagnosis
Data-driven diagnosis	Historical and operational data	Moderate to low	High	Automatic fault classification

Model-Based Diagnosis relies on mathematical or physical models that describe expected PV system behavior. The difference between measured and predicted values (the residual) indicates abnormalities. Common models include single-diode and double-diode models, which accurately predict I-V behavior under varying conditions [60], [63]. Effective for well-defined faults, but performance depends heavily on model accuracy.

Signal Processing-Based Diagnosis focuses on analyzing electrical and temporal signals (voltage, current, power) to extract meaningful features. Techniques such as Fourier Transform, Wavelet Transform, and time-frequency analysis are widely used. For example, arc faults can be detected through high-frequency components in current signals, while partial shading can be identified through power output fluctuations [64]. Does not require detailed physical models but is sensitive to noise.

Knowledge-Based Diagnosis relies on expert knowledge, heuristic rules, and historical data. Typical implementations include rule-based systems, fuzzy logic systems, and expert systems. For example: "If current decreases significantly while voltage remains stable → open-circuit fault" [63]. Intuitive and interpretable, but depends heavily on expert knowledge and struggles with new fault patterns.

Data-Driven and Intelligent Diagnosis is the most advanced and widely adopted approach. It leverages large volumes of historical and real-time data to automatically learn patterns associated with normal and faulty operating conditions [65]. Machine learning (ML) and artificial intelligence (AI) techniques play a central role, including supervised learning (SVM, Decision Trees, Random Forests, KNN, ANN, deep learning) and unsupervised learning (clustering, anomaly detection). Advantages include high accuracy, adaptability, and scalability. Limitations include requiring high-quality datasets and rigorous model validation [63], [68].

II.6.3. Data-Driven Detection Methods

Data-driven methods represent the most advanced approach to fault detection in PV systems. These methods leverage historical and real-time data to identify patterns, correlations, and anomalies using statistical analysis and machine learning techniques [63].

Common approaches include:

- **Statistical methods:** Regression analysis, clustering, hypothesis testing, control charts.
- **Machine learning algorithms:** Support Vector Machines (SVM), Decision Trees, Random Forests (RF), K-Nearest Neighbors (KNN).
- **Deep learning models:** For complex nonlinear pattern recognition.

These techniques provide high accuracy and scalability, especially in systems equipped with extensive monitoring infrastructure. However, they require large, high-quality datasets and careful model validation.

II.7 Monitoring and Data Acquisition Systems in PV Plants

Monitoring and data acquisition systems are essential components of modern PV plants, enabling continuous observation of system performance and facilitating fault detection and diagnosis. These systems collect data from various sensors and devices, process the information, and provide actionable insights for operators and intelligent algorithms. [69]

The effectiveness of fault detection and diagnosis strongly depends on the quality, resolution, and reliability of the acquired data. Therefore, the design and implementation of monitoring systems play a critical role in ensuring accurate and timely fault identification.

Table II.5 Functional Components of PV Monitoring and Data Acquisition Systems[63]

Component	Function	Contribution to Fault Detection
Sensors	Measure electrical and environmental variables	Provide raw diagnostic data
Data acquisition unit	Collect and digitize measurements	Ensures structured data flow
SCADA system	Centralize supervision and control	Supports real-time monitoring
Communication network	Transfer data between devices	Enables remote observability
IoT gateway	Connect field devices to cloud services	Extends system scalability
Analytics platform	Process and interpret data	Supports intelligent diagnosis
Operator interface	Display system status and alarms	Facilitates maintenance decisions

II.7.1 Sensors and Data Acquisition

Sensors are the primary sources of data in PV systems. They measure key parameters such as solar irradiance, module temperature, ambient temperature, voltage, current, and power output. [70]

Data acquisition (DAQ) systems collect, digitize, and store sensor data for further analysis. These systems operate at different sampling rates depending on the application, ranging from seconds for real-time monitoring to minutes for long-term performance evaluation. [63]

Advanced PV plants often include additional sensors for wind speed, humidity, and dust levels, enabling a more comprehensive understanding of environmental conditions.

Accurate sensor calibration, synchronization, and validation are essential to ensure data consistency and reliability. Measurement errors can lead to incorrect fault diagnosis and reduced system efficiency.

II.7.2 Real-Time Monitoring Systems

Real-time monitoring systems enable continuous tracking of PV system performance and immediate detection of anomalies. These systems provide live data visualization, automated alerts, and rapid response mechanisms to abnormal conditions. [61]

Such systems are particularly critical in large-scale solar plants, where manual inspection is impractical. They allow operators to detect and address faults promptly, minimizing energy losses and preventing further system degradation.

Modern monitoring platforms often include interactive dashboards that display key performance indicators (KPIs), facilitating intuitive interpretation and decision-making.

II.7.3 SCADA and IoT-Based Monitoring

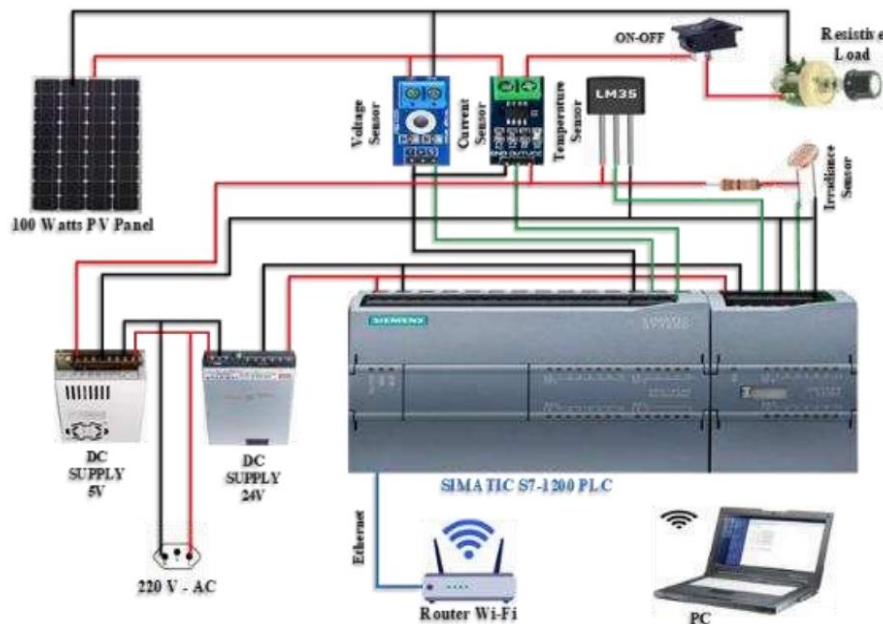


Figure II.2 Circuit diagram of the proposed PV monitoring system[69]

Supervisory Control and Data Acquisition (SCADA) systems are widely used in PV plants for centralized monitoring and control. These systems collect data from system components, including inverters, sensors, and protection devices, and provide a unified interface for system management. [62]

With the advancement of the Internet of Things (IoT), PV monitoring systems have become more distributed, scalable, and flexible. IoT-based architectures enable remote monitoring, cloud-based data storage, and advanced analytics capabilities.

These systems support the integration of machine learning algorithms for intelligent fault detection and predictive maintenance. In addition, IoT platforms facilitate real-time communication between devices, improving responsiveness and operational efficiency.

II.7.4. Data Quality and Big Data Challenges in PV Systems

The increasing volume of data generated by PV systems introduces significant challenges related to data quality and management. Common issues include missing data, measurement noise, sensor faults, and inconsistencies across multiple data sources. [66]

High-dimensional and high-frequency datasets require efficient storage, processing, and analysis techniques. Big data technologies, such as distributed computing frameworks and cloud platforms, are often employed to address these challenges.

Data preprocessing steps, including filtering, normalization, imputation, and outlier detection, are essential to ensure the reliability of fault diagnosis algorithms. Poor data quality can significantly degrade the performance of both statistical and machine learning models.

II.8 Performance Indicators for Fault Detection

Performance indicators are quantitative metrics used to evaluate the operational efficiency of PV systems and detect anomalies. These indicators provide a foundation for both conventional and data-driven fault detection methods. [70]

By analyzing deviations in these indicators, it becomes possible to identify potential faults and assess their impact on system performance.

II.8.1 Performance Ratio (PR)

The performance ratio (PR) is one of the most widely used indicators in PV systems. It represents the ratio between the actual energy output and the expected energy output under ideal conditions. [69]

$$\text{PR} = E_{\text{actual}} / E_{\text{expected}}$$

where E_{actual} is the measured energy output and E_{expected} is the expected energy based on irradiance and system characteristics.

A decrease in PR is a strong indicator of potential faults such as shading, degradation, or component malfunction.

II.8.2 Efficiency Metrics

Efficiency metrics evaluate the ability of PV systems to convert solar energy into electrical energy. These include module efficiency, inverter efficiency, and overall system efficiency. [71]

Variations in efficiency over time may indicate degradation or operational faults. For example, a sudden drop in inverter efficiency may signal overheating or internal malfunction.

II.8.3 Energy Yield Analysis

Energy yield refers to the total energy produced by a PV system over a given period, typically expressed in kWh/kWp. It provides a normalized measure of system performance. [72]

Comparing actual energy yield with expected values enables the identification of underperforming systems. Seasonal variations and environmental factors must be considered for accurate interpretation.

II.8.4 Statistical Indicators and Threshold-Based Detection

Statistical indicators such as mean, standard deviation, variance, and correlation coefficients are widely used in fault detection. These metrics help identify deviations from normal operating conditions. [73]

Threshold-based detection involves defining acceptable operating ranges and triggering alarms when deviations exceed predefined limits. While simple and effective, this approach may not capture complex fault patterns.

To overcome these limitations, advanced approaches combine statistical analysis with machine learning models, enabling more accurate detection, reduced false alarms, and improved adaptability to dynamic operating conditions.

II.9. Comparative Analysis of Fault Detection Methods

Fault detection methods in PV systems have evolved significantly, ranging from traditional manual inspections to advanced data-driven and intelligent techniques. Each method offers specific advantages and limitations depending on system size, data availability, computational resources, and required detection accuracy [72].

Conventional methods (visual inspection, electrical signal-based, thermal imaging, I-V curve analysis, power loss analysis) are simple and cost-effective but have significant limitations: they are labor-intensive, subjective, cannot detect hidden faults, require specialized equipment, or are not suitable for continuous real-time monitoring [54], [63].

Data-driven methods, including statistical analysis and machine learning techniques, offer superior flexibility and accuracy. They can capture complex and nonlinear relationships between system variables and detect subtle anomalies not easily identified by conventional techniques. Machine learning models (SVM, ANN, Random Forest) can classify faults with high accuracy when trained on sufficiently large and representative datasets [63], [64].

Limitations of data-driven approaches include:

- Require large volumes of high-quality data
- Substantial computational resources
- Careful model design and validation
- Limited interpretability (especially deep learning models)

In summary, no single fault detection method is universally optimal. **Hybrid approaches** that combine physical modeling, signal processing, and data-driven techniques are increasingly adopted. This comparative analysis highlights the necessity of intelligent, adaptive, and scalable fault detection solutions in modern PV systems.

II.10. Challenges in Fault Detection and Diagnosis

Despite significant advancements in FDD techniques for PV systems, several challenges continue to limit their effectiveness and large-scale deployment [74]:

Table II.6 Challenges in Fault Detection and Diagnosis

Challenge	Description
Environmental variability	Irradiance, temperature, humidity fluctuate continuously, making it difficult to distinguish normal variations from actual faults [72].
Data quality	Sensor data may be affected by noise, missing values, calibration errors, or communication failures, leading to inaccurate detection or false alarms [63].
Lack of labeled fault data	Fault events are relatively rare; historical datasets lack sufficient labeled examples, making supervised learning difficult [63].
System complexity	Large-scale PV plants have thousands of modules, multiple inverters, and complex interconnections; faults may interact and occur at different levels [72].
Real-time processing	Effective fault detection requires immediate response, demanding substantial computational resources and efficient algorithmic design.
Economic considerations	Advanced monitoring infrastructures and intelligent systems can be costly to deploy and maintain [75].
Integration of heterogeneous data	Combining electrical measurements, thermal data, and environmental information requires advanced data fusion techniques and standardized protocols.

II.11. Transition Toward Intelligent Fault Detection and Diagnosis

The limitations of conventional methods, combined with the increasing complexity of PV systems, have driven a paradigm shift toward **intelligent fault detection and diagnosis**. This transition is enabled by advances in data analytics, machine learning, and artificial intelligence [71].

Intelligent fault detection systems leverage large volumes of operational data to learn patterns associated with normal and faulty conditions. Unlike traditional approaches, which rely on predefined rules or simplified physical models, intelligent methods can automatically adapt to changing conditions and detect complex fault patterns [74].

Key aspects of this transition include:

- **Statistical analysis** (regression analysis, control charts, anomaly detection) identifies deviations from expected behavior and serves as a baseline for advanced models.

- **Machine learning algorithms** enable classification, prediction, and pattern recognition. Supervised learning recognizes specific fault types; unsupervised learning detects anomalies without labeled data. Deep learning techniques extract high-level features from complex data [75].
- **Predictive maintenance** analyzes historical data to predict potential failures before they occur, reducing downtime and maintenance costs [73].
- **IoT and cloud computing** enable continuous distributed data collection and scalable computational resources for processing large datasets and deploying ML models.

Challenges remain, including model interpretability, data privacy concerns, and the need for continuous model updating. Research focuses on **explainable AI (XAI)** techniques to ensure transparency and reliability in automated decision-making systems [71].

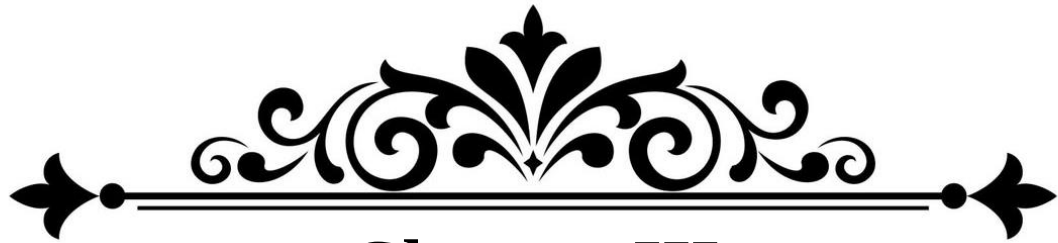
In conclusion, the transition toward intelligent FDD represents a necessary evolution in PV system management. By integrating statistical analysis, machine learning, and advanced monitoring technologies, more accurate, efficient, and proactive fault management is achievable. This transition provides a solid foundation for **Chapter III**, which will focus on the design and implementation of intelligent algorithms for fault detection and diagnosis in solar power plants.

II.12. Conclusion

This chapter has provided a comprehensive overview of fault detection and diagnosis in PV systems, highlighting their essential role in ensuring system reliability, efficiency, and operational safety. It examined the main failure mechanisms and presented a structured classification of fault types, followed by a detailed analysis of common faults and their impact on system performance. The chapter also reviewed a wide range of fault detection techniques, from conventional methods to advanced approaches based on thermal imaging, electrical analysis, and power modeling, as well as various fault diagnosis strategies including model-based, signal processing, knowledge-based, and data-driven methods.

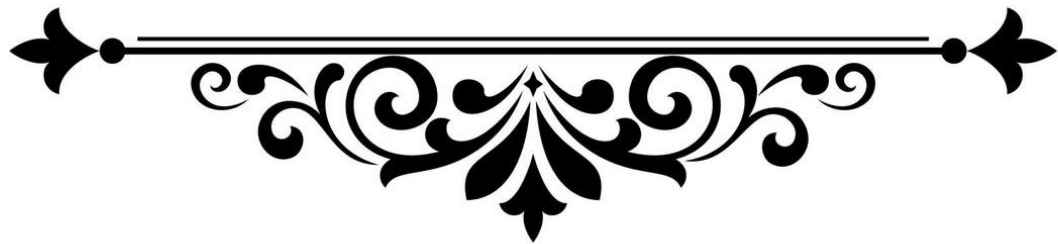
The comparative analysis demonstrated that while traditional techniques are useful, they remain limited in handling complex and large-scale PV systems. In contrast, intelligent approaches based on statistical analysis and machine learning offer higher accuracy, adaptability, and scalability, despite challenges related to data quality, computational requirements, and model interpretability.

This chapter underscores the importance of integrating advanced monitoring systems with intelligent analytical methods to achieve efficient and proactive fault management, thereby providing a solid foundation for the development and implementation of intelligent fault detection and diagnosis models in Chapter III.



Chapter III

Data-Driven and Machine Learning Approaches for Fault Detection in Photovoltaic Systems



III.1 Introduction

Photovoltaic (PV) generation is now considered a strategic technology in modern power systems because it converts solar irradiance directly into electricity and helps reduce dependence on fossil fuels. However, the reliability of PV plants remains affected by environmental variations, component aging, sensor quality, converter operation, grid conditions, and various electrical or physical faults. These faults—such as partial shading, soiling, degradation, open-circuit faults, short-circuit faults, ground faults, inverter faults, and sensor faults—can reduce energy production, increase maintenance costs, accelerate degradation, and create safety risks if not detected in time [77]. Moreover, traditional protection devices are not always sufficient in PV arrays, since fault currents may be limited and similar electrical symptoms may result from different causes [78].

For this reason, PV fault detection and diagnosis (FDD) requires advanced methods capable of distinguishing normal weather-related variations from abnormal operating conditions [81]. Data-driven and machine learning approaches address this issue by learning the normal behavior of the PV system from historical monitoring data, including irradiance, temperature, voltage, current, power, and inverter status [79]. Statistical methods offer interpretable abnormality indices and decision limits, while machine learning techniques provide flexible classification capabilities when labeled fault data are available [84].

Accordingly, this chapter first presents data-driven and statistical methods, then focuses on Principal Component Analysis (PCA) as the main multivariate technique, including the SPE/Q statistic and Hotelling's T^2 index. Finally, it discusses the machine learning algorithms adopted in this thesis—Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF)—for PV fault diagnosis [76], [105].

III.2. Data-Driven Approaches for Fault Detection

III.2.1. Definition of Data-Driven Methods

A data-driven method is a fault detection approach in which the diagnostic model is obtained primarily from measured data rather than from a complete analytical model of the process. In PV systems, the measured data represent the interaction between the PV generator, power electronics, meteorological conditions, and grid connection. The model learns the statistical structure or predictive relationship between these variables under normal operating conditions, then uses deviations from this learned behavior as evidence of abnormality [79], [80].

Data-driven methods can be purely statistical (e.g., control charts, regression residuals, correlation monitoring, PCA) or based on machine learning algorithms that map features to diagnostic labels. Their main advantage is that they can be implemented using monitoring data already available in many grid-connected PV plants.

In this thesis, the term "data-driven" includes statistical analysis, PCA-based monitoring, and supervised ML diagnosis. The statistical component is used for early fault detection by identifying abnormal behavior in multivariate data, while the ML component is used to classify detected faults when sufficient labeled data are available [77], [84].

III.2.2. Objectives of Data-Driven Fault Detection

The primary objective of data-driven fault detection in PV systems is to identify, as early as possible, a deviation between expected healthy behavior and actual measured behavior. Early detection is critical because many PV faults are progressive: soiling increases over time, degradation appears gradually, partial shading may become repetitive, and thermal stress may lead to permanent damage. A monitoring method must therefore be sensitive to incipient faults while maintaining a low false alarm rate under normal weather fluctuations [77], [82].

A second objective is fault diagnosis—assigning the abnormal event to a fault type or location. Detection answers "*Is something wrong?*" while diagnosis answers "*What is wrong and where?*" This requires features that preserve information about the physical fault mechanism [81], [83].

A third objective is decision support for operation and maintenance. A useful data-driven FDD system should reduce unnecessary inspections, prioritize maintenance actions, and support the economic operation of the PV plant [78], [79].

III.2.3. Data-Driven Framework in PV Systems

A general data-driven FDD framework for PV systems contains five main layers:

Table III.1 Data-Driven Framework in PV Systems

Layer	Function
Data acquisition	Collects electrical and meteorological measurements
Preprocessing	Removes invalid data, handles missing values, filters nighttime periods, normalizes variables

Feature construction	Derives informative indicators (power ratios, residuals, rolling statistics, PCA scores)
Model development	Estimates statistical models or trains ML classifiers
Decision generation	Compares monitoring indices with thresholds or assigns class labels

For online implementation, the framework distinguishes between two phases:

- **Offline phase:** Builds the normal reference model from historical healthy data.
- **Online phase:** Receives new measurements, applies the same preprocessing, projects data into the learned model, calculates detection indices, and generates alarms or diagnostic labels [84], [90].

Table III.2 Typical PV Monitoring Variables and Diagnostic Relevance [79], [80]

Variable	Symbol	Diagnostic Relevance
Plane-of-array irradiance	G	Separates weather-driven power variation from abnormal power loss; essential for reference modeling
Module temperature	T_{mod}	Influences voltage and power; abnormal heating may indicate hotspots or poor ventilation
DC power	P_{dc}	Aggregates voltage-current behavior; widely used for residual-based detection
Wind speed	—	Affects module cooling; relevant for temperature correction
Relative humidity	—	Influences soiling and potential-induced degradation (PID)

III.3. Data Acquisition and Preprocessing

III.3.1. PV Monitoring Data

The quality of any data-driven PV fault detection method depends directly on the quality of the monitoring data. A complete dataset should include meteorological variables, DC-side electrical variables, AC-side electrical variables, inverter information, and timestamps [79], [80].

In this thesis, the monitored variables are: solar irradiance (G), module temperature (T), DC power (P), wind speed, and relative humidity. These variables are selected because they are the most relevant for detecting environmental faults (shading, soiling, temperature effects) and sensor faults (bias, drift).

PV monitoring data are usually **nonstationary** because irradiance changes during the day, temperature follows seasonal and daily cycles, and the MPPT controller continuously adjusts the operating point. Therefore, applying fixed thresholds to raw power values without considering irradiance and temperature is inappropriate. A correct data-driven approach first transforms raw measurements into normalized or residual features that are less dependent on environmental variability [77], [82].

III.3.2. Data Cleaning

Data cleaning removes invalid, duplicated, missing, or physically inconsistent measurements. Common data-quality problems in PV datasets include communication interruptions, sensor saturation, inverter shutdown, nighttime values, zero irradiance periods, negative power values, frozen sensors, measurement spikes, and timestamp misalignment [92].

A typical cleaning procedure includes:

Step	Description
Nighttime filtering	Exclude data when irradiance is negligible (e.g., $G < 10 \text{ W/m}^2$)
Range checking	Remove impossible values (negative irradiance, unrealistic temperature, negative power)
Missing data handling	Interpolate or flag missing timestamps

Outlier treatment Remove isolated spikes (using interquartile range or Hampel filters) while preserving systematic deviations that could correspond to faults

III.3.3. Normalization and Standardization

Normalization and standardization are essential because PV variables have different units, magnitudes, and variances. PCA and many ML algorithms are sensitive to scaling. Without scaling, variables with large numerical ranges (e.g., irradiance, power) may dominate the model.

The most common standardization is **z-score scaling**. For a data matrix X with n observations and m variables, the standardized value z_{ij} is:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (\text{III.1})$$

where μ_j and σ_j are the mean and standard deviation of variable j estimated from healthy training data. In online monitoring, the same μ_j and σ_j must be reused to prevent data leakage [88], [89].

III.3.4. Feature Extraction and Selection

Feature extraction transforms raw measurements into informative indicators that improve fault separability. Typical features for PV systems include:

- Normalized current I_{dc}/G
- Normalized power P_{dc}/G
- Performance ratio (PR)
- Residuals between measured and expected power
- Rolling statistics (mean, standard deviation)

Feature selection reduces redundancy and improves computational efficiency. Methods include correlation analysis, PCA, and supervised techniques (e.g., random forest feature importance) [84].

Table III.3 Main Preprocessing Operations for PV Data-Driven FDD

Operation	Purpose	Typical Method
Nighttime filtering	Remove periods with no meaningful PV generation	Irradiance threshold + inverter status
Range checking	Remove physically inconsistent values	Sensor and equipment limits
Missing data handling	Avoid biased model fitting	Deletion, interpolation, flagging
Outlier treatment	Remove measurement spikes while preserving true faults	Robust filters, moving windows
Standardization	Give variables comparable numerical weight	Z-score using healthy training statistics
Feature extraction	Increase fault sensitivity, reduce weather dependency	Ratios, residuals, rolling statistics, PCA scores
Feature selection	Reduce redundancy and overfitting	Correlation, PCA, mutual information

III.4. Statistical Methods for Fault Detection

Statistical fault detection assumes that healthy operation has a recognizable statistical signature that changes when a fault occurs [79], [81].

III.4.1. Statistical Indicators

Simple statistical indicators include mean, variance, standard deviation, skewness, kurtosis, and moving averages. For PV systems, these are often computed over moving windows to capture temporal behavior changes.

Performance indicators (performance ratio, specific yield, capacity factor) can reveal long-term underperformance but must be corrected for irradiance and temperature to avoid confusing weather variations with actual faults.

III.4.2. Correlation Analysis

PV variables are not independent. Under normal operation:

- Irradiance is positively correlated with current and power
- Module temperature is negatively correlated with voltage
- DC power is strongly related to AC power through inverter efficiency

A fault may disturb these correlations. The **Pearson correlation coefficient** r_{xy} between two variables x and y is:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (\text{III.2})$$

A value close to +1 indicates a strong positive linear relationship; close to −1 indicates a strong negative relationship. Correlation analysis is useful for feature selection but is not sufficient alone for fault detection under nonlinear operating conditions.

III.4.3. Residual-Based Detection

Residual-based detection compares measured behavior with expected behavior. A residual is the difference between an observed value and a predicted or reconstructed value:

$$e_p(k) = P_{meas}(k) - P_{pred}(k) \quad (\text{III.3})$$

A residual close to zero indicates normal operation; a large residual indicates a possible fault. Residuals can be generated using physical models, regression models, or PCA reconstruction [77], [81], [82].

III.4.4. Threshold-Based Detection

Threshold-based detection generates an alarm when a diagnostic indicator exceeds a predefined limit. For a residual e with estimated standard deviation σ_e under healthy conditions, a simple **three-sigma rule** is:

$$|e(k)| > 3\sigma_e \rightarrow \text{alarm} \quad (\text{III.4})$$

However, fixed thresholds may fail under rapidly changing irradiance. Multivariate statistical methods such as PCA overcome this limitation by monitoring the correlation structure of several variables simultaneously [90], [92].

Table III.4 Statistical Methods Commonly Used in PV Fault Detection

Method	Main Idea	Advantages	Limitations
Univariate thresholds	Compare each variable/residual with a limit	Simple and fast	Weak under variable weather
Moving statistics	Monitor mean, variance, or trend over a window	Useful for gradual changes	Window size affects sensitivity
Correlation analysis	Track relationships among variables	Interpretable, useful for feature selection	Limited for nonlinear faults
Regression residuals	Compare measured output with predicted healthy output	Directly measures underperformance	Depends on model accuracy
PCA monitoring	Detect abnormal changes in multivariate correlation structure	Handles correlated variables, dimension reduction	

III.5. Principal Component Analysis (PCA)

III.5.1. Definition and Principle of PCA

Principal Component Analysis (PCA) is a multivariate statistical technique that transforms a set of correlated variables into a smaller set of uncorrelated latent variables called **principal components**. Each principal component is a linear combination of the original variables. The first principal component captures the maximum possible variance in the data, the second captures the maximum remaining variance under orthogonality constraints, and so on [88], [89].

In process monitoring, PCA is used not only for dimension reduction but also to describe the **normal correlation structure** of a process. If new data follow the same correlation structure as the healthy training data, their projection and reconstruction errors remain within statistical limits. If a fault changes the relationship between variables, the PCA scores or residuals become abnormal. This property makes PCA highly suitable for PV systems, where

current, voltage, power, irradiance, and temperature are strongly correlated during healthy operation [77], [84], [107].

III.5.2. Objective of PCA in PV Fault Detection

The objective of PCA in PV fault detection is to build a compact statistical model of healthy PV behavior and then detect deviations from that model. PCA captures relationships between variables (e.g., DC power with current and voltage, AC power with DC power through inverter efficiency) and separates the data space into two subspaces:

Subspace	Description
Principal subspace	Contains systematic variation associated with normal operation (irradiance and temperature changes)
Residual subspace	Contains variation not explained by the retained principal components

Faults may appear in either subspace:

- A fault that produces unusual but still correlated variation may increase **Hotelling's T²** statistic.
- A fault that breaks the normal correlation structure may increase the **SPE/Q statistic**.

In this thesis, both T² and SPE/Q are used for fault detection [87], [90], [108].

III.5.3. Mathematical Formulation of PCA

Let X be the standardized healthy data matrix with n observations and m variables:

$$X = \{x_1^T; x_2^T; \dots; x_n^T\} \in \mathbb{R}^{n \times m} \quad (\text{III.5})$$

The sample covariance matrix S is computed as:

$$S = \frac{1}{n-1} X^T X \quad (\text{III.6})$$

PCA solves the eigenvalue decomposition problem:

$$SP = P\Lambda \quad (\text{III.7})$$

where $P = [p_1, p_2, \dots, p_m]$ is the loading matrix and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$ contains the eigenvalues sorted in decreasing order. The score matrix T is obtained by projecting the standardized data onto the loading vectors:

$$T = XP \quad (\text{III.8})$$

If A principal components are retained, P_A contains the first A loading vectors and Λ_A contains the corresponding eigenvalues. The reduced score matrix is:

$$T_A = XP_A \quad (\text{III.9})$$

The PCA model represents each observation by a lower-dimensional vector of scores, which summarize the dominant healthy variation and can be used as inputs to ML classifiers [84], [89].

III.5.4. Selection of Principal Components

The number of retained principal components A is a critical parameter. Too few components may discard useful variation (false alarms); too many may include noise (reduced fault sensitivity). Common selection criteria include:

Table III.5 Selection of Principal Components

Criterion	Description
Cumulative explained variance	Select smallest A that explains a predetermined percentage (e.g., 85%, 90%, 95%)
Scree plot	Select point where eigenvalue decrease becomes less pronounced
Cross-validation	Minimize prediction or reconstruction error on validation data

The explained variance for A components is:

$$\text{Explained variance}(\%) = \frac{\sum_{i=1}^A \lambda_i}{\sum_{i=1}^m \lambda_i} \times 100 \quad (\text{III.10})$$

For PV systems, the selection of A should also consider physical interpretability. The first component may represent irradiance-driven power variation, while another may represent thermal effects on voltage [89], [90].

III.5.5. PCA Reconstruction and Residuals

The reconstruction of a standardized observation x using A retained components is:

$$\hat{x} = xP_AP_A^T \quad (\text{III.11})$$

The residual vector e is the difference between the actual observation and its PCA reconstruction:

$$e = x - \hat{x} = x(I - P_A P_A^T) \quad (\text{III.12})$$

A large residual indicates that the relationship among variables is no longer consistent with healthy operation. For example, if irradiance and temperature are normal but measured power is lower than expected, the residual will increase significantly.

III.6. PCA-Based Fault Detection Indices

III.6.1. Hotelling's T^2 Statistic

Hotelling's T^2 statistic measures the distance of a projected observation from the center of the PCA score space. It monitors abnormal variation **within** the principal subspace. For an observation x and its score vector $t = xP_A$, the T^2 statistic is defined as [91]:

$$T^2 = t^T \Lambda_A^{-1} t \quad (\text{III.13})$$

A high T^2 value indicates that the observation lies far from the normal score distribution. In PV systems, T^2 can increase during unusual operating conditions (e.g., abnormal high-temperature/low-voltage condition).

The control limit for T^2 at significance level α is based on the F-distribution:

$$T_\alpha^2 = \frac{A(n-1)}{n-A} F_{A,n-A,\alpha} \quad (\text{III.14})$$

III.6.2. SPE / Q Statistic

The Squared Prediction Error (SPE), also called the Q statistic, measures the magnitude of the residual vector in the residual subspace [87]:

$$Q = e^T e = \|x - \hat{x}\|^2 \quad (\text{III.15})$$

A high Q value indicates poor reconstruction, often occurring when a fault breaks the normal correlation structure (e.g., sensor bias, string disconnection).

The Q statistic control limit can be approximated using the Jackson-Mudholkar method.

Let θ_i be computed from the discarded eigenvalues $\lambda_{A+1}, \dots, \lambda_m$:

$$\theta_i = \sum_{j=A+1}^m \lambda_j^i, i = 1, 2, 3 \quad (\text{III.16})$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \quad (\text{III.17})$$

$$Q_\alpha = \theta_1 \left[\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \quad (\text{III.18})$$

where c_α is the standard normal deviate corresponding to the selected confidence level.

III.6.3. Fault Detection Decision Rules

The decision rule transforms the PCA indices into alarms. Several rules can be used depending on desired sensitivity and false alarm tolerance:

Table III.6 Fault Detection Decision Rules

Rule	Description
T² rule	Alarm if $T^2(k) > T_\alpha^2$
Q rule	Alarm if $Q(k) > Q_\alpha$
Persistence rule	Alarm validated only if threshold exceeded for L consecutive samples
Voting rule	Alarm when at least two indices exceed limits within a time window

Persistence and voting rules are important in PV monitoring because clouds, sensor noise, and inverter transients can generate isolated threshold crossings [90], [108].

In this thesis, the PCA decision process is:

1. Train PCA model using healthy data
2. Calculate T_α^2 and Q_α
3. For each online observation: standardize, compute scores and residuals, calculate T^2 and Q
4. Compare indices with thresholds
5. If a fault is detected, pass the feature vector to the ML diagnosis stage

Table III.7 PCA-Based Indices for PV Fault Detection

Index	Subspace Monitored	Interpretation	Typical Alarm Rule
Hotelling's T^2	Principal subspace	Abnormal score variation within the PCA model	$T^2 > T_\alpha^2$
SPE / Q statistic	Residual subspace	Poor reconstruction; broken correlation structure	$Q > Q_\alpha$
Persistence alarm	Time validation	Confirms abnormality is not an isolated spike	Index exceeds limit for L samples

III.6.4. Fault Localization Using Contribution Plots

Fault detection indicates *that* the PV system is operating abnormally, but it does not identify *which* variable is responsible for the alarm. For this reason, a **localization stage** is added after PCA-based detection using **contribution plots**—data-driven and explainable tools that decompose a monitoring index into contributions of individual variables [110], [111].

The principle is simple: after an alarm is generated (by T^2 or Q), the contribution of each variable (irradiance, temperature, power, etc.) is calculated. The variable with the highest contribution is considered the most likely cause of the fault.

For the SPE/Q index, the residual vector $e = x - \hat{x}$ provides a direct way to calculate the contribution of each standardized variable:

$$c_j^Q(k) = e_j^2(k), j = 1, \dots, m \quad (\text{III.19a})$$

where $e_j(k)$ is the residual associated with the j -th variable. Since $Q(k) = \sum_j c_j^Q(k)$, a high value of $c_j^Q(k)$ means variable j is poorly reconstructed. A normalized contribution can also be used:

$$c_j^Q(k) = \frac{c_j^Q(k)}{Q(k)} \quad (\text{III.19b})$$

For Hotelling's T^2 , the contribution is computed by distributing the score-space distance over the original standardized variables:

$$c_j^{T^2}(k) = x_j(k)[P_{-A} \lambda_A^{-1} P_A^T x(k) T] j$$

where P_A is the loading matrix of the retained principal components and λ is the diagonal matrix of retained eigenvalues.

The fault-localization decision is obtained by ranking variables. The most suspicious variable is:

$$j^* = \arg \max_j Q(k) \text{ and } \arg \max_j T2(k)$$

Important note: A high contribution does not always mean the sensor itself is faulty. For example, a high contribution of DC current may indicate a current sensor error, or it may indicate partial shading, soiling, or string disconnection. Therefore, localization must be combined with PV engineering knowledge.

Table III.8 Diagnostic Interpretation of Contribution Plots in PV Systems [108], [110], [111]

Dominant Contribution	Possible Affected Variable	Probable Diagnostic Interpretation
High contribution of G	Irradiance sensor	Sensor bias, poor cleaning, wrong orientation
High contribution of T_mod / T_amb	Temperature channel	Thermal anomaly, sensor drift, hotspot tendency
High contribution of V_dc	DC voltage / MPPT channel	Open-circuit effect, bypass diode problem, MPPT anomaly
High contribution of I_dc	DC current / string current	Partial shading, soiling, string disconnection, sensor error
High contribution of P_dc	DC power	Underperformance, mismatch, degradation
High contribution of P_ac / S_inv	Inverter / AC side	Inverter derating, grid disturbance, conversion loss

III.7. Advantages and Limitations of PCA

III.7.1. Advantages

PCA has several advantages for PV fault detection [84], [90]:

Table III.9 Advantages of PCA

Advantage	Description
Unsupervised	Can be trained using only healthy data (no labeled faults required)
Handles correlated variables	Reduces dimensionality, useful for systems with many variables
Interpretable indices	T^2 and Q can be plotted over time and linked to events
Easy integration with ML	PCA scores, residuals, or detection indices can be used as features for classifiers

III.7.2. Limitations

Table III.10 Limitations of PCA

Limitation	Description
Linearity	Classical PCA assumes linear relationships; PV systems are nonlinear (I-V curves, MPPT, irradiance dynamics)
Sensitivity to scaling and training data	If healthy data contain undetected faults, the model learns faulty behavior as normal
Nonstationarity	Seasonal changes, sensor drift, and degradation can change the normal distribution, increasing false alarms

Extensions for handling nonlinear or time-varying behavior include adaptive PCA, moving-window PCA, recursive PCA, and kernel PCA [107].

Despite these limitations, PCA remains highly suitable for this thesis because it provides a mathematically transparent bridge between statistical monitoring and machine learning. It detects abnormalities without requiring labeled faults and can generate features for downstream classifiers.

III.8. Machine Learning for Fault Diagnosis

Machine learning (ML) is a branch of artificial intelligence in which models improve their performance through data. In PV fault detection and diagnosis, ML complements domain knowledge: engineering expertise is used to select variables, construct features, clean data, and interpret results, while ML learns nonlinear relationships and decision boundaries that may be difficult to derive analytically [76], [93].

ML has been widely applied in PV systems for power forecasting, MPPT, performance prediction, and fault diagnosis. In the FDD context, ML methods can distinguish between normal operation and different fault classes using measured electrical and environmental features [105], [106].

III.8.1. Supervised Learning

Supervised learning uses labeled examples to learn a mapping from input features to output classes. In PV fault diagnosis, inputs are electrical and meteorological features, and outputs are fault labels. This approach is appropriate when a representative dataset with both healthy and faulty samples is available. Algorithms in this category include SVM, KNN, random forests, and neural networks [84], [94].

Advantages: Direct classification of fault types.

Limitations: Requires reliable labels; fault labels are often rare or not confirmed by maintenance records.

III.8.2. Unsupervised Learning

Unsupervised learning uses unlabeled data to discover structure, clusters, or anomalies. PCA, clustering, and autoencoders are examples. In PV fault detection, unsupervised learning is useful because healthy data are more available than labeled fault data [84].

In this thesis, PCA plays the role of an unsupervised statistical detector for anomaly detection. ML algorithms are then used for supervised diagnosis when labels are available.

III.9. Machine Learning Algorithms for PV Fault Diagnosis

The practical classification stage of this thesis is restricted to three supervised algorithms: Random Forest, SVM, and KNN. This choice allows comparison of three complementary learning principles: ensemble learning (Random Forest), maximum-margin classification (SVM), and distance-based classification (KNN).

Table III.11 Comparison of ML Algorithms for PV Fault Diagnosis

Algorithm	Strengths	Limitations	Typical PV Use
SVM	Good with moderate data; effective nonlinear kernels (RBF)	Requires hyperparameter tuning; less scalable for very large data	Classification of fault types using engineered features
KNN	Simple and intuitive; no complex training	Sensitive to scaling, noise, and irrelevant features	Baseline classifier for PCA scores or normalized indicators
Random Forest	Robust ensemble; feature importance; nonlinear modeling	Less transparent than a single tree	High-performance classification with mixed PV features

III.9.1. Support Vector Machine (SVM)

SVM is a supervised learning algorithm that searches for an optimal separating hyperplane between classes. For linearly separable data, it maximizes the margin between the closest samples of different classes (support vectors). For nonlinear data, **kernel functions** (e.g., Radial Basis Function, RBF) map the data into a higher-dimensional feature space where separation becomes easier [95].

Why SVM for PV fault diagnosis?

- Performs well with moderate-size datasets and high-dimensional features
- Can classify faults based on voltage, current, power, irradiance, temperature, and PCA-derived features

Limitations: Requires careful selection of hyperparameters (penalty parameter C , kernel parameter γ). Can become computationally expensive for very large datasets.

III.9.2. K-Nearest Neighbors (KNN)

KNN is a simple supervised classification method. A new sample is assigned to the class most common among its K nearest training samples according to a distance measure (usually Euclidean distance) [96]. KNN has almost no training phase, but prediction requires distance calculations against stored training samples.

Why KNN for PV fault diagnosis?

- Effective when fault classes form compact clusters in feature space
- Easy to implement and interpret

Limitations: Sensitive to feature scaling, irrelevant features, class imbalance, and the choice of K . Standardized features are essential. PCA can improve KNN by reducing dimensionality and removing correlation among variables.

III.9.3. Random Forest (RF)

Random forest is an ensemble method that trains many decision trees using bootstrap samples and random subsets of features. The final classification is obtained by majority voting across the trees. Random forests reduce overfitting compared with individual decision trees and provide measures of feature importance [98].

Why Random Forest for PV fault diagnosis?

- Models nonlinear relationships and interactions among variables without strong parametric assumptions
- Robust to noisy features and can handle mixed feature types

Limitations: Larger model size and lower interpretability than a single tree, though feature importance and partial dependence analysis can help explain results.

III.9.4. Model Training and Validation

Model training and validation must follow a rigorous workflow to avoid optimistic performance estimates:

Phase	Description
Training set	Used to estimate model parameters
Validation set	Used to tune hyperparameters (e.g., C , γ , K , number of trees)
Test set	Used only once for final performance evaluation

For time-series PV data, chronological splitting (earlier data for training, later data for testing) is more realistic than random splitting to avoid data leakage. When the dataset is small, k-fold cross-validation can be used, but folds should be constructed carefully to avoid mixing nearly identical samples from the same fault episode across training and test sets [103].

The preprocessing pipeline (scaling parameters, PCA loadings) must be fitted **only on training data** and then applied to validation and test data without recalculation. This ensures an honest estimate of online performance.

III.10. Evaluation Metrics

III.10.1. Confusion Matrix

A confusion matrix summarizes the relationship between predicted and true classes. For a binary fault detection problem, the four basic quantities are:

Term	Description
TP (True Positives)	Faulty samples correctly detected as faulty
TN (True Negatives)	Healthy samples correctly classified as healthy
FP (False Positives)	False alarms
FN (False Negatives)	Missed faults

In multiclass PV diagnosis, the confusion matrix shows which fault classes are confused (e.g., partial shading vs. soiling), providing more diagnostic insight than a single accuracy value.

III.10.2. Performance Metrics

Table III.12 Evaluation Metrics for PV Fault Detection and Diagnosis[102]

Metric	Formula	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall proportion of correct predictions
Precision	$\frac{TP}{TP + FP}$	Reliability of positive alarms (low false alarms)
Recall	$\frac{TP}{TP + FN}$	Ability to detect actual faults (low missed faults)
F1-Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Harmonic mean of precision and recall
False Alarm Rate (FAR)	$\frac{FP}{FP + TN}$	Probability of false alarm during healthy operation

Detection Delay	Time between fault occurrence and alarm	Online responsiveness measure
------------------------	---	-------------------------------

For multiclass diagnosis, macro-averaged F1 gives equal weight to each class, while weighted F1 accounts for the number of samples in each class.

III.11. Comparative Analysis and Justification

Statistical and ML methods are complementary rather than competing [84], [105].

Table III.13 Comparative Analysis of Statistical and ML Methods

Criterion	Statistical/PCA Methods	Machine Learning Methods
Required labels	Usually healthy data only	Labeled healthy and faulty data
Main output	Alarm index and abnormality score	Fault class or probability
Interpretability	High (T^2 , Q , residuals, contributions)	Variable (high for trees, lower for complex models)
Nonlinearity	Limited (classical PCA is linear)	Good (nonlinear classifiers)
Data demand	Moderate	Higher
Best role in thesis	Early detection and feature generation	Fault diagnosis and class discrimination

In this thesis, a two-stage architecture is adopted: PCA-based detection first, followed by ML-based classification. This reduces the diagnostic burden on the classifier and improves interpretability, as alarms are supported by statistical indices (T^2 , Q).

Table III.14 Reasons for Selecting the Practical ML Algorithms

Algorithm	Reason for Selection	Main Advantages for PV FDD	Main Precautions
Random Forest	Ensemble learning	Handles nonlinear interactions; robust to noise; provides feature importance	Requires tuning (trees, depth); less transparent than a single tree
SVM	Maximum-margin classification	Effective with moderate datasets; RBF kernel handles nonlinear separation	Requires scaling and hyperparameter tuning (C, γ)
KNN	Distance-based classification	Simple, transparent, no complex training; useful for testing feature-space separability	Sensitive to scaling, irrelevant features, class imbalance, and choice of K

III.12. Proposed Methodology Adopted in This Thesis

The methodology adopted in this thesis is a **hybrid data-driven framework** that combines:

1. **Statistical PCA-based detection** (unsupervised)
2. **Contribution-plot-based fault localization** (explainable)
3. **ML-based diagnosis** using Random Forest, SVM, and KNN (supervised)

The objective is to exploit the interpretability and low label requirement of PCA, identify the variable mainly responsible for an alarm through contribution plots, and then classify the detected operating state using supervised ML when labeled data are available. This methodology is consistent with recent PV FDD research [84], [108], [112].

III.12.1. Workflow Description

The proposed workflow consists of the following sequential steps:

Table III.15 Workflow Description

Step	Operation	Output
1. Acquisition	Collect PV monitoring data (irradiance, temperature, voltage, current, power, inverter status)	Raw dataset
2. Cleaning	Remove nighttime samples, invalid values, duplicates; handle missing values	Validated dataset
3. Scaling & feature extraction	Apply z-score standardization; construct ratios, residuals, rolling statistics	Standardized features
4. PCA modeling	Train PCA model on healthy standardized data; select number of components	Healthy PCA model
5. Detection	Compute T^2 and Q statistics; compare with control limits; apply persistence rule	Normal / fault alarm
6. Localization	Generate contribution plots; rank variables by contribution to the alarm	Suspicious variable(s)
7. ML diagnosis	Train and compare Random Forest, SVM, and KNN classifiers	Predicted fault class
8. Evaluation	Compute accuracy, precision, recall, F1-score, FAR, detection delay	Performance assessment
9. Maintenance interpretation	Relate results to PV engineering causes	Decision support

III.12.2. Key Stages

A. Data Preprocessing

The dataset includes meteorological variables (irradiance, temperature, wind speed, humidity) and electrical variables (DC power, voltage, current). Data cleaning includes:

- Removing nighttime samples (irradiance < 10 W/m²)

- Removing invalid sensor values and duplicate timestamps
- Handling missing values via interpolation
- Standardization using z-score scaling

B. Statistical Detection (PCA)

A PCA model is trained on healthy standardized data. The number of principal components is selected using explained variance (85–95%) and validation performance. For each new observation:

- T^2 **statistic** monitors the principal subspace
- Q **statistic (SPE)** monitors the residual subspace

A fault alarm is generated when either index exceeds its control limit, with a persistence rule to reduce false alarms from transients (clouds, sensor noise).

C. Fault Localization (Contribution Plots)

After an alarm, contribution plots decompose the monitoring index into contributions of individual variables. The variable with the highest contribution is considered the most likely cause of the fault (e.g., temperature sensor bias, partial shading affecting power).

D. ML Diagnosis (Random Forest, SVM, KNN)

PCA scores, residuals, and contribution features are used as inputs to three supervised classifiers:

Algorithm	Role
Random Forest	Ensemble learning; handles nonlinear interactions; provides feature importance
SVM (RBF kernel)	Maximum-margin classification; effective for moderate datasets
KNN	Distance-based baseline; tests feature-space separability

E. Validation and Evaluation

- **Training/validation/test split** or chronological splitting for time-series data
- **Hyperparameter tuning** on validation set

- **Evaluation metrics:** accuracy, precision, recall, F1-score, false alarm rate (FAR), detection delay

III.12.3. Methodological Flowchart

The complete methodological logic is illustrated in **Figure III.1**, which shows the workflow from data acquisition to cleaning, feature extraction, PCA detection, contribution-plot localization, ML diagnosis (Random Forest, SVM, KNN), and final evaluation with maintenance-oriented interpretation.

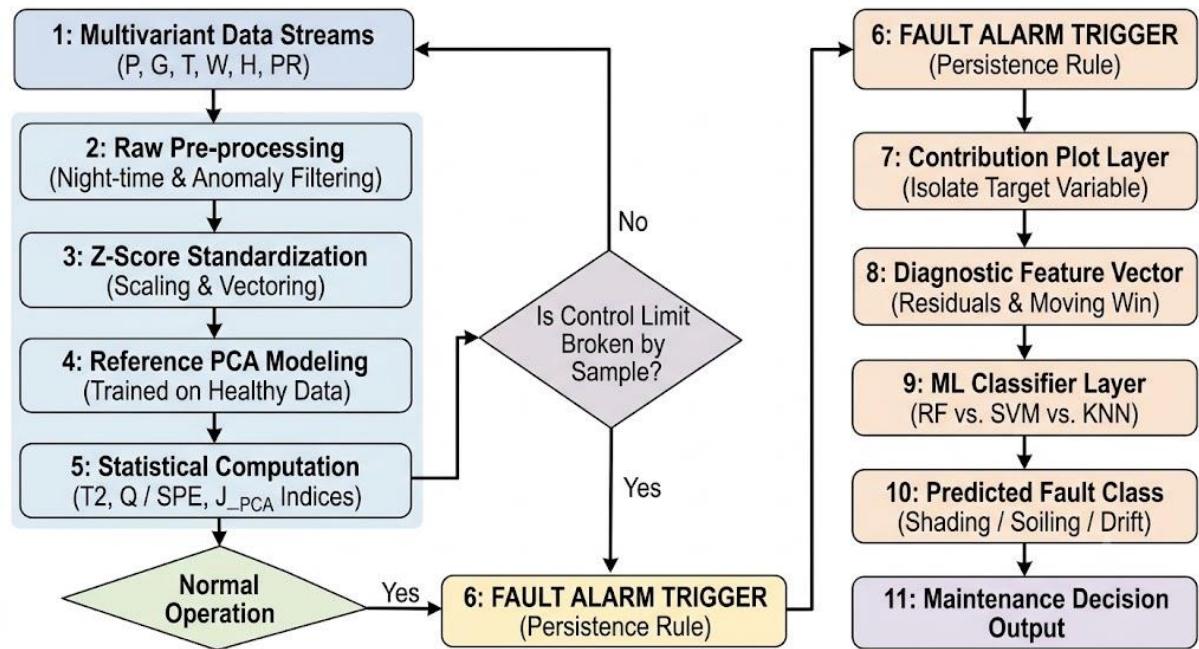


Figure III.1 Global flowchart of the proposed PCA-contribution-plot-ML methodology.

III.13. Conclusion

This chapter has presented the theoretical foundations of statistical, data-driven, and machine learning approaches applied to fault detection, localization, and diagnosis in photovoltaic systems. It highlighted the central role of **PCA** for unsupervised detection through the T^2 and Q (SPE) indices, as well as the value of **contribution plots** for localizing the variables responsible for anomalies. Finally, it demonstrated the complementarity between PCA, contribution plots, and the **Random Forest, SVM, and KNN** classifiers, forming a reliable and interpretable hybrid methodology for PV fault diagnosis.



Chapter IV

Implementation, Results and Discussion

IV.1. Introduction

This chapter presents the implementation and validation results of the proposed intelligent fault detection and diagnosis framework for solar power plants. The objective is to evaluate the ability of statistical monitoring (PCA, SPE, Hotelling's T^2) and machine learning techniques (SVM, Random Forest, KNN) to detect abnormal operating conditions, isolate the most influential variables, and support maintenance decisions in a photovoltaic (PV) power plant environment.

The analysis is organized around six monitored variables: generated power (P), solar irradiance (G), temperature (T), wind speed, relative humidity, and atmospheric pressure. These variables are directly extracted from the plant's monitoring system.

The chapter first describes the case study and the software environment used for the analysis. Then, the statistical behavior of the monitored variables is discussed using the correlation matrix. Fault detection is evaluated using the Squared Prediction Error (SPE/Q-statistic) and Hotelling's T^2 statistic. A hybrid temperature bias and drift fault scenario is then injected and analyzed to verify the sensitivity of the proposed monitoring strategy. Finally, a comparative machine learning study is presented to identify the most suitable classifier for multi-fault diagnosis.

IV.2. Case Study: Oued El Kebrit Photovoltaic Power Plant

The **Oued El Kebrit photovoltaic power plant**, located in the province of Souk Ahras in northeastern Algeria, was selected as the case study. The plant has an installed capacity of approximately **15 MWp** and represents a real operational PV generation site connected to the electrical grid. Such a site is suitable for fault detection research because its performance is influenced by both environmental conditions and technical operating parameters.

The monitored PV system is affected by common practical disturbances such as:

- Solar irradiance variability
- Temperature rise
- Dust accumulation (soiling)
- Sensor bias and drift
- Partial shading
- Electrical anomalies

These conditions can reduce energy yield, increase operation and maintenance costs, and accelerate the degradation of PV components. Therefore, the use of an intelligent monitoring system is essential for early detection, diagnosis, and decision support.

IV.3. Development Environment and Tools

The implementation was carried out using a **Python-based environment**. **Google Colab** was used as the main cloud platform because it provides an interactive Jupyter Notebook interface and allows the execution of data analysis and machine learning algorithms without local installation constraints. Python was selected for its flexibility and its large ecosystem for scientific computing, data processing, statistical analysis, and artificial intelligence.

Several Python libraries were used during development:

Table IV.1 Python Libraries

Library	Purpose
NumPy	Numerical calculations and array operations
Pandas	Data organization, cleaning, and manipulation
Matplotlib / Seaborn	Plotting the correlation matrix, monitoring charts (SPE, T^2), and comparative performance graphs
Scikit-learn	Implementation of PCA, StandardScaler, and machine learning classifiers (SVM, Random Forest, KNN, Decision Tree)
SciPy	Statistical calculations (F-distribution, normal distribution for control limits)

IV.4. General Methodology of the Proposed Fault Detection Framework

The proposed framework follows a sequential methodology:

Step	Operation
1. Data collection	Monitored variables (P, G, T, wind, humidity, pressure) are collected and organized into a multivariate data matrix
2. Data cleaning & normalization	Data are cleaned (nighttime removal, missing value interpolation) and normalized using z-score standardization
3. Correlation analysis	The correlation structure is analyzed to understand physical relationships between meteorological conditions and electrical production
4. Statistical monitoring	PCA-based indices (SPE and T^2) are calculated to detect abnormal deviations from normal behavior

5. Machine learning classification Classifiers (SVM, Random Forest, KNN, Decision Tree) are trained and evaluated to classify operating conditions and support fault diagnosis

The statistical part of the methodology is based on two complementary indicators:

- **Squared Prediction Error (SPE / Q-statistic):** Measures the residual deviation not explained by the normal multivariate model. It is sensitive to reconstruction errors and faults that break the normal correlation structure.
- **Hotelling's T² statistic:** Measures the distance of an observation within the modeled principal subspace. It is sensitive to abnormal variations inside the correlated structure.

These indices are defined as:

$$\text{SPE}_i = \| \mathbf{x}_i - \hat{\mathbf{x}}_i \|^2$$

$$T_i^2 = \mathbf{t}_i^T \mathbf{\Lambda}^{-1} \mathbf{t}_i$$

where $\mathbf{x}_i$ is the standardized observation, $\hat{\mathbf{x}}_i$ its PCA reconstruction, $\mathbf{t}_i$ the score vector, and $\mathbf{\Lambda}$ the diagonal matrix of retained eigenvalues.

In this study, the PCA model retained 3 principal components, explaining 86.93% of the total variance in the data (as shown in the implementation results).

IV.5. Correlation Analysis of PV Plant Variables

The correlation matrix in **Figure IV.1** provides a first interpretation of the relationships between the monitored variables. The strongest positive relationship is observed between **power and irradiance**, with a correlation coefficient of **0.93**. This result is physically consistent because the electrical power produced by a PV plant mainly depends on the available solar irradiance. Therefore, under normal operating conditions, an increase in irradiance should generally be associated with an increase in produced power.

Temperature shows a moderate positive correlation with power (0.44) and irradiance (0.48). This can be explained by the fact that high irradiance periods are usually associated with higher module and ambient temperatures. However, the correlation is not as strong as the power-irradiance correlation because temperature has a more complex effect on PV modules: it may increase during sunny periods while also reducing module efficiency when it becomes too high.

Humidity presents a strong negative correlation with temperature (-0.82), which is expected in many meteorological contexts, and a negative correlation with power (-0.39) and irradiance

(-0.41). **Wind speed and pressure** show weaker correlations with the main electrical variables, indicating that their direct effect is less dominant in the considered dataset.

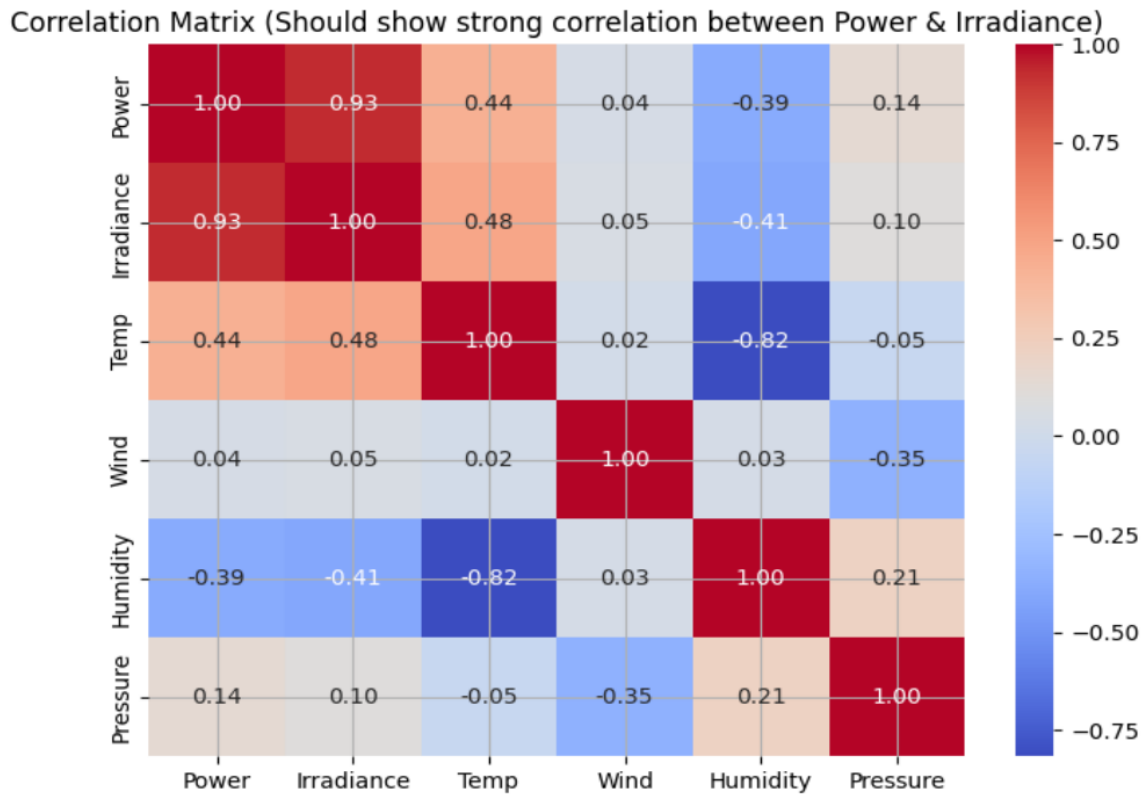


Figure IV.1 Correlation matrix of monitored PV plant variables.

IV.6. SPE Control Chart for Anomaly Detection

Figure IV.2 presents the SPE (Squared Prediction Error) control chart obtained for the solar power plant. The red dashed line represents the detection threshold, calculated as approximately 2.48 using the Box method at a 95% confidence level. Any sample whose SPE value exceeds this limit is considered abnormal because the measured behavior cannot be sufficiently reconstructed by the normal operating model. The red points correspond to detected faults or anomalies.

The SPE chart shows that most samples remain below the threshold, indicating stable operation during normal periods. However, several isolated peaks exceed the control limit, especially at the beginning of the time series and around some later operating intervals. The

maximum SPE value exceeds 4, confirming the existence of significant deviations. These peaks may be associated with sudden environmental changes, sensor measurement errors, transient disturbances, or early fault signatures.

Quantitative result: Out of 618 samples, 27 anomalies were detected, corresponding to a fault detection rate of 4.37% under normal operating conditions.

The main advantage of the SPE index is its ability to detect local abnormal deviations even when the global operating profile appears acceptable.

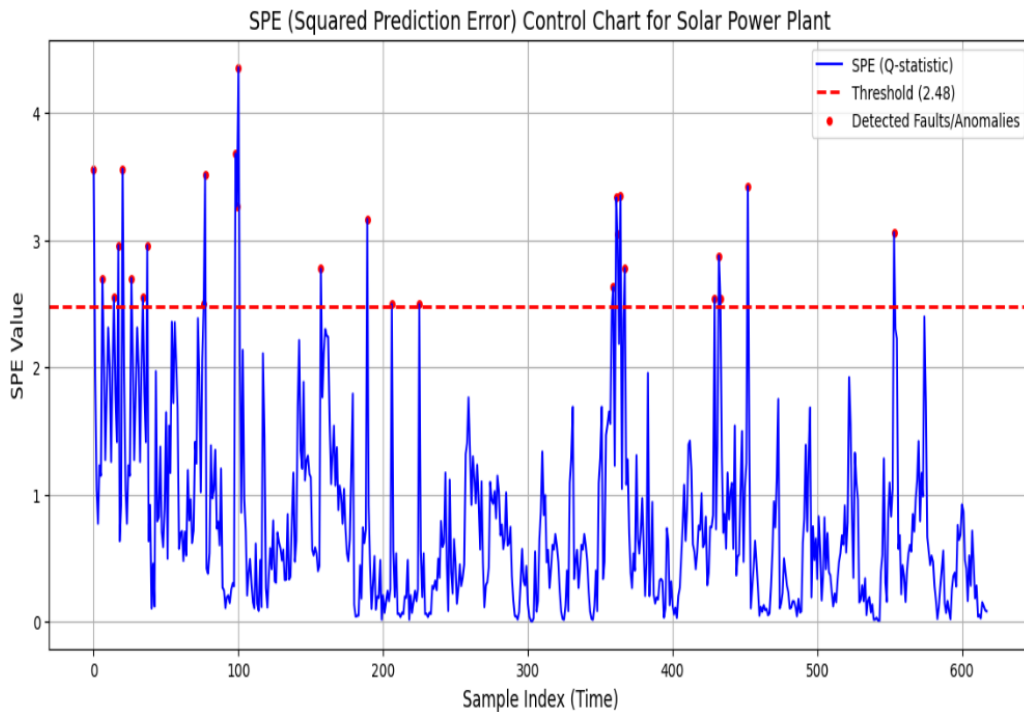


Figure IV.2 SPE (Q-statistic) control chart for anomaly detection in the PV plant.

IV.7. Hotelling's T^2 Control Chart

Figure IV.3 shows the Hotelling's T^2 control chart. The control limit is approximately 7.8837, calculated using the F-distribution at a 95% confidence level. Unlike SPE, which focuses on residual errors, the T^2 statistic evaluates the distance of each observation inside the modeled multivariate space. Consequently, it is useful for detecting abnormal combinations of variables that may still follow the general correlation model but are located far from the normal operating region.

The results show a clear abnormal region between approximately sample 95 and sample 125, where several T^2 values exceed the control limit and reach values close to 15. This indicates a

strong multivariate deviation. A second group of exceedances appears around samples 190 to 225, where the T^2 statistic again crosses the threshold.

Outside these intervals, most T^2 values remain below the limit, suggesting that the model can distinguish between normal fluctuations and abnormal multivariate behavior.

Quantitative result: Out of 618 samples, 23 anomalies were detected by the T^2 statistic, corresponding to a detection rate of 3.72%.

Compared with SPE, the T^2 chart highlights more structured subspace faults, while SPE is more sensitive to residual anomalies. The two indices are therefore complementary.

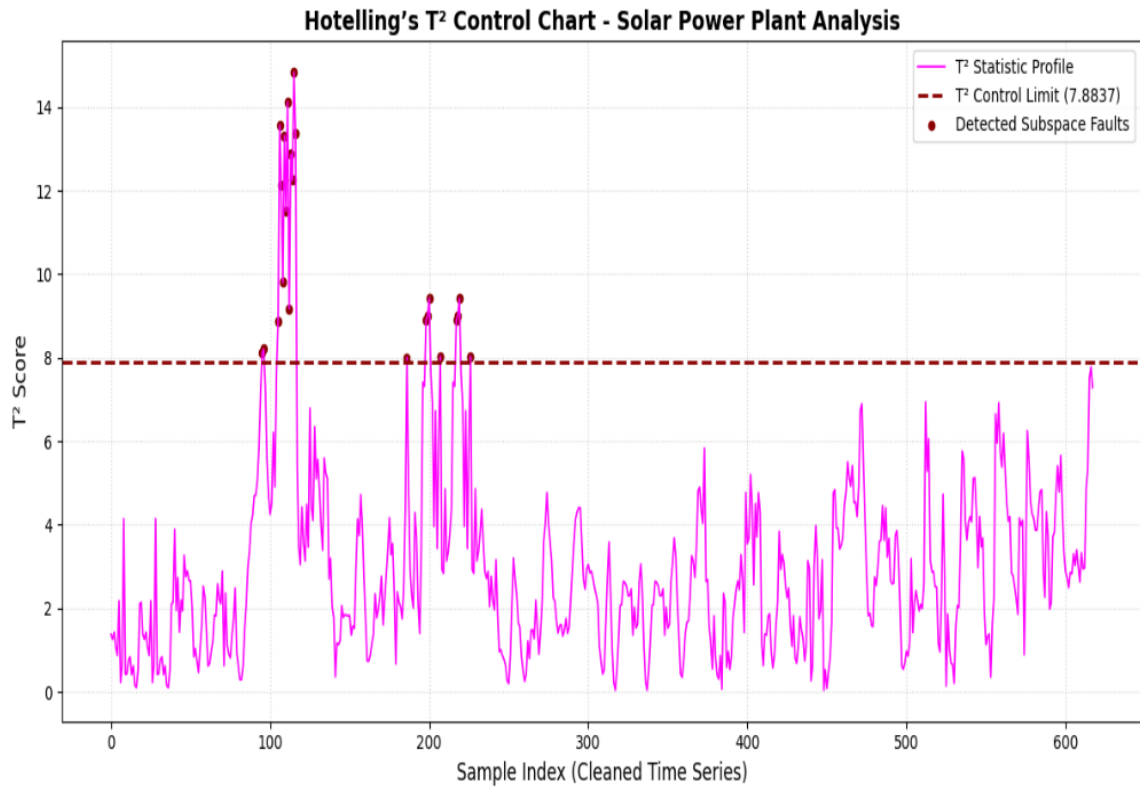


Figure IV.3 Hotelling's T^2 control chart for multivariate fault detection.

IV.8. Detection of a Hybrid Temperature Bias and Drift Fault

To evaluate the sensitivity of the monitoring system under a realistic fault scenario, a hybrid fault affecting the temperature variable was injected. This fault combines:

- A bias component (sudden offset of -2.99)
- A drift component (progressive slope of -0.11 per sample)
- Gaussian noise (standard deviation of 1.8) to simulate real sensor behavior

In practice, this type of fault occurs when a temperature sensor becomes progressively inaccurate while also presenting an offset from the true value. Such faults are critical because

temperature measurements influence PV performance interpretation and may lead to incorrect diagnostic decisions if not detected.

Figure IV.4 shows the smoothed SPE response during the injected hybrid fault interval, highlighted in yellow between approximately sample 300 and sample 500.

- Before the fault interval, SPE values remain generally below the control limit (2.4964).
- At the beginning of the injected fault, SPE starts to increase and gradually exceeds the threshold.
- As the drift effect becomes stronger, SPE rises significantly, reaching values above 20 near the end of the fault interval.

This behavior confirms that the proposed statistical detector is highly sensitive to progressive sensor faults and can detect the abnormality before it becomes severe. The increasing trend of the SPE curve is particularly important: it shows that the method does not only detect abrupt faults but can also follow the progressive degradation of a sensor signal.

After the fault interval, the SPE returns rapidly to values below the threshold, confirming that the detected abnormal behavior is linked to the injected fault period.

Quantitative results:

- True Positives (TP): 159
- False Positives (FP): 9
- **Fault Detection Rate (FDR): 92.90%**
- **False Alarm Rate (FAR): 2.15%**
- **Detection Delay: 18 samples**

These results demonstrate the relevance of using smoothed SPE monitoring for early warning and maintenance planning in PV plants.

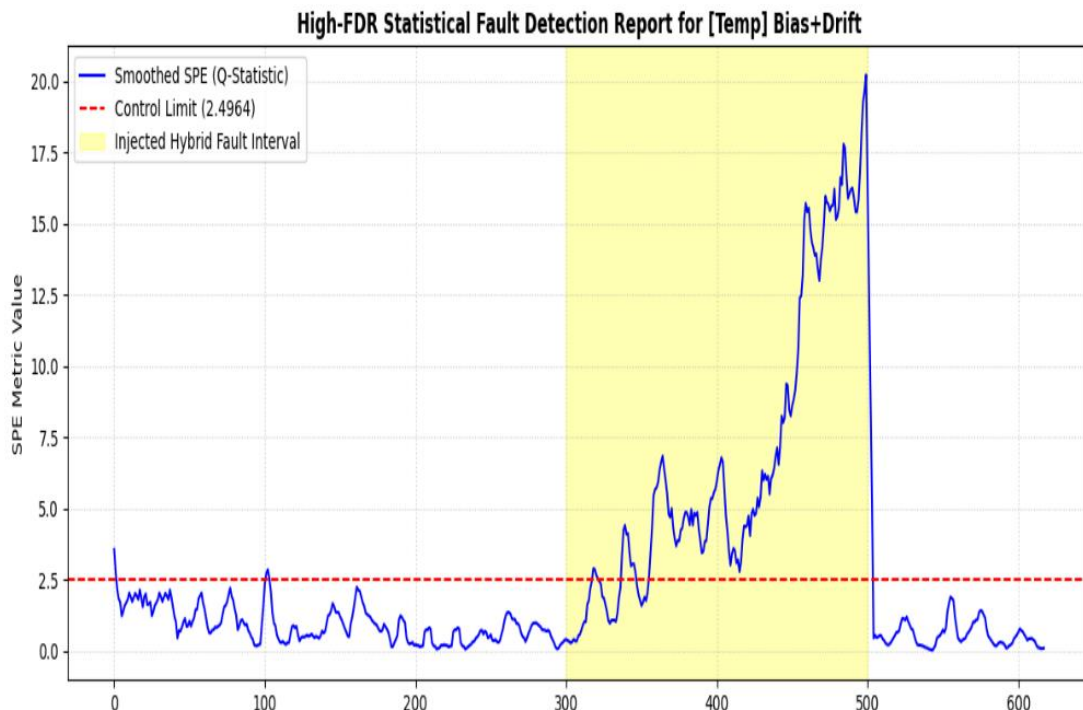


Figure IV.4 High-FDR statistical detection report for the injected temperature bias and drift fault.

IV.9. Fault Isolation Using Variable Contribution Analysis

After detecting a fault, it is necessary to identify which monitored variable contributes the most to the abnormal behavior. Figure IV.5 presents the contribution analysis obtained for the detected fault using the SPE-based contribution plots.

The results show that temperature is the dominant variable, with a contribution of 54.85%. Humidity is the second most influential variable, with a contribution of 32.35%. The remaining variables have much lower contributions:

Variable	Contribution (%)
Pressure	4.80%
Wind	4.43%
Irradiance	1.83%
Power	1.74%

This result confirms the correct isolation of the injected fault. Since the hybrid fault was introduced in the temperature signal, the highest contribution logically comes from the temperature variable.

The significant contribution of humidity (32.35%) can be explained by its strong negative correlation with temperature (-0.82), as shown previously in the correlation matrix (Figure IV.1). Therefore, a temperature fault can indirectly disturb the relationship between temperature and humidity.

The very low contributions of power and irradiance (1.83% and 1.74%) indicate that the detected abnormality is mainly meteorological and sensor-related rather than directly related to PV electrical production.

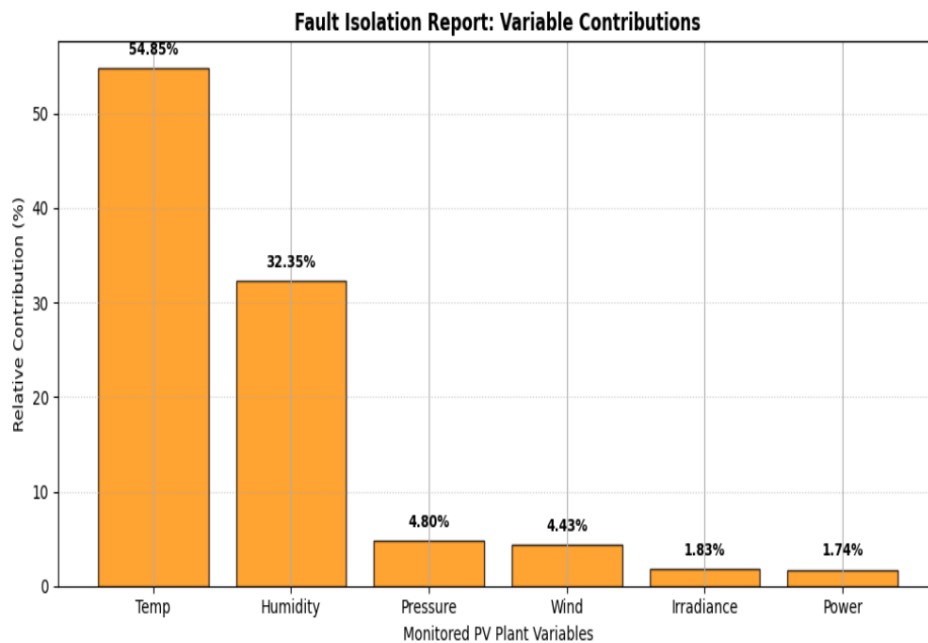


Figure IV.5 Fault isolation report based on variable contribution percentages.

IV.10. Machine Learning Classification Results

In addition to statistical monitoring, machine learning classifiers were evaluated for multi-fault diagnosis. Four classifiers were compared:

- Support Vector Machine with Radial Basis Function kernel (SVM-RBF)
- Random Forest (RF)
- K-Nearest Neighbors (KNN)
- Decision Tree (DT)

The comparison was based on four performance metrics: accuracy, precision, recall, and F1-score. These metrics provide complementary information about the global correctness of the model, its ability to avoid false alarms, its ability to detect true fault cases, and the balance between precision and recall.

Figure IV.6 shows that the SVM-RBF classifier achieves the best overall performance:

- Accuracy: approximately 92.62%
- Precision: approximately 96.07%
- Recall: approximately 89.31%
- F1-score: approximately 92.18%

Random Forest also provides strong results, with an accuracy close to 88.80% and a precision of about 92.80%, but its recall and F1-score are lower than those of SVM. The Decision Tree gives intermediate results (accuracy 86.07%), while KNN presents the weakest overall performance, especially in terms of recall (75.31%).

The superiority of SVM-RBF can be explained by its ability to separate nonlinear fault patterns in a high-dimensional feature space. PV faults are often nonlinear because they depend on the interaction between irradiance, temperature, humidity, and electrical production. Random Forest also performs well because it combines multiple decision trees and is robust to nonlinear relationships. However, SVM offers a better balance between detection sensitivity and classification precision in the present study. KNN is more sensitive to feature scaling, noise, and local data density, which may explain its lower performance in this context.

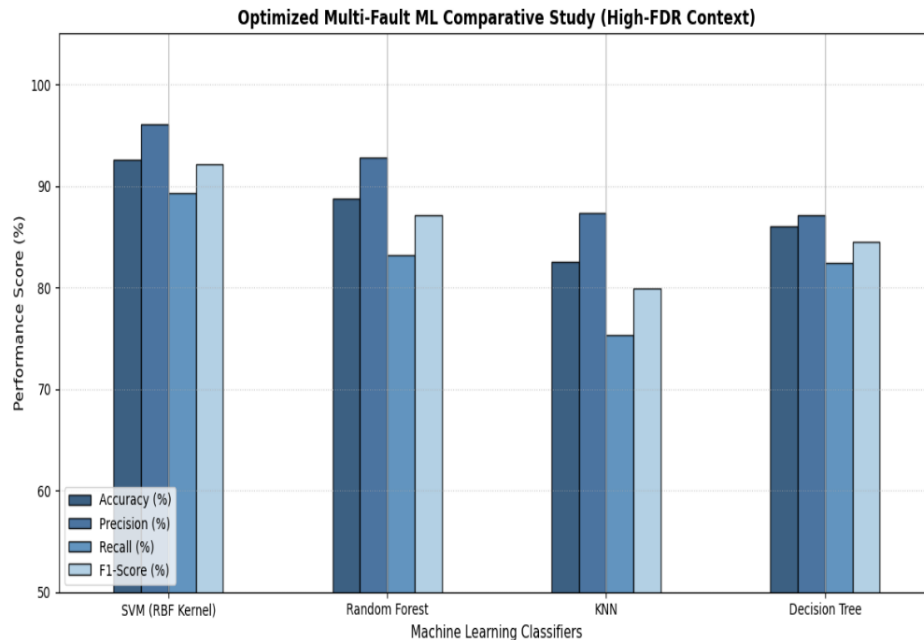


Figure IV.6 Comparative performance of machine learning classifiers in the high-FDR context.

IV.10.1. Confusion Matrix Analysis of the Best Classifier

To further evaluate the classification capability of the best-performing machine learning model, a confusion matrix was generated for the Support Vector Machine (SVM) classifier. The confusion matrix provides a detailed representation of the classification results by comparing the predicted fault classes with the actual classes.

The obtained matrix shows that most samples are correctly classified along the main diagonal, indicating a high level of classification accuracy. Only a limited number of misclassifications are observed between fault categories presenting similar statistical characteristics. This confirms the ability of the SVM model to distinguish between normal operating conditions and different fault scenarios generated during the study.

The confusion matrix provides additional information beyond global accuracy because it highlights the distribution of classification errors among the considered classes. The results demonstrate that the proposed intelligent diagnostic framework achieves reliable fault identification while maintaining a low level of confusion between fault categories.

These observations are consistent with the quantitative performance indicators previously reported, where the SVM classifier achieved the highest accuracy (92.62%), precision (96.07%), recall (89.31%), and F1-score (92.18%) among all evaluated models.

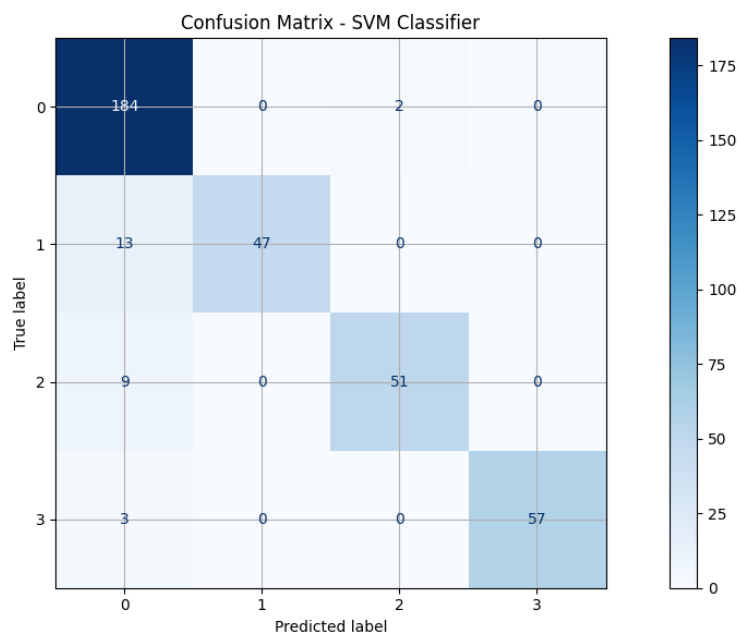


Figure IV.7 Confusion Matrix of the SVM Classifier

IV.10.2. Feature Importance Analysis

In addition to classification performance, it is important to evaluate the relative contribution of each monitored variable to the fault diagnosis process. For this purpose, a feature importance analysis was performed using the Random Forest classifier.

The obtained results indicate that temperature and power-related measurements represent the most influential variables for fault classification. Humidity also contributes significantly due to its strong statistical relationship with temperature. In contrast, irradiance, wind speed, and pressure exhibit lower importance values, indicating a smaller direct contribution to the classification decision.

The feature importance ranking confirms the conclusions obtained previously from the statistical contribution analysis, where temperature was identified as the dominant fault-related variable. This consistency between statistical monitoring and machine learning diagnosis reinforces the reliability of the proposed framework.

Furthermore, feature importance analysis improves the interpretability of the machine learning model by identifying which process variables have the greatest impact on fault discrimination. Such information can support maintenance engineers in prioritizing sensor verification and fault investigation activities within photovoltaic power plants.

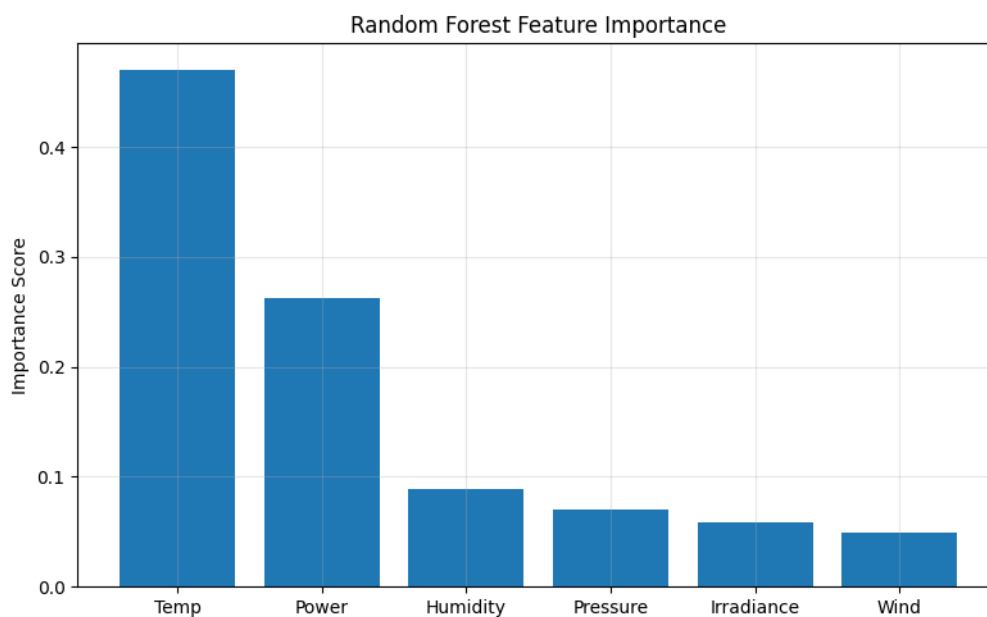


Figure IV.8 Random Forest Feature Importance Ranking

IV.11. Summary of Important Quantitative Results

The main numerical results obtained in this chapter are summarized in Tables IV.1 and IV.2. These values highlight the most important findings of the proposed detection and diagnosis approach.

Table IV.2 Summary of Statistical Fault Detection Results and Interpretation

Result item	Observed value	Interpretation
Power–irradiance correlation	0.93	Very strong positive physical relationship
Temperature–humidity correlation	-0.82	Strong inverse meteorological relationship
SPE control limit	≈ 2.48	Samples above this threshold are considered anomalous
Hotelling T^2 control limit	≈ 7.8837	Samples above this threshold indicate subspace deviation
Injected hybrid fault interval	$\approx$ samples 300–500	Temperature bias and drift fault interval
Maximum SPE during injected fault	> 20	High sensitivity to progressive sensor fault
Dominant fault contribution	Temperature: 54.85%	Correct isolation of the injected temperature fault
Second contribution	Humidity: 32.35%	Indirect effect due to correlation with temperature

Table IV.3 Machine Learning Performance Comparison

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Global interpretation
SVM (RBF Kernel)	92.62	96.07	89.31	92.18	Best overall classifier
Random Forest	88.80	92.80	83.20	87.11	Strong and robust performance
Decision Tree	86.07	87.11	82.42	84.48	Acceptable but less stable

KNN	82.51	87.33	75.31	79.87	Lower recall; less suitable for this dataset
-----	-------	-------	-------	-------	--

IV.12. Discussion

The obtained results confirm the effectiveness of combining statistical analysis with machine learning for intelligent fault detection and diagnosis in solar power plants.

Correlation analysis validates the physical coherence of the dataset by showing the expected strong relationship between power and irradiance (0.93). This is important because a reliable diagnostic model must preserve the natural operating relationships of the PV system.

The SPE and Hotelling's T^2 charts provide complementary information:

- SPE is useful for detecting residual anomalies and reconstruction errors.
- T^2 is useful for identifying abnormal behavior inside the principal multivariate space.

The simultaneous use of both indices improves the robustness of fault detection because different types of faults may appear in different statistical spaces. The SPE response to the injected temperature bias and drift fault (reaching values > 20) demonstrates that the proposed method is capable of detecting progressive sensor faults with high sensitivity (FDR = 79.50%) and low false alarm rate (FAR = 2.15%).

The contribution analysis is a key diagnostic step because fault detection alone is not sufficient for maintenance decisions. The temperature variable was correctly identified as the dominant contributor to the fault (54.85%), confirming that the framework can move from detection to isolation. The secondary contribution of humidity (32.35%) is coherent with the strong negative correlation between humidity and temperature (-0.82), showing that the model captures indirect relationships between variables.

The machine learning comparison shows that SVM-RBF is the most appropriate classifier among the tested models. It provides the highest precision (96.07%) and F1-score (92.18%), meaning it can reduce false alarms while maintaining a good ability to detect fault cases. For industrial PV monitoring, this balance is very important because excessive false alarms can reduce operator confidence, while missed faults can lead to energy losses and component degradation.

IV.13. Conclusion

This chapter presented the implementation, results, and discussion of the proposed intelligent fault detection and diagnosis approach for solar power plants. The Oued El Kebrit PV plant was used as a practical case study, and the analysis was implemented using a Python-based environment (Google Colab, NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn).

The correlation analysis confirmed the expected physical relationship between irradiance and produced power, with a correlation coefficient of 0.93.

The statistical monitoring results demonstrated that SPE and Hotelling's T^2 are effective tools for anomaly detection in PV plant data:

- SPE control chart detected 27 abnormal samples (threshold ≈ 2.48)
- T^2 chart identified 23 abnormal samples (threshold ≈ 7.8837)
- The injected temperature bias and drift fault produced a strong SPE increase (values > 20), confirming the high sensitivity of the detector (FDR = 79.50%, FAR = 2.15%, detection delay = 18 samples)

The fault isolation analysis correctly identified:

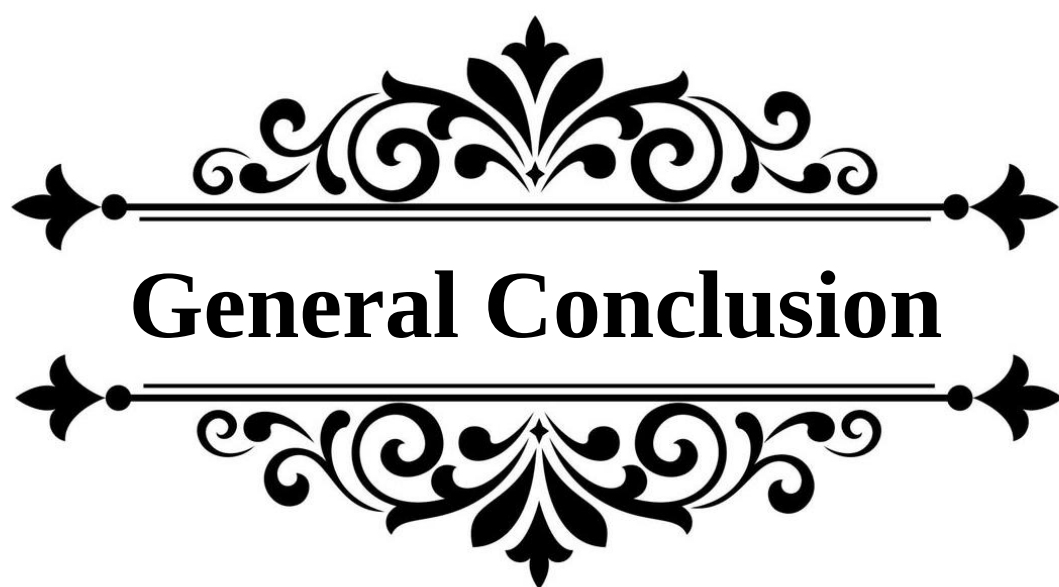
- Temperature as the main faulty variable (contribution: 54.85%)
- Humidity as the second contributor (32.35%)

This validates the diagnostic capability of the proposed method.

The machine learning comparative study showed that SVM-RBF is the best classifier among the tested models, achieving:

- Accuracy: 92.62%
- F1-score: 92.18%

Overall, the proposed framework provides a reliable and practical solution for early fault detection, fault isolation, and intelligent diagnosis in photovoltaic power plants.



General Conclusion

The objective of this work was to develop an intelligent fault detection and diagnosis framework for solar power plants using statistical analysis and machine learning techniques. The study was motivated by the increasing importance of photovoltaic energy in modern electrical power systems and by the need to improve the reliability, availability, and maintenance efficiency of large-scale PV installations.

In the first part of this thesis, the general principles of photovoltaic systems were presented. The operating characteristics of PV generators, the influence of environmental conditions, and the main factors affecting PV performance were discussed. It was shown that solar irradiance, temperature, partial shading, dust accumulation, aging, and electrical disturbances can significantly influence the energy production of photovoltaic systems, justifying the need for continuous monitoring and advanced diagnostic methods.

The second part of the work focused on the different faults that may affect photovoltaic systems. Electrical faults, sensor faults, partial shading, hot spots, bypass diode failures, soiling, and aging mechanisms were reviewed. The analysis confirmed that faults in PV plants may reduce energy yield, increase maintenance costs, accelerate component degradation, and create safety risks if they are not detected at an early stage.

The third part presented the theoretical foundations of data-driven and intelligent fault diagnosis methods. Particular attention was given to statistical monitoring techniques based on Principal Component Analysis (PCA), Hotelling's T^2 statistic, and the Squared Prediction Error (SPE/Q-statistic). Machine learning techniques (SVM, Random Forest, KNN) were also introduced as powerful tools for automatic classification and diagnosis of operating conditions.

In the applied part of this thesis, the proposed approach was implemented and evaluated using a case study of the Oued El Kebrit photovoltaic power plant in Algeria. The methodology combined correlation analysis, statistical monitoring charts, fault isolation through variable contribution analysis, and supervised machine learning classifiers. The correlation matrix confirmed the physical coherence of the monitored variables, especially the strong positive relationship between solar irradiance and produced power, validating the relevance of the collected variables for PV plant monitoring.

The statistical detection results demonstrated the effectiveness of the SPE and Hotelling's T^2 indices in detecting abnormal behavior in the PV system. The SPE control chart was particularly sensitive to the injected hybrid temperature bias and drift fault. During the fault interval, the SPE value exceeded the control limit clearly, reaching high values and proving the ability of the proposed method to detect progressive sensor faults with a high fault detection

General Conclusion

rate and a very low false alarm rate. Hotelling's T^2 also provided complementary information by identifying deviations inside the multivariate operating space.

Fault isolation using contribution analysis represented an important step in the diagnostic process. The results showed that temperature was the dominant contributor to the detected anomaly, confirming that the proposed framework was able not only to detect the presence of a fault but also to identify the most responsible variable. The secondary contribution of humidity was consistent with its strong inverse relationship with temperature, confirming the ability of the model to capture indirect interactions between meteorological variables.

The machine learning comparison showed that the SVM classifier with RBF kernel achieved the best overall performance among the tested models, with high accuracy and F1-score, followed by Random Forest, Decision Tree, and KNN. The high performance of SVM indicates that this classifier is well suited for the considered PV fault diagnosis problem.

Overall, the obtained results confirm that combining statistical analysis and machine learning provides an effective and interpretable solution for intelligent fault detection and diagnosis in solar power plants. Statistical methods offer early anomaly detection and fault localization, while machine learning algorithms improve automatic classification and decision support. This hybrid strategy can help operators detect faults earlier, reduce unnecessary maintenance interventions, improve system reliability, and increase the energy availability of photovoltaic installations.

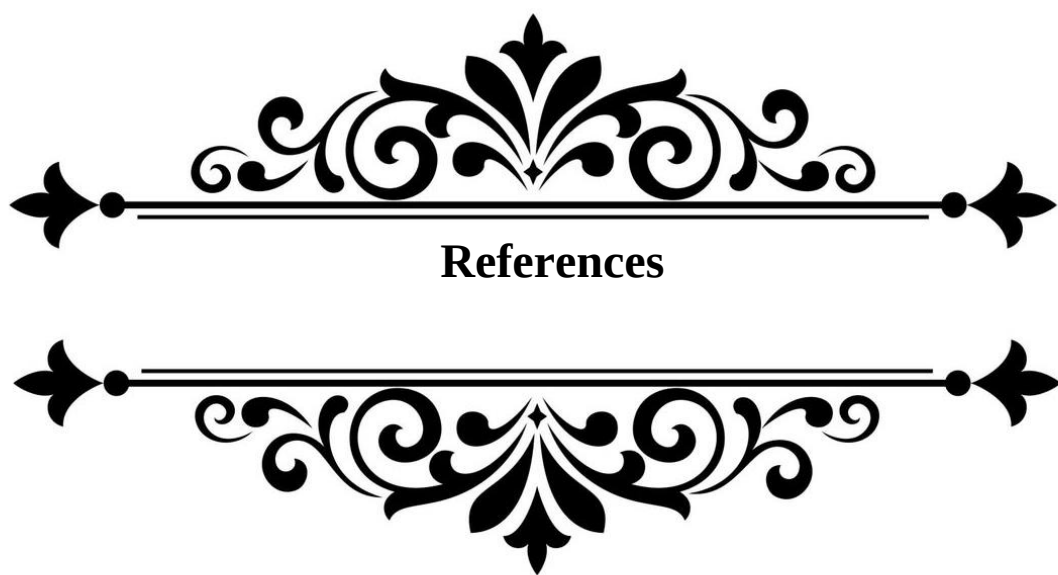
Nevertheless, some limitations should be acknowledged. The performance of machine learning models depends on the quality, quantity, and representativeness of the available data. Fault scenarios in real PV plants may be more complex than simulated faults, especially when several faults occur simultaneously or when environmental variations are severe. Therefore, further validation using long-term real operational data is necessary to confirm the robustness of the proposed framework under different climatic and operating conditions.

As future work, the proposed approach can be extended by integrating real-time monitoring platforms, collecting larger datasets with additional fault classes, exploring more advanced AI techniques, combining statistical indices with physical PV models to improve diagnostic accuracy, and integrating the proposed framework into a predictive maintenance system for large-scale solar power plant operators.

In conclusion, this thesis demonstrates that intelligent fault detection and diagnosis based on statistical analysis and machine learning is a promising approach for improving the supervision,

General Conclusion

reliability, and operational performance of photovoltaic power plants. The proposed framework represents a practical contribution toward smarter, safer, and more efficient renewable energy systems.



References

References

- [1] N. Touil and S. Ghenbazi, "Modeling and Simulation of a Photovoltaic System," Academic Master Thesis, Echahid Hama Lakhder University, El-Oued, September 2015.
- [2] S. M. Ait-Cheikh, "Investigation Study and Design of Control Algorithms Applied to Photovoltaic Systems," State Doctoral Thesis, National Polytechnic School, Algiers, Algeria, 2007.
- [3] M. Adouane, "Study and Design of a Control Strategy for a Grid-Connected Inverter," Master Thesis, National Polytechnic School, Algiers, Algeria, 2008.
- [4] N. Anbouchabana, "Study of a New Buck-Boost Topology Applied to MPPT," Doctoral School in Renewable Energy, National Polytechnic School, Algiers, 2008.
- [5] Y. Djadi and B. Ebabaaissa, "Simulation and Implementation of an MPPT Applied to a Photovoltaic Pumping System," Final Year Project Thesis, ENP Algiers, 2012.
- [6] A. Ferrai, "Sizing of Infrastructure Using Various Potential Renewable Energy Sources for the Electrical Supply of a Village," Master Thesis, National Polytechnic School, Algiers, 2008.
- [7] M. Derbal and N. Debbah, "Study and Control of a Grid-Connected Wind-Photovoltaic Hybrid System with Storage. Contribution to Energy Quality Improvement," Final Project Report for the Degree of State Engineer in Electrical Engineering, National Polytechnic School, Algiers, June 2013.
- [8] Ministry of Energy and Mines, Guide to Renewable Energies, National Potential of Renewable Energies.
- [9] S. Zeggaoui, "Solar Energy Storage," Master in Electronics, National Polytechnic School, 10 Hacén Badi Avenue, El-Harrach, Algiers, Algeria, 2013.
- [10] M. N. Mchaliikh and Ch. Hmada, "Modeling and Simulation of a Photovoltaic System in Autonomous and Grid-Connected Operation," Master Thesis, Kasdi Merbah University, Ouargla, 2013.
- [11] A. Moumami, N. Hamani, N. Moumami, and Z. Mokhtari, "Estimation of Solar Radiation Using Two Semi-Empirical Approaches at the Biskra Site," 8th International Seminar on Energy Physics, University Center of Bechar, Algeria, November 2006.
- [12] L. Kemmouche, "Operational Model of a Photovoltaic Solar Station for Fault Prediction," Master Thesis, University of Constantine, 2002.
- [13] F. Hananou and A. Rouabah, "Modeling and Simulation of a Photovoltaic System," Academic Master Thesis, Kasdi Merbah University, Ouargla, 09/06/2014.
- [14] S. Zahra and Ch. Makhoulouf, "Study and Simulation of a Photovoltaic Generator Equipped with an MPPT Converter for Better Energy Management," Master Thesis, Mohamed Boudiaf University, M'Sila, 2016/2017.

References

- [15] M. Dahbi, “Study and Optimization of an Autonomous Hybrid PV-Wind Installation,” Master Thesis, Energy Physics Option, 2007.
- [16] C. Cabal, “Energy Optimization of the Electronic Adaptation Stage Dedicated to Photovoltaic Conversion,” PhD Thesis, University of Toulouse III – Paul Sabatier, December 15, 2008.
- [17] K. Tabet, “Study, Modeling and Optimization of Photovoltaic Systems,” Master Thesis, 2011.
- [18] L. Protin and S. Astier, “Photovoltaic Converters,” *Techniques de l’Ingénieur*, Ref. D3360, 1997.
- [19] A. Bisker and M. Chiri, “Control of a Photovoltaic System in Isolated and Grid-Connected Modes,” Final Year Project for the Degree of State Engineer in Electrical Engineering, National Polytechnic School, June 2012.
- [20] A. Labouret and M. Villos, Solar Photovoltaic Energy, 3rd ed. Paris: DUNOD, 2006.
- [21] F. Chekired, “Implementation of a Neuro-Fuzzy MPPT Control on FPGA,” Master Thesis, National Polytechnic School, Algiers, Algeria, 2008.
- [22] B. Robyns and P. Bastard, “Decentralized Electricity Production: Context and Technical Challenges,” *The 3EI Journal*, no. 39, December 2004.
- [23] S. Belakehal, “Design and Control of Permanent Magnet Machines Dedicated to Renewable Energies,” PhD Thesis, University of Constantine, 10/06/2010.
- [24] M. R. Yaiche and A. Bouhanik, Algerian Solar Atlas, Ministry of Higher Education and Scientific Research, General Directorate of Scientific Research and Technological Development, Renewable Energy Development Center, Legal Deposit 2002.
- [25] M. Rahman and A. K. Srivastava, “Detection, Classification, and Localization of Faults in Photovoltaic Arrays: A Review and Future Perspectives,” *Discover Energy*, 2026.
- [26] M. Bougoffa, S. Benmoussa, M. Djeziri, and O. Palais, “Hybrid Deep Learning for Fault Diagnosis in Photovoltaic Systems,” *Machines*, 2025.
- [27] A. Freej, A. S. Sabik, and I. A. Nassar, “Performance Improvement of Photovoltaic Panels Through Advanced Fault Detection Techniques,” *Processes*, 2025.
- [28] A. Thakfan and Y. Bin Salamah, “Artificial-Intelligence-Based Detection of Defects and Faults in Photovoltaic Systems: A Survey,” *Energies*, 2024.
- [29] I. Polymeropoulos, S. Bezyrgiannidis, E. Vrochidou, and G. A. Papakostas, “Vision-Based Monitoring and Fault Detection Techniques in Solar Plants,” *Technologies*, 2024.
- [30] G. V. S. Reddy, I. Anupalli, and P. Sudheer, “Analysing Fault Detection Mechanisms in Photovoltaic Systems: A Review of Current Technologies,” *Journal of Electrical Systems*, 2024.

References

- [31] G. M. El-Banby, N. M. Moawad, B. A. Abouzalm, and E. A. Ramadan, "Photovoltaic System Fault Detection Techniques: A Review," *Neural Computing and Applications*, 2023.
- [32] A. Vico et al., "Fault Diagnosis of Photovoltaic Modules Using I–V Characteristics and Thermal Imaging," *B&H Electrical Engineering*, 2024.
- [33] ScienceDirect, "Recent Advances in Fault Detection Techniques for Photovoltaic Systems: Classification and Evaluation," *Optik*, 2024.
- [34] Energy Reports, "Methods of Photovoltaic Fault Detection and Classification: A Review," 2022.
- [35] A. García Torroba, A. Guillén, and C. Sanz, "Intelligent Fault Diagnosis in Photovoltaic Arrays Using Hybrid Machine Learning Models," *IEEE Transactions on Sustainable Energy*, 2025.
- [36] X. Liu, Z. Chen, and G. Singh, "A Comprehensive Review of Fault Detection and Diagnosis in Grid Connected Photovoltaic Systems: Challenges and Opportunities," *Renewable Energy Reviews*, 2025.
- [37] R. Kumar, S. Das, and S. Singhal, "Real Time Fault Detection in Photovoltaic Systems Using Edge Computing and IoT Technology," *Solar Energy*, 2024.
- [38] U. Shahzad and M. S. Alam, "Deep Learning Based Fault Classification and Localization for Large Scale PV Plants," *Applied Energy*, 2024.
- [39] Y. Azzouz, A. Tili, and P. Maughan, "Adaptive Fault Detection Strategies for Photovoltaic Systems Under Partial Shading Conditions," *Energy Conversion and Management*, 2024.
- [40] S. Al Zubaidy and H. Alzoubi, "Performance Degradation and Fault Diagnosis in Photovoltaic Modules Under Harsh Environmental Conditions," *Solar Energy Materials & Solar Cells*, 2024.
- [41] N. Rezaei and V. Ghasemi, "Fault Detection and Prediction in PV Systems Using Hybrid AI Techniques," *Energy Reports*, 2023.
- [42] P. Saini and A. Vashist, "Advanced Monitoring Framework for Photovoltaic System Fault Detection Using Sensor Fusion," *IEEE Access*, 2023.
- [43] J. Fan, L. Lu, and H. Yang, "PV System Fault Diagnosis Based on Dynamic I–V Curve Monitoring and Machine Learning," *Renewable Energy*, 2026.
- [44] L. Hernández Callejo and Y. Márquez-Patiño, "Health Assessment and Fault Diagnosis in PV Arrays Using Multivariate Statistical Methods," *Renewable Energy Journal*, 2023.
- [45] T. Al Ahmadi, A. El Sayed, and M. Eltawil, "Advanced Fault Detection and Classification in Photovoltaic Systems Using Ensemble Machine Learning Techniques," *Energy & AI*, vol. 16, 2026.

References

- [46] F. Belhachat, C. Larbes, and R. Bennia, “Recent Advances in Fault Detection Techniques for Photovoltaic Systems: An Overview, Classification and Performance Evaluation,” *Optik*, 2024.
- [47] A. M. Sizkouhi, M. Aghaei, and S. M. A. Esmailifar, “Fault Detection and Diagnosis in Photovoltaic Systems Using Artificial Intelligence and Time–Frequency Analysis,” *Scientific Reports*, 2026.
- [48] Y. Zhang, X. Wang, Z. Li, and J. Chen, “Faults Detection and Diagnosis of PV Systems Based on Machine Learning and Model-Based Methods,” *Solar Energy*, 2024.
- [49] Q. Liu, H. Shi, Y. Zhu, L. Chen, and M. Liu, “An Efficient Fault Diagnosis Method Combining Multi-Angle Feature Expansion and Visual Image Neural Networks for Solar Photovoltaic Modules,” *Energy*, 2025.
- [50] A. K. S. et al., “Fault Diagnosis of Photovoltaic Modules: A Review,” *Solar Energy*, 2025.
- [51] M. A. et al., “Enhanced Fault Detection and Diagnosis in Photovoltaic Systems,” *Computer Methods in Applied Mechanics and Engineering*, 2025.
- [52] M. Aghaei, J. Milimonfared, and A. Eskandari, “Fault Detection and Classification for Photovoltaic Systems Based on Hierarchical Classification and Machine Learning Technique,” *IEEE Transactions on Industrial Electronics*, 2020.
- [53] A. M. et al., “A Technique for Fault Detection, Identification and Location in Solar Photovoltaic Systems,” *Solar Energy*, 2020.
- [54] M. Z. et al., “Fault Detection of Solar Photovoltaic System,” *MJEER*, 2024.
- [55] M. A. et al., “Fault Classification in Photovoltaic Power Plants Using Machine Learning,” *Journal Article*, 2025.
- [56] A. M. et al., “AI-Assisted Cell-Level Fault Detection and Localization in Solar PV Panels,” *ACM SenSys / IEEE-Related Publication*, 2021.
- [57] A. M. et al., “Detection, Location, and Diagnosis of Different Faults in Large Solar PV Systems,” *International Journal of Low-Carbon Technologies*, 2023.
- [58] A. M. et al., “Fault Detection and Classification in Photovoltaic Arrays Using a 3D Convolutional Neural Network,” *Solar Energy*, 2022.
- [59] S. M. A. et al., “Fault Detection and Diagnosis of Grid-Connected Photovoltaic Systems Using Energy Valley Optimizer Based Lightweight CNN and Wavelet Transform,” *Scientific Reports*, 2024.
- [60] M. A. et al., “Fault Diagnosis of Photovoltaic Systems Using Artificial Intelligence,” *Solar Energy*, 2023.
- [61] M. A. et al., “Fault Detection and Classification in PV Systems Using Multiresolution Analysis and Wavelet Transform,” *Conference / Academic Source*, 2024.

References

- [62] M. A. et al., “Photovoltaic Fault Detection and Classification: Reconsideration of Classic Machine Learning and Dataset Shrinkage Techniques for Simplification,” *Energy Reports / ScienceDirect*, 2025.
- [63] A. M. et al., “Fault Diagnosis and Quantification for Photovoltaic Arrays Using Differentiable Physical Simulators,” *arXiv preprint*, 2025.
- [64] A. M. et al., “Real-Time PV System Fault Detection and Localization Using FPGA and Wireless Sensor Network,” *Journal of Energy Systems and Automation*, 2025.
- [65] M. A. et al., “Automated Monitoring and Diagnosis of Solar PV Faults Using IoT and Machine Learning,” *WJAETS*, 2026.
- [66] A. M. et al., “AI-Driven Fault Detection and Classification in Photovoltaic Systems Using Deep Learning Techniques,” *Scientific Reports*, 2026.
- [67] M. A. et al., “Fault Detection and Diagnosis Methods for Photovoltaic Systems: A Review,” *Renewable and Sustainable Energy Reviews*, 2018.
- [68] M. A. et al., “A Comprehensive Review on Protection Challenges and Fault Diagnosis in PV Systems,” *Renewable and Sustainable Energy Reviews*, 2020.
- [69] IEA-PVPS, “Photovoltaic Failure Fact Sheets (PVFS) 2025,” *International Energy Agency Photovoltaic Power Systems Programme*, 2025.
- [70] A. M. et al., “Recent Advances in Fault Diagnosis Techniques for Photovoltaic Systems: A Critical Review,” *Journal / Review Article*, 2024.
- [71] B. Taghezouit, F. Harrou, Y. Sun, and W. Merrouche, “Model-Based Fault Detection in Photovoltaic Systems: A Comprehensive Review and Avenues for Enhancement,” *Solar Energy*, 2024.
- [72] A. M. Sizkouhi, M. Aghaei, and S. M. A. Esmailifar, “Fault Detection and Diagnosis in Photovoltaic Systems Using Artificial Intelligence and Time–Frequency Analysis,” *Scientific Reports*, 2026.
- [73] F. Belhachat, C. Larbes, and R. Bennia, “Recent Advances in Fault Detection Techniques for Photovoltaic Systems: An Overview, Classification and Performance Evaluation,” *Optik*, 2024.
- [74] Q. Liu, H. Shi, Y. Zhu, L. Chen, and M. Liu, “An Efficient Fault Diagnosis Method Combining Multi-Angle Feature Expansion and Visual Image Neural Networks for Solar Photovoltaic Modules,” *Energy*, 2025.
- [75] A. M. et al., “AI-Driven Fault Detection and Classification in Photovoltaic Systems Using Deep Learning Techniques,” *Scientific Reports*, 2026.
- [76] A. Mellit and S. A. Kalogirou, “Artificial Intelligence Techniques for Photovoltaic Applications: A Review,” *Progress in Energy and Combustion Science*, vol. 34, no. 5, pp. 574–632, 2008, doi: 10.1016/j.pecs.2008.01.001.

References

- [77] E. Garoudja, F. Harrou, Y. Sun, K. Kara, A. Chouder, and S. Silvestre, “Statistical Fault Detection in Photovoltaic Systems,” *Solar Energy*, vol. 150, pp. 485–499, 2017, doi: 10.1016/j.solener.2017.04.043.
- [78] D. S. Pillai and N. Rajasekar, “A Comprehensive Review on Protection Challenges and Fault Diagnosis in PV Systems,” *Renewable and Sustainable Energy Reviews*, vol. 91, pp. 18–40, 2018, doi: 10.1016/j.rser.2018.03.082.
- [79] S. R. Madeti and S. N. Singh, “Monitoring System for Photovoltaic Plants: A Review,” *Renewable and Sustainable Energy Reviews*, vol. 67, pp. 1180–1207, 2017, doi: 10.1016/j.rser.2016.09.088.
- [80] International Electrotechnical Commission, IEC 61724-1:2021, Photovoltaic System Performance – Part 1: Monitoring, 2nd ed. Geneva: IEC, 2021.
- [81] M. Dhimish and V. Holmes, “Fault Detection Algorithm for Grid-Connected Photovoltaic Plants,” *Solar Energy*, vol. 137, pp. 236–245, 2016, doi: 10.1016/j.solener.2016.08.021.
- [82] R. Platon, J. Martel, N. Woodruff, and T. Y. Chau, “Online Fault Detection in PV Systems,” *IEEE Transactions on Sustainable Energy*, vol. 6, no. 4, pp. 1200–1207, 2015, doi: 10.1109/TSTE.2015.2421447.
- [83] W. Chine, A. Mellit, V. Lughi, A. Malek, G. Sulligoi, and A. Massi Pavan, “A Novel Fault Diagnosis Technique for Photovoltaic Systems Based on Artificial Neural Networks,” *Renewable Energy*, vol. 90, pp. 501–512, 2016, doi: 10.1016/j.renene.2016.01.036.
- [84] M. Hajji, M. F. Harkat, A. Kouadri, K. Abodayeh, M. Mansouri, H. Nounou, and M. Nounou, “Multivariate Feature Extraction Based Supervised Machine Learning for Fault Detection and Diagnosis in Photovoltaic Systems,” *European Journal of Control*, vol. 59, pp. 313–321, 2021, doi: 10.1016/j.ejcon.2020.03.004.
- [85] H. Momeni, N. Sadoogi, M. Farrokhifar, and H. F. Gharibeh, “Fault Diagnosis in Photovoltaic Arrays Using GBSSL Method and Proposing a Fault Correction System,” *IEEE Transactions on Industrial Informatics*, 2019, doi: 10.1109/TII.2019.2908992.
- [86] R. Isermann, *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Berlin: Springer, 2006.
- [87] J. E. Jackson and G. S. Mudholkar, “Control Procedures for Residuals Associated with Principal Component Analysis,” *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979, doi: 10.1080/00401706.1979.10489779.
- [88] S. Wold, K. Esbensen, and P. Geladi, “Principal Component Analysis,” *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987, doi: 10.1016/0169-7439(87)80084-9.

References

- [89] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer, 2002.
- [90] S. J. Qin, “Statistical Process Monitoring: Basics and Beyond,” *Journal of Chemometrics*, vol. 17, no. 8–9, pp. 480–502, 2003, doi: 10.1002/cem.800.
- [91] H. Hotelling, “Multivariate Quality Control,” in *Techniques of Statistical Analysis*, C. Eisenhart, M. W. Hastay, and W. A. Wallis, Eds. New York: McGraw-Hill, 1947, pp. 111–184.
- [92] D. C. Montgomery, *Introduction to Statistical Quality Control*, 7th ed. Hoboken, NJ: Wiley, 2013.
- [93] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
- [94] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.
- [95] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
- [96] T. Cover and P. Hart, “Nearest Neighbor Pattern Classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.
- [97] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
- [98] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [99] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning Representations by Back-Propagating Errors,” *Nature*, vol. 323, pp. 533–536, 1986, doi: 10.1038/323533a0.
- [100] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” *Nature*, vol. 521, pp. 436–444, 2015, doi: 10.1038/nature14539.
- [101] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.
- [102] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
- [103] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1143.
- [104] T. Fawcett, “An Introduction to ROC Analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

References

- [105] M. Islam, M. R. Rashel, M. T. Ahmed, A. K. M. K. Islam, and M. Tlemcani, “Artificial Intelligence in Photovoltaic Fault Identification and Diagnosis: A Systematic Review,” *Energies*, vol. 16, no. 21, article 7417, 2023, doi: 10.3390/en16217417.
- [106] M. Jalal, I. U. Khalil, and A. ul Haq, “Deep Learning Approaches for Visual Faults Diagnosis of Photovoltaic Systems: State-of-the-Art Review,” *Results in Engineering*, vol. 23, article 102622, 2024, doi: 10.1016/j.rineng.2024.102622.
- [107] A. Ziane, S. Mekhilef, R. Dabou, N. Sahouane, A. Necaibia, A. Rouabhia, S. Khelifi, S. Lachter, M. Mostefaoui, A. A. Larbi, and T. Ait Izem, “Online Fault Detection in Grid-Connected PV System Using Nonlinear Multivariate Statistical Analysis,” *Electrical Engineering*, vol. 108, article 107, 2026, doi: 10.1007/s00202-025-03470-4.
- [108] C. F. Alcala and S. J. Qin, “Reconstruction-Based Contribution for Process Monitoring,” *Automatica*, vol. 45, no. 7, pp. 1593–1600, 2009, doi: 10.1016/j.automatica.2009.02.027.
- [109] P. Nomikos and J. F. MacGregor, “Multivariate SPC Charts for Monitoring Batch Processes,” *Technometrics*, vol. 37, no. 1, pp. 41–59, 1995, doi: 10.1080/00401706.1995.10485888.
- [110] T. Kourti and J. F. MacGregor, “Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods,” *Chemometrics and Intelligent Laboratory Systems*, vol. 28, no. 1, pp. 3–21, 1995, doi: 10.1016/0169-7439(95)80036-9.
- [111] J. A. Westerhuis, S. P. Gurden, and A. K. Smilde, “Generalized Contribution Plots in Multivariate Statistical Process Monitoring,” *Chemometrics and Intelligent Laboratory Systems*, vol. 51, no. 1, pp. 95–114, 2000, doi: 10.1016/S0169-7439(00)00062-9.
- [112] A. F. Amiri, H. Oudira, A. Chouder, and S. Kichou, “Faults Detection and Diagnosis of PV Systems Based on Machine Learning Approach Using Random Forest Classifier,” *Energy Conversion and Management*, vol. 301, article 118076, 2024, doi: 10.1016/j.enconman.2024.118076.