الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research

Faculty of Sciences and Technology
Electrical Engineering Department

Thesis

Presented to obtain the

Degree of Doctorate 3rd Cycle LMD

in Electrical Engineering

Option: Electrical Control

Presented by:

Slimani Hassina

Industrial Diagnosis of Photovoltaic Systems Using Metaheuristics

In front of jury composed of :

NABIL BENHADDA	Professeur	Abbes Laghrour – Khenchela University	Chairman
OUAHIBA CHOUHAL	MCA	Abbes Laghrour – Khenchela University	Supervisor
YASSINE BEDDIAF	Professeur	Abbes Laghrour – Khenchela University	Co-Supervisor
DAOUD REZZAK	MCA	Abbes Laghrour – Khenchela University	Examiner
OUAHAB KADRI	Professeur	Batna2– Mostefa Ben Boulaïd University	Examiner
DJAMEL SAKRI	MCA	Oum El Bouaghi University	Examiner

Academic year: 2025/2026

Acknowledgment

First and foremost, I extend my sincere gratitude to Allah Almighty for His guidance and support and for granting me strength, patience and determination. His assistance has been the decisive factor in the completion of this scientific thesis**.**

It is with great pleasure that I extend my deepest gratitude and profound appreciation to Professor Chouhal Ouahiba, my thesis supervisor, for her continuous efforts, precise guidance and constructive feedback, which played a pivotal role in shaping this research and presenting it in the desired academic form.

I am also sincerely grateful to Professor Mahdaoui Rafik for his valuable co-supervision and for the breadth of knowledge and intellectual rigor he generously shared, which had a significant impact on the development of my research trajectory.

My heartfelt thanks are due to Professor Beddiaf Yassine, Director of the Doctoral Program, for his essential support and guidance, which contributed greatly to the success of this work.

I would also like to express my deepest gratitude and appreciation to the esteemed members of the discussion committee: Professor Sakri DJamal from the University of Oum El Bouaghi and Professor Kadri Ouahab from the University of Batna, for kindly agreeing to examine this thesis despite their numerous commitments and the burdens of travel. I am also grateful for the valuable scientific insights and constructive feedback they will provide, which will contribute to improving this work and enhancing its academic and methodological quality.

I would also like to extend my sincere thanks to Professor Daoud Razzak from the University of Khenchela for the valuable scholarly observations and thoughtful guidance he will offer, which will help further develop this work and strengthen its academic standard.

Furthermore, I would like to express my profound gratitude to the Chairman of the discussion committee, Professor Benhadda Nabil from the University of Khenchela, for kindly accepting to preside over the committee, and for thfor the effort he will devote to evaluating this thesis and enriching it with his insightful remarks.

To the soul of my late father may God have mercy upon him whose constant prayers were a source of strength and comfort, I dedicate my sincerest loyalty and eternal gratitude for his enduring influence.

To my beloved mother, my supportive husband and my dear children, I owe my deepest appreciation for their patience, sacrifices and unwavering encouragement throughout the years of this research, steadfastly enduring the challenges and difficulties that accompanied this journey.

Finally, I would like to express my heartfelt thanks and appreciation to my dear brothers and sisters for their continuous encouragement and moral support, which had a tangible positive impact on the successful completion of this work.

Slimani Hassina

شكر وعرفان

أولًا، أتقدّم بخالص امتناني إلى الله تعالى على توفيقه وعونه، وما أمدّني به من قوة وصبر وعزيمة، فكان عونه السبب في إتمام هذه الرسالة العلمية.

ويسعدني أن أعرب عن بالغ امتناني وعظيم تقديري للأستاذة شوحل وهيبة، المشرفة على أطروحتي، لما بذلت من جهد متواصل، وما قدّمته من توجيهات دقيقة وملاحظات بنّاءة أسهمت بفاعلية في صياغة هذا البحث وإخراجه بالصورة العلمية المرجوّة.

كما أتقدّم بجزيل الشكر للأستاذ مهداوي رفيق على مشاركته القيّمة في الإشراف، وما أبداه من سعة علم، ورصانة فكر، كان لها الأثر البالغ في تطوير مساري البحثي.

كما أخص بالشكر الأستاذ بضياف ياسين، مدير مشروع الدكتوراه، لما قدّمه من دعم وتوجيهات أساسية ساعدت على إنجاح هذا العمل.

كما لا يفوتني أن أتقدّم بأسمى عبارات الشكر وعظيم الامتنان إلى السادة أعضاء لجنة المناقشة الموقّرين: الأستاذ ساكري جمال من جامعة أم البواقي، والأستاذ قادري وهاب من جامعة باتنة، على تفضّلهم الكريم بقبول مناقشة هذه الأطروحة، رغم ما يرافق ذلك من التزامات ومشقّة التنقّل، وعلى ما سيقدّمونه من ملاحظات علمية دقيقة وتوجيهات بنّاءة تُسهم في تحسين هذا العمل والارتقاء بقيمته العلمية والمنهجية.

كما أتوجّه بخالص الشكر إلى الأستاذ داود رزاق من جامعة خنشلة، على ما سيقدّمه من ملاحظات علمية قيّمة وتوجيهات رصينة من شأنها أن تُسهم في تطوير هذا العمل وتعزيز مستواه العلمي.

ويطيب لي كذلك أن أتوجّه بخالص الامتنان إلى رئيس لجنة المناقشة، الأستاذ بن حدة نبيل من جامعة خنشلة، على تفضّله برئاسة اللجنة، وما سيبذله من جهد في تقييم هذه الأطروحة وإغنائها بملاحظاته السديدة.

وإلى روح والدي، رحمه الله تعالى، الذي كان بدعائه المستمرّ سندًا وعونًا لي، أتوجّه بخالص الوفاء والاعتراف بفضله الذي لن يزول.

أما والدتي الغالية، وزوجي الكريم، وأبنائي الأحبة، فلهم منّي أسمى معاني العرفان لما أبدوه من صبر وتضحيات وتشجيع متواصل طيلة سنوات هذا البحث، متحمّلين بثبات ما رافق هذه المسيرة من صعوبات وتحديات.

كما أخص إخوتي وأخواتي الأعزاء بخالص الشكر والتقدير، لتشجيعهم المستمر ودعمهم المعنوي طوال سنوات هذا البحث، وهو ما كان له أثر إيجابي ملموس في استكمال هذا العمل بنجاح.

سليماني حسينة

Abstract

Photovoltaic (PV) systems are increasingly recognized as effective and sustainable solutions to meet the global demand for clean energy. However, their efficiency and performance are often compromised by various technical faults, including partial shading, cell aging, hotspots and short or open-circuit issues, among other operational challenges. Consequently, the development of intelligent and interpretable fault diagnosis methods is crucial to ensure the reliability and sustainability of these systems.

This thesis presents innovative approaches that integrate intelligent optimization techniques and include metaheuristic algorithms, with explainable artificial intelligence (XAI) to enhance diagnostic accuracy and modeling efficiency. Initially, a hybrid signal filtering method was proposed to improve data quality and reduce noise, thereby strengthening the robustness of fault diagnosis using Grey Wolf Optimization (GWO) to evaluate algorithmic effectiveness. Furthermore, an Adaptive variant of GWO (AGWO) was developed for feature selection, significantly improving model accuracy and reducing execution time compared to conventional methods.

To further advance classification capabilities, this research introduces the Discrete Grey Wolf Optimization (DGWO) algorithm, which adapts classical GWO for discrete search spaces in rule-based classification. The proposed algorithm demonstrated high performance in terms of accuracy, robustness and interpretability, producing results that are both reliable and actionable for decision-makers.

The findings confirm that combining intelligent optimization techniques, particularly meta-heuristic algorithms, with explainable models provides a promising pathway for the development of advanced diagnostic systems. This approach ultimately enhances the efficiency and reliability of PV technologies while laying a solid foundation for real-world applications in monitoring, fault detection and data-driven decision-making.

Keywords: Fault diagnosis in PV systems, Grey Wolf Optimizer, explainable classification, DGWO , feature selection, AGWO , signal quality enhancement.

ملّخـص

تُعتبر أنظمة الطاقة الشمسية الفوتوفولتية (PV) حلولًا فعّالة ومستدامة لتلبية الطلب العالمي على الطاقة النظيفة. ومع ذلك، فإن كفاءتها وأدائها غالبًا ما تتأثر بمختلف الأعطال التقنية، بما في ذلك التظليل الجزئي، شيخوخة الخلايا، البقع الساخنة، ومشكلات الدائرة القصيرة أو المفتوحة، إلى جانب تحديات تشغيلية أخرى. لذلك، يُعد تطوير أساليب تشخيص ذكية وقابلة للتفسير أمرًا بالغ الأهمية لضمان موثوقية واستدامة هذه الأنظمة.

تقدم هذه الأطروحة مقاربات مبتكرة تدمج تقنيات التحسين الذكية، وتشمل الخوارزميات الميتاهيوستية مع الذكاء الاصطناعي القابل للتفسير (XAI) بهدف تعزيز دقة التشخيص وكفاءة النمذجة. في البداية، تم اقتراح طريقة هجينة لتصفية الإشارة لتحسين جودة البيانات وتقليل الضوضاء، وبالتالي تعزيز متانة التشخيص باستخدام تحسين الذئب الرمادي (GWO) لتقييم فعالية الخوارزمية. علاوة على ذلك، تم تطوير نسخة تكيفية من (AGWO) GWO لاختيار الميزات، مما حسّن بشكل كبير دقة النموذج وخفض وقت التنفيذ مقارنة بالطرق التقليدية.

لتعزيز قدرات التصنيف بشكل أكبر، قدمت هذه الدراسة خوارزمية تحسين الذئب الرمادي المنفصل(DGWO) ، التي تكيف خوارزمية GWO الكلاسيكية لتعمل في فضاءات البحث المنفصلة للتصنيف القائم على القواعد. وأظهرت الخوارزمية أداءً عاليًا من حيث الدقة، والمتانة، وقابلية التفسير، مما ينتج نتائج موثوقة وقابلة للتطبيق العملي لصناع القرار.

تؤكد النتائج أن دمج تقنيات التحسين الذكية، لا سيما الخوارزميات الميتاهيوستية ، مع النماذج القابلة للتفسير يمثل طريقًا واعدًا لتطوير أنظمة تشخيص متقدمة. ويسهم هذا النهج في تعزيز كفاءة وموثوقية تقنيات الطاقة الشمسية الفوتوفولتية، مع توفير أساس قوي للتطبيقات العملية في المراقبة، واكتشاف الأعطال، واتخاذ القرارات المعتمدة على البيانات.

الكلمات المفتاحية :تشخيص الأعطال في أنظمة الطاقة الشمسية، تحسين الذئب الرمادي ، التصنيف القابل للتفسير،DGWO ، اختيار الخصائص، AGWO ، تحسين جودة الإشارة.

Résumé

Les systèmes photovoltaïques (PV) sont de plus en plus reconnus comme des solutions efficaces et durables pour répondre à la demande mondiale en énergie propre. Cependant, leur efficacité et leurs performances sont souvent compromises par diverses défaillances techniques, telles que l'ombrage partiel, le vieillissement des cellules, les points chauds, ainsi que des problèmes de court-circuit ou de circuit ouvert, entre autres défis opérationnels. Par conséquent, le développement de méthodes de diagnostic intelligentes et interprétables est essentiel pour garantir la fiabilité et la durabilité de ces systèmes.

Cette thèse propose des approches innovantes qui intègrent des techniques d'optimisation intelligente, incluant des algorithmes métaheuristiques, avec l'intelligence artificielle explicable (XAI) afin d'améliorer la précision du diagnostic et l'efficacité de la modélisation. Dans un premier temps, une méthode hybride de filtrage du signal a été proposée pour améliorer la qualité des données et réduire le bruit, renforçant ainsi la robustesse du diagnostic par l'utilisation de l'optimisation par loup gris (GWO) pour évaluer l'efficacité de l'algorithme. De plus, une variante adaptative du GWO (AGWO) a été développée pour la sélection des caractéristiques, améliorant significativement la précision du modèle et réduisant le temps d'exécution par rapport aux méthodes conventionnelles.

Pour renforcer davantage les capacités de classification, cette étude introduit l'algorithme Discrete Grey Wolf Optimization (DGWO), qui adapte le GWO classique aux espaces de recherche discrets pour la classification basée sur des règles. L'algorithme proposé a démontré de hautes performances en termes de précision, de robustesse et d'interprétabilité, produisant des résultats à la fois fiables et exploitables par les décideurs.

Les résultats confirment que la combinaison des techniques d'optimisation intelligente, en particulier des algorithmes méta-heuristiques, avec des modèles explicables constitue une voie prometteuse pour le développement de systèmes de diagnostic avancés. Cette approche permet d'améliorer l'efficacité et la fiabilité des technologies photovoltaïques tout en fournissant une base solide pour des applications réelles en surveillance, détection de défaillances et prise de décision basée sur les données.

Mots-clés : Diagnostic des défauts dans les systèmes photovoltaïques, Optimiseur de loups gris, classification explicable, DGWO , sélection de caractéristiques, AGWO , amélioration de la qualité du signal.

Table of contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Tables

Acronyms

AC	Alternating Current
ABC	Artificial Bee Colony Algorithm
AGWO	Adaptive Grey Wolf Optimizer
ANNS	Artificial Neural Networks
ACO	Ant Colony Optimization Algorithm
AUC	Area Under the Curve
ADMC201	Advanced Digital Measurement and Control 201
BCO	Bee Colony Optimization Algorithm
BN	Bayesian Network
BPNN	Backpropagation Neural Network
Boost	Step-up DC-DC Converter
Buck	Step-down DC-DC Converter
C class	Classification Category
C0	Classification Category 0
C1	Classification Category 1
C2	Classification Category 2
C3	Classification Category 3
CA	Classification Accuracy
CBR	Case-Based Reasoning
CEC	California Energy Commission
CO_2	Carbon Dioxide
CSP	Concentrated Solar Power
Ćuk converter	DC-DC Converter Topology (Inverted Output, Low Ripple)
D	Distance Vector Between Prey and a Wolf Maxiter
DE	Differential Evolution
DT	Decision Tree
DGWO	Discrete Grey Wolf Optimization Algorithm
DC	Direct Current
DC–AC	Direct Current to Alternating Current (Inverter)
DC–DC	Direct Current to Direct Current (Converter)
DSP	Digital Signal Processing
DE	Differential Evolution Algorithm
EA	Evolutionary Algorithm

EEA	Expert-Guided Evolutionary Algorithm
EMO	Electromagnetic Field Optimization Algorithm
EP	Evolutionary Programming Algorithm
ES	Evolution Strategies Algorithm
EO	Equilibrium Optimizer
ES	Evolution Strategy
E-GWO	Enhanced Grey Wolf Optimizer
F0	Normal Operating State
F1	String Fault
F2	String-to-Ground Fault
F3	String-to-String Fault
FCS	Fuzzy Control System
FA	Firefly Algorithm
FLS	Fuzzy Logic Systems
FP	Number of Negative Cases Incorrectly Predicted as Positive
FN	Number of Positive Cases Incorrectly Predicted as Negative.
Fw(t)	The Filtered Signal Obtained with a Window of Size W
Flyback	Isolated DC-DC Converter Topology
GA	Genetic Algorithm
G	Global Solar Irradiance
GSA	Gravitational Search Algorithm
GW	Gigawatt
GWO	Grey Wolf Optimization Algorithm
HGS	Human-Guided Search Algorithm
HILO	Human-in-the-Loop Optimization Algorithm
Vpvfn(t)	The Original Noisy Signal Voltage
IGBT	Insulated Gate Bipolar Transistor
IGWO	Improved Grey Wolf Optimizer
IEC	Interactive Evolutionary Computation Algorithm
IRENA	International Renewable Energy Agency
INCC	Incorrectly Covered Cases
ICC	Correctly Covered Cases
KNN	K-Nearest Neighbors
Kmeans	K-means Clustering
kw	Kilowatt – Unit of Power

kV	Kilovolt – Unit of Voltage
kWh/m²	Kilowatt-Hour Per Square Meter
KBD	Knowledge-Based Diagnosis
LTI	Innovative Technologies Laboratory (LTI), EA 3899,France
LightGBM	Light Gradient Boosting Machine
MAF	Moving Average Filter
ML	Machine Learning
MPPT	Maximum Power Point Tracking
MPP	Maximum Power Point
MPC	Model Predictive Control
MSE	Mean Squared Error
MW	Megawatt
MWp	Megawatt-Peak
MTOE	Million Tons of Oil Equivalent
N_s	Number of Series Cells
N	N-type Semiconductor Material
N	The Total Number of Observations or Data Points
NB	Naive Bayes
NN	Neural Network
P	P-type Semiconductor Material
P&O	Perturb & Observe (MPPT)
P_DC	Total DC Power Output
P_max	Standard Test Conditions Maximum Power (W)
P_mean	Mean Power Output
PN	Junction Formed by Joining P and N
PSO	Particle Swarm Optimization
PCA	Principal Component Analysis
Prec	Precision
PV	Photovoltaic System
QEA	Quantum-Inspired Evolutionary Algorithm
RL	Reinforcement Learning
ROC	Receiver Operating Characteristic
RES	Renewable Energy Sources
SA	Simulated Annealing
SMC-PID	Sliding Mode Control PID

STA	Super Twisting Algorithm
SVMs	Support Vector Machines
SHAP	Shapley Additive Explanations (SHAP)
TP	Number of Positive Cases Correctly Predicted.
TN	Number of Negative Cases Correctly Predicted.
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boostin
Y_{actual}	The true value of the Dependent Variable for the i-th Observation
$Y_{prdicted}$	The Value Predicted by the Model for the i-th Observation

Symbols

α	Alpha Wolf (Leader)
β	Beta Wolf (Second Best)
δ	Delta Wolf (Third Best)
ω	Omega Wolf (Remaining Wolves)
Ω	Ohm – Unit of Resistance
°C	Degrees Celsius
F(t)	Filtered Signal at Time t
I	Current (A)
I0	Diode Saturation Current
I1	Current of PV String 1
I2	Current of PV String 2
I3	Current of PV String 3
I4	Current of PV String 4
I5	Current of PV String 5
I6	Current of PV String 6
I1max	Maximum Current 1
I2max	Maximum Current 2
I3max	Maximum Current 3
Itotalmax	Maximum Total Current
I1min	Minimum Current 1
I2min	Minimum Current 2
I3min	Minimum Current 3
Itotalmin	Minimum Total Current
I1VAR	Current Variance 1

I2VAR	Current Variance 2
I3VAR	Current Variance 3
Id	Diode Current
I-V	Current–Voltage Characteristic
J/K	Joules Per Kelvin
K	Boltzmann Constant
Range 1	The Difference Between the Maximum and Minimum Current Values for String 1
Range 2	The difference Between the Maximum and Minimum Current Values for String 2
Range 3	The difference between the top and bottom current values for string 1
Range 4	The Difference Between the Top and Bottom Current Values for String 2
q	Electron Charge
V	Voltage
V_DC	Aggregate DC Voltage
Vdcmean	Mean DC Voltage
Vdcmax	Maximum DC Voltage
Vdcmin	Minimum DC Voltage
Voc	Open-Circuit Voltage
VT	Thermal Voltage
W	Window Size
W/m²	Watts Per Square Meter
Y[t]	Noisy Signal at Time t
Y[t−k]	Signal Value at Time (t−k)
T	Temperature
T(K)	Absolute Temperature
X	Grey Wolf's Position Vector
Xp	Prey Position Vector

General Introduction

The global energy sector is undergoing major transformations driven by the depletion of fossil fuel resources, market volatility, and the increasing impacts of climate change. These challenges highlight the urgent need for sustainable energy systems that can meet rising demand while minimizing environmental impacts. In this context, renewable energy has emerged as a strategic solution, contributing to energy diversification, supply security, reduced greenhouse gas emissions, and sustainable economic growth. Despite diversification efforts, fossil fuels still dominate global energy consumption, accounting for about 86% in 2024 [1]. This dependence intensifies environmental and climate challenges, making energy efficiency and cleaner alternatives key priorities. At the same time, renewable energy has experienced significant growth, reaching 28.2% of global electricity generation in 2021 with a capacity of 3,064 GW, a 9% increase from 2020 [2]. This capacity further rose to around 3,870 GW by the end of 2023, marking a 13.9% increase from 2022 [3]. Although growth rates vary across countries, this trend confirms the crucial role of renewables in long-term energy security.

In this context, Algeria has strong renewable energy potential, particularly in solar power, due to its favorable geographic and climatic conditions. This represents a key opportunity to support the national energy transition, diversify energy production, and strengthen the country's position in the renewable energy sector regionally and internationally.

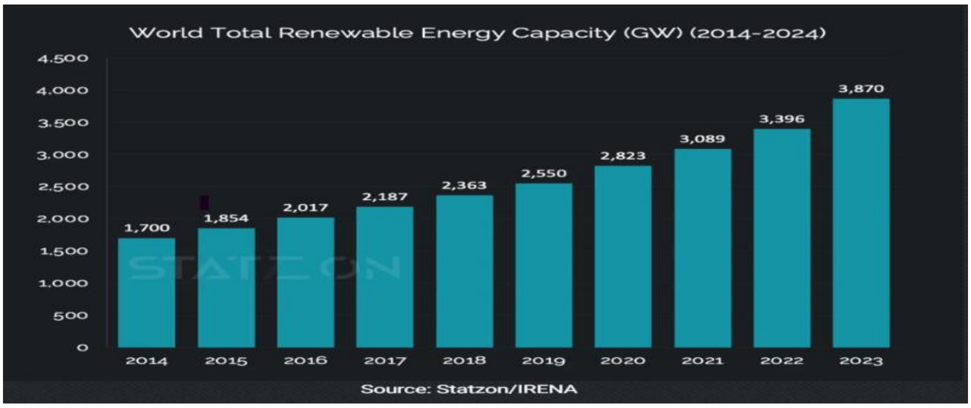


Figure.0.1 World Total Renewable Energy Capacity (GW) from 2014 to 2023

Algeria's Renewable Energy Strategy and Solar Potential

Algeria's renewable energy strategy focuses on diversifying its energy mix through large-scale development of renewable sources, particularly solar energy, due to the country's high solar irradiation levels in the south. The national target is to reach 22,000 MW of renewable capacity by 2030, alongside reducing CO_2 emissions [1][2][3].

The program has evolved through three phases: a pilot phase (2011–2014), a deployment phase (2015–2020), and a scale-up phase (2021–2030), which emphasizes grid integration and large solar and wind projects. Between 2015 and 2018, several PV plants (2–40 MW) and a major 343 MWp project were implemented, improving energy access and grid stability, as illustrated in Table 0.1 and Figure 0.2.

However, photovoltaic systems still face operational challenges such as equipment failure, degradation, shading, and grid integration issues, which affect overall performance and efficiency.

Table 0.1. Planned Installed Capacities of Renewable Energy Technologies in Algeria (2015–2030), in MW

Unit: MW	1ère phase 2015-2020	2ème phase 2021-2030	TOTAL
PV energy	3 000	10 575	13 575
Wind energy	1 010	4 000	5 010
CSP	-	2000	2 000
Cogeneration	150	250	400
Biomass	360	640	1 000
Geothermic	05	10	15
TOTAL	4 525	17 475	22 000

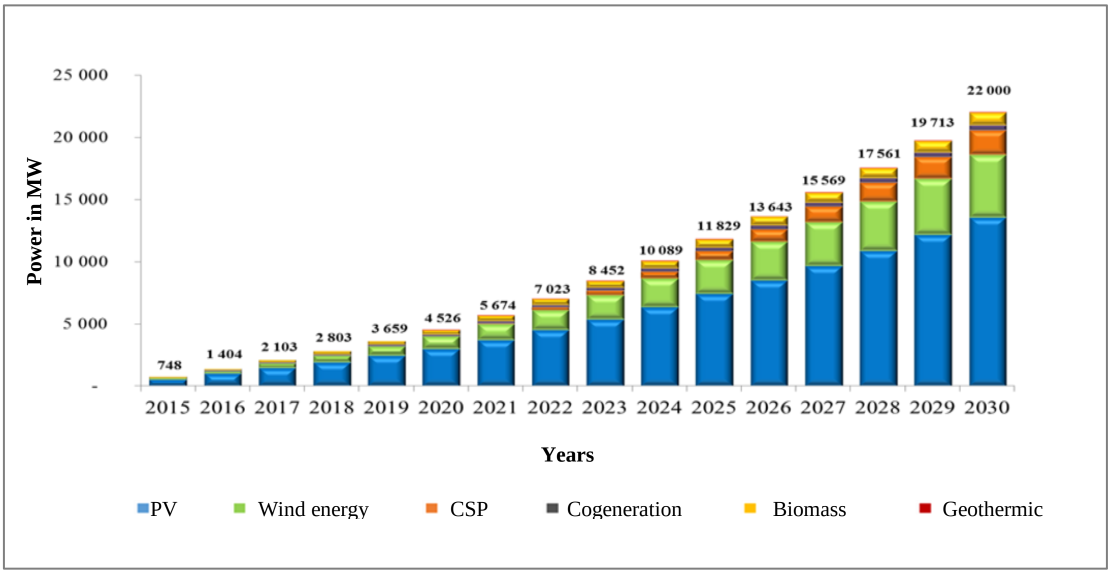


Figure.0.2 Evolution of Renewable Energy Development in Algeria (2015-2030)

Between 2015 and 2018, Algeria expanded its solar capacity through the deployment of photovoltaic plants (2 – 40 MW) across several regions, improving grid stability and electricity access in remote areas. A major 343 MWp PV project was also launched, marking a key step in the country's renewable energy development and long-term sustainability efforts.

However, despite these advancements, operational challenges persist. PV systems, especially in harsh or remote environments, are prone to technical issues such as inverter failures, panel degradation, shading, and grid connection problems, which can negatively impact performance, reliability, and economic efficiency.

Problem Formulation and Motivation

Photovoltaic (PV) solar energy systems have become one of the most widely adopted renewable energy technologies due to their economic viability and their significant role in reducing carbon emissions. However, these systems remain highly susceptible to operational and technical faults caused by environmental conditions, component degradation, and system malfunctions, leading to reduced efficiency and long-term reliability issues. Therefore, fault diagnosis is essential to ensure optimal performance and operational resilience. Nevertheless, traditional fault detection methods are limited in handling the nonlinear, complex, and dynamic nature of PV systems.

In this context, metaheuristic algorithms represent a promising solution due to their strong capability to explore complex and uncertain search spaces and provide near-optimal solutions for challenging optimization and diagnostic problems. This research aims to leverage these capabilities to develop an effective framework for fault detection and diagnosis in PV systems, thereby improving their operational efficiency and performance.

Despite significant progress in this field, an important research gap still exists, as most existing models whether based on machine learning or metaheuristic approaches operate as "black-box" systems with limited interpretability, which restricts their adoption in industrial applications that require transparency and traceability.

Contributions and Originality of the Thesis

The main contributions and originality of this thesis are summarized as follows:

1. Hybrid Signal Filtering Technique

A novel hybrid method combining the Moving Average Filter with the GWO is proposed for signal filtering and smoothing in PV systems. The GWO algorithm is used as an intelligent optimization strategy to tune filter parameters by minimizing the MSE, thereby enhancing signal quality and integrity.

2. Feature Selection Using Adaptive GWO

An adaptive Grey Wolf Optimizer (AGWO) was proposed to enhance feature selection by balancing exploration and exploitation. It was integrated with XGBoost, LightGBM, and CatBoost, leading to improved performance in terms of accuracy, reliability, and efficiency. SHAP was further used to interpret feature contributions in error classification, improving model interpretability and transparency.

3. Development of DGWO Algorithm

A new variant of the Grey Wolf Optimizer (DGWO) is proposed to handle discrete search spaces in classification tasks by using symbolic rule representations instead of continuous vectors. It is evaluated in terms of accuracy, robustness, and stability, and compared with traditional classification methods.

Thesis Structure Overview:

- Chapter 1: Presents the theoretical background of PV systems, including solar energy conversion principles, PV cell types, Maximum Power Point Tracking (MPPT), system classifications, and fault types affecting PV systems.
- Chapter 2: Introduces fault diagnosis fundamentals in PV and industrial systems, describing diagnostic stages and reviewing both conventional and advanced methods, forming a foundation for intelligent optimization-based diagnosis.
- Chapter 3: Covers metaheuristic optimization methods, focusing on the GWO. It explains its principles, advantages, and applications, including integration with a Moving Average Filter to enhance PV signal quality and reduce noise.
- Chapter 4: Develops interpretable diagnostic models through dataset exploration and feature analysis. AGWO is combined with gradient boosting-based classifiers (XGBoost, LightGBM, CatBoost), followed by Soft Voting ensemble learning to improve diagnostic performance and reliability.
- Chapter 5: Presents an advanced interpretable diagnostic framework using the DGWO algorithm integrated with Explainable AI techniques, aiming to generate transparent and accurate rule-based fault diagnosis for PV systems.

This thesis presents an integrated AI-based diagnostic framework that combines advanced machine learning techniques with state-of-the-art optimization algorithms. The proposed approach enhances fault detection and prediction accuracy in photovoltaic systems, while also improving feature selection, signal filtering precision, and overall decision-making support.

Chapter 1: Photovoltaic System Architecture and Components

1.1 Introduction

PV cell technology is one of the fundamental technologies in the field of solar energy conversion, enabling the direct transformation of solar radiation into electrical energy based on the photovoltaic effect. Over the past decades, photovoltaic systems have attracted increasing attention due to their capability to provide sustainable, low-carbon energy solutions, as well as their significant role in meeting the growing global demand for energy.

This section presents a comprehensive overview of photovoltaic technologies, with emphasis on fundamental operating principles, material properties and the key factors influencing system performance. The design and operational configurations of photovoltaic systems are also examined, highlighting their classification according to grid-connection modes, system scale and application domains.

Furthermore, common faults and operational challenges affecting photovoltaic systems are analyzed, as these issues have a direct impact on system efficiency, reliability and operational lifespan. This systematic analysis contributes to a holistic understanding of the technical and operational aspects of photovoltaic systems, thereby supporting rigorous performance evaluation and providing a solid foundation for system assessment and improvement.

1.2 Photovoltaic Cell Technology and Energy Conversion

A PV cell is the fundamental unit in solar energy conversion systems, enabling the transformation of solar radiation into electrical energy through the PV effect, first discovered by physicist Alexandre Edmond Becquerel in 1839. This effect involves the direct conversion of light energy into electrical energy within a semiconducting material, where incident photons generate charge carriers (electrons and holes) and induce their motion under light excitation.

PV cells based on PN junctions are constructed using semiconductor materials that are doped to form two adjacent regions: an N-type region, rich in free electrons and a P-type region, characterized by a high concentration of holes. At the interface between these two regions, an internal electric field is established, which prevents the free diffusion of charge carriers. This results in the formation of a depletion region, which plays a key role in separating photo-generated carriers and facilitating the generation of an electric current when light radiation strikes the cell. Photons with sufficient energy excite electrons from the valence band to the conduction band, generating electron-hole pairs. Under the influence of the internal electric field of the PN junction, these charge carriers are separated: electrons are drawn toward the N region. At the same time, holes migrate toward the P region(Figure.1.1) [4], [5].This movement of charge carriers generates an electric current when an external circuit is connected to the cell terminals. This process underlies electricity generation in PV cells, whose current-voltage characteristic resembles that of a silicon diode, with the emergence of a photocurrent proportional to the intensity of incident radiation [6].

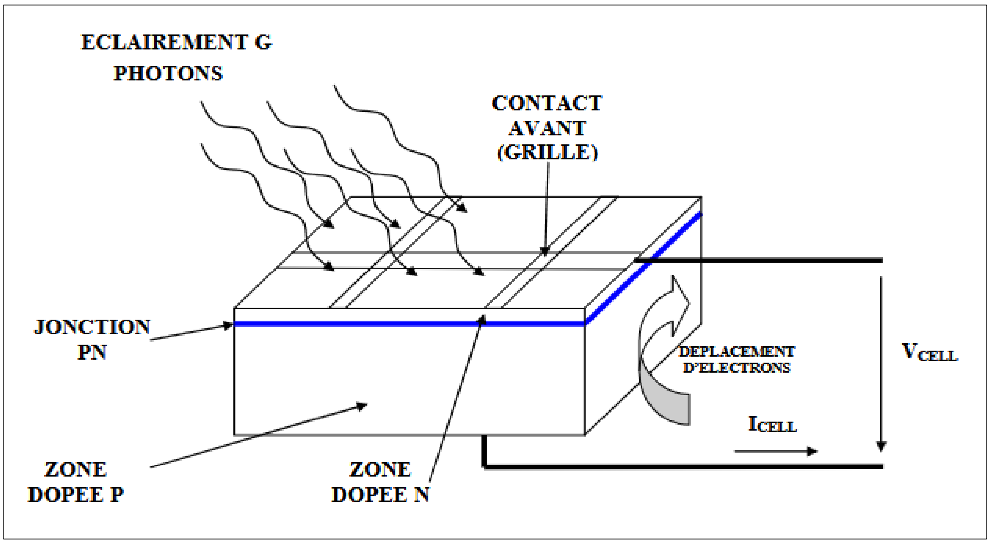
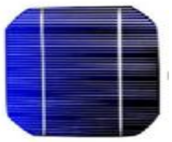


Figure.1.1 Cross-Section of a PV Cell [5]

1.3 Various Photovoltaic Cell Technologies

PV cells are classified into three main categories: monocrystalline silicon, polycrystalline silicon and thin-film cells. Each technology presents distinctive characteristics in terms of efficiency, manufacturing cost and adaptability to various climatic conditions, which influence their selection for different energy applications. The following table provides a detailed comparative analysis of the technical specifications and performance associated with these different cell technologies (Table.1.1).

Table 1.1. Comparaison of PV Technologies [7],[8]

Technology	**Monocrystalline**	**Polycrystalline**	**Thin-Film**
Cell / **Specifacations**			
Panel			
Efficiency (%)	Ranges between 16% and 24%, with a performance models can achieve efficiencies of up to 21.5% to 24%	13%–18% lower efficiency than monocrystalline due to lower silicon purity	~10% Lower efficiency compared to mono-crystalline and polycrystalline panels
Cost	Due to its high cost of production, it is one of the most expensive PV technologies	More cost due to simpler manufacturing process and less silicon waste.	Generally lower cost due to simple manufacturing processes and use of less expensive materials
Advantages	- High efficiency, even in cloudy conditions - Long lifespan (25 to 30 years) - Superior energy output	- Lower cost compared to Monocrystalline - Good heat tolerance (maintains stable efficiency above 25°C) - Economical alternative	-Lightweight and flexible - Performs well in low light conditions - Suitable for unconventional surfaces
Disadvantages	- High cost - Susceptibility to shading, -Performance decreases at high temperatures	- Lower efficiency - Less uniform appearance (blue-tinted surface) -Shorter lifespan: Typically 10 to 15 years -Reduced performance in heat	- Lower efficiency - Shorter lifespan (10 to 15 years) -Not as efficient in high temperatures or direct sunlight.
Applications	- Residential and commercial installations - Limited-space applications requiring high efficiency	-Suitable for residential and commercial solar installations. - Ideal for limited-space environments	- Ideal for portable and mobile solar solutions - Curved surfaces (boats, vehicles, unique building structures)

1.4 Photovoltaic Cell Performance Characteristics

The performance of a PV cell is determined by several key parameters, which can be derived from its current–voltage (I–V) characteristic curve. The most important parameters are the open-circuit voltage (Voc) and the short-circuit current (Isc).

These parameters are highly influenced by environmental factors, such as solar irradiance, temperature and the cell's material composition. Figure.1.2 illustrates the typical current–voltage behavior of a PV cell.

➢ Key Characteristics of Photovoltaic Cells

- **Open-Circuit Voltage (Voc)**

This is the maximum voltage the cell can produce when no external load is connected. It is measured directly across the cell terminals and typically ranges from 0.5 to 0.8 V.

- **Short-Circuit Current (Isc)**

The maximum current a PV cell can generate when its terminals are short-circuited.

- **Maximum Power Point Voltage and Current (Vmp and Imp)**

These values correspond to the operating point where the cell delivers its maximum power. They are selected to maximize the product of voltage and current .

On the (I–V) curve, V_{oc} is observed at the intersection with the voltage axis, I_{sc} at the intersection with the current axis and the maximum power point (V_{mp} , I_{mp}) corresponds to the point where the cell delivers its highest power output, as expressed in Equation (1.1) and shown in Figure 1.2.

$$P_{max} = V_{mp} \times I_{mp} \tag{1.1}$$

Where :

Pmax**:** Maximum power of the solar panel (W)

Vmp: Voltage at maximum power (V)

Imp: Current at maximum power (A)

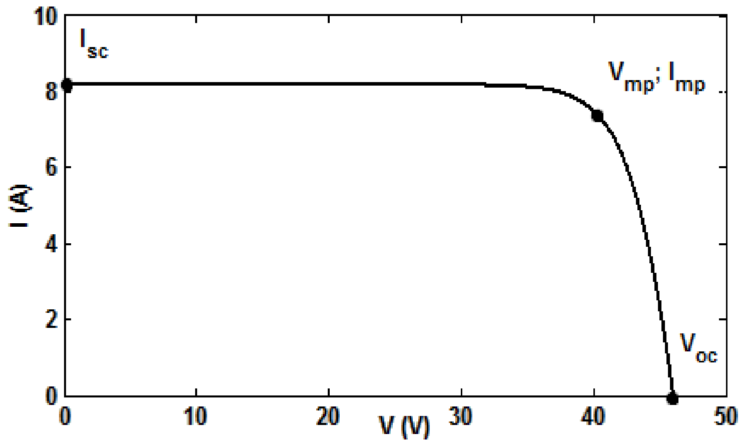


Figure.1.2 (I-V) Characteristic [9]

1.5 Design and Operational Principles of a Photovoltaic Systems

PV system is a technology designed to convert solar energy into usable electrical power through the operation of PV cells, which function based on the PV effect. The energy conversion process involves a series of interdependent technical stages that require precise synchronization and functional integration among system components to ensure optimal efficiency and high reliability in electricity production. The PV array, composed of solar modules, captures sunlight and converts it into direct current, which is then transmitted through wiring and junction boxes that interconnect the modules and link them with other system components (Figure.1.3).

In systems where it is required, an inverter converts DC into alternating current to ensure compatibility with various electrical loads. Energy storage devices, typically batteries, are incorporated to provide a stable power supply during periods of low solar irradiation, thereby maintaining continuous operation of the loads. Various protection mechanisms, such as bypass diodes, blocking diodes and other protective devices, are implemented to preserve the system's reliability and operational performance. The overall efficiency and performance of any PV system depend on the precise integration of its components and an understanding of its dynamic behavior, enabling maximum utilization of solar energy while ensuring a stable electricity supply to the connected loads[10].

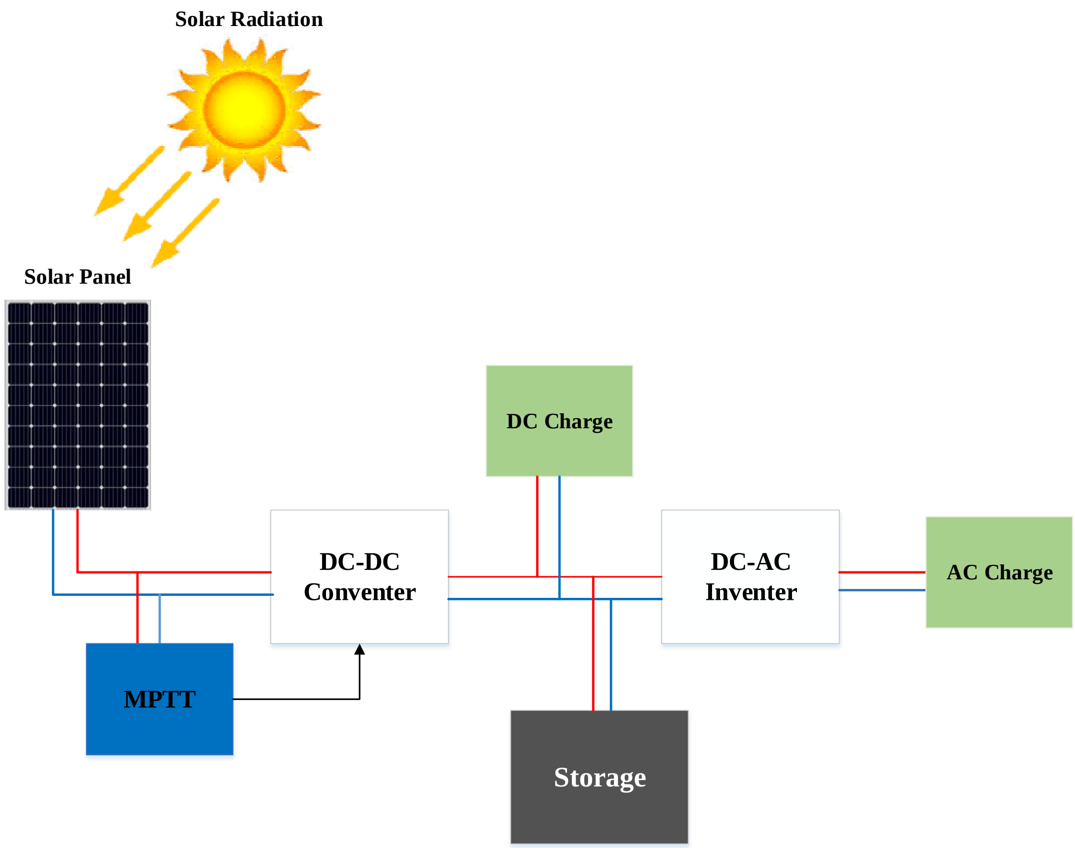


Figure.1.3 Comprehensive Architecture of a PV System

1.5.1 Photovoltaic Generator

A PV generator is an energy system designed to convert solar radiation into direct current electricity through the PV effect. It consists of PV cells, which are semiconductor-based components that absorb photons to induce electron flow.

These cells are assembled into modules, where the generated electrical energy can be either utilized immediately or stored, depending on system requirements (Figure.1.4).

To enhance performance, PV cells are typically connected in series, increasing the voltage while maintaining a constant current. A standard module generally comprises between 36 and 72 cells, although this configuration may vary based on specific applications[11],[12].

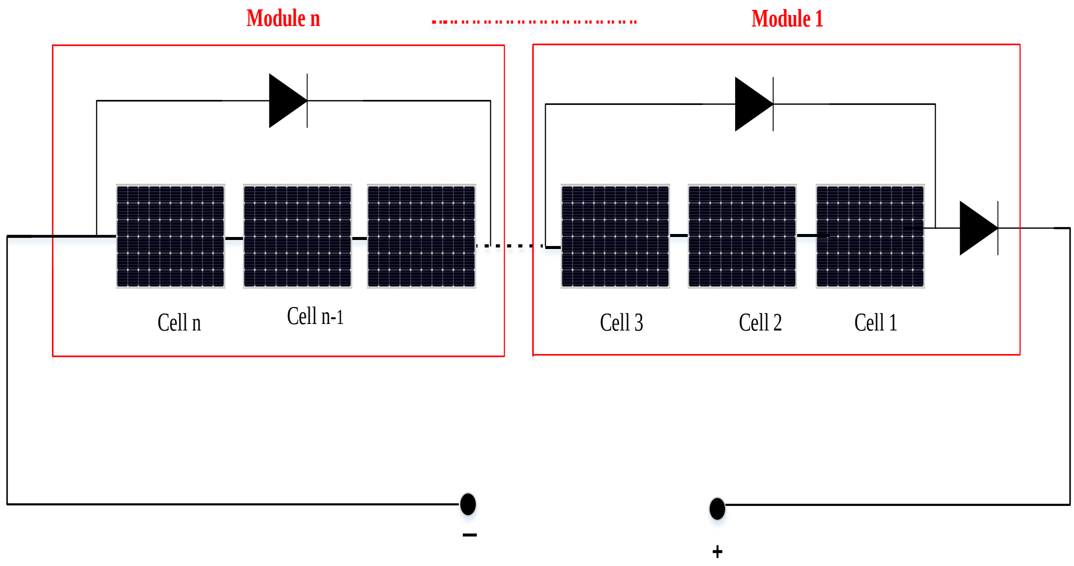


Figure.1.4 PV module

1.5.2 Photovoltaic String

A PV string comprises a series connection of modules engineered to deliver a specified voltage output. To prevent reverse current flow that may compromise system operation, the string incorporates protective diodes, utilizing bypass and blocking diodes as protective elements to safeguard the system's continuity and performance [11] . The structural configuration of a PV module is illustrated in Figure.1.5.

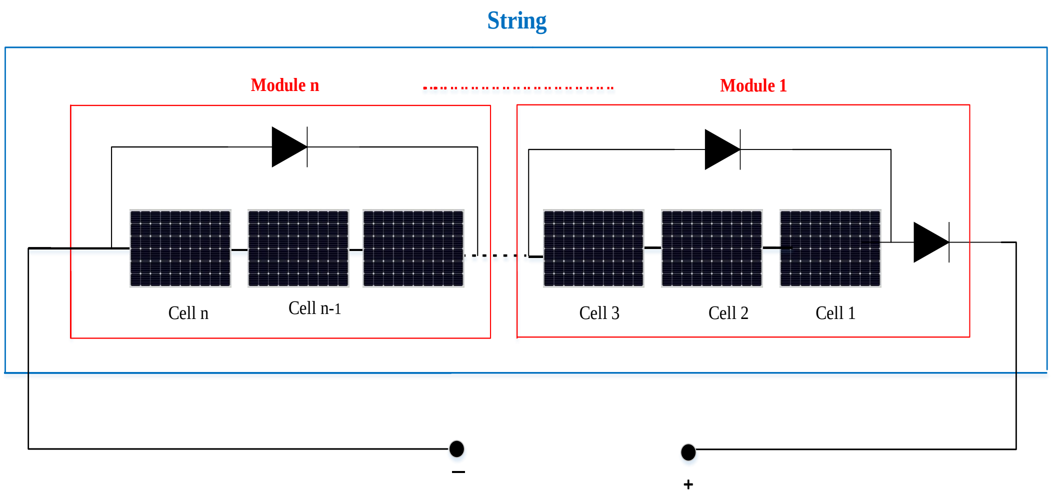


Figure.1.5 PV string

1.5.3 Photovoltaic Field

A PV field is composed of multiple PV strings, each containing an identical number of modules. These strings are connected in parallel to increase the overall current of the system, thereby enabling the achievement of the required electrical power output for the installation (Figure.1.6). This parallel connection facilitates an even distribution of electrical load across the different strings, enhancing the overall operational efficiency of the PV system[11].

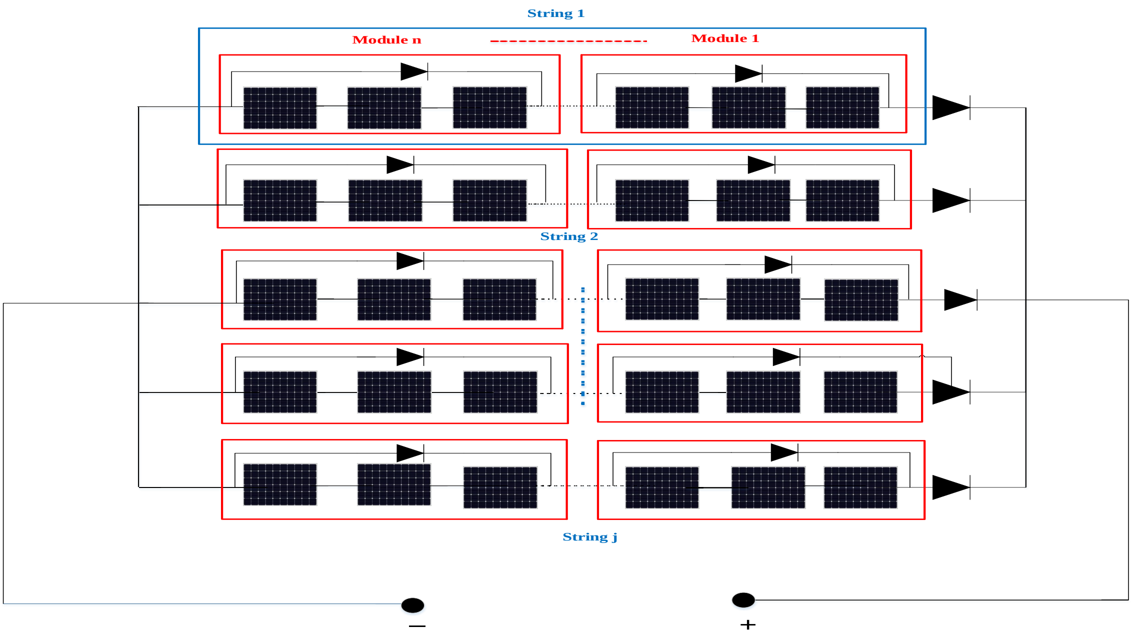


Figure.1.6 Series-Parallel PV Field

1.5.4 DC-DC Converters and Maximum Power Point Tracking in Photovoltaic Systems

DC-DC converters play a fundamental role in PV systems by regulating the output voltage and ensuring that the PV generator operates at its MPP. This regulation is achieved through MPPT algorithms, which continuously adjust the operating conditions to maximize energy extraction under varying solar irradiance and temperature.

These converters utilize various topologies to regulate the relationship between the input and output voltages. Standard configurations include Buck, Boost and Buck-Boost converters, each designed to meet specific operational requirements [13]. Additionally, advanced topologies such as Ćuk, SEPIC and Flyback offer efficient bidirectional voltage conversion, enhanced design flexibility, improved signal stability and reduced ripple. The Flyback converter, in particular, provides electrical isolation, making it well suited for renewable energy applications and industrial electronics [14].

The selection of a suitable DC-DC converter topology is Subject to on the electrical characteristics of the PV array and the voltage specifications of downstream system components. Efficient DC-DC conversion not only improves the overall energy conversion efficiency but also ensures smooth integration with subsequent elements such as inverters and energy storage units.

The MPPT functionality embedded in the converter controller dynamically adjusts the operating point of the PV array, which exhibits a nonlinear (I–V) characteristic (Figure.1.7). MPPT algorithms perform real time tracking by measuring voltage and current, continuously updating the operating conditions to follow the MPP under fluctuating irradiance and temperature [15].

Various MPPT techniques have been proposed in the literature, each differing in convergence speed, tracking accuracy, sensor dependency and implementation complexity [16],[17],[18]. The selection of a specific MPPT algorithm depends on system-level trade-offs between performance and cost.

Therefore, DC-DC converters, in conjunction with embedded MPPT control, serve as key components in the electrical architecture of PV systems, ensuring optimal energy conversion and compatibility with downstream elements such as inverters and storage devices.

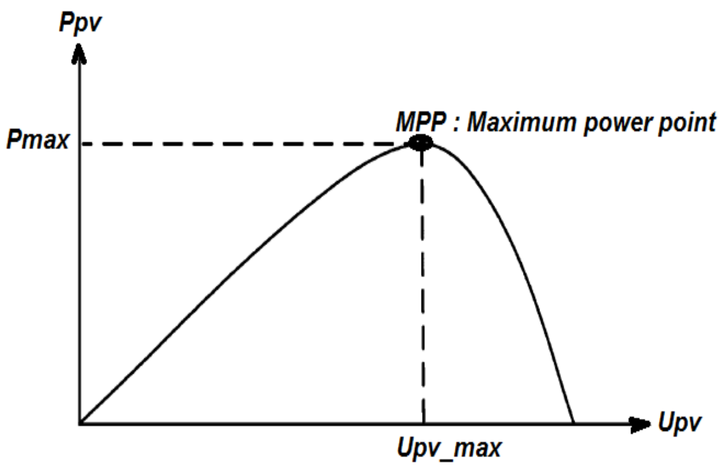


Figure.1.7 (I-V) Curve of a PV field in Normal Operation [15]

1.5.5 DC-AC Conversion and Inverter Topologies in Photovoltaic Systems

This stage involves converting the direct current electrical energy supplied by the DC-DC converter into alternating current that is compatible with the utility grid in terms of voltage, frequency and phase. This ensures seamless integration of the generated power into the electrical network.

The conversion process requires the use of inverters, whose design and configuration vary according to the size of the PV system and its specific operational requirements. Three types of inverters are used in PV power systems [11],[19],[20] Modular inverters، central inverters، and string inverters (Figure.1.8).

- **Modular Inverters**

The modular architecture connects an individual inverter to each PV module in large-scale installations. The outputs of these inverters are combined in parallel on the AC side, helping to minimize energy losses caused by partial shading and enhancing overall energy production.

- **Centralized Inverters**

The Centralized inverters, by contrast, are designed to handle the total DC output from a PV array and convert it into AC using a single high-capacity unit. This topology is particularly suited for high-power applications, where cable losses are reduced by connecting a large number of PV modules in series.

- **String Inverters**

The String inverters, which are the most widely used, are configured to process the output of a string of PV modules, typically eight or more connected in series. This configuration provides a favorable balance between installation cost and system performance. Figure 1.8 illustrates the topology of grid-connected PV inverters.

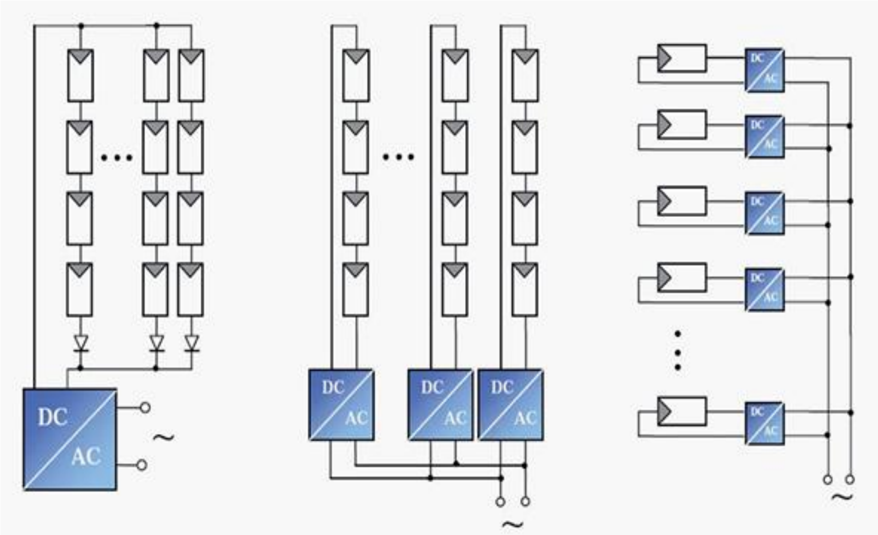


Figure.1.8 Typology of Grid-Connected PV Inverters [20]

1.5.6 Energy Storage in the Energy Transition

Energy storage systems are a crucial component in PV systems, as they play a vital role in mitigating fluctuations in solar energy production. They store excess electricity generated during periods of high solar irradiance and supply it during times of low production or increased demand. This supports power supply stability and enhances the overall system performance. Moreover, these systems allow for greater use of locally generated energy, reducing reliance on the conventional power grid and contributing to system flexibility and long-term sustainability[21].

1.5.7 Role and Function of Protection Diodes in Photovoltaic Systems

In PV systems, protection diodes are critical components that ensure reliability and operational efficiency. Two primary types of diodes are utilized: the bypass diode and the blocking diode, each serving a distinct function in optimizing system performance and preventing electrical faults[11].

- **Bypass Diode**

The bypass diode is connected in reverse polarity and in parallel with a group of PV cells. It is used to mitigate the adverse effects of partial shading or localized faults (Figure.1.9).
Under normal operating conditions, the diode remains non-conductive, allowing current to flow uninterrupted through the cells. However, when a cell or group of cells becomes shaded or

damaged, the diode provides an alternative path for the current, thereby preventing harmful reverse voltage on the affected cells, reducing electrical losses and ensuring the continuous generation of PV power [22], [23].

- **Blocking Diode**

The blocking diode is installed at the output of each PV string, ensures unidirectional current flow by allowing energy generated by the PV modules to flow toward the load or storage units, while simultaneously preventing reverse current from external sources such as batteries.

This function becomes particularly crucial in two scenarios: first, to prevent reverse currents arising from voltage imbalances between parallel-connected strings, which could otherwise cause open-circuit conditions or system damage; and second, to avoid battery discharge during periods of no generation, such as nighttime as shown in Figure.1.10.

The integration of protection diodes in PV systems constitutes a fundamental strategy for enhancing system reliability, optimizing energy utilization and minimizing performance degradation due to unwanted current flows. This approach ultimately contributes to extended component lifespan and stable system operation [22].

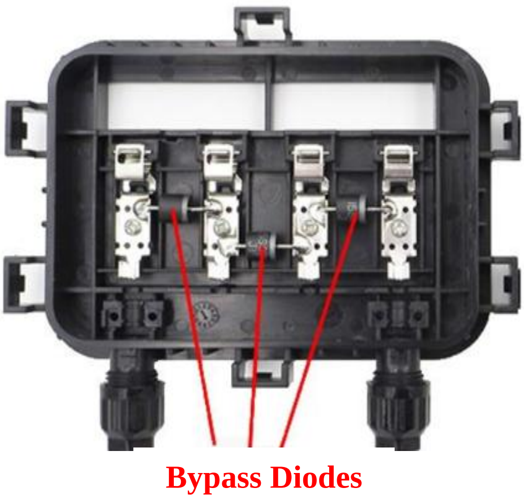


Figure.1.9 Bypass Diodes in a PV Junction Box [23]

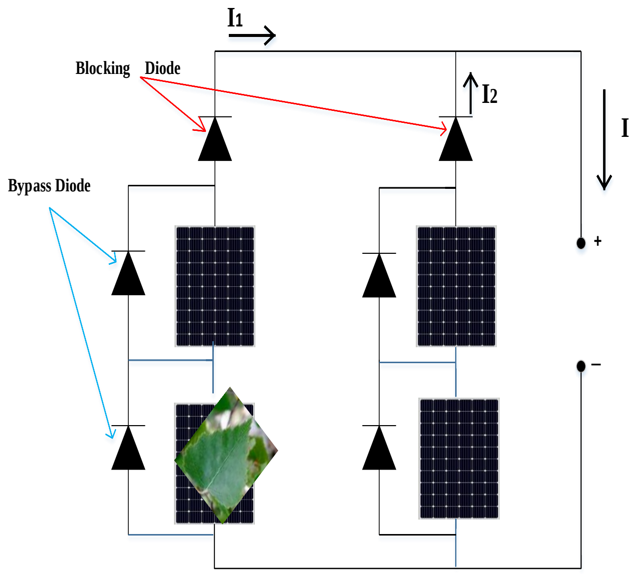


Figure.1.10 Configuration of a PV System with Bypass and Blocking Diodes

1.6 Classification of Photovoltaic Systems

PV systems are classified into three main categories based on their operational mechanisms and interaction with the electrical grid: stand-alone systems, grid-connected systems and hybrid systems. The nature of energy demand and surrounding environmental conditions primarily determines this classification.

- **Stand-Alone Photovoltaic Systems**

The Stand-alone PV systems operate independently of the public electricity grid and are particularly suitable for remote areas without access to the grid. These systems supply electrical loads such as lighting and water pumps directly with power [24] or store the generated energy in batteries for later use during periods of reduced or absent solar irradiation [25].

- **Grid-Connected Photovoltaic Systems**

Grid-connected PV systems are designed to operate in parallel with the public electricity grid. The electricity generated can be fed directly into the grid when production exceeds local consumption [26]. The central component of this configuration is the inverter, which converts the direct current produced by the PV modules into alternating current compatible with grid voltage and frequency standards. For safety, the system is designed to disconnect automatically during power outages to prevent back feeding.

- **Hybrid Photovoltaic Systems**

hybrid PV systems integrate multiple energy sources, such as solar, wind, diesel generators and battery storage, with flexible operational modes that allow switching between grid-connected and off-grid functions, depending on real-time demand and system conditions[27], [28].
This configuration ensures a continuous power supply despite solar intermittency; thereby enhancing the integration of renewable energy sources and supporting stable, adaptive energy management (see Figure.1.11).

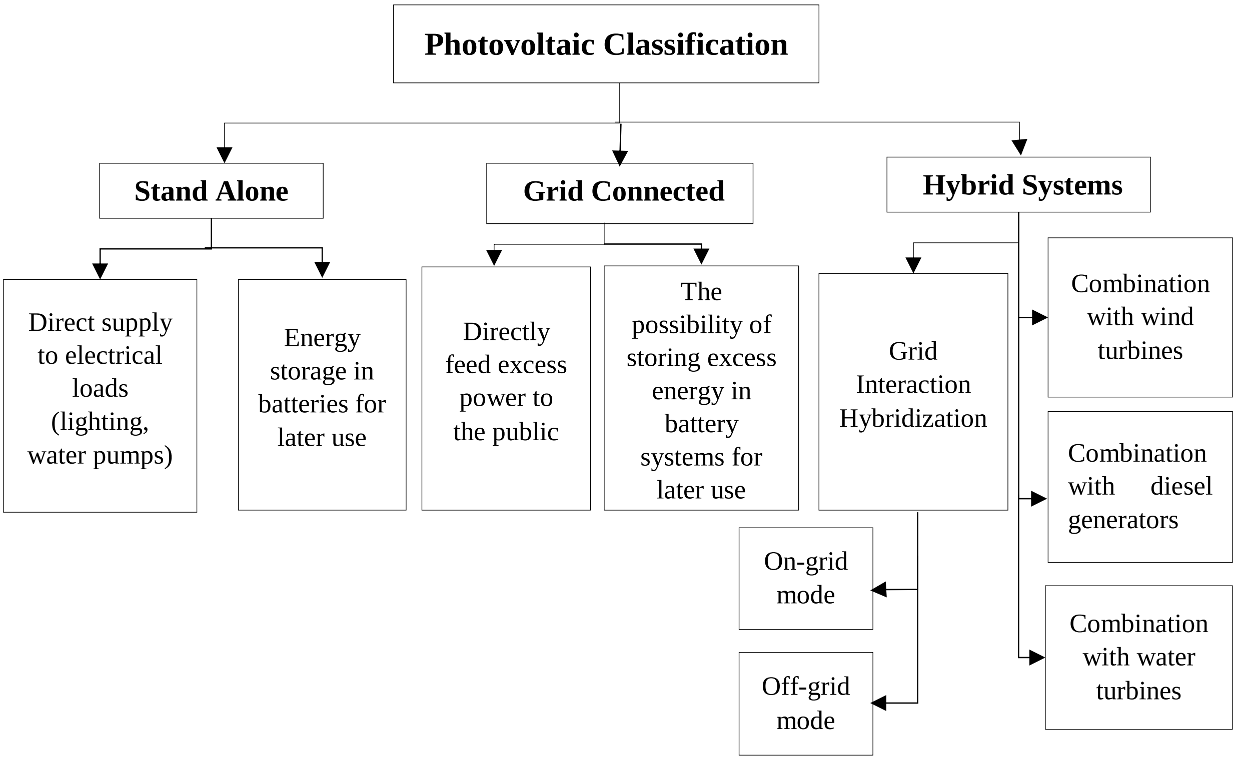


Figure.1.11 PV Systems' Classifications

1.7 Faults in Photovoltaic Systems

Despite the notable progress in the development of PV systems and their increasing adoption as a reliable source of clean energy, these systems remain susceptible to various types of faults during operation. Such faults often result from unfavorable climatic conditions or from design and installation errors, leading to a decline in overall system performance and energy conversion efficiency. Faults occurring in PV systems directly undermine their reliability and operational stability over both short and long-term operation. In certain instances, such faults can result in significant energy production losses or even complete system shutdowns. These faults are diverse in nature, encompassing electrical, mechanical and environmental issues, which necessitates a structured classification approach to facilitate accurate detection and effective mitigation. Accordingly, the PV system is divided into two principal sections: the direct current side and the alternating current side. Each section comprises specific components that are susceptible to particular types of faults, potentially leading to a decline in the overall efficiency and performance of the system [29], [30], (See Figure. 1.12).

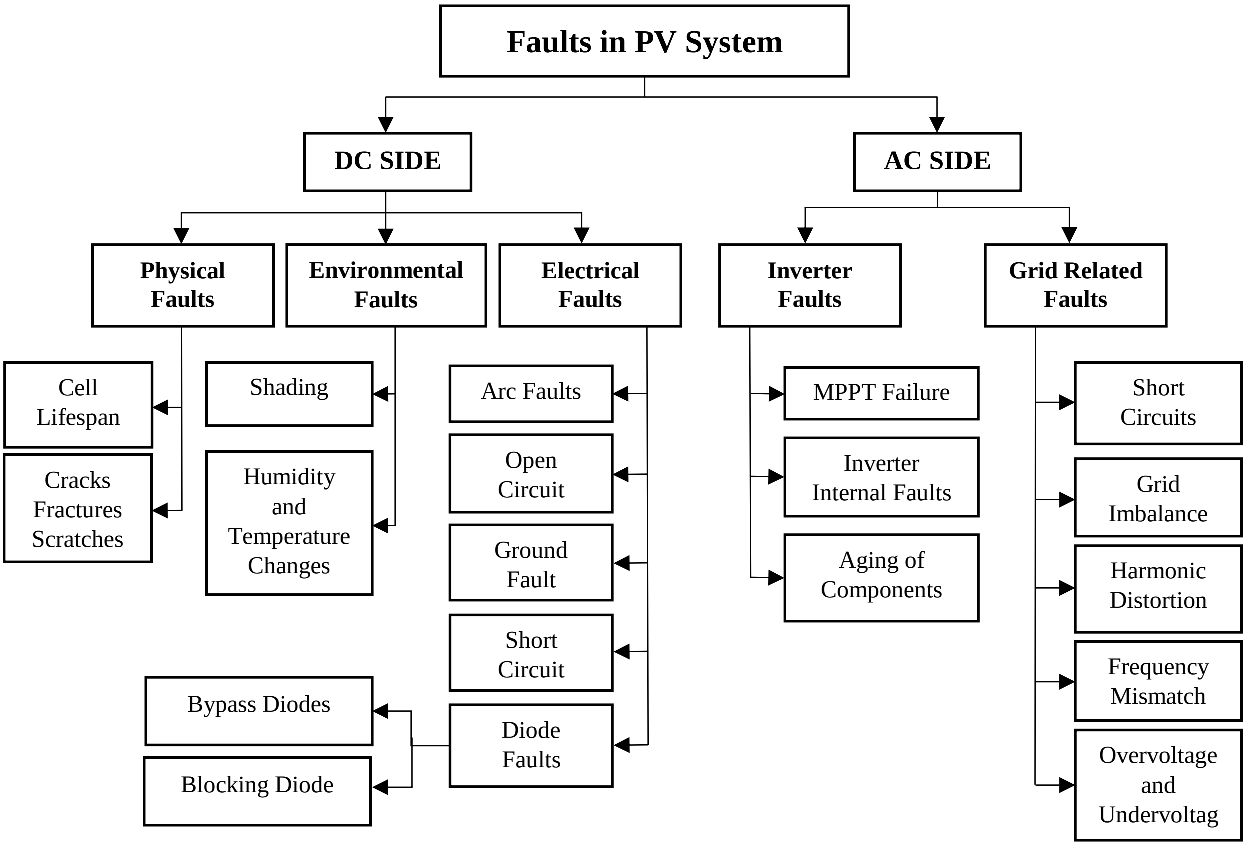


Figure.1.12 Difference Faults in PV System

1.7.1 Direct Current Side

The direct current side represents the initial stage of a PV system, where the electricity generated by the solar panels is processed before being converted into alternating current via the inverter. This section contains essential components that work together to capture, convert and distribute solar energy. However, it is susceptible to a range of operational issues, which are classified into three main categories of faults or malfunctions [31], [29].

A. Physical Faults

These faults are consider permanent because they are typically irreparable and cannot be restore to their original state. Correction is only achievable through the replacement of the affected components. Some examples of such faults include**:**

- **Cracks, Fractures, Scratches**

Cracks, fractures and surface scratches are types of physical damage that may occur due to natural phenomena such as high winds or hailstorms and in some cases, may originate from manufacturing defects. These imperfections compromise the integrity of the PV cell structure, disrupting the internal current flow and resulting in a permanent decline in energy conversion efficiency.

- **Cell Lifespan**

Over time, PV cells undergo natural aging, resulting in the gradual degradation of encapsulation and protective materials intended to shield them from environmental stressors such as moisture, ultraviolet radiation and dust. This ongoing deterioration weakens the protective function of these layers and contributes to a measurable decline in the cell’s energy conversion efficiency.

B. Environmental Faults

Environmental faults in PV systems refer to disruptions caused by external environmental factors that negatively influence system performance and efficiency. These faults are categorize as follows:

- **Shading**

Shading may be transient, caused by factors such as dust accumulation, fallen leaves, bird droppings, or surrounding vegetation whose shadows vary with time due to solar movement or wind (see Figure 1.13), or permanent (or quasi-permanent), resulting from fixed structures such as trees or buildings (Figure 1.14).

Shading leads to the formation of localized hot spots, creating thermal imbalances due to energy concentration in small regions of the photovoltaic cells, which results in elevated temperatures. These conditions reduce system performance, accelerate material degradation and may cause irreversible damage or complete failure of PV modules (Figure 1.15).

Figure.1.13 Vegetation Induced Partial Shading on PV Modules [12]

Figure.1.14 Permanent Shading in PV Systems [32]

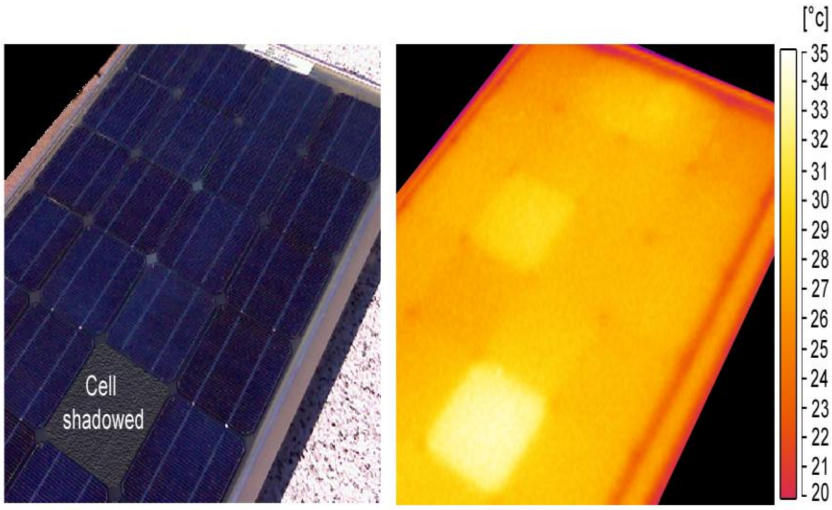


Figure.1.15 Discrepancy Due to Cell Shading [33]

- **Humidity and Temperature Changes**

High humidity accelerates the corrosion of internal conductive pathways and other sensitive components within the PV system. In addition, significant temperature fluctuations induce repeated thermal expansion and contraction of system materials, which can lead to the formation of cracks in solar cells. This structural degradation negatively affects the system's operational efficiency and reduces its overall lifespan.

C. Electrical Faults

PV systems are susceptible to various types of electrical faults that can adversely affect their functionality and long-term durability. We can classify these electrical malfunctions into several main categories:

- **Ground Fault**

A ground fault is a type of electrical malfunction occurring in PV systems characterized by an unintended connection between a conductor and the ground. This issue may arise due to various factors, including the gradual corrosion of insulating materials that separate electrical conductors from the ground, which can result in unintended conduction. Additionally, improper installation practices or incorrect cable connections can cause accidental contact between conductors and the ground. Dielectric damage may also lead to current leakage to the ground. These undesirable currents can compromise the performance of the PV system, pose risks of electric shock to individuals and potentially result in fires.

- **Short Circuit**

A short circuit occurs when two conductors with different voltages come into contact, creating a path of low resistance that allows a significant current to flow. This excessive current can exceed the capacity of the wiring and other components, resulting in intense heating, system damage and potentially causing a fire. The causes of short circuits include the degradation of insulation materials over time, exposure to moisture, overheating, or mechanical damage such as cracking or tearing.

- **Open Circuit**

An open circuit is a type of fault defined by the interruption or disconnection in the path of current flow between conductors, which obstructs the flow of electric current. One primary cause of open circuits is exposure to high temperatures, which can induce high currents or environmental factors, such as humidity, that lead to corrosion over time. Additionally, excessive and repetitive bending of the conductors can result in broken or interrupted electrical connections.

- **Arc Fault**

An arc fault is a malfunction in the electrical system that occurs when the insulation around the wires is compromised, allowing an electrical arc to form. This damage allows the current to jump between conductors through the air, generating extremely high heat and electrical sparks, which can lead to a fire.

D. Diode Faults

Faults in diodes, whether open-circuit or short-circuit, can adversely affect the performance and efficiency of PV systems. In the case of an open-circuit fault, bypass diodes fail to allow current to bypass damaged cells, thereby reducing chain efficiency and causing energy losses. Similarly, open-circuit faults in the blocking diode disrupt the flow of energy from solar cells to batteries, impairing system functionality (Figure 1.16) . On the other hand, short circuit faults in bypass diodes result in the current bypassing even functional cells, thereby diminishing energy output and increasing the risk of hotspots, which can lead to permanent cell damage. In an anti-return diode, short circuit faults allow batteries to discharge back into solar panels during periods of low illumination, such as at night, potentially overheating the diode and compromising the system's overall performance.

Figure.1.16 Diode Fault [34]

1.7.2 Alternating Current Side

Several types of faults may arise on the alternating current side as shown in Figure.1.16, with the most significant being the following[31] ,[29]:

A. Grid-related Faults

- **Grid Imbalance**

Grid imbalances arise due to variations in frequency and voltage, uneven distribution of electrical loads across phases in a three-phase system, or the presence of harmonic distortions in the electrical waveform. These factors can compromise the stability and efficiency of the grid.

- **Short Circuits**

Short circuits within AC circuits can disrupt the entire system, potentially causing severe damage to components such as adapters, cables and other electrical devices.

- **Harmonic Distortion**

This issue arises from a mismatch between the AC waveform generated by the inverter and that of the electrical grid, resulting in degraded electricity quality and operational inefficiencies.

- **Overvoltage and Under Voltage**

Overvoltage, where the voltage exceeds permissible levels, can result in damage to connected devices. Conversely, under-voltage, when the voltage falls below the required thresholds, can hinder proper device functionality and compromise system performance.

- **Frequency Mismatch**

Discrepancies between the frequency of the electrical current produced by the system and that of the grid, often due to inverter malfunctions, can lead to significant inefficiencies. This mismatch frequently results in the generation of harmonics, further reducing the overall system performance.

B. Inverter faults

- **Internal Faults in the Inverter**

Internal issues may arise in the inverter's components, such as malfunctioning fans, which can lead to overheating at unsafe levels or electrical problems [35],[36]. These faults often cause the inverter to shut down, resulting in the system failing to generate the expected amount of electricity.

- **Failure of Maximum PowerPoint Tracking**

This occurs when the inverter cannot track and achieve the maximum power point of solar energy, resulting in inefficient operation of the PV system and a reduction in energy output [37],[38].

- **Component Aging**

Over time, the efficiency of the inverter's components decreases due to wear and tear, which increases the likelihood of malfunctions. This gradual degradation can lead to frequent failures and a significant reduction in overall system performance.

1.8 Conclusion

This chapter has provided a comprehensive review of PV systems, beginning with the fundamental cell technologies and their various types, followed by the core design principles and diverse operating modes. It further presented a structured classification of PV system configurations and highlighted the primary faults that can impact system efficiency and reliability. Through this comprehensive presentation, a coherent conceptual framework has been established, facilitating a clear understanding of operational challenges and highlighting the critical role of effective fault diagnosis in maintaining system performance. Building on this foundation, the subsequent chapter focuses on the essential terminologies and methodological stages involved in fault detection and analysis. It also offers an overview of recent advancements that have enhanced both the efficiency and accuracy of diagnostic techniques. These developments contribute to the formulation of more robust and reliable diagnostic strategies, thereby supporting the safe and efficient operation of PV and industrial systems.

Chapter 2: photovoltaic Diagnosis Methods

2.1 Introduction

Based on the fundamental principles of PV systems, this chapter focuses on analyzing the methods and techniques employed in fault diagnosis to ensure reliable and efficient operation, while preventing performance degradation and premature aging of components due to undetected faults. It begins by presenting the theoretical foundations of fault diagnosis, analysis and characterization in industrial systems, followed by a structured review of the diagnostic process and its sequential stages, from fault detection to decision-making. Various diagnostic approaches are discussed, including model-based, data-driven and knowledge-based methods. The chapter concludes with the hybrid approach, which will be applied in the subsequent practical chapters. This approach combines ML algorithms with human expertise, leveraging the strengths of both to enhance diagnostic accuracy and provide results that are interpretable and reliable.

2.2 Diagnostic Concepts and Terminology

In order to establish a clear understanding of the diagnostic processes applied to PV and industrial systems, it is essential first to define the fundamental terms and concepts used throughout this chapter. Diagnostic terminology serves as the backbone for effective communication and accurate interpretation of fault detection results.

- **Diagnostic**

A breakdown refers to a failure within a system, which may arise from abnormal conditions or the malfunction of system components. In some cases, the system may continue to operate, but the results may be unsatisfactory or inaccurate. To prevent such issues before they occur, diagnostic methods serve to identify faults, determine their sources, monitor their progression and, if necessary, repair or replace the damaged components.

- **Prediction**

Prediction is the process of anticipating a failure before it occurs and estimating the remaining operational lifespan of a system. This process relies on models that analyze available data based on the validity of assumptions, historical information and current conditions. The goal is to make informed decisions for the future and prepare for anticipated events in order to prevent system failure.

- **Fault**

A fault is defined as any abnormal deviation in the performance of a system or the behavior of one of its components. It is often the result of factors such as improper system design, irregular or insufficient maintenance, or excessive operational stress. Faults may manifest in various forms, including unexpected variations in sensor readings, reduced system efficiency, or the sudden failure of specific components.

- **Failure**

Failure refers to the condition in which a system or one of its components loses the ability to perform its intended functions in accordance with the required specifications or standards. It often results in a partial or complete disruption of system operations and necessitates immediate intervention to restore normal functionality.

- **Malfunction**

This concept refers to recurrent deviations from the expected system performance due to operational errors, inconsistent interactions among system components, or faults in software or electronic circuits. Such malfunctions often lead to inaccurate outcomes or unstable performance, requiring prompt analysis and corrective action to restore system reliability.

2.3 Diagnostic Process in Photovoltaic and Industrial Systems

The diagnostic process is a fundamental element in ensuring the optimal performance and reliability of both PV solar energy systems and industrial systems. It plays a crucial role in fault detection, root cause analysis and decision-making, thereby maintaining continuous and efficient system operation. The diagnostic approach follows a structured methodology, generally divided into several key stages, including data acquisition, signal processing, fault detection and systematic decision-making strategies. These stages facilitate a comprehensive understanding of system behavior and enable the timely implementation of corrective measures to mitigate performance degradation. In the case of PV systems, fault diagnosis is particularly significant due to the complexity of environmental factors affecting energy generation. Various diagnostic methods are employed for fault identification and localization, providing a scientific framework for analyzing critical failures and understanding their impact on overall system performance. Similarly, in industrial systems, failure diagnosis follows a structured approach to ensure operational efficiency and system reliability. While no single diagnostic method can address all possible failure scenarios, a systematic and data-driven approach enhances fault detection accuracy and enables effective mitigation strategies across different application domains as represent in Figure.2.1.

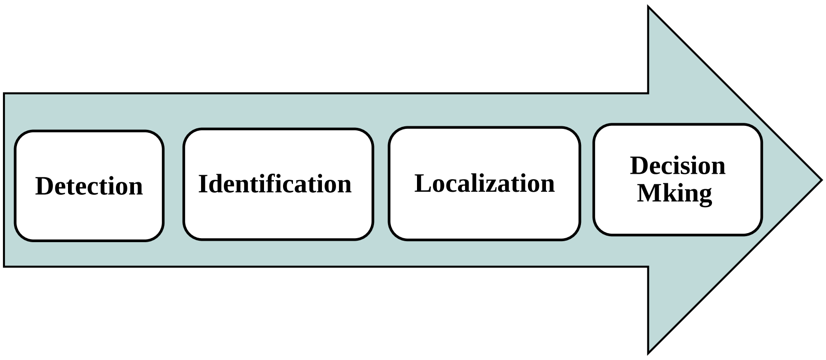


Figure.2.1 Steps of Failure Diagnosis in Industrial Systems

- **Detection**

The first step involves determining whether the system is operating normally or if a failure has occurred. This allows for the immediate identification of any deviation from the optimal state of the system, thereby signaling the presence of a malfunction.

- **Localization**

Once a problem is recognized, it is crucial to identify the source of the anomaly with accuracy. This step involves identifying the specific components (such as sensors, actuators, etc.) responsible for the failure, thus enabling targeted intervention.

- **Identification**

The identification phase aims to assess the extent of the failure and analyze its evolution over time. It provides a deeper understanding of the impact of the malfunction on the system and helps in considering appropriate corrective actions.

- **Decision-Making**

Finally, after identifying the defect, decisions must be should regarding the actions to restore the system's performance. This step helps minimize the negative effects of the failure and ensures the system's long-term proper functioning.

2.4 Classification of Photovoltaic Fault Diagnosis Approaches

In recent years, the rapid development of renewable energy systems and advanced industrial technologies has significantly increased the demand for accurate and reliable fault diagnosis. The growing complexity of system architectures, the integration of intelligent controllers and the need for continuous operation have made early fault detection essential for maintaining system performance, minimizing unexpected failures and ensuring operational safety.

Effective diagnostic processes rely on a thorough understanding of system behavior, clear definitions of key terminologies and a systematic approach to fault identification, analysis and interpretation. Accordingly, this chapter presents an integrated conceptual framework for diagnostic methodologies applied in PV and industrial systems. It begins with fundamental definitions that establish the theoretical foundation of fault diagnosis, followed by a structured description of diagnostic stages and their role in detecting deviations from normal operating conditions. Furthermore, this chapter reviews modern trends in diagnostic techniques, including advanced mathematical, structural and causal reasoning, data-driven methods and artificial intelligence-based approaches.

These developments have significantly enhanced fault prediction capabilities, improved diagnostic accuracy and accelerated response times, making fault diagnosis a critical component of intelligent and autonomous industrial systems. Diagnostic methods in PV and industrial systems can generally be categorized into four main paradigms: model-based diagnosis, data-driven diagnosis, knowledge-based methods and hybrid approaches, as represented in Figure 2.2.

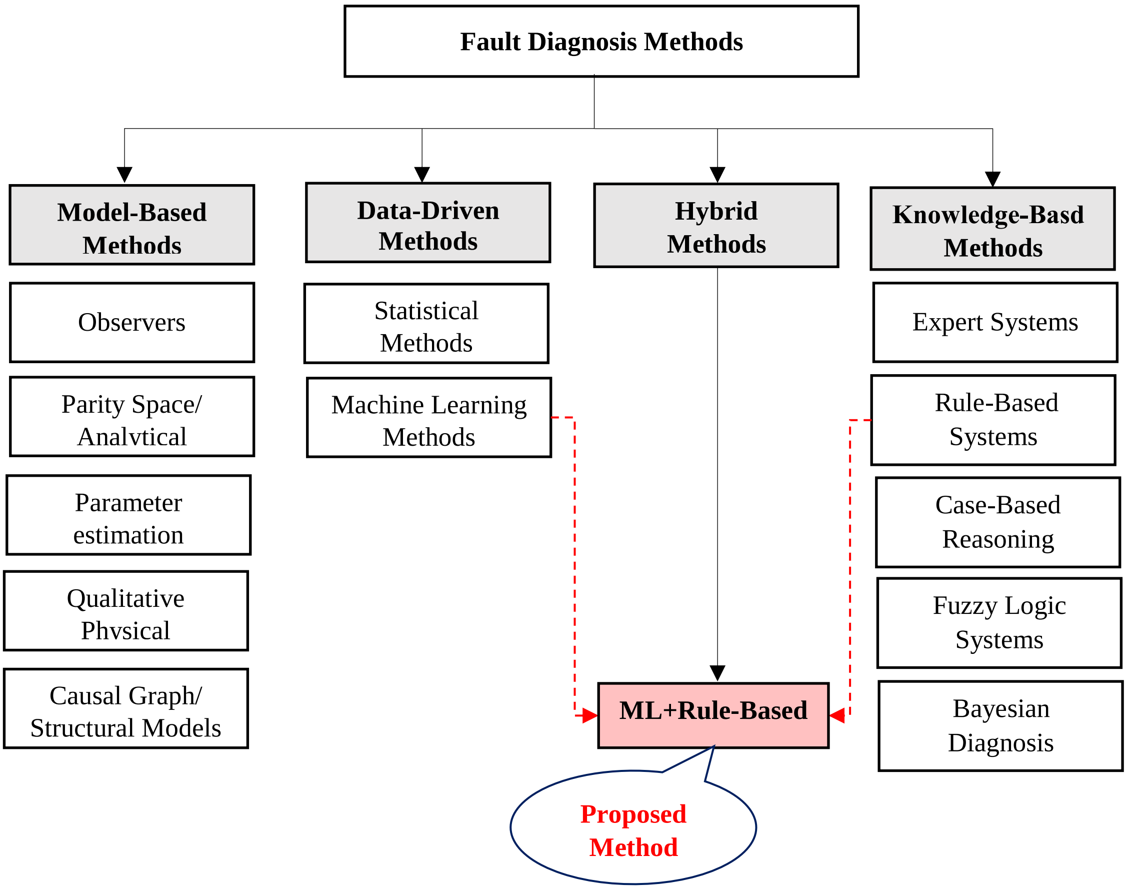


Figure.2.2 Classification of Industrial Fault Diagnosis Methods

2.4.1 Model-Based Fault Diagnosis

Model-based diagnosis relies on constructing a mathematical or physical representation of the system that accurately describes its dynamic behavior under nominal operating conditions[39]. By comparing the expected outputs generated by the model with real measurements, discrepancies referred to as residuals are computed and analyzed to identify and isolate faults. This diagnostic model includes different types of approaches:

A. Observers

An observer is a mathematical construct used to estimate internal states that cannot be measured directly, based on a dynamic model and measurement data[40]. It generates the system's predicted state and compares it with actual responses to produce a residual signal that reflects the level of agreement between the model and reality. Any significant deviation in this signalserves as a clear indicator of a fault, making observers an effective tool for early fault detection related to internal states, such as sensor and actuator faults or changes in system properties.

B. Parity Space / Analytical Redundancy

This approach relies on deriving analytical redundancy relations from the system model, which hold true only under normal operating conditions. A residual vector is formed by comparing the expected relations with actual measurements. Any deviation in these residuals is considered evidence of a fault in one of the system elements. This methodology is suitable for fault detection without the need to know the internal states, relying solely on available inputs and measurements[41].

C. Parameter Estimation

This method focuses on dynamically estimating the system’s physical or operational parameters using a mathematical model and measurement data[42]. The estimated values are compared with the nominal expected values of the system. Any significant change in these parameters is typically interpreted as resulting from a fault or gradual degradation of components, making this approach suitable for detecting slow faults and monitoring changes in physical properties over time.

D. Qualitative Physical Models

Qualitative physical models represent causal and structural relationships between physical variables without relying on precise numerical equations[43]. System behavior is described using general rules and concepts such as increase/decrease trends and causal directions. This approach allows the detection of abnormal performance patterns even in the absence of a detailed mathematical model, making it suitable for complex or nonlinear systems that are difficult to model quantitatively.

E. Causal Graph / Structural Models

These models represent the internal structure of a system as a causal network linking components and their interactions. The fault source can be inferred by tracing influence paths within the causal graph. This methodology enables the use of cause-and-effect reasoning to identify the fault location and its propagation through the system structure, making it suitable for large and complex systems that rely on interconnected and interdependent processes[44].

2.4.2 Data-Driven Fault Diagnosis

Data-Driven Diagnosis represents a contemporary diagnostic paradigm that derives knowledge directly from operational data, rather than relying on explicit physical or mathematical models of the system[45],[46]. This approach is particularly advantageous for complex systems. By leveraging both historical and real-time operational data, data-driven methods enable reliable detection, identification and prediction of abnormal operating conditions. Within this approach, a variety of methods exist, including statistical methods and machine learning-based techniques, which enhance diagnostic accuracy and reliability.

A. Statistical Methods

Statistical diagnostic methods are based on modeling and analyzing the statistical properties of operational data to detect structural changes or deviations from normal behavior[47]. These methods include:

- Principal Component Analysis , which reduces data dimensionality while preserving the most informative features, thereby facilitating the detection of abnormal patterns.
- Statistical classifiers, which apply probabilistic decision-making techniques to distinguish normal operating conditions from faulty ones.

These methods are characterized by their strong capability to handle high-dimensional data and to extract accurate diagnostic information without requiring an explicit physical model of the system.

B. Machine Learning

Machine Learning represents a fundamental component of quantitative data-driven diagnosis, as it focuses on the development of predictive models trained on historical or real time data to infer the operational state of a system[48],[49]. Machine learning techniques are generally classified into two main paradigms: supervised learning and unsupervised learning, as illustrated in Figure.2.3.

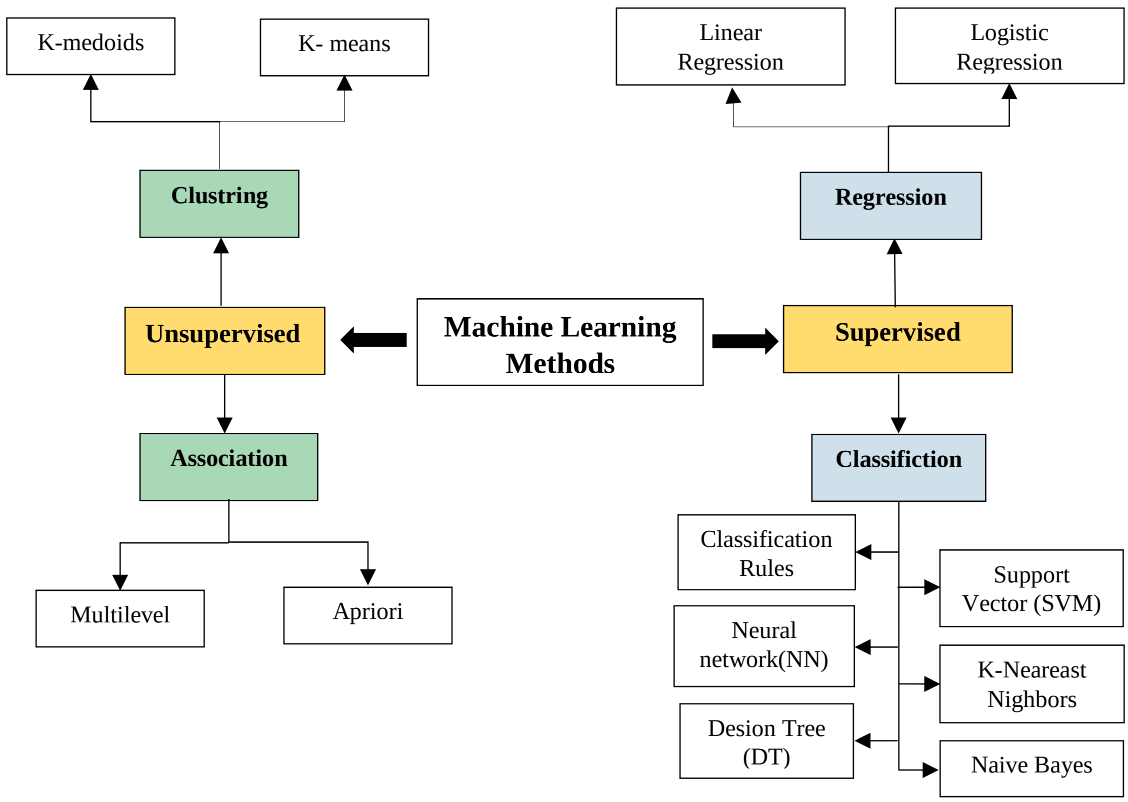


Figure.2.3 Machine Learning Methods [48],[49]

1. Supervised Learning

Supervised data mining is applied when labeled datasets are available and the objective is to predict predefined outcomes[50]. This approach is closely related to machine learning, as it relies on guided learning from annotated examples. It involves deriving conceptual descriptions of classes from pre-classified data instances. The two principal branches of supervised data mining are classification and regression.

1.1. Classification

Classification is a fundamental supervised data mining technique that assigns data instances to predefined categories[51]. The process involves constructing a predictive model by analyzing a labeled training dataset, allowing the model to learn underlying patterns and relationships. Once trained, the model can be applied to categorize new, unseen instances based on the acquired knowledge. Classification thus provides a systematic framework for differentiating between distinct data classes or concepts, enabling accurate predictions and informed decision making in various data-driven applications .

There are several algorithms used for classification, including the following:

- **Naive Bayes**

NB is a supervised learning algorithm based on Bayes' theorem, assuming conditional independence among features given the class label[52].

- **Support Vector Machines**

SVMs construct an optimal separating hyperplane by maximizing the margin between classes, using kernel functions to model non-linear decision boundaries[53].

- **Decision Tree**

DTs are intuitive and widely used data mining techniques that generate explicit classification rules[54]. They can effectively handle heterogeneous data, missing values and non-linear relationships. A decision tree has a flowchart-like structure, where nodes represent attribute tests, branches represent outcomes and leaves represent class labels. One of their main advantages is the ease of interpreting the generated rules.

- **Neural Networks**

NNs are computational models inspired by the structure of the human brain. They use statistical learning methods to identify patterns in data[55]. Although they achieve high performance, their models are often complex, less interpretable and sensitive to data representation. They are particularly effective in pattern recognition and tasks involving noisy or continuous data.

- **Classification Rules**

Rule based classification is a technique that uses a set of human-readable IF–THEN rules to assign class labels to data instances. Each rule consists of two main parts: the antecedent (IF part), which defines one or more conditions and the consequent (THEN part), which specifies the resulting class label[56],[57].

The antecedent typically includes multiple terms, each formed by a feature, a comparison operator and a value. When multiple terms are present, they are usually connected using logical operators such as AND. A data instance is classified when it satisfies the conditions of a specific rule. Rule-based classifiers are valued for their transparency and interpretability, as the generated rules can be easily understood and validated by domain experts.

- **k-Nearest Neighbors**

KNN classifies instances based on similarity to their nearest neighbors, grouping data points according to user-defined distance or similarity measures[58].

1.2. Regression

Regression methods predict a continuous dependent variable using one or more independent variables. Common types include linear, multiple, polynomial, nonparametric and robust regression[59].

- **Logistic Regression**

Logistic regression estimates the probability of categorical outcomes based on predictor variables[60].

- **Linear Regression**

Linear regression models the linear relationship between dependent and independent variables and is valued for its interpretability[61].

2. Unsupervised Learning

Unsupervised learning focuses on discovering patterns without labeled data, primarily through clustering and association rule mining[50].

2.1 Clustering

Clustering group's similar objects together while maximizing differences between groups. Common algorithms include K-means and K-medoids[62].

- **K-means**

K-means partitions data into K clusters by minimizing the within-cluster sum of squared distances between data points and their corresponding cluster centroids, which are computed as the mean of points in each cluster. [63].

- **K-medoids**

K-medoids partitions data into K clusters by selecting actual data points (medoids) as cluster centers and minimizing the sum of dissimilarities between points and their assigned medoids, making it more robust to outliers than K-means. [64].

2.2. Association

Association rule mining identifies frequent relationships and co-occurring items in large datasets, widely used in marketing and data analysis[65].

- **Apriori**

Apriori is an algorithm used to discover frequent item sets and relationships among variables[66].

- **Multilevel Association**

Multilevel association rules use multiple minimum support and confidence thresholds across different levels of abstraction.

2.4.3 Knowledge-Based Fault Diagnosis

Knowledge-Based Diagnosis (KBD) is an approach to fault diagnosis that relies on explicitly represented domain knowledge and expert experience to analyze system behavior and identify faults or abnormal operating conditions[67]. Rather than depending solely on mathematical models or large volumes of data, this approach exploits expert knowledge encoded in logical rules, structured knowledge representations, or reasoning mechanisms.
A key characteristic of knowledge-based diagnosis is its interpretability, as the diagnostic decisions can be logically explained through the underlying rules or reasoning paths. Furthermore, this approach is particularly suitable for systems that are difficult to model accurately, operate under limited or incomplete data, or involve complex cause–effect relationships.

A. Expert systems

Expert systems are an advanced form of rule-based systems designed to emulate the reasoning process of human experts in fault diagnosis[68].They typically consist of a knowledge base containing expert rules and facts and an inference engine that applies logical reasoning to derive diagnostic conclusions. Expert systems are widely used in industrial and engineering applications where explainable and reliable diagnosis is required .

B. Rule-based diagnostic systems

Rule-based diagnostic systems use a set of logical IF–THEN rules to associate observed symptoms with possible faults. Each rule represents expert knowledge that links specific conditions or measurements to a diagnostic conclusion[69]. This type of system is especially effective when fault patterns are well understood and can be clearly articulated by domain experts.

C. Fuzzy Logic Systems

Fuzzy knowledge-based diagnostic systems extend classical rule-based systems by incorporating fuzzy logic to manage uncertainty and imprecision[70]. Instead of using crisp thresholds, these systems employ linguistic variables (e.g., "high", "moderate", "low") and fuzzy membership functions to represent expert knowledge more realistically. This approach is particularly suitable for fault diagnosis in complex systems where measurements are noisy or boundaries between normal and faulty states are not well defined.

D. Case-Based Reasoning

Case-Based Reasoning diagnosis is founded on the principle that similar problems have similar solutions[71]. In this approach, new fault situations are diagnosed by comparing them with previously solved cases stored in a case library. The most similar past case is retrieved and adapted to fit the current situation, making this approach particularly effective in systems with rich historical fault records.

E. Bayesian Diagnosis

Bayesian Knowledge-Based Diagnosis: Methods that utilize Bayesian Networks or similar probabilistic graphical models to represent expert knowledge about causal relationships and uncertainty between symptoms and faults[72].

These methods integrate conditional probabilities and prior knowledge to infer the likelihood of fault states, making them especially useful in systems with incomplete or uncertain information.

2.4.4 Hybrid Fault Diagnosis Method

Within the state-of-the-art of advanced fault diagnosis methodologies, the proposed model adopts a hybrid approach that combines ML techniques with rule-based systems, leverag ingthestrengths of both [73]. This approach is particularly suitable for industrial applications, where the choice of diagnostic method depends on factors such as system complexity, data availability, interpretability requirements, computational resources and safety criticality.
Highly complex or non-linear systems benefit from data-driven methods capable of uncovering hidden patterns, while safety-critical applications require the transparency and determinism offered by rule-based approaches. Although rule classification can be implemented solely within ML frameworks, relying exclusively on data-driven techniques presents challenges, including the need for extensive datasets to extract accurate rules and the difficulty of interpreting some automatically inferred rules. By integrating rules derived from human expertise, the hybrid model enhances reliability, enables rapid and effective handling of known faults and provides logically interpretable results. Recognizing that no single diagnostic method is universally optimal, the hybrid approach capitalizes on the synergy of multiple techniques to overcome individual limitations. Machine learning contributes its ability to analyze large and complex datasets, uncover subtle patterns and adapt to new conditions, while rule-based systems provide clear reasoning, incorporate accumulated expert knowledge and satisfy interpretability and safety requirements.Through this integration, the hybrid model aims to achieve state-of-the-art diagnostic accuracy and reliability while offering a flexible and interpretable framework, laying the groundwork for its practical evaluation in Chapter Five.

2.5 Conclusion

This chapter has presented a comprehensive and structured overview of fault diagnosis methodologies applied to photovoltaic and industrial systems. By first introducing the fundamental concepts and terminology of system diagnosis, the chapter established a clear understanding of the diagnostic process, including fault detection, isolation and evaluation stages. This systematic framework highlights the critical role of diagnosis in ensuring the reliability, safety and operational continuity of PV installations.
The analysis of diagnostic approaches revealed that model-based methods offer strong physical interpretability and analytical rigor but suffer from scalability issues and reduced effectiveness when applied to complex, nonlinear and large-scale PV systems. Data-driven methods,

particularly those based on machine learning, demonstrated superior adaptability and fault detection accuracy under diverse operating conditions; however, their black-box nature often limits transparency and trust in industrial decision-making. Knowledge-based methods, including expert systems and rule-based reasoning, provide high interpretability and domain relevance but are constrained by knowledge acquisition complexity and limited generalization capability.

To overcome the individual limitations of these approaches, hybrid diagnostic methods have emerged as a promising paradigm, combining physical insight, data adaptability and expert knowledge. Nevertheless, existing hybrid frameworks still face challenges in achieving an optimal balance between diagnostic accuracy, computational efficiency and interpretability.

Overall, this chapter highlights a critical research gap: the need for intelligent diagnostic frameworks that are both high-performing and explainable, particularly for industrial PV systems characterized by dynamic behavior and high-dimensional data. This observation justifies the adoption of metaheuristic optimization techniques, which offer powerful search capabilities for complex problem spaces, and their integration with XAI models.

Consequently, the concepts and limitations discussed in this chapter provide the theoretical foundation and motivation for the subsequent chapters, which focus on the development of optimized, interpretable and robust diagnostic approaches based on Grey Wolf Optimization variants and explainable classification strategies for photovoltaic fault diagnosis.

Chapter 3: Metaheuristic Optimization

3.1 Introduction

Metaheuristic algorithms play a pivotal role in addressing complex problems that are challenging to solve using conventional methods, particularly large-scale or nonlinear problems. These algorithms are inspired by natural phenomena and biological processes, such as natural selection, swarm behavior and collective intelligence, enabling the exploration of vast search spaces and the extraction of near-optimal solutions even under limited information or incomplete knowledge of the problem's initial conditions.

This chapter begins by presenting the foundations of metaheuristic algorithm classification, offering an overview of various algorithm types, their characteristics and operational mechanisms. Such classification helps in identifying the most suitable approaches for different problem domains. The focus then shifts to the GWO , examining its working principles and methodology. GWO is chosen for its simplicity of implementation, rapid convergence, high efficiency in finding optimal solutions and ability to handle complex and multi-objective problems, making it particularly suitable for practical optimization tasks.

Finally, the chapter presents a practical application example, illustrating the use of GWO to enhance signal quality in PV energy systems. In this example, the Moving Average Filter is integrated with GWO to optimize the filter parameters, aiming to minimize MSE between the original and filtered signals. This integration maintains essential signal characteristics while effectively reducing unwanted noise, thereby improving system efficiency, stability and reliability.

3.2 Classification of Metaheuristic Algorithms

The classification of metaheuristic algorithms is fundamentally grounded in several methodological criteria that characterize the structural and behavioral properties of each algorithm. A primary basis for this classification is the source of inspiration, which distinguishes nature-inspired algorithms such as evolutionary and swarm intelligence methods from non-nature-inspired approaches that rely on purely heuristic or mathematical search strategies[68]. Another essential criterion concerns the number of solutions processed during the optimization process; some algorithms operate on a single solution and iteratively refine its neighborhood, whereas others employ a population-based framework, enabling broader exploration of the search space. Additionally, metaheuristics are classified according to their reliance on deterministic versus stochastic mechanisms, with most contemporary approaches incorporating probabilistic components to enhance global exploration and avoid premature convergence. A further distinction is made based on the trajectory of the search, differentiating algorithms that follow a single search path from those that utilize multiple parallel trajectories. Finally, classification also considers whether the algorithm performs local search, focusing on the immediate vicinity of current solutions, or global search, exploring the entire solution domain. Collectively, these criteria provide a comprehensive foundation for analyzing metaheuristic algorithms and determining their suitability for addressing complex optimization problems.

Figure.3.1 illustrates the primary classifications, which are widely recognized in the literature:

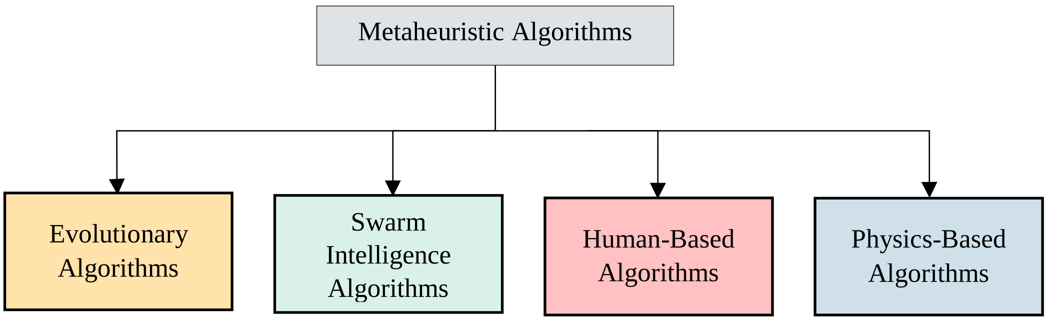


Figure.3.1 Metaheuristic Algorithm Classification[74]

A. Evolutionary Algorithms

Evolutionary Algorithms (EAs) are a class of metaheuristic techniques inspired by the principles of natural evolution, including selection, reproduction and mutation [74],[75]. They operate by generating a population of candidate solutions and iteratively improving it through selection, crossover and mutation to achieve increasingly optimal solutions. Common types of EAs include Genetic Algorithms (GA)[76], which rely on natural selection and genetic inheritance and are widely used in optimization, classification and clustering; Evolutionary Programming (EP)[77],[78], a stochastic method primarily applied to function optimization; Evolution Strategies (ES)[79], known for their robustness in handling noisy fitness functions and effectiveness in continuous optimization; and Differential Evolution (DE)[80], which uses differential mutation and is recognized for its efficiency and reliability in solving complex problems. EAs have been successfully applied in diverse fields such as engineering, finance, biology and computer science, particularly for problems characterized by high dimensional search spaces, nonlinearity and uncertain or imprecise fitness evaluations.

B. Physics-Based Algorithms

Physics-based metaheuristic algorithms constitute optimization techniques inspired by physical principles, such as energy, force and motion, to guide the search for optimal solutions in complex problem spaces [74],[81]. Prominent examples include the Gravitational Search Algorithm (GSA), which models candidate solutions as particles interacting via gravitational forces [82], Electromagnetic Field Optimization (EMO), where solutions behave as charged particles influenced by electromagnetic interactions [83] and Quantum-inspired Evolutionary Algorithms (QEA), which represent solutions as quantum states manipulated through quantum operators [84]. These algorithms have demonstrated high efficacy in engineering design, scheduling and image processing, particularly for high-dimensional and multi-objective optimization problems.

C. Human-Based Algorithms

Human-based metaheuristic algorithms are a class of optimization techniques that integrate human intelligence, knowledge and experience into the search for optimal solutions[74],[85]. These algorithms leverage expert guidance to improve the efficiency and quality of the optimization process. Notable examples include the Expert-Guided Evolutionary Algorithm (EEA)[86], which combines evolutionary processes with expert knowledge to steer the search toward promising solutions; the Human Guided Search (HGS)[85] algorithm, in which human experts direct the search toward favorable regions of the solution space; Interactive Evolutionary Computation (IEC)[86], which incorporates feedback from human evaluators to refine candidate solutions; and Human in the Loop Optimization (HILO), where humans interact with the algorithm in real-time to provide guidance and evaluation[87]. By incorporating human insight, these approaches have been shown to enhance the effectiveness of optimization, often achieving high quality solutions more efficiently than purely automated techniques.

D. Swarm Intelligence Algorithms

Swarm intelligence is a branch of artificial intelligence inspired by the collective behavior of social animals such as ants, bees, birds and wolves. It involves decentralized, self-organized systems in which individual agents interact with each other and with their environment to achieve common goals. Swarm-based algorithms are generally founded on the principles of decentralization, self-organization and adaptation, allowing the system to respond dynamically to environmental changes and emergent patterns.Well-known examples include Ant Colony Optimization (ACO)[73],[88], based on ant foraging; Particle Swarm Optimization (PSO)[89], inspired by bird flocking; Bee Colony Optimization (BCO)[90] and the Artificial Bee Colony (ABC) algorithm[91], modeled on honeybee foraging; the Firefly Algorithm (FA)[92], derived from the flashing communication of fireflies; and the GWO , which simulates the social hierarchy and cooperative hunting strategies of grey wolves[93] . These algorithms have been effectively applied in engineering, finance, biology and computer science, particularly for complex optimization problems that are difficult to solve using traditional methods[94][95]. After reviewing the major swarm intelligence algorithms and their underlying principles including those inspired by the behavior of ants, birds, bees, fireflies and other social organisms the significant role of these approaches in producing effective solutions to complex optimization problems becomes evident. Within this diverse landscape, the GWO stands out as one of the most modern and promising algorithms, owing to its simplicity of implementation, rapid convergence and high efficiency in achieving near optimal solutions even in highly complex environments. Building on these advantages, the following section provides a detailed examination of the GWO Algorithm, its operational mechanisms, fundamental components and the characteristics that make it particularly suitable for a wide range of practical optimization applications [96],[97].

3.3 Grey Wolf Optimization Algorithm

The Grey Wolf Optimization Algorithm , developed by Seyedali Mirjalili and his team in 2014 [93] is a nature-inspired metaheuristic designed to address complex optimization problems through stochastic search processes. The algorithm emulates the leadership hierarchy and cooperative hunting behavior observed in grey wolf packs. In this model, wolves are classified into four hierarchical roles: Alpha (α), which represents the decision-making leaders; Beta (β), the advisors to the Alpha; Delta (δ), responsible for assisting in coordination and hunting; and Omega (ω), the lowest ranking members tasked with following other; as illustrated in Figure.3.2

This hierarchical structure enhances the balance between exploration and exploitation during the search process.

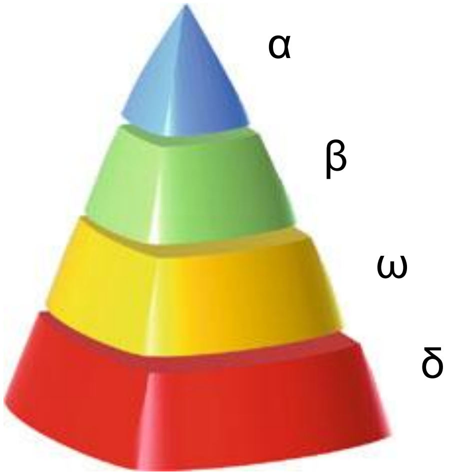


Figure.3.2 Hierarchy of Grey Wolves (Dominance Decrease From Top Down)

Due to its effectiveness, GWO has gained significant traction across various engineering domains [93],[95] and [98]. In particular, it has been widely employed in numerous PV applications [96], such as MPPT and solar parameter estimation [97] and in the estimation of solar cell parameters [99]. Modified versions of the original GWO have been proposed to improve convergence speed and tracking stability under challenging conditions such as partial shading [100]. One such variant, the Enhanced Grey Wolf Optimization (E-GWO), was developed to optimize MPPT performance under variable shading patterns. It demonstrated a dynamic response time improvement and over 2% increase in tracking efficiency compared to the standard GWO [101]. Another notable variant, the Improved GWO (IGWO), achieved high accuracy in estimating the parameters of a four-diode PV cell model, outperforming traditional metaheuristics by reducing error margins and enhancing system reliability [102]. This version notably reached a response time of 0.01 seconds and a tracking efficiency of 99.89%, surpassing conventional algorithms [103]. Building upon these enhancements, hybrid approaches have emerged by integrating GWO with advanced control techniques. For instance, the combination of IGWO and a Sliding Mode Control Proportional Integral Derivative (SMC-PID) controller significantly improved real-time identification of the optimal operating voltage and enhanced dynamic performance under fluctuating environmental conditions [104].

Additionally, integrating GWO with Particle Swarm Optimization (PSO) proved effective for optimizing the sizing and allocation of PV systems, yielding a 42.2% reduction in power losses and improved voltage stability [105].
Moreover, coupling GWO with the Super Twisting Algorithm (STA) further improved operational efficiency and transient response beyond what conventional controllers could achieve [106]. Ultimately, these advancements culminated in the development of a hybrid PV system incorporating GWO based techniques. This integrated system achieved a reduction in energy losses exceeding 87%, enhanced voltage regulation and economic feasibility with a projected payback period of six years [107].

3.4 Modeling the Hunting Behavior of Grey Wolves

The GWO Algorithm is inspired by the hierarchical social structure and hunting behaviors of grey wolves, which are essential for solving complex optimization problems. This algorithm models the hunting behavior of grey wolves through a series of stages, including exploration, circling the prey and attacking the prey. These stages, each representing different hunting tactics, are systematically integrated into the algorithm to balance exploration and exploitation in the solution search process [93].

A. Exploration Phase : Chasing, Approaching and Tracking the Prey

The first phase, known as the exploration phase, involves the wolves updating their positions relative to the prey. The movement behavior is governed by $|A|$:when $|A| < 1$, wolves move closer to the prey (exploitation); when $|A| > 1$, they move away from it (exploration), thereby expanding the search space and avoiding suboptimal solutions. The parameter (a) a gradually decreases from 2 to 0 over the iterations to control the range of $|A|$. The values of A, C, r_1 and r_2 are generated randomly within the interval [0, 1]. These components introduce diversity into the search process and help prevent the algorithm from becoming trapped in local optima [86],[93], thereby maintaining its effectiveness in locating the global solution, as illustrated in Figure.3.3.A and Figure.3.4.

B. Encircling Phase: Pursuit, Harassment and Encircling the Prey

As the parameter (a) decreases and approaches zero, the wolves transition into the circling phase, a critical stage in the optimization process where their positions are fine-tuned to move closer to the optimal solution This phase mimics the natural behavior of grey wolves as they surround and harass their prey through coordinated circular movements. In the context of the algorithm, this behavior translates into iterative positional adjustments, allowing the search agents to gradually refine the quality of the solution and reduce the distance to the global optimum. The search becomes more focused within (a) confined region of the solution space, enhancing convergence and improving the algorithm's precision. This stage leads to a steady approach toward the final solution, as illustrated in Figure.3.3.B, C, D and Figure.3.4.

The following equations represents the behavior of an encircling grey wolf [87]:

$$D = |C . Xp(t) - X(t)| \tag{3.1}$$

$$X(t+1) = Xp(t) - A.D \tag{3.2}$$

$$A = 2a.r1 - a \tag{3.3}$$

$$C = 2.r2 \tag{3.4}$$

$$a = 2 - (2t / \text{Maxiter}) \tag{3.5}$$

C. Attacking Phase: Convergence and Prey Capture

The attack phase represents the final stage of the GWO , during which the positions of the Alpha, Beta and Delta wolves are utilized to guide the Omega wolves toward regions closer to the optimal solutions.

The Alpha wolves act as leaders, relying on the most successful search experiences, while the Beta and Delta wolves assist in reinforcing guidance and maintaining balance throughout the position update process. During this phase, the wolves' movement dynamics gradually evolve, resulting in an effective convergence toward; the global optimum, ensuring that the best possible solution is achieved [93], as presented in Figure.3.3.E.

The positions of the wolves are updated using mathematical relationships that reflect the hunting behavior [93] .

$$X(t+1) = (X1 + X2 + X3)/3 \tag{3.6}$$

X_1, X_2 and X_3 is calculated by :

$$X1 = X_{\alpha} - A1.(D_{\alpha}) \tag{3.7}$$

$$X2 = X_{\beta} - A1.(D_{\beta}) \tag{3.8}$$

$$X3 = X_{\delta} - A1.(D_{\delta}) \tag{3.9}$$

$D_{\propto}$, D_{β} and D_{δ} Calculated by:

$$D_{\alpha} = |C1 . X_{\alpha} - X| \tag{3.10}$$

$$D_{\beta} = \left|C2 . X_{\beta} - X\right| \tag{3.11}$$

$$D_{\delta} = |C3 . X_{\delta} - X| \tag{3.12}$$

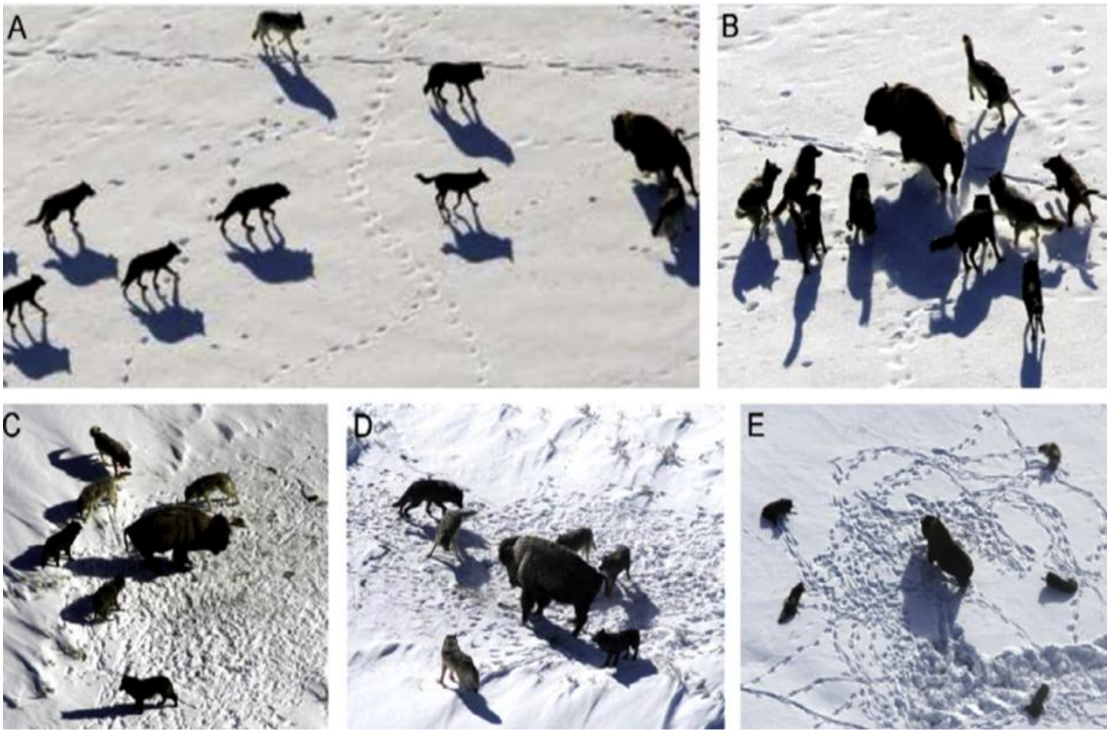


Figure.3.3 Grey Wolf Predatory Phases: (A) Tracking, (B) Pursuit, (C) Encirclement, (D) Immobilization, (E) Attack[93]

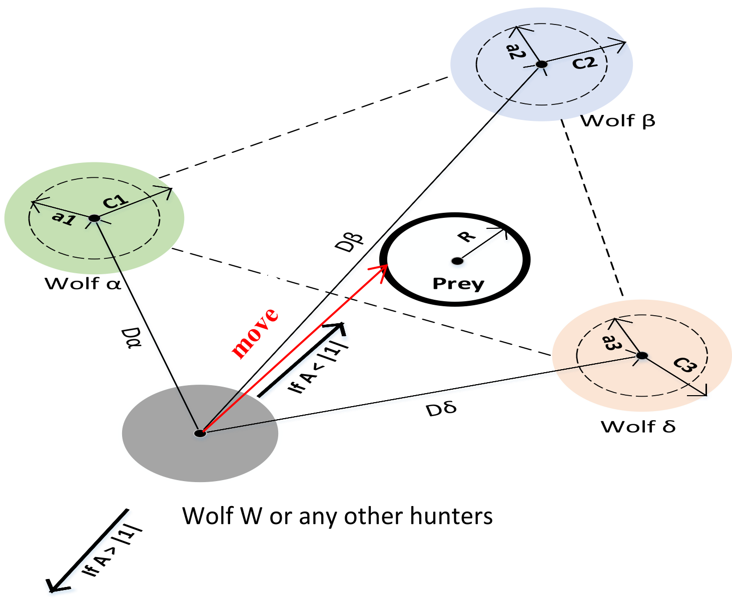


Figure.3.4 Attack or Search for Prey and Update GWO Position

A flowchart of the GWO algorithm is presented in Figure.3.5, summarizing its main stepsinitialization, position updating, hunting behavior and termination to facilitate understanding and practical implementation.

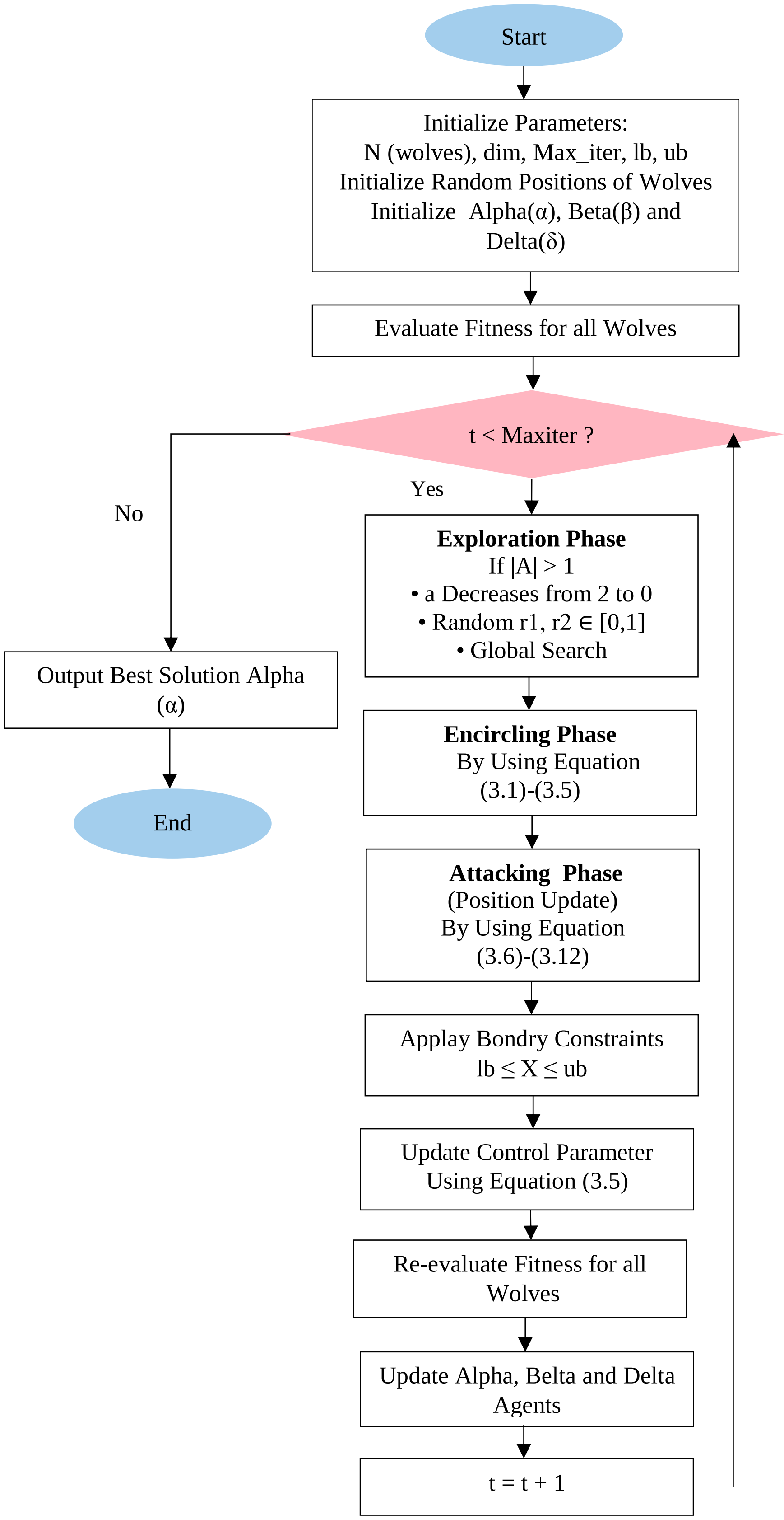


Figure.3.5 Flowchart of the GWO Algorithm

3.5 Practical Implementation of the Grey Wolf Optimizer for Signal Enhancement in Photovoltaic Systems

Building upon the theoretical foundations and operational principles of the GWO discussed earlier, this section aims to demonstrate the practical applicability of the algorithm in real-world optimization tasks. With the continuous advancement of PV energy systems, maintaining high-quality electrical signals has become crucial for ensuring system stability and reliable data processing. Electrical noise and harmonics can adversely affect power quality and overall system performance, emphasizing the need for effective signal enhancement techniques.

In this context, a practical methodology is presented based on the Moving Average Filter (MAF) for signal enhancement, whose performance is optimized using the GWO algorithm. The MAF is a widely used smoothing technique that reduces high-frequency noise by averaging signal values over a sliding window. However, selecting an appropriate window size is critical: small windows may inadequately suppress noise, while larger windows may distort essential signal characteristics.

To overcome this limitation, GWO a nature inspired, intelligent metaheuristic optimization algorithm was employed to determine the optimal window size. The optimization objective is to minimize the mean squared error (MSE) between the original and filtered signals, ensuring effective noise reduction while preserving the key features of the signal.

This approach integrates the simplicity of conventional filtering with the adaptability of intelligent optimization, resulting in improved signal integrity without compromising critical information. The proposed system was implemented and evaluated in the MATLAB/Simulink environment under various realistic operating conditions to validate its effectiveness.

3.5.1 Experimental Setup and Data Acquisition

The experimental work was carried out at the LTI Laboratory (EA 3899) of the Institute of Technological University of Soissons, France, which specializes in advanced control and renewable energy systems. The setup involved a PV cell as the primary energy source, with voltage and current sensors integrated for real time performance monitoring.

Signal waveforms were analyzed using an oscilloscope and processed digitally via the ADMC201 coprocessor connected to a DSP unit, responsible for controlling the DC-DC converter. All measurement data were collected and processed by a computer for further analysis and evaluation (see Figure.3.6).

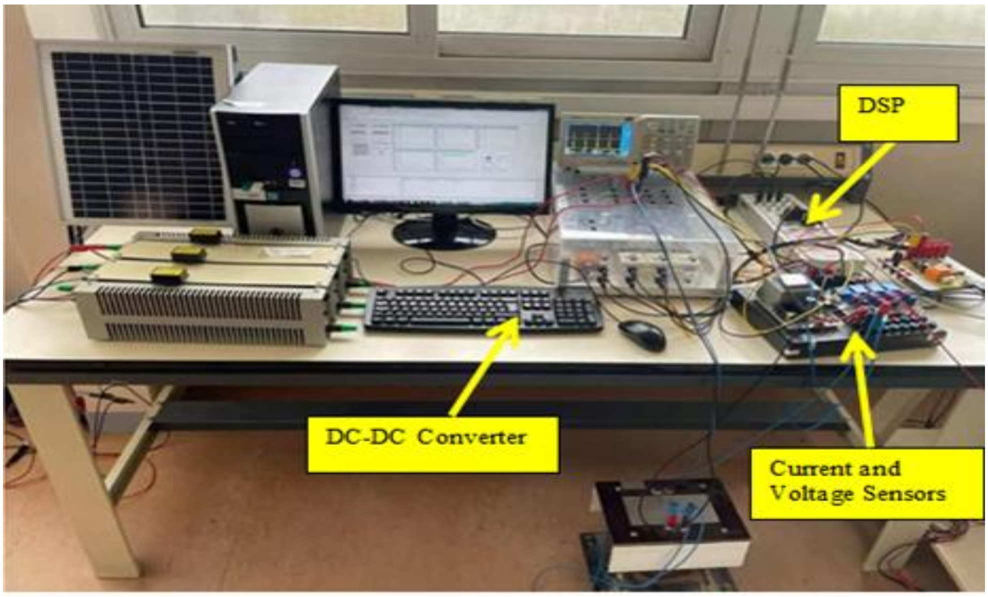


Figure.3.6 Experimental Setup of in Loop Devices

1. Moving Average Filter

The moving average filter is a widely utilized technique due to its computational simplicity and effectiveness in noise reduction. It operates by calculating the arithmetic mean of data points within a fixed-size window, which helps in mitigating random fluctuations and smoothing out noise in the signal.

This method is extensively applied in various engineering and scientific fields, offering an efficient means of improving signal quality and enhancing data accuracy. Its ease of implementation and ability to reduce unwanted noise make it an essential tool in signal processing applications [108].

$$F(t) = \frac{1}{W} \sum_{k=0}^{W-1} y[t-k] \tag{3.13}$$

2. Evaluation Metric

- MSE was adopted due to its widespread use in fault prediction studies and its effectiveness in detecting high-impact discrepancies. MSE is calculated as the average of the squared differences between the actual and predicted values, according to Equation (1.1), providing a precise measure of model deviation. In this study, MSE was employed to evaluate the performance of the filtering technique by comparing the original signal with the filtered output, enabling an objective assessment of the filter's ability to reduce noise while preserving the signal's fine details.

$$\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} (y_{\mathrm{actual}} - y_{\mathrm{predicted}})^2 \tag{3.14}$$

- To evaluate the effectiveness of the optimization strategy in signal processing, specifically for filtering and harmonic enhancement, a structured methodology was implemented, as illustrated in Figure 3.7.

- Initially, a moving average filter was applied using a preliminary window size of 3000 to filter the output voltage signal from the PV systems (see Figure 3.9).

- To ensure an objective and reliable evaluation of filtering quality, the mean square error was selected as the objective function.

- To provide an objective and reliable evaluation of filtering quality, the mean square error was selected as the objective function. The MSE quantifies the difference between the original and filtered signals by squaring the deviations and averaging them over the entire signal:

$$\mathrm{F_{OBJ}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{t}=1}^{\mathrm{N}} (\mathrm{Vpvfn(t)} - \mathrm{Fw(t)})^2 \tag{3.15}$$

Where $\mathrm{Vpvfn(t)}$ represents the original noisy signal voltage and $\mathrm{Fw(t)}$ denotes the filtered signal obtained with a window of size W. The squared difference is preferred over the simple difference to ensure all values are positive and to penalize larger deviations more strongly, thereby allowing a clear distinction between effective and ineffective filtering solutions.

- The GWO algorithm was then employed to dynamically adjust the window size W and determine the optimal configuration that minimizes the MSE. In this process:

- Each wolf represents a candidate solution corresponding to a specific window size.
- The MSE is evaluated for each candidate solution.
- Solutions are ranked according to their MSE values, with the lowest MSE designated as the alpha (α) solution.
- Other candidate solutions update their positions based on the locations of the top-ranking wolves (α , β, δ), with the MSE recalculated at each iteration to guide convergence.
- The process continues until convergence, at which point the alpha solution represents the optimal window size, achieving the minimum MSE and the highest filtering quality.

By employing MSE as the objective function, the methodology provides a systematic and quantitative measure of filtering performance, guiding the GWO algorithm to efficiently explore the solution space, discard suboptimal window sizes and converge toward the window size that ensures the most effective noise reduction while preserving the original characteristics of the signal (see Figure.3.7).

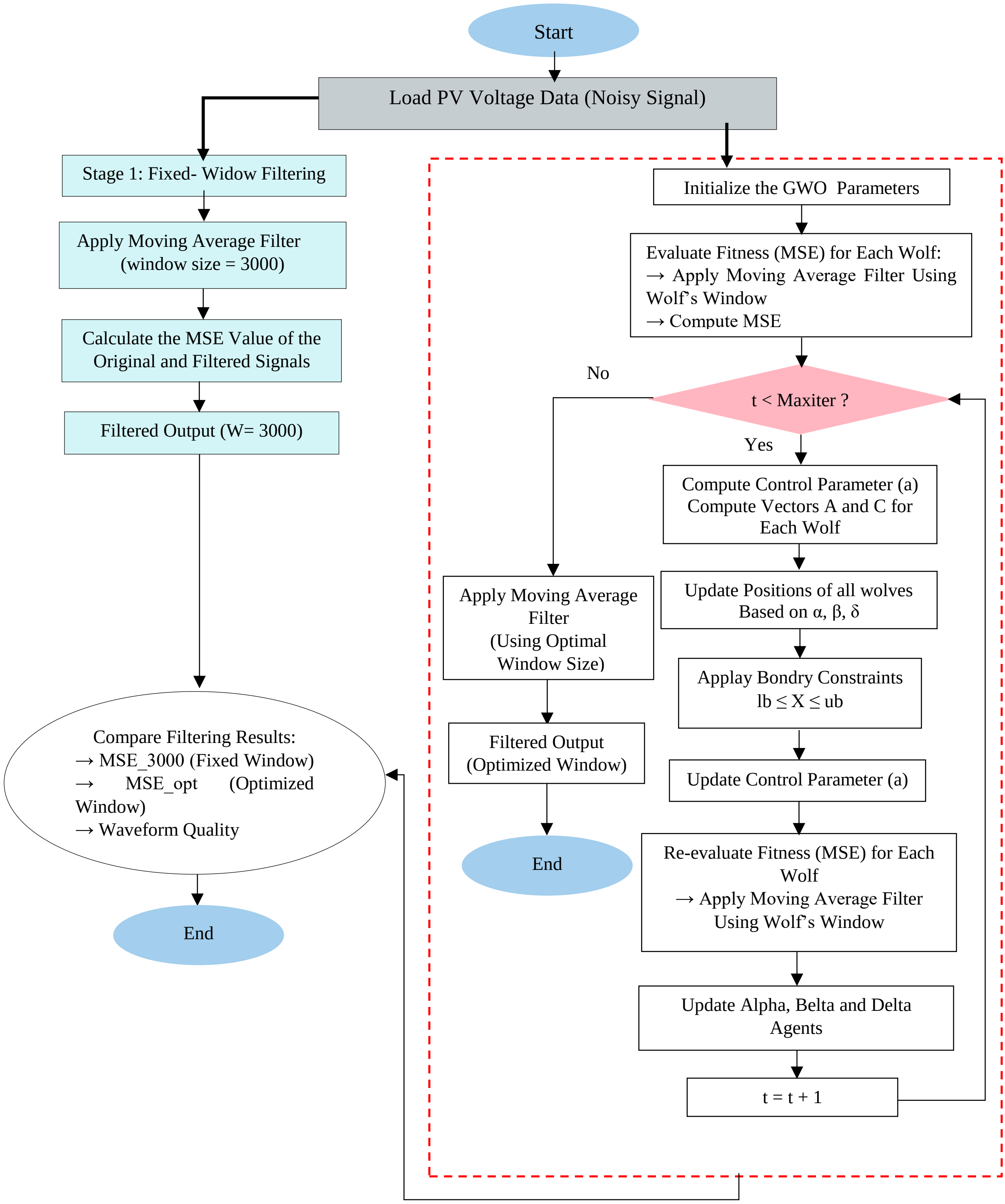


Figure.3.7 Filtering Optimization Methodology Using GWO

3.5.2 Performance Evaluation

The filtering results obtained with a window size of 3000 demonstrated strong performance, effectively smoothing the noise present in the voltage curve (see Figure. 3.9).

In contrast, the GWO optimizer was successful in identifying an optimal window size within the defined search space, selecting it based on minimizing the mean square error a reliable metric used to assess the efficiency of the filtering process (see Figure. 3.10).

Interestingly, the voltage curve filtered with the GWO selected window size still contained more residual noise (Figure. 3.10) when compared to the curve filtered using the fixed window size of 3000 (Figure. 3.9). This discrepancy can be explained as follows:

- The GWO algorithm selected a window size of 880, which, being smaller, tends toretain more of the signal’s fine details. While this allows for better preservation of the original signal’s structure, it also permits more noise to pass through. Conversely, the larger window size of 3000 is more effective at reducing noise, but may unintentionally smooth out key signal components, making the data harder to interpret and potentially reducing overall system accuracy.

- To validate this reasoning, we compared the MSE values between the original and filtered signals across both window sizes. The analysis revealed that the optimized window size (880) achieved a lower MSE (approximately 29.2574), indicating better fidelity to the original signal.

  In comparison the fixed size of 3000 resulted in a higher MSE (around 35.5759) , reflecting a trade-off between noise reduction and detail preservation (see Figure. 3.11).These findings highlight the importance of selecting an appropriate window size tailored to the specific application, balancing noise suppression with signal integrity.

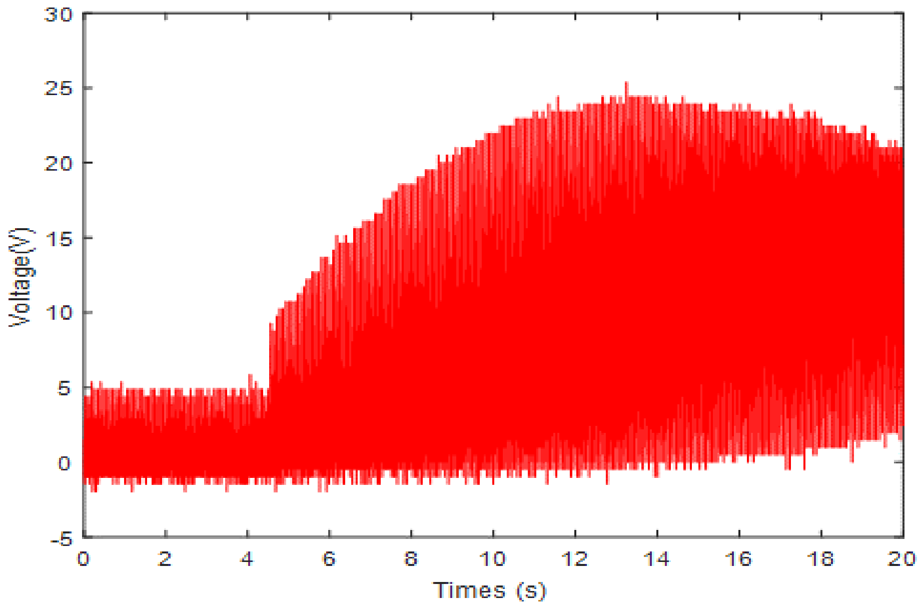


Figure.3.8 PV System Output Voltage Curve with Noise

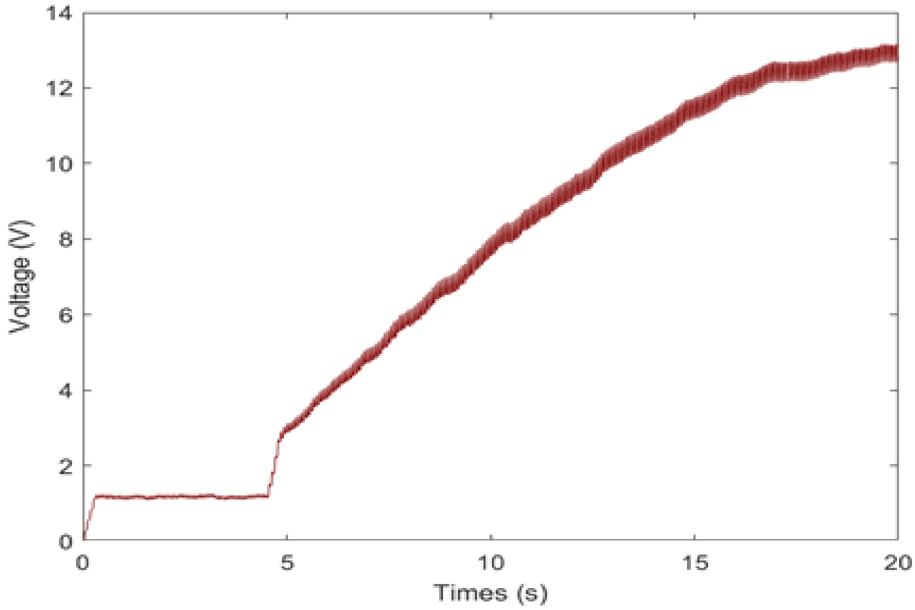


Figure.3.9 Filtered Signal with Window Size = 3000

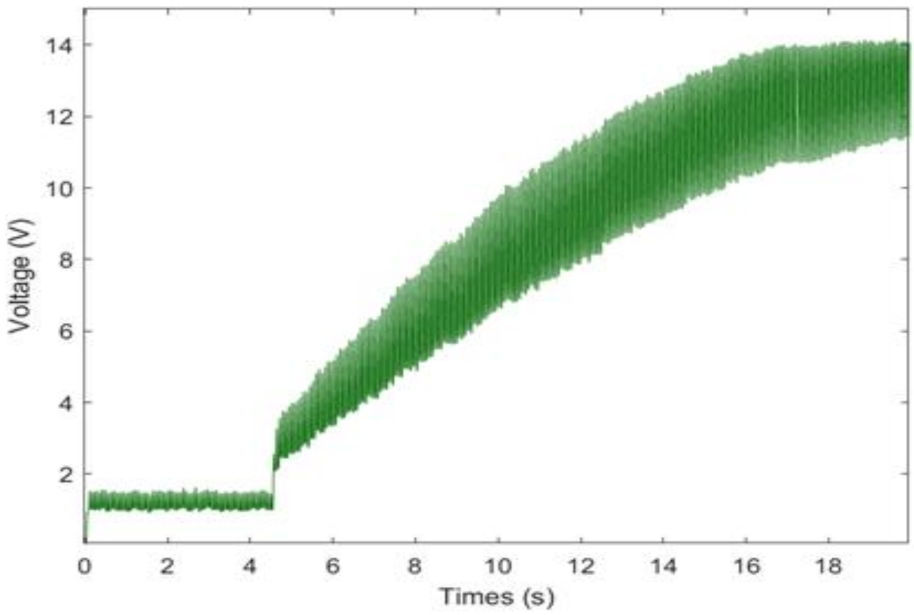


Figure.3.10 Voltage Curve After Using a Window Sized Filter Determined by the GWO Algorithm

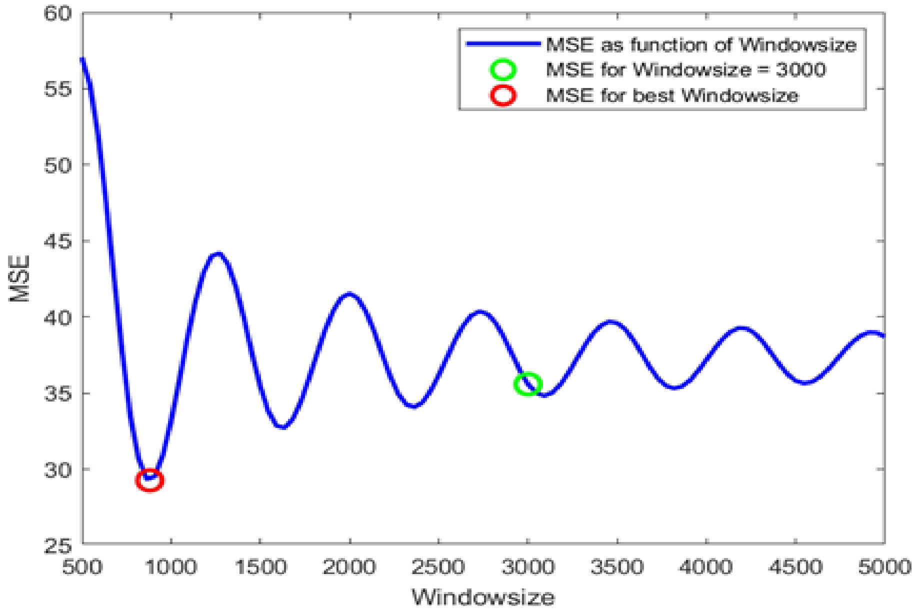


Figure.3.11 MSE Comparison Between the Original and Filtered Signal in Terms of Window Size

3.5.3 Convergence Analysis of the GWO Algorithm Based on MSE Evolution

To assess the convergence performance of the GWO algorithm, the evolution of the objective function, represented by the MSE, was analyzed over successive iterations (see Figure 3.12). Each iteration reflects an adaptive adjustment of the wolves' positions in the search space, leading to a gradual refinement of candidate solutions and demonstrating the algorithm's ability to achieve consistent optimization.

During the initial iterations, a sharp reduction in MSE is observed, mainly due to the random initialization of wolves in the search space. At this stage, the algorithm rapidly explores the solution domain and identifies promising regions.

As iterations progress, the rate of MSE reduction decreases, illustrating a transition from exploration to exploitation. This behavior indicates that the algorithm is approaching convergence, with diminishing variations in the solution updates. Eventually, the MSE curve reaches a near-constant value (approximately 29.2574), indicating convergence toward a near-optimal solution, where further improvements become minimal.

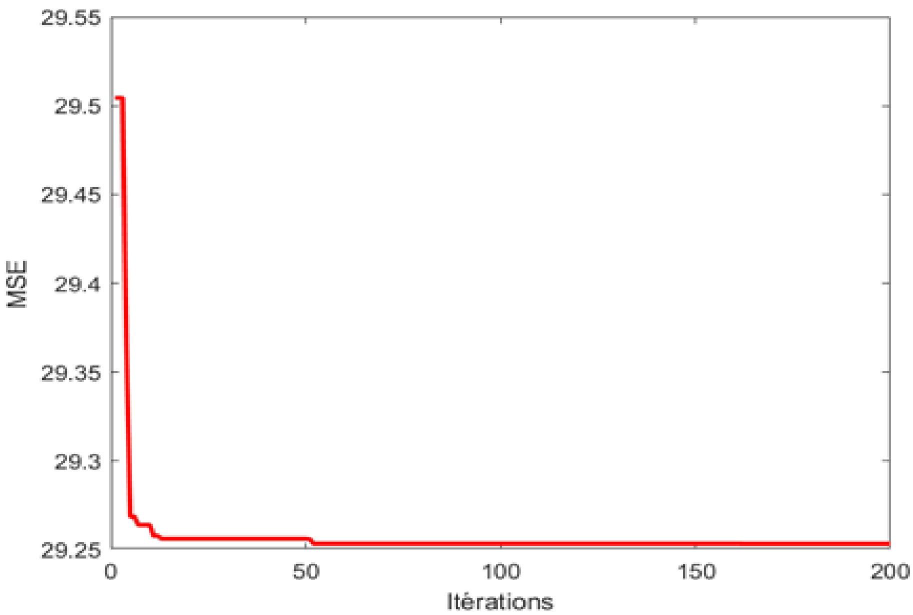


Figure.3.12 Tracking the Performance of the Algorithm and the Progression of the Objective Function with Iterations

3.6 Conclusion

This chapter has reviewed metaheuristic algorithms and their role in addressing complex optimization problems, with a particular focus on the GWO due to its simplicity, fast convergence and high efficiency in finding optimal solutions for multi-objective problems. The practical application of GWO in enhancing signal quality within PV systems demonstrated its effectiveness in noise reduction while preserving essential signal characteristics, thereby improving system efficiency, stability and reliability.

This application also highlighted the algorithm’s potential to optimize system performance through parameter tuning in intelligent systems.The findings of this chapter provide a solid foundation for extending Grey Wolf based approaches in the subsequent chapter, which will focus on the AGWO for Feature Selection, aiming to enhance data quality and model inputs and to develop more intelligent and accurate decision-making systems.

Chapter 4: Optimized Feature Selection for Photovoltaic Systems Using AGWO and Machine Learning

4.1 Introduction

Building upon the methodological insights and optimization strategies presented in the previous chapter, which addressed signal quality enhancement and noise mitigation in photovoltaic systems using GWO, this chapter focuses on the development of an intelligent and interpretable diagnostic framework, with particular emphasis on data exploration and analysis.

The proposed framework relies on data generated from simulated fault scenarios and aims to achieve a comprehensive understanding of the dataset structure, as well as an assessment of feature relevance. This stage serves as a critical foundation for subsequent diagnostic modeling by supporting informed feature analysis and enhancing transparency in machine learning–based decision-making. In this context, an improved variant of GWO is employed in conjunction with gradient-boosted decision tree classifiers to support feature selection and data analysis. Subsequently, the Shapley Additive Explanations (SHAP) framework is applied to quantify the contribution of each feature to the final predictions, thereby facilitating the development of a transparent and interpretable diagnostic model.

This chapter provides a methodological foundation for evaluating the proposed diagnostic approach, focusing on the development of a reliable and interpretable model based on feature and data analysis.

4.2 Foundations of Intelligent Feature Engineering

Recent studies have reported a significant increase in the adoption of AI techniques and data driven analytical platforms across diverse application domains, including science, industry, medicine and energy systems [109],[110] and [111]. In parallel, advances in intelligent sensing technologies have resulted in the generation of large scale, complex and high-dimensional datasets, which introduce major challenges related to computational efficiency, predictive accuracy and model interpretability. Consequently, the development of advanced data processing and analysis techniques has become essential for extracting meaningful information from such data [112], [113].

Feature selection is a critical process for improving the performance, robustness and generalization capability of machine learning models by reducing dimensionality, eliminating redundant or irrelevant features and decreasing computational complexity. In this context, metaheuristic optimization algorithms have demonstrated strong effectiveness for feature selection problems due to their ability to efficiently explore large, nonlinear and high dimensional search spaces when combined with machine learning frameworks [114]–[119]. Among these algorithms, the GWO [93] has gained considerable attention because of its simple structure, ease of implementation and balanced exploration–exploitation behavior. However, the standard GWO may suffer from premature convergence and reduced performance in complex optimization scenarios.

To overcome these limitations, several enhanced GWO variants have been proposed. These include nonlinear adaptation strategies for the control parameter (a), inspired by Particle Swarm Optimization (PSO), as well as opposition-based learning techniques to improve global search capability and convergence behavior [120].

Moreover, improved GWO based approaches have shown superior performance in clustering and high-dimensional optimization tasks. For instance, enhanced GWO clustering algorithms demonstrated higher clustering accuracy and lower error rates compared to conventional methods [121], while the CBRGWO algorithm effectively addressed high-dimensional and multimodal optimization problems by integrating Gaussian Barebone, random selection, chaotic mechanisms and dynamic parameter control, outperforming state-of-the-art algorithms on IEEE CEC 2017 benchmark functions and real-world engineering applications [122]. Additional improvements, such as EGWO and MDM-GWO , further enhanced convergence speed, solution accuracy and population diversity compared to the original GWO [123], [124].

Based on previous studies on metaheuristic optimization algorithms and intelligent feature selection, AGWO was applied to select the most informative features from high dimensional photovoltaic datasets and their importance was evaluated using XGBoost. Subsequently, three ensemble classifiers (XGBoost, LightGBM and CatBoost) were trained and their outputs aggregated via a soft voting strategy, enhancing classification stability and providing insights into feature relevance. Finally, the SHAP framework was employed to assess the contribution of each feature to the final predictions, facilitating the development of a transparent and interpretable diagnostic model. This approach demonstrates how intelligent feature selection, combined with ensemble classification and XAI techniques, can improve model accuracy, robustness and interpretability.

4.3 Improving Grey Wolf Exploration via Adaptive Control of Parameter (a)

The exploration capability of the GWO algorithm is crucial for avoiding premature convergence and ensuring a comprehensive search of the solution space. In the standard GWO the adaptive control coefficient (a) is linearly decreased from 2 to 0 over iterations to balance exploration and exploitation. However, this linear reduction may lead to insufficient exploration and entrapment in local optima.

To overcome this limitation, the proposed approach introduces a dynamic oscillatory adjustment of (a) using a cosine-based formulation. The coefficient oscillates periodically between positive and negative values, enhancing the algorithm's ability to escape local optima. Additionally, the quality of the best solution (best_score) is incorporated in the calculation as (2−best_score), providing dynamic control over (a) and allowing the exploration intensity to adapt based on the optimization progress.

The adaptive coefficient is calculated as follows:

$$a = \cos(\frac{t}{Maxiter} \cdot (2 - best\_score)) \tag{4.1}$$

Low (best_score) values make (2−best_score) larger, increasing (a) and promoting broader exploration. In contrast, high (best_score) values reduce (2−best_score), decreasing (a) to focus on exploitation. This dynamic adjustment ensures effective control of the search trajectory, maintaining a balance between global exploration and local refinement throughout the optimization process.

4.4 AGWO-Driven Feature Selection and Model Training

In this section, the AGWO algorithm is employed to identify the most relevant features within the photovoltaic power plant dataset. This includes describing the available feature characteristics, applying the algorithm to extract the significant features and utilizing them to train ensemble classification models, thereby providing a systematic and structured framework for data handling and model evaluation based on the selected features.

4.4.1 Feature Descriptions of the Dataset

- **Modeling of a 250 kW Photovoltaic Power Plant**

A 250 kW grid-connected PV power plant was modeled to capture its operational characteristics and analyze the performance of medium-scale solar energy systems [125]. The model comprises an array of series parallel connected PV modules, forming a configuration that realistically reflects the behavior of actual installations. The PV array is interfaced with the grid through a three-level inverter equipped with MPPT control to ensure optimal energy extraction, along with a step-up transformer to match the voltage levels required by the grid. Environmental variability, particularly solar irradiance and ambient temperature, was incorporated into the model as key parameters affecting PV module efficiency.

Several operating scenarios were simulated, including the fault-free condition as well as common fault cases such as string faults and string-to-ground faults, as illustrated in Figure .4.1. These scenarios provide a representative dataset that serves as a reference for fault studies and evaluating the performance of fault detection techniques in PV systems [125].

- **Simulation of Fault Conditions and Data Preparation for PV Systems**

To represent the dynamic operational characteristics of PV systems under realistic conditions, a simulation framework was developed encompassing both healthy and faulty states as . Four representative fault scenarios were considered :

- **Normal Operation (F0):** Represents baseline performance without faults, serving as a reference for evaluating fault impacts.
- **String Fault (F1):** Modeled by disconnecting the first string, resulting in partial power loss and enabling analysis of performance degradation due to open-circuit conditions.
- **String-to-Ground Fault (F2):** Simulated by connecting the first string directly to ground.
- **String-to-String Fault (F3):** Modeled by creating a direct connection between the first and second strings.

Faults were introduced at 0.2 seconds after the system reached steady-state operation. The total simulation time was 0.4 seconds, covering both pre-fault and post-fault stages, allowing detailed analysis of transient responses.

To enhance realism, probabilistic variations were applied to three critical parameters: solar irradiance (100 –1000 W/m²), ambient temperature (10 –35°C) and fault resistance (1–2000 Ω). These variations improve the representativeness of the dataset for real-world PV operating conditions. The final dataset consisted of 600 training samples and 100 testing samples.

Training data were collected during the interval (0.2 – 0.4 s) to capture post-fault dynamics, while testing data were extracted from (0.1 – 0.3 s), encompassing both the transient phase (0.1 – 0.2 s) and the fault initiation period (0.2 – 0.3 s). The training dataset was distributed as follows: 100 healthy cases (16.7%), 153 string disconnections (25.5%), 149 string-to-ground faults (24.8%) and 198 inter-string faults (33%), as summarized in Table 4.1.

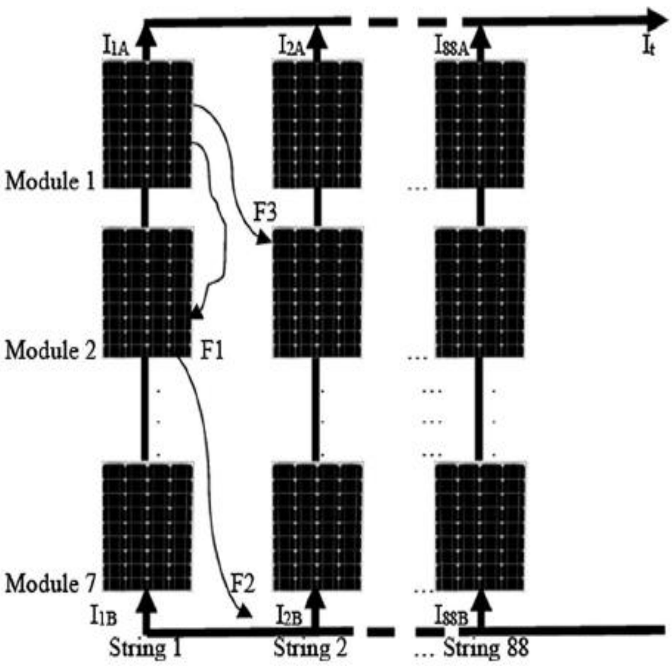


Figure.4.1 Create Training and Testing on PV Failure [125]

Table 4.1. Distribution of 600 Samples in the Training Dataset [125]

Description	Number of Cases	Percentage (%)	Class #
Normal state (F0)	100	16.67 %	1
String fault (F1)	153	25.50 %	2
String-to-ground fault (F2)	149	24.83 %	3
String to string fault (F3)	198	33.0	4

4.4.2 Structural and Dynamic Complexity of the Dataset and Its Implications for Data-Driven Modeling

The dataset, as characterized in the previous subsection, demonstrates pronounced structural and dynamic complexity. This complexity arises directly from the interplay of multiple operational scenarios, environmental factors and transient behaviors observed in the data. The main features contributing to this complexity are summarized below:

- **High-Bimensional and Multi-Source Feature Space**

Each data sample is represented by 30 numerical features extracted from distributed electrical measurements at both string and plant levels, in addition to environmental parameters such as solar irradiance (100 – 1000 W/m²) and ambient temperature (10 – 35 °C). This heterogeneous composition results in a high-dimensional feature space governed by strong nonlinear dependencies, which cannot be effectively characterized using simplified form mathematical models.

- **Temporal Dynamics and Fault Induced Transient Responses**

The dataset incorporates measurements from both pre-fault and post-fault intervals within a total simulation time of 0.4 s, with fault initiation occurring at 0.2 s. The presence of short-duration transient responses following fault occurrence generates overlapping electrical patterns, particularly under high-resistance fault conditions (up to 2000 Ω), thereby increasing the difficulty of fault discrimination.

- **Multiple Fault Scenarios with Overlapping Electrical Signatures**

Four operating classes are included in the dataset: normal operation(F0) and three fault types ; string fault (F1) , string-to-ground fault (F2) and string-to-string fault (F3). Despite their distinct physical origins, these faults can generate partially overlapping electrical patterns under certain operating and environmental conditions, making accurate fault classification challenging.

- **Class Imbalance Among Data Categories**

The dataset comprises 600 training samples with a non-uniform class distribution, where string-to-string faults represent the dominant class (33%), while fault-free conditions account for only

16.7% of the data. This imbalance introduces an additional challenge for learning algorithms, which may otherwise become biased toward the more frequent classes.

- **Probabilistic Variations and Operational Uncertainty**

Wide probabilistic variations in key parameters, including solar irradiance, ambient temperature and fault resistance (1–2000 Ω), were intentionally introduced to reflect realistic PV operating conditions. However, these variations increase uncertainty and overlap between fault patterns, limiting the effectiveness of mathematical models for fault detection.

- **Interdependence Between Electrical and Environmental Variables**

Electrical variables are strongly coupled with environmental conditions, fault type and fault occurrence time, resulting in a highly correlated feature structure. Consequently, no single variable can be reliably interpreted in isolation from the overall operational context of the PV system. Based on the above analysis, it is clear that the structural and dynamic complexity of the dataset imposes fundamental limitations on the effectiveness of mathematical modeling approaches. In contrast, these characteristics provide an ideal foundation for the application of data-driven machine learning methodologies, which will be systematically investigated and addressed in this chapter and in Chapter Five of this thesis .

4.4.3 AGWO-Based Feature Selection and Ensemble Classification

An intelligent and interpretable diagnostic framework for photovoltaic systems is developed with a focus on data exploration and feature analysis, aiming to achieve accurate, reliable, and transparent fault diagnosis using machine learning techniques, as illustrated in Figure 4.2.

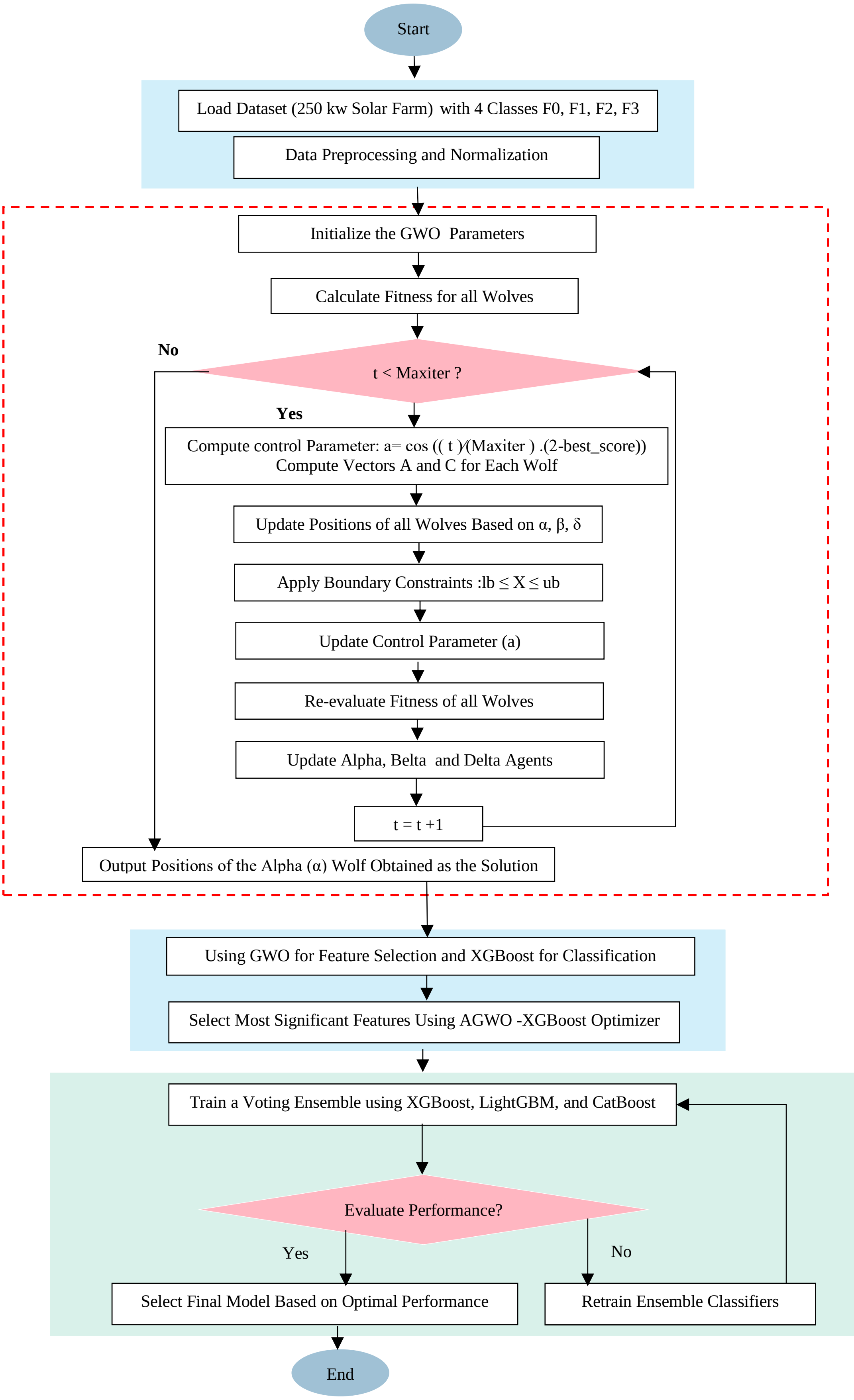


Figure.4.2 The Proposed Method AGWO

A. Adjustment of Parameter (a) in Experiments

During experimental implementation, the improved calculation of (a) was applied directly within the GWO framework to guide feature selection. The dynamic adjustment ensured a balance between exploration and exploitation throughout the iterative optimization, enhancing the identification of high-quality features.

B. Integration with XGBoost

AGWO was integrated with the XGBoost model to evaluate and rank feature importance efficiently. XGBoost is a gradient boosting framework widely used in machine learning for its high performance, ability to handle heterogeneous data and robustness to overfitting[126]. By combining AGWO with XGBoost, the framework leverages the optimization strength of AGWO to identify the most informative features, while XGBoost quantitatively measures the contribution of each feature to the predictive task. This integration not only streamlines the feature selection process but also ensures that the selected features have a demonstrable impact on the model's accuracy and generalization capability.

C. Classification and Ensemble Learning

Following feature selection, three advanced gradient boosting classifiers XGBoost, LightGBM and CatBoost were independently trained on the reduced feature set.

- XGBoost is a scalable and efficient gradient boosting framework that iteratively builds decision trees, optimizing a regularized objective function to reduce overfitting and improve predictive accuracy [126].

- LightGBM is designed for faster training and lower memory usage. It uses a histogram-based algorithm and leaf-wise tree growth, which allows it to handle large datasets efficiently while maintaining high accuracy[127].

- CatBoost excels at handling categorical features automatically without extensive preprocessing, reducing bias from one-hot encoding and improving model robustness[128].

- Ensemble Learning: After training, the outputs of the three classifiers were combined using a probability based voting mechanism. This ensemble approach leverages the strengths of each model, mitigates individual weaknesses and often yields superior overall predictive performance[129]. In the context of Machine Learning, such ensembles are effective for improving generalization, reducing variance and achieving more reliable classification outcomes.

D. Interpretability Layer: SHAP-Based Explanation

To enhance the transparency and reliability of the ensemble classification framework, a SHAP-based interpretability layer was implemented as a post-hoc mechanism to explain model predictions [130]. Since XGBoost, LightGBM and CatBoost are considered black-box models, this layer interprets feature contributions without altering their internal structures or interfering with the model's operation. SHAP is grounded in cooperative game theory, assigning an importance value to each feature based on its contribution to the model's output.

In this study, Tree SHAP was employed due to its computational efficiency and suitability for tree-based ensemble models. The interpretability layer facilitates both global and local explanations: SHAP statistics rank features by their overall influence on classification outcomes, while instance-level values clarify individual predictions. This integration ensures high predictive performance alongside transparency, supporting informed decision-making and model validation.

E. Evaluation and Benchmarking

The proposed approach was assessed using standard metrics including precision, recall, accuracy and F1-score. Additionally, AGWO based feature selection was compared with conventional GWO and PSO under identical experimental conditions to demonstrate the superiority of the proposed enhancements.

4.5 AGWO Performance Evaluation for Feature Selection and Classification

1. Accuracy and Computational Efficiency Assessment

The AGWO Algorithm successfully selected nine salient features out of thirty, considered the most important among all features, providing a concise and informative representation of the data as shown in Figure.4.3.

This selection significantly enhanced the classification performance of the XGBoost model by reducing computational complexity and minimizing noise from uninformative features.

The improvement is attributed to the adjustment of the control coefficient, which is based on the balanced interaction between the iteration progress t, the total number of iterations Maxiter and the quality of the best solutions discovered best _score. Additionally, the cosine oscillation function plays a crucial role in modulating this coefficient, enabling the algorithm to periodically and evenly alternate its search behavior between exploration and exploitation.

This mechanism prevents convergence to local solutions and improves mobility within the feature space, enhancing the algorithm's ability to select an informative and diverse feature set.

Subsequently, the selected features were evaluated using ensemble voting mechanisms, which increased the classification accuracy to 96% .

To evaluate the classification performance of the proposed model, several key metrics were employed, including Precision, Recall, F1-score, Accuracy, and Support metrics .

These measures collectively provide a comprehensive understanding of model behavior across all categories [131].

The calculation of these metrics is based on the confusion matrix, which provides a structured representation of classification outcomes by comparing predicted labels with actual labels. The confusion matrix summarizes the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), forming the foundation for quantifying the effectiveness and reliability of the classification model. This approach is applicable to both binary-class and multi-class classification tasks [132].

Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive and is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.2}$$

Recall (also called Sensitivity) measures the proportion of correctly identified positive instances among all actual positive instances, defined as:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4.3}$$

The F1-score represents a balanced combination of Precision and Recall, defined as:

$$\mathrm{F1} - \mathrm{Score} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{4.4}$$

Accuracy evaluates the overall correctness of the model's predictions across all categories and is computed as:

$$\mathrm{Accurcy(CA)} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{4.5}$$

Specificity measures the proportion of correctly identified negative instances among all actual negative instances and is calculated as:

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4.6}$$

Support refers to the number of true instances for each class in the dataset, representing how many samples actually belong to each category. It provides an indication of the class distribution within the dataset and helps in understanding the relative representation of each class during performance evaluation.

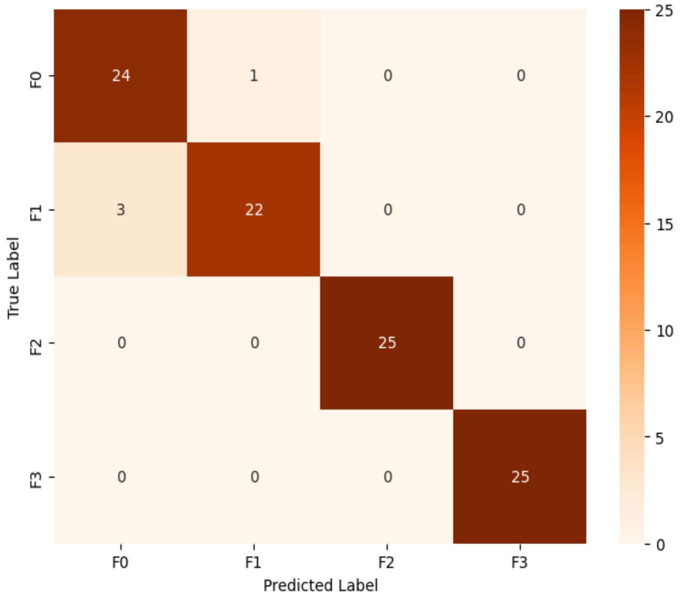
The proposed model demonstrates a well-balanced performance across these metrics for all categories, representing a significant improvement over the single-model performance, as shown in Table 4.2 and Table 4.3

Table 4.2. Feature Selection and Classification Performance of AGWO

Faults	Precision	Recall	F1-score	Support
F0 Normal state (Class 0)	0.89	0.96	0.92	25
F1 (Class 1)	0.96	0.88	0.92	25
F2 (Class 2)	1.00	1.00	1.00	25
F3 (Class 3)	1.00	1.00	1.00	25

Overall Accuracy = 0.96

Table 4.3. Confusion Matrix For Ensemble Model With AGWO Feature Selection



2. Feature Level Evaluation of AGWO Selected Features Using SHAP Analysis

➢ Interpretability Analysis Via SHAP Framework

To evaluate the effectiveness of the AGWO-based feature selection, the impact of features across different fault classes was analyzed using the SHAP framework within an ensemble model integrating XGBoost, LightGBM, and CatBoost. This analysis aims to provide a deeper understanding of each feature's contribution to the classification decision process and to identify the most influential features for each fault class.

The results presented in Table (4.4) and Figure (4.3) indicate that feature contributions are not uniformly distributed, but are instead concentrated in a small number of highly discriminative features, while the remaining features exhibit relatively lower contributions. Specifically, Range 1 shows the highest impact on the prediction of F0 (Normal state) with a SHAP value of (2.372502), highlighting its strong ability to distinguish the normal condition from other fault classes. In contrast, Range 3 demonstrates a strong association with both F1 and F2, with SHAP values of (2.745177 and 2.135832, respectively), indicating its effectiveness in discriminating between these fault categories. Similarly, Range 4 exhibits the most significant influence on F3, achieving a high SHAP value of (4.058128), making it the most dominant feature for this class.

On the other hand, the remaining features, including I1MIN, I2MAX, I4, I4MAX, Itotalmin1, and Pdcmean1, show relatively low SHAP values across all classes, indicating limited contributions to the classification process compared to the main features.

Furthermore, Figure (4.3) clearly illustrates the dominance of the three key features (Range 1, Range 3, and Range 4) in driving the model's decisions, while the influence of the remaining features is marginal.

Overall, these results confirm the effectiveness of the AGWO algorithm in selecting a compact set of highly discriminative features, thereby enhancing model interpretability while maintaining strong classification performance.

Table 4.4 Impact of Features Across Classes Using an Ensemble of XGBoost, LightGBM, and CatBoost

	F0 (Normal state)	**F1**	**F 2**	**F3**
Range 1	2.372502	1.148701	0.249354	0.215931
Range 3	0.696981	2.745177	2.135832	0.471339
Range 4	0.620910	0.198895	1.877211	4.058128
I1MIN	0.027432	0.019842	0.024731	0.031255
Ittalmin1	0.022392	0.016497	0.009925	0.011771
I4MAX	0.008111	0.067721	0.006346	0.034300
I4	0.008069	0.008456	0.006835	0.007462
I2MAX	0.008010	0.013547	0.006046	0.008713
Pdcmean1	0.006656	0.010398	0.007199	0.004898

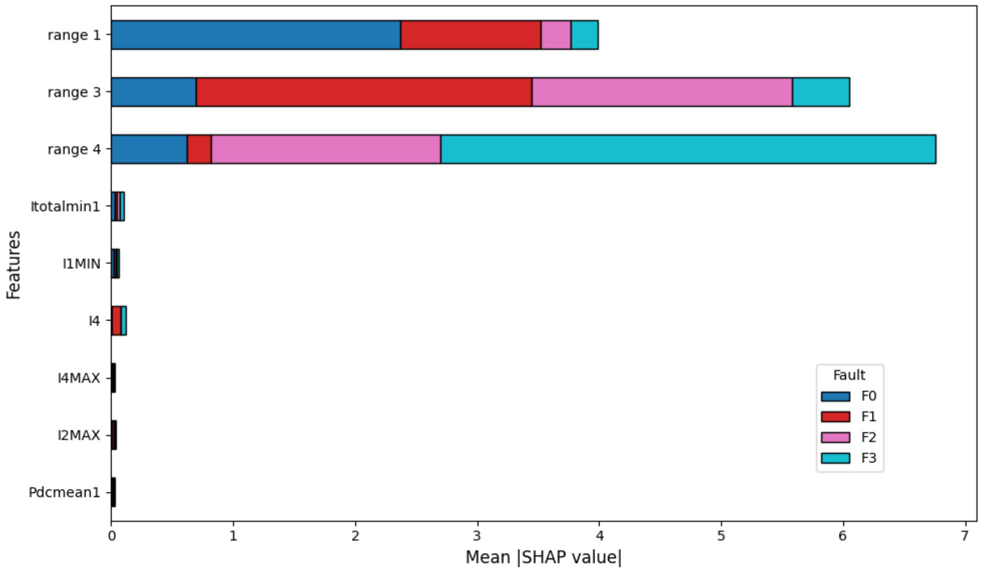


Figure 4.3. Impact of the Most Important AGWO -Selected Features Across Classes Using SHAP Analysis

3. Comparison of the Efficiency and Accuracy of the AGWO Algorithm with Classical GWO and PSO

The traditional GWO algorithm selected only four features out of thirty (I1, range 1, range 3, and range 4), resulting in a significant reduction in data dimensions. However, this limited set of features negatively affected the performance of the XGBoost classifier, as the accuracy of the ensemble voting method did not exceed 72% (Table 4.5, Figure 4.5).

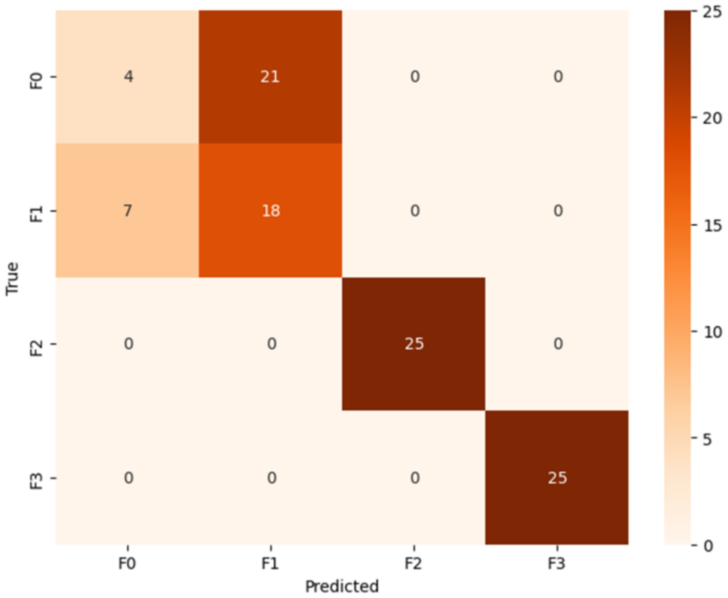
The confusion matrix (Table 4.6) confirmed the limited ability to distinguish between different fault types. In addition, GWO required 115.53 seconds to reach convergence (Figure 4.6), indicating slow execution and low efficiency compared to the AGWO algorithm.

Table 4.5. Feature Selection and Classification Performance of GWO

Faults	Precision	Recall	F1-score	Support
F0 Normal state (Class 0)	0.36	0.16	0.22	25
F1 (Class 1)	0.46	0.72	0.56	25
F2 (Class 2)	1.00	1.00	1.00	25
F3 (Class 3)	1.00	1.00	1.00	25

Overall Accuracy = 0.72

Table 4.6. Confusion Matrix for Ensemble Model with GWO Feature Selection



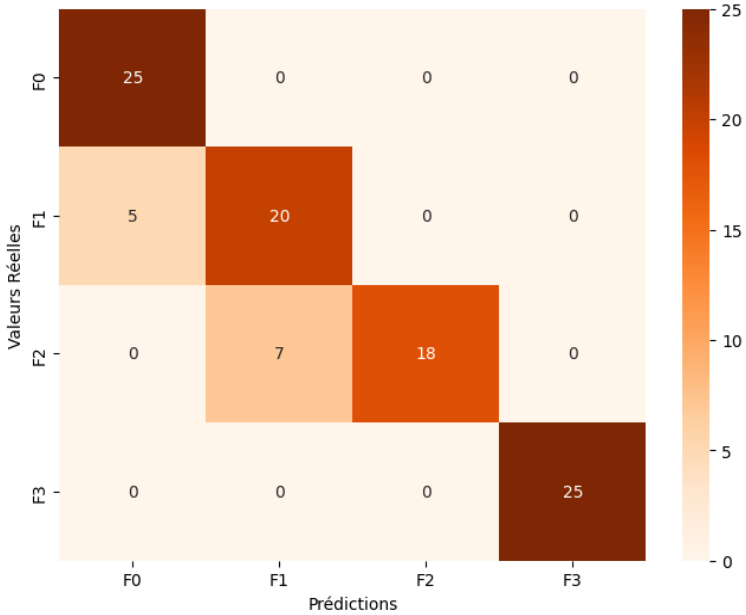
In contrast, the particle swarm optimization (PSO) algorithm selected a much larger set of features (18 out of 30), including I2, I1MAX, I2MAX, I2MIN, I3, I4, I3MIN, I4MAX, I5, I6, Itotal1, Itotalmin1, Vdcmax1, IR, range 1, range 2, range 3, and range 4, outperforming GWO (Figure 4.4). This broader feature representation improved the performance of the XGBoost classifier, increasing accuracy to 88%. Although this improvement was moderate, PSO showed a better ability to achieve balanced classification across different faults (Table 4.7, Table 4.8); with an average execution time of 68.87 seconds (Figure 4.6). Nevertheless, AGWO proved superior to both GWO and PSO, achieving the highest classification accuracy of 96% (Figure 4.5) and the shortest execution time of 49.94 seconds (Figure 4.6). This confirms its efficiency in achieving an optimal balance between high accuracy and computational performance.

Table 4.7. Feature Selection and Classification Performance of PSO

Faults	Precision	Recall	F1-score	Support
F0 Normal state (Class 0)	0.83	1.00	0.91	25
F1 (Class 1)	0.74	0.80	0.77	25
F2 (Class 2)	1.00	0.72	0.84	25
F3 (Class 3)	1.00	1.00	1.00	25

Overall Accuracy = 0.88

Table 4.8. Confusion Matrix for Ensemble Model with PSO Feature Selection



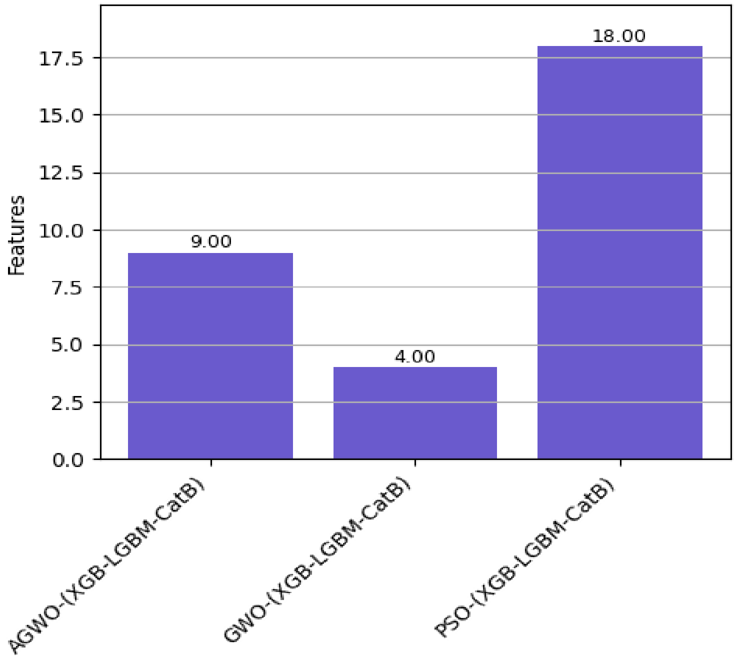


Figure.4.4 Number of Selected Features per Model

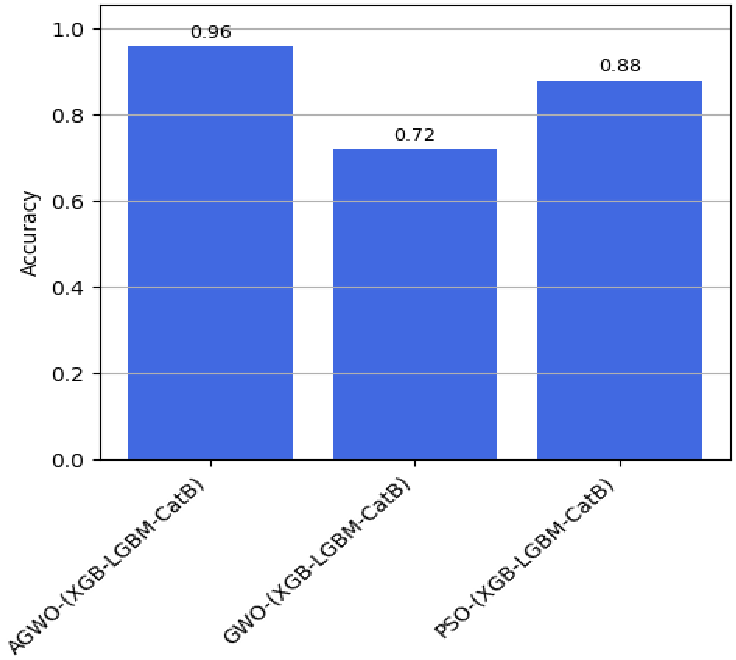


Figure.4.5 Accuracy Comparison

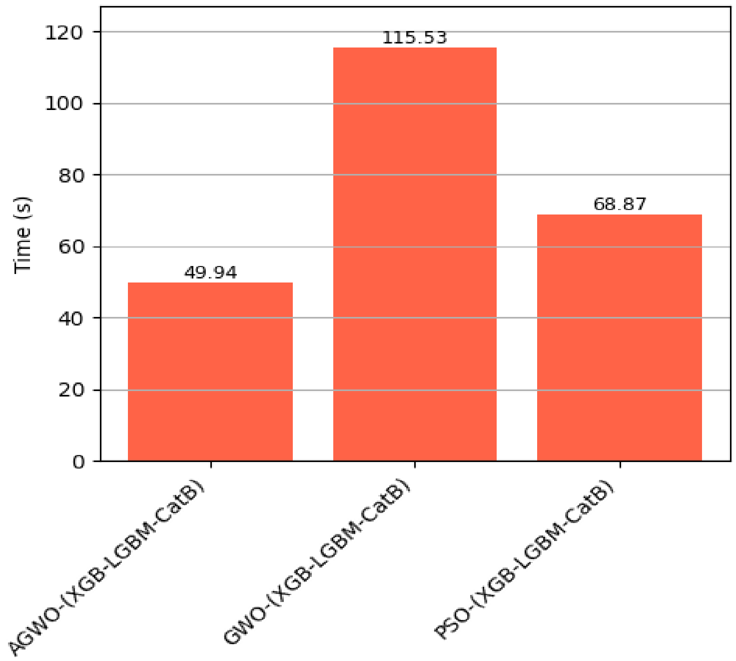


Figure.4.6 Execution Time per Model

4. Stability and Robustness of Selected Feature Subsets

To evaluate the stability of the AGWO Algorithm in feature selection, as well as the consistency of its results and their impact on classification performance, multiple experiments were conducted over 30 runs, each using 50 agents and 100 iterations, with varying random seed values to account for the stochastic nature of the algorithm. All experiments were performed on the same training and test datasets, following the same methodology as in the previous experiments. In each run, the quality of the selected features was assessed by measuring classification accuracy, while also recording the selection frequency of each feature across all runs to identify the most stable and influential features in enhancing the model's performance.

- **Feature Selection Consistency and Model Robustness**

Twelve new features (I1, I1VAR, I2, I2MIN, I2VAR, I3, I3MIN, I5, I6, IR, T, and Vdcmin1) were identified within the set of selected features across thirty independent iterations. Although the selection frequencies of these features varied between iterations, the consistent selection of specific features indicates that they provide complementary information that significantly contributes to improving the model's classification performance, particularly considering the stochastic nature of the AGWO algorithm. Furthermore, features such as range_1, range_3, range_4, and I2VAR demonstrated strong stability across all thirty iterations, highlighting their essential role in supporting reliable predictive performance. Conversely, other features, including I1MIN, I3, I3MIN, and T, appeared with moderate frequency (between 11 and 14 times), suggesting that while their selection was less consistent, they offer additional predictive support that enhances overall model accuracy (Figure 4.7). Regarding the original features, eight of them remained consistently selected across all iterations, reflecting high stability and confirming their critical contribution to the predictive reliability of the model. Conversely, some original features, such as Itotalmin1, were not selected in any iteration, confirming that the feature selection process was driven by statistical relevance and predictive utility rather than prior inclusion. Overall, the model achieved a mean classification accuracy of 0.8923 with a low standard deviation of 0.0926, indicating that the combination of complementary new features and consistently stable original features yields a robust and reliable predictive model capable of capturing complex and nonlinear relationships in photovoltaic system datasets. (See

Figure 4.7). These results confirm the algorithm's ability to identify a stable and influential subset of features, which enhances classification accuracy and supports the robustness and reliability of the model.

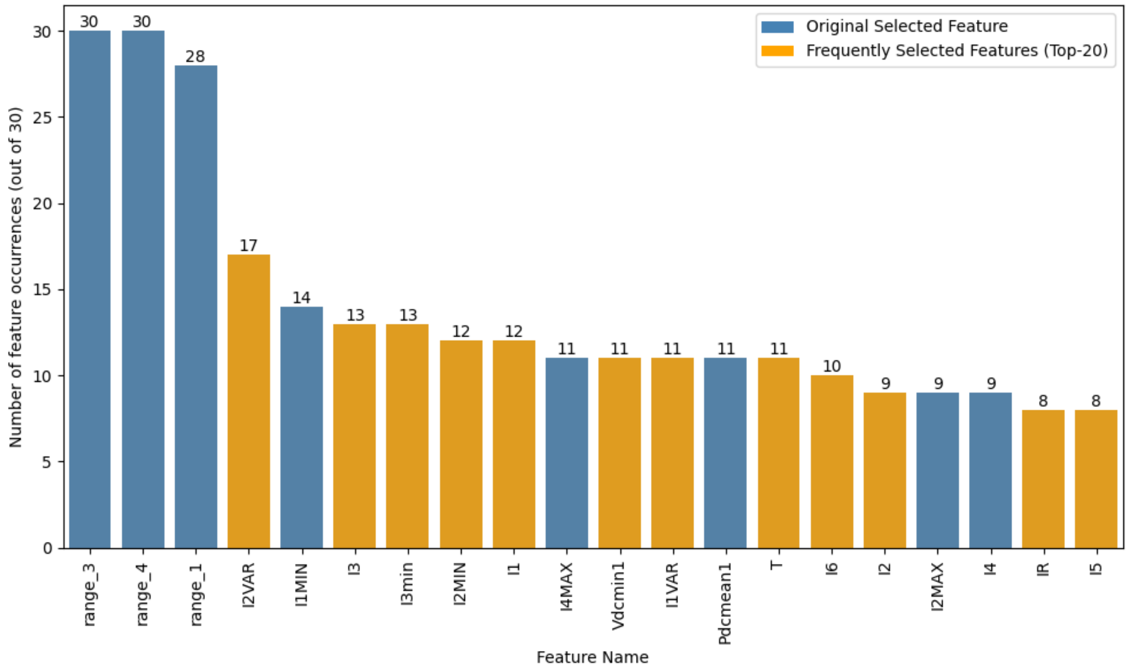


Figure.4.7 Number of Feature Appearances (out of 30 Iteration)

- **Stable Feature Set Analysis**

Analysis of the Venn diagram indicate that the original group consists of nine characteristics, with only one characteristic, Itotalmin1, absent, reflecting its instability across repeated experiments. The intersection between the two groups consists of eight common features (I1MIN, I2MAX, I4, I4MAX, Pdcmean1, range1, range3 and range4) that are key contributors to the robustness of the model, as shown in Figure.4.8.These results show that the model has a high degree of stability in feature selection, relying on a set of consistent variables that have proven their repeatability and reliability across multiple experiments.

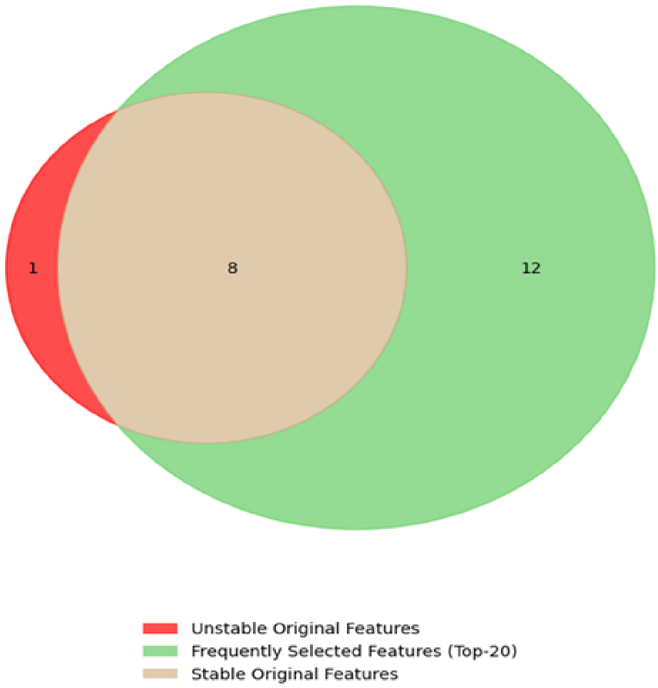


Figure.4.8 Venn Diagram of Feature Selection

- **Analysis of Feature Stability and Reliability Using Heatmaps**

The heatmap in Figure .4.9 illustrates the frequency distribution of the selected original features across thirty independent experiments, where the color gradient reflects the strength and stability of each feature within the model. Darker colors represent higher frequencies, indicating the stability and reliability of these features throughout the selection process across different experiments. This distribution serves as an important indicator of the consistency of the features within the model and helps distinguish between core features with significant impact and less stable ones.The results also demonstrate the effectiveness of the AGWO in selecting the most stable and reliable features. These selected features contributed to achieving a high classification accuracy of 96%, highlighting the model's ability to rely on the most consistent core features to deliver robust and reliable performance across different experiments.

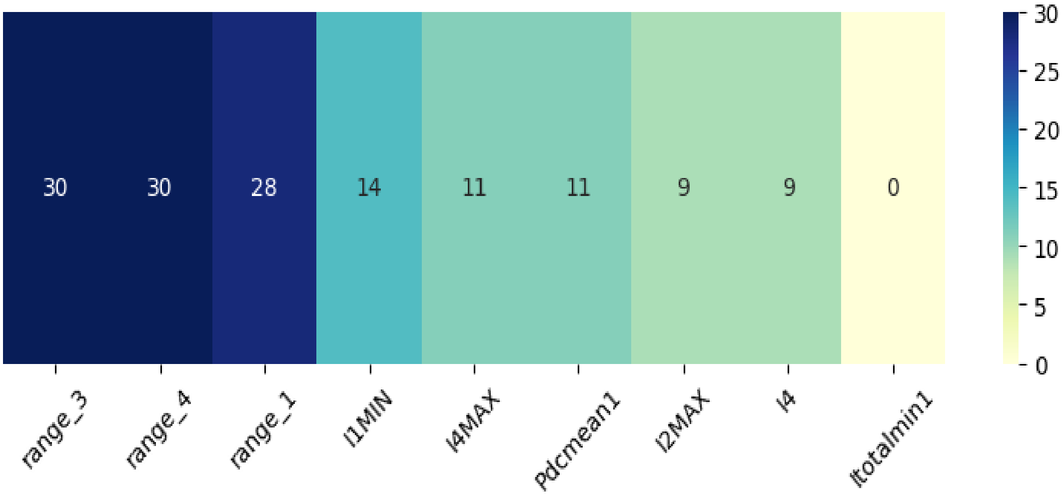


Figure.4.9 Heatmap of Original Feature Frequencies

4.6 Conclusion

The obtained results demonstrate that the proposed AGWO algorithm effectively selects a significant subset from thirty features, providing a concise and informative representation of the data while reducing computational complexity and noise from irrelevant features.

The adaptive control of the parameter a enables an effective balance between exploration and exploitation, mitigating the risk of local minima and enhancing the diversity of selected features. SHAP analysis confirmed that the chosen features have a clear impact on the classification of each class, highlighting AGWO 's ability to capture the essential signals in the data. Compared to conventional methods such as GWO and PSO, AGWO outperformed these approaches, achieving the highest classification accuracy (96%) while maintaining the shortest execution time (49.94 seconds), reflecting its efficiency in balancing predictive performance and computational cost. Furthermore, stability analyses across multiple experiments showed that AGWO maintains a consistent selection of features, with a stable set of key variables, reinforcing the reliability of the model and its consistent performance. These findings indicate that AGWO provides an effective and reliable feature selection strategy, enhancing both predictive accuracy and operational efficiency for high-dimensional datasets. Building on these promising results, the next chapter focuses on the development of DGWO , which integrates explainable principles to combine high predictive performance with transparent diagnostic decision-making.

Chapter 5: Rule-Based Classification for Photovoltaic Fault Diagnosis Using DGWO

5.1 Introduction

Given the high efficiency demonstrated by the GWO in feature selection and classification performance, as reviewed in the previous chapter through AGWO , most machine learning techniques still face a fundamental challenge: their "black-box" nature, where high predictive accuracy is often achieved at the expense of transparency and interpretability in decision-making. To address this limitation, an enhanced variant of the GWO , known as the DGWO , has been developed. DGWO is specifically designed for discrete classification tasks and is integrated with XAI principles, enabling the extraction of clear and interpretable diagnostic rules. This diagnostic approach leverages a search mechanism inspired by the social behavior of wolves, allowing it to uncover the underlying relationships between system variables and fault conditions, while overcoming traditional limitations such as the difficulty of modeling complex system behaviors and premature convergence. The model was evaluated using a dataset simulating faults in PV systems within the MATLAB/Simulink environment, ensuring a systematic and rigorous assessment of diagnostic performance.

Overall, this work provides a comprehensive methodological foundation for deriving transparent diagnostic rules that enhance system reliability and support informed decision-making in the monitoring and diagnosis of PV systems, effectively balancing predictive accuracy with interpretability.

5.2 Photovoltaic Fault Diagnosis Through Supervised Classification

The objective of fault diagnosis in PV systems is to associate electrical features with known operating conditions (see Chapter 2, p. 37). In conventional PV systems, typically consisting of a single series or a limited number of modules, with simple inverters and basic control strategies, the system behavior could be effectively simplified and accurately modeled using physics-based mathematical models. These models were not merely theoretical tools; they were capable of predicting system performance under varying operating conditions, making them highly useful for practical applications and early fault detection in simple PV systems.

With the development of modern PV systems, which incorporate multiple strings, multiple inverters, advanced control strategies and dynamic environmental effects, a high-dimensional complexity challenge has emerged. Traditional mathematical models are no longer capable of accurately predicting system behavior due to complex interactions among electrical, environmental and operational variables. As highlighted in Chapter 4 (p. 71), this study utilizes data from a 250 kW PV system [125] characterized by high structural and dynamic complexity, rendering conventional physics-based modeling insufficient for precise fault diagnosis. The dataset exhibits high-dimensional features, transient behaviors, overlapping fault signatures and probabilistic variations, which significantly complicate the development of a comprehensive mathematical model capable of capturing all operating conditions. Consequently, relying solely on mathematical models for accurate fault diagnosis is extremely challenging and this complexity often negatively impacts diagnostic accuracy.

To address this challenge, a data-driven approach has been adopted, allowing the system to extract complex patterns and interactions directly from measured data without relying on simplifications that might compromise diagnostic accuracy. The nature of the dataset high dimensional, rich and diverse makes conventional modeling techniques inadequate for effective analysis.

Within a supervised classification framework, the system learns a mapping function from a labeled training dataset containing electrical variables such as current, voltage and their deviations, with the goal of detecting faults and classifying them by type. Although regression models are widely used for PV system analysis and modeling and remain valuable for predicting continuous operational variables and performance estimation, their numerical outputs limit direct applicability for fault diagnosis. Converting these outputs into discrete diagnostic states requires additional processing, such as thresholding or decision rules, which can increase uncertainty, particularly in complex and dynamic operating environments.

In contrast, supervised classification provides a more suitable framework for generating clear and decisive diagnostic outputs, enabling direct identification of fault types and supporting the implementation of corrective actions. Practically, this approach allows continuous and precise monitoring of industrial PV systems, immediate detection of faults and classification according to type and severity, facilitating rapid operational decisions, minimizing energy loss and maintaining grid stability and efficiency. Furthermore, the data-driven approach exhibits high adaptability to environmental and operational variations, a critical feature for achieving reliable and sustainable performance in large and complex industrial installations.

5.3 Optimization of PV Fault Classification Using DGWO

Using the previously extracted features, the GWO algorithm was employed as a metaheuristic optimization approach to enhance fault classification accuracy in PV systems. This algorithm seeks to improve the rule extraction process by identifying the optimal values of critical parameters that directly influence mode performance, thereby increasing its accuracy, reliability and interpretability [93]–[107]. The effectiveness of GWO has been demonstrated across a wide range of optimization applications [118]–[124], highlighting its suitability for addressing complex classification problems [133]–[139].

To adapt classical GWO for classification tasks operating within discrete search spaces, the DGWO was developed. Unlike its continuous counterpart, DGWO relies on symbolic rule representations instead of numerical vectors, making it particularly suitable for rule-based classification systems. Each candidate solution (wolf) is encoded as a discrete set of rules and the population is guided through the search space using a modified position update mechanism that preserves the hierarchical leadership model (α, β, δ) of the original GWO . Furthermore, discrete recombination and mutation operators are incorporated to enhance exploration and maintain diversity among candidate solutions.

In the context of the DGWO engine used in this chapter, the optimization process does not aim to search for numerical weights, but rather to discover the most accurate and concise symbolic logic that minimizes misclassification. This clear-box logic is particularly valuable for PV fault diagnosis, as it enables engineers to understand the reasoning behind each detected fault, thereby improving system trust and facilitating faster and more effective maintenance interventions.

Figure 5.1 illustrates the architecture of the DGWO based diagnostic system for PV faults. This system contributes to overall efficiency improvement and accelerates convergence when addressing complex optimization challenges.

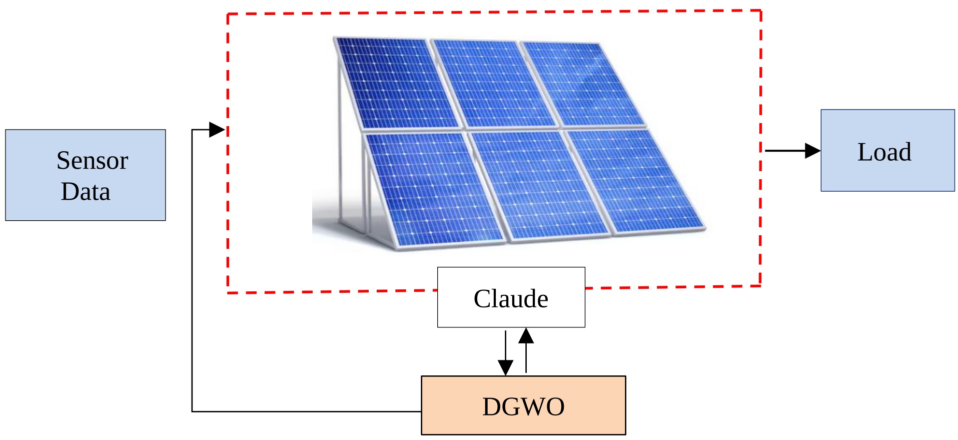


Figure.5.1 Structure of the DGWO Based System for Diagnosing PV Faults

Figure.5.2 presents the overall structure of the proposed DGWO based framework, highlighting its modular and systematic design.

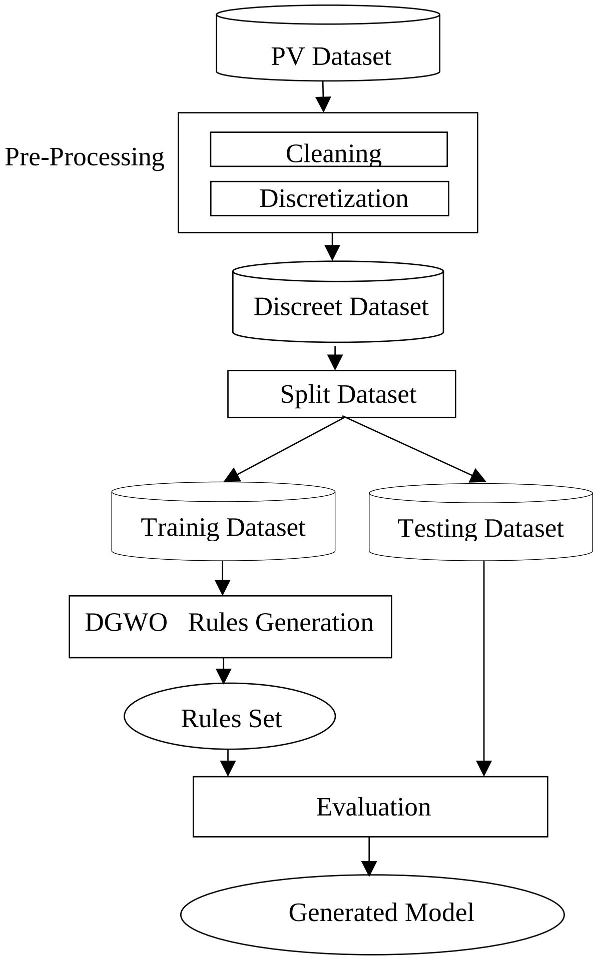


Figure.5.2 Proposed DGWO Method

5.3.1 Structure of Rule-Based Classification

Unlike advanced classification methods such as Neural Networks or Support Vector Machines, which are often considered "black-box" models with opaque decision-making, Rule-Based Classification employs a symbolic approach that emphasizes transparency and interpretability. This methodology is part of XAI, allowing experts to understand, audit and validate the models effectively.

A rule-based classifier consists of independent logical units called IF–THEN rules, each encoding explicit diagnostic knowledge. Each rule comprises two essential components:

- **Antecedent (Condition)**

The antecedent specifies a set of logical constraints applied to input features. In real-world applications, these constraints are typically composite, involving multiple attributes (e.g., current, voltage, or derived performance metrics) combined using logical operators such as AND or OR. Each composite condition represents a precise diagnostic criterion, defining the circumstances under which a particular outcome is inferred.

- **Consequent (Conclusion)**

The consequent indicates the resulting class, state, or decision that follows when the antecedent conditions are satisfied. It represents the explicit knowledge output of the rule and provides a direct, interpretable link between input variables and the classification outcome.
Example of a composite rule:
IF Attribute X = Value 1 AND Attribute Y = Value 2 AND Attribute Z = Value 3 **THEN** the class is [Class Label].

5.3.2 DGWO Fitness Function

Within the framework of the DGWO algorithm, each wolf represents a candidate solution, specifically a potential diagnostic rule. These solutions are evaluated using a fitness function, which serves as a central evaluator of solution quality based on a support-based relevance measure as defined in (5.1) [140], [141], considering accuracy, generalizability and simplicity. Wolves are ranked according to their performance, with the Alpha wolf representing the current best solution and serving as the ultimate target the algorithm seeks to achieve, while the remaining wolves, including Beta, Delta and the followers, act as assistants by progressively adjusting their positions and solutions to approach the Alpha. Through this adaptive mechanism, candidate rules are gradually improved, becoming increasingly accurate, reliable, generalizable and structurally simple, thereby enhancing the effectiveness of DGWO in extracting precise and efficient diagnostic rules.
Accordingly, the fitness function is designed to prioritize rules that achieve high classification accuracy, strong generalizability across instances and structural simplicity, as formalized in (5.2).

$$Support\ metric = \frac{ICC}{TI} \tag{5.1}$$

$$F(R) = \frac{TI}{(TI - ICC) * (TI + INCC\ )} \tag{5.2}$$

where each term is defined as follows:

R : The candidate diagnostic rule (the " wolf ").

TI (Total Instances) : The total number of observations in the training set.

ICC (Instances Correctly Covered) : The number of observations in the training set that satisfy the rule’s conditions and belong to the correct fault class.

INCC (Instances Incorrectly Covered) : The number of observations in the training set that satisfy the rule’s conditions but belong to a different fault class (false positives).

5.3.3 Data Discretization

The dataset employed in this study was partitioned into training and testing subsets. The training dataset comprises 600 instances representing both fault conditions and normal operation, while the testing dataset consists of 100 instances [125] . To ensure data quality, rows containing highly anomalous values (e.g., rows 71, 148, 319) as well as rows with incomplete data (e.g., rows 14, 27) were removed, as reported in Table 5.1.

Prior to applying the DGWO Algorithm, the data were discretized into intervals based on categorical features. The discretization process was further refined by adjusting the values of the last four columns (range 1 – range 2 – range 3 – range 4).

This adjustment was not performed

arbitrarily but was instead derived from a careful analysis of both the training and testing datasets, which revealed that these intervals were significant relative to the remaining data.

Moreover, the primary objective of this study is to generate a set of classification rules that are efficient in terms of both size and number. The size criterion ensures that only critical Variable with direct influence are retained, while irrelevant variables are excluded, thereby yielding results that emphasize solely the factors most strongly associated with the studied fault.

Table 5.1. Discretized Features of the PV Frame Dataset

Range	Interval 1	Interval 2	Interval 3	Interval 4
I1	< (-4.5074)	(-4.5074) - (-1.1168)	(-1.1168) -(2.2738)	> (2.2738)
I2	<(1.62149)	(1.62149 - 3.04446)	(3.04446 - 4.46743)	>(4.46743)
I1MAX	< (1.876247)	(1.876247 - 3.169064)	(3.169064 - 4.461882)	>(4.461882)
I1MIN	< (-15.0132)	(-15.0132) - (-8.1244)	(-8.1244) - (-1.2356)	> (-1.2356)
I1VAR	<(45.7125)	(45.7125)-(91.425)	(91.425)-(137.1375)	>(137.1375)
I2MAX	<(1.926653)	(1.926653)-(3.252935)	(3.252935)-(4.579218)	>(4.579218)
I2MIN	<(1.321445)	(1.321445)-(2.79963)	(2.79963)-(4.277815)	>(4.277815)
I2VAR	<(0.190158)	(0.190158)-(0.380315)	(0.380315)-(0.570473)	>(0.570473)
I3	<(1.912674)	(1.912674)-(3.221783)	(3.221783)-(4.530891)	>(4.530891)
I4	<(1.620377)	(1.620377)-(2.976385)	(2.976385)-(4.332392)	>(4.332392)
I3max	<(1.919827)	(1.919827)-(3.234918)	(3.234918)-(4.550009)	>(4.550009)
I3min	<(1.878742)	(1.878742)-(3.171028)	(3.171028)-(4.463314)	>(4.463314)
I3var	<(0.03523)	(0.03523)-(0.07046)	(0.07046)-(0.10569)	>(0.10569)
I4MAX	<(1.877475)	(1.877475)-(3.17115)	(3.17115)-(4.464825)	>(4.464825)
I4MIN	<(1.610297)	(1.610297)-(2.961532)	(2.961532)-(4.312766)	>(4.312766)
I5	<(1.865541)	(1.865541)-(3.139827)	(3.139827)-(4.414114)	>(4.414114)
I6	<(1.865541)	(1.865541)-(3.139827)	(3.139827)-(4.414114)	>(4.414114)
I total	<(449.2375)	(449.2375)-(472.505)	(472.505-495.7725	>(495.7725)
I totalmax1	<(472.125)	(472.125)-(488.03)	(488.03)-(503.935)	>(503.935)
I totalmin1	<(424.227447)	(424.227447)-(455.374965)	(455.374965)-(486.522482)	>(486.522482)
Vdcmean1	<(485.298969)	(485.298969)-(499.982505)	(499.982505)-(514.666041)	>(514.666041)
Vdcmax1	<(489.839388)	(489.839388)-(504.389592)	(504.389592)-(518.939796)	>(518.939796)
Vdcmin1	<(481.442992)	(481.442992)-(493.571327)	(493.571327)-(505.699662)	>(505.699662)
Pdcmean1	<(84.190451)	(84.190451)-(143.983634)	(143.983634)-(203.776817)	>(203.776817)
IR	<(328)	(328)-(552)	(552)-(776)	>(776)
T	<(16.25)	(16.25)-(22.5)	(22.5)-(28.75)	>(28.75)
Range1	<(0.00193)	(0.00193)-(0.1069)	(0.1069)-(5.3141)	>(5.3141)
Range2	<(0.00193)	(0.00193)-(0.1069)	(0.1069)-(5.3141)	>(5.3141)
Range3	<(4.25)	(-4.25)-(-0.0662)	(-0.0662)-(0)	>(0)
Range4	<(0)	(0)-(3.10E-14)	(3.10E-14)-(0.0621)	>(0.0621)

5.3.4 DGWO Rule Generation

The proposed rule generation framework for fault diagnosis is based on DGWO , which provides a structured and interpretable search mechanism inspired by the social hierarchy and cooperative hunting behavior of grey wolf packs. Figure 5.3 illustrates the overall flowchart of the proposed DGWO -based rule-based fault diagnosis approach.

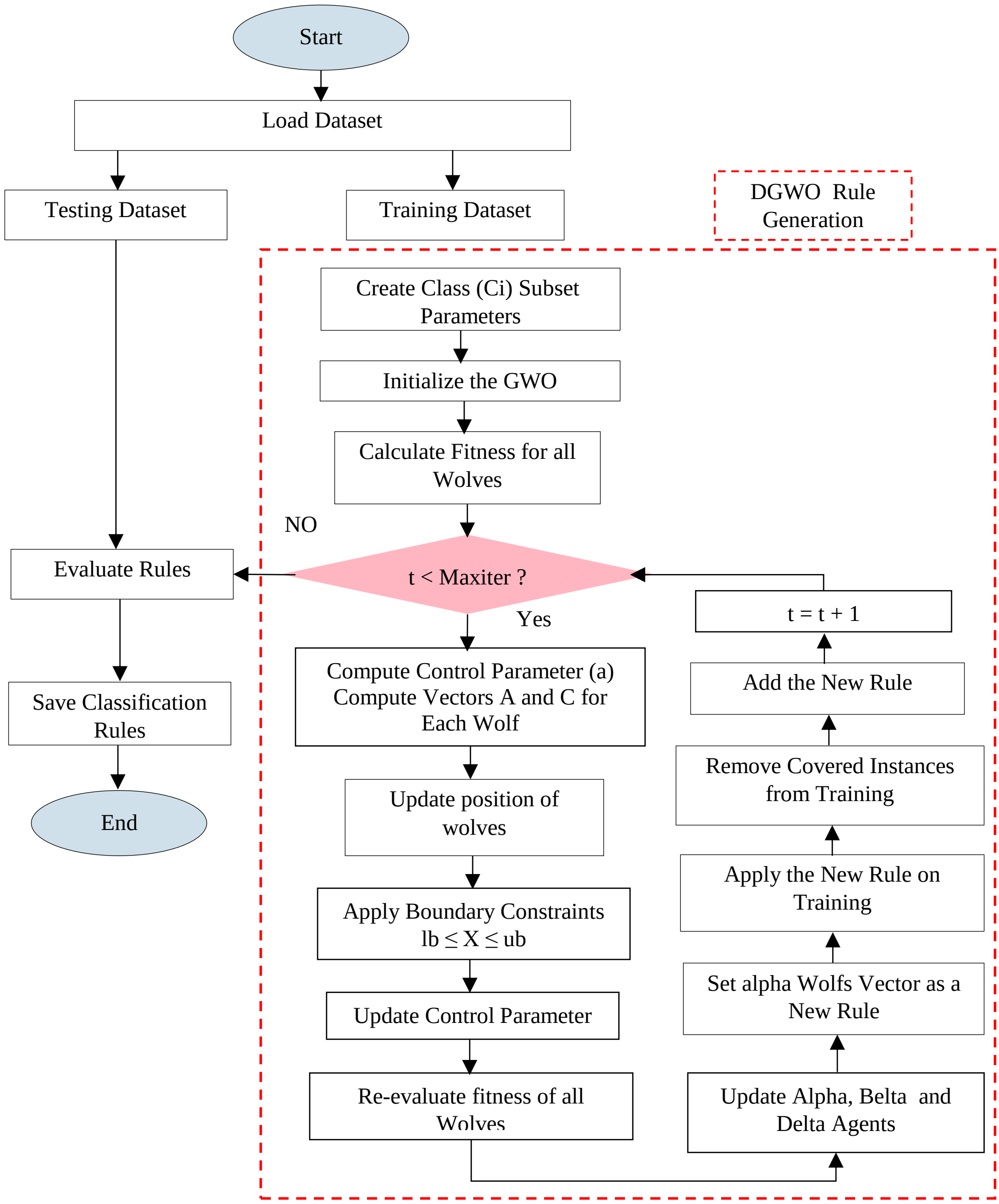


Figure.5.3 Flowchart of the Proposed Rule-Based Fault Diagnosis Approach

The diagnostic process starts with loading the dataset and partitioning it into training and testing subsets, followed by a discretization stage applied to continuous features to enable rule-based reasoning. Within the resulting discrete rule space, each wolf encodes a candidate classification rule composed of logical antecedents defining feature intervals and a consequent associated with a specific fault class Ci. The DGWO algorithm emulates the Alpha–Beta–Delta leadership hierarchy to guide the search toward optimal diagnostic rules. Specifically, the Alpha wolf denotes the rule with the highest fitness value in the current iteration, achieving an effective trade-off between diagnostic accuracy and interpretability, while the Beta and Delta wolves represent the second- and third-best rules, respectively, thereby preserving solution diversity and mitigating premature convergence to suboptimal or overly specialized rules. By computing logical distances between the current population and the three leading wolves, DGWO updates the antecedent boundaries of the remaining rules, ensuring a balanced exploration–exploitation process. Over successive iterations, rule antecedents are progressively refined to identify the most discriminative feature regions associated with different fault patterns, such as partial shading or short-circuit conditions, ultimately enhancing the robustness and reliability of the diagnostic system.Building upon this conceptual framework, the practical rule generation process isimplemented through a sequential and systematic pipeline, summarized as follows:

- **Initialization**

A population of wolves is randomly initialized, where each wolf encodes a candidate classification rule associated with a specific fault class Ci, following the discretization of continuous features.

- **Iterative Search**

  The discrete rule space is explored iteratively under the guidance of the Alpha–Beta–Delta hierarchy, with candidate rules updated according to a fitness function designed to balance classification accuracy and interpretability.

- **Rule Extraction and Data Reduction**

The rule represented by the Alpha wolf is selected as a final diagnostic rule and added to the rule set. Subsequently, the training samples covered by this rule are removed from the dataset to reduce redundancy and improve extraction efficiency.

- **Repetition**

The DGWO algorithm is re-applied to the remaining data to discover additional rules that capture other diagnostically significant fault patterns.

- **Termination**

The process terminates once sufficient coverage of all fault classes is achieved or when no meaningful patterns can be further extracted, ensuring the construction of a comprehensive, accurate and interpretable diagnostic model.

5.4 Evaluation of DGWO Performance and Diagnostic Results

A. Optimization Efficiency and Parameter Sensitivity

The performance of the DGWO algorithm was evaluated under varying numbers of wolves and iterations, as summarized in Table 5.2. The results demonstrate clear trends regarding both optimization efficiency and parameter sensitivity. It is evident that the best performance was achieved with 110 wolves and 10 iterations, yielding a classification accuracy of 0.99 and a Configurations with fewer wolves and iterations (e.g., 20 wolves and 5 iterations, 50 wolves and 8 iterations) resulted in moderate performance, with accuracy and precision around 0.81 and 0.75, respectively. This suggests that a limited population or insufficient iterations may hinder the algorithm from effectively exploring the search space, thus reducing its ability to accurately classify the PV system data.

Interestingly, increasing the number of wolves beyond the optimal configuration (e.g., 120, 135, 180, or 200 wolves) did not improve performance. In some cases, accuracy and precision even declined (e.g., 120 wolves and 15 iterations: accuracy 0.76, precision 0.62). This indicates that excessively large populations can introduce redundancy and computational overhead, which may reduce the algorithm's efficiency and overall performance.

Overall, these findings emphasize the sensitivity of DGWO to parameter settings. Optimal performance is achieved by balancing the number of wolves and iterations to maximize accuracy while minimizing computational cost. Selecting inappropriate configurations can significantly degrade performance, highlighting the importance of systematic parameter tuning in practical applications.

Table 5.2. DGWO Performance Under Varying Numbers of Wolves and Iterations

Number of Wolves	Number of Iterations	Accuracy	Precision
20	5	0.81	0.75
50	8	0.81	0.75
110	**10**	**0.99**	**1.00**
120	15	0.76	0.62
135	25	0.81	0.75
180	30	0.78	0.65
200	50	0.82	0.70

B. Quantitative Validation of Diagnostic Results

Table 5.3 presents the set of rules generated by the DGWO algorithm, which are used to classify the operating states of the PV system into four categories: the normal state F0 (Class 0) and three fault conditions F1, F2, and F3 (Class 1, Class 2, and Class 3, respectively). Each rule is formulated in an If–Then logical structure based on specific ranges of extracted feature values, with each rule corresponding to a single class.

The number of terms varies from one rule to another. Specifically, Rules 1, 2, and 3 rely on two conditions, whereas Rule 4 is based on a single condition. This variation reflects the DGWO algorithm's ability to simplify rules when the distinctive features of a class are clear and well defined.

In terms of performance, Rule 4 (Class 3) achieved the best results, covering all instances (198/198) with zero misclassifications, thus attaining perfect accuracy. This indicates a strong and distinctive pattern for this class within the feature space, allowing it to be represented by a simple and highly efficient rule.

In contrast, Rule 1 (Class 0) showed the lowest relative accuracy, correctly covering 94 out of 100 cases. This reduction is attributed to overlaps between normal-state data and some fault conditions, which is a common challenge in data-driven fault diagnosis systems.

Rules 2 (Class 1) and 3 (Class 2) demonstrated balanced performance, achieving high coverage with good accuracy, confirming the effectiveness of the selected features in distinguishing these fault types.

Overall, these findings highlight the robustness of the proposed rule-based model in accurately classifying both normal and faulty conditions.

Table 5.3. Resulting Rules

Rule #	Generated rules	Class	Number of terms	Number of instances correctly covered	Number of instances not correctly covered	rule's accuracy
01	If range1 in range (0.00193)-(0.1069) and range3 in range >(0) then Normal mode (F0)	Class0	2	94/100	0	16 %
02	If range1 in range (0.1069)-(5.3141) and range3 in range > (0) then Default1 (F1)	Class1	2	144/146	1	24%
03	If range3 in range (-4.25)-(-0.0662) and range4 in range <(0) then Default2 (F2)	Class2	2	143/144	0	24%
04	If range4 in range >(0.0621) then Default3(F3)	Class3	1	198/198	0	34%

Evaluation Results:

Accuracy = 0.99

Precision = 1.0

Correctly classified instances = 96 / 97.96

Figure 5.4 presents a conceptual diagram of a rule-based classification mechanism for the PV system array, based on rules extracted using the DGWO algorithm. The system analyzes operational features of the array to accurately diagnose its condition and generate an optimized set of decision rules (Rule 1– Rule 4). These rules classify the system into four categories (C0–C3), where C0 represents the normal operating state (F0) with no faults, and C1–C3correspond to distinct fault conditions (F1–F3). The diagram illustrates the sequential stages of data processing, from feature extraction to the final diagnostic decision, highlighting the effectiveness of the DGWO algorithm in distinguishing normal operation from various fault scenarios.

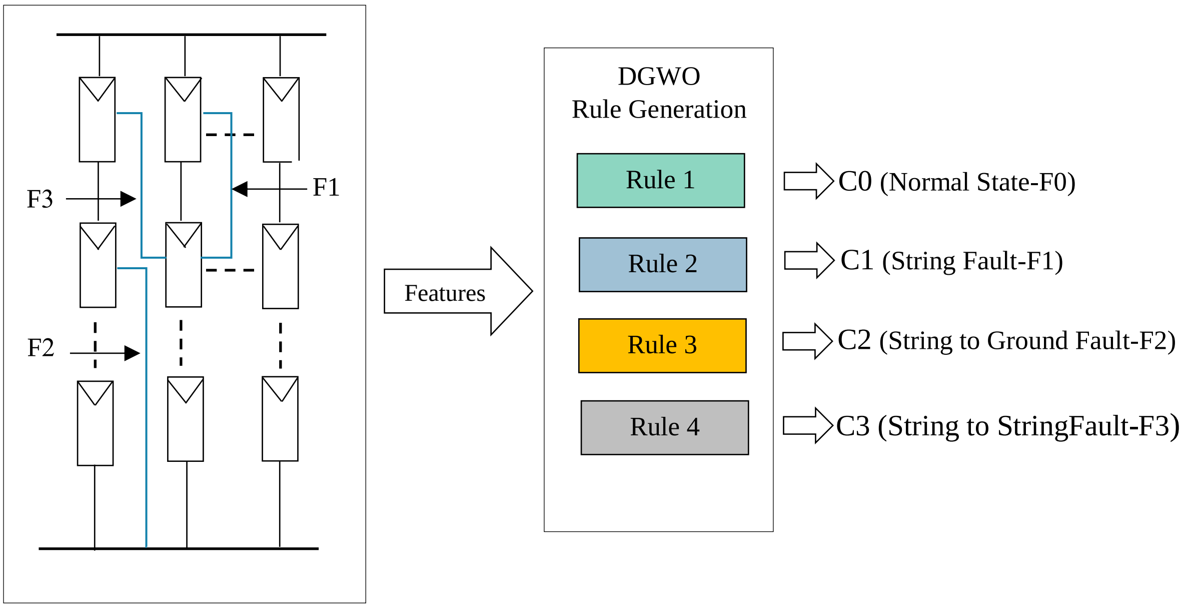


Figure.5.4 Rule-Based Classification Diagram for Photovoltaic System Fault Diagnosis

C. Diagnostic Results Based on the Extracted Rule Set

The diagnostic performance of the proposed rule-based model, applied to the PV system is presented in this section. Table 5.4 summarizes the values of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), as well as accuracy and precision for each class. The results indicate that classes C0, C1 and C3 achieved perfect classification performance, with no false positives or false negatives, resulting in an accuracy and precision of 1.00 for each class. In contrast, class C2 exhibited two false negative cases, which slightly reduced the accuracy to 0.98, while precision remained perfect at 1.00, as evaluated using Equations (4.2) and (4.5) defined in Chapter 4.

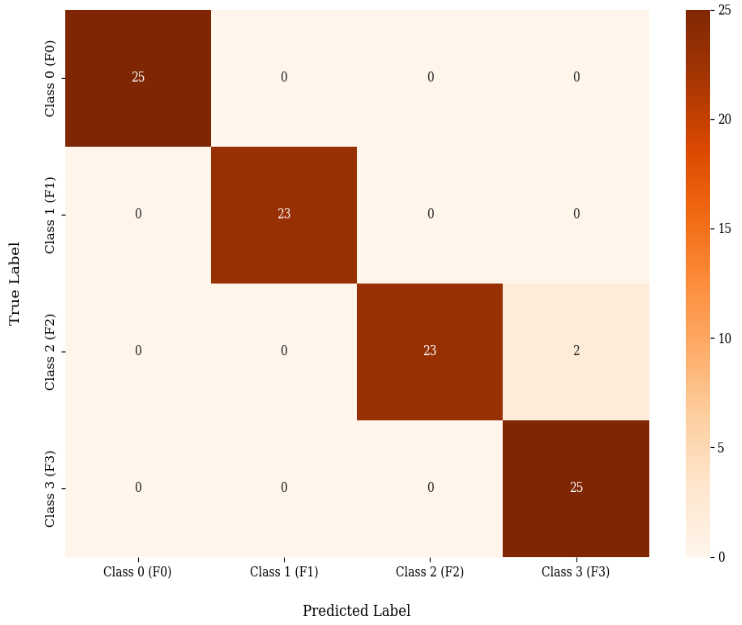
Furthermore, the confusion matrix presented in Table 5.5 provides a comprehensive overview of correct and incorrect predictions across all classes, confirming the high reliability of the rule-based classifier, with minor errors limited exclusively to class C2.

Overall, these findings demonstrate that the rules extracted using the DGWO algorithm enable accurate and reliable classification of the solar energy system conditions, effectively handling all classes with only a minimal number of misclassifications.

Table 5.4 Model Performance

Classes	TP	FP	TN	FN	Accuracy	Precision
Class C0 (F0)	25	0	73	0	1.0	1.00
Class C1 (F1)	23	0	75	0	1.0	1.00
Class C2 (F2)	23	0	73	2	0.98	1.00
Class C3 (F3)	25	0	73	0	1.0	1.00

Table 5.5. Analysis of Classifier Performance Using Confusion Matrix



> **Example of DGWO Rule and Explainability Analysis**

Field measurements of the electric current were conducted on a single PV string to evaluate its operational performance and identify potential faults. The measured currents ranged from (−1 A) to (2.5 A).

The internal current variation within the string (Range1) was calculated as:

$$\text{Range1} = I1_{Max} - I1_{Min} = 2.5-(-1) = 3.5 \;\; \text{A}$$

The deviation from the parallel strings (Range3) was measured at 0.12 A. Using the DGWO algorithm, an optimized diagnostic rule for faults in category C0 (Fault Type F1) was derived as follows:

If Range1 ∈ (0.1069 , 5.3141) and Range3 > 0 then the fault type is F1

Comparing the measured values with this optimized rule, the calculated Range1 value of 3.5 A lies within the specified interval, and Range3 is positive. Since both conditions are satisfied, the analyzed PV string is classified as exhibiting a fault of type F1.

The XAI framework interprets this classification by linking the measured parameters to the diagnostic decision. The elevated internal variation indicates significant current non-uniformity among modules, suggesting degraded, partially shaded, or malfunctioning units. The positive deviation from parallel strings confirms that the anomaly is localized to the examined string. This explicit connection ensures transparency and enables maintenance teams to identify defective modules precisely, implement targeted corrective actions, enhance operational efficiency, and reduce energy losses.

D. Graphical Analysis of Discriminative Power (ROC/AUC)

Figure.5.5 presents the results of the receiver operating characteristic (ROC) curve and area under the curve (AUC) metrics, indicating that the model effectively distinguishes between positive and negative instances across all four classes (C0, C1, C2 and C3).

The precision and AUC scores for C0, C1 and C3 are nearly perfect, approaching 1.00, while those for C2 are 0.98 and 0.96, respectively . These values substantially exceed the random classification threshold (AUC = 0.5). Analysis of Classifier Performance Using Confusion Matrix illustrated in Table 5.6 confirms the model's strong classification performance, its robust discriminative power across multiple categories and its high predictive accuracy, supporting its applicability in complex classification tasks.

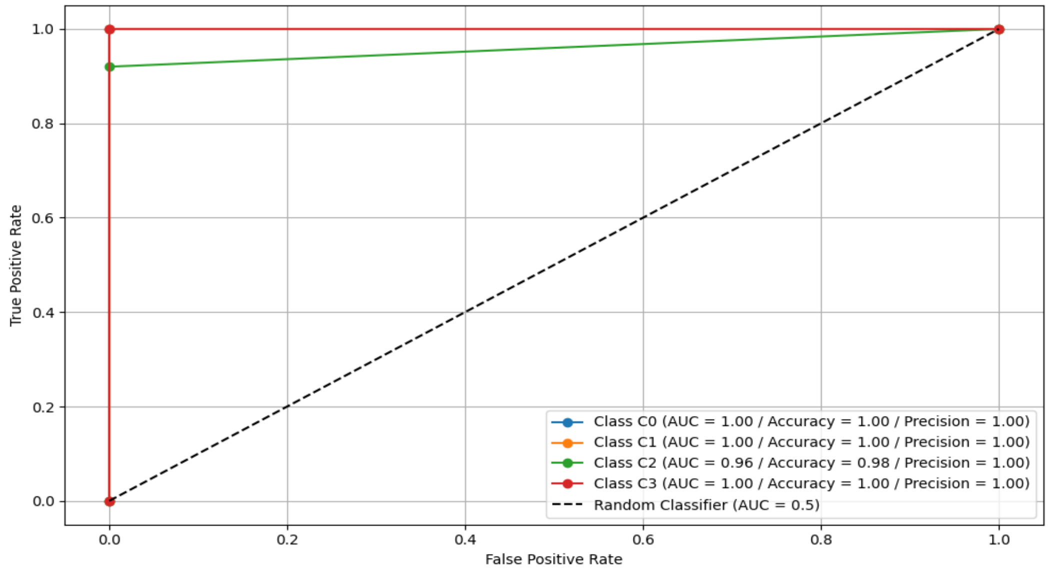


Figure.5.5 Multi-Class ROC Curves with AUC, Accuracy and Precision

5.5 Comparative Benchmarking and Industrial Value

To assess the effectiveness of the proposed DGWO algorithm, a comprehensive comparative evaluation was conducted using two benchmark experiments. The first experiment compared DGWO with standard machine learning classifiers reported in the literature[125], while the second focused on rule-based models implemented within the WEKA environment.

A. Comparison with Standard Machine Learning Classifiers

The DGWO algorithm was benchmarked against widely used classifiers such as AdaBoost and Random Forest, as well as regression-based models. Although AdaBoost achieved 95% accuracy with an F1-score of 94.9% and Random Forest reached 81% accuracy with an F1-score of 80.6%, these ensemble methods function as black boxes, offering limited interpretability. Regression models (Lasso and Ridge) demonstrated lower performance (76% and 71% accuracy, respectively)[125], as shown in Table 5.7 and Figure 5.6.

In contrast, DGWO achieved 99.48% accuracy, an F1-score of 98.96%, 100% precision, 97.96% recall and 100% specificity. Beyond predictive performance, DGWO provides full transparency by generating explicit if–then rules, allowing engineers to understand the exact decision-making logic a critical advantage for high-stakes industrial applications such as PV fault diagnosis.

B. Benchmarking Against WEKA-Based Rule Classifiers

DGWO was also compared with traditional rule-based classifiers available in WEKA, including CN2, ZenoR, Oner and Decision Table. DGWO outperformed all these models in accuracy and interpretability, producing a compact and generalized set of decision rules.

While CN2 (50% accuracy) and Decision Table (73.5% accuracy) were able to generate rules, their predictive performance remained substantially lower than DGWO . ZenoR and Oner failed to produce meaningful rules or achieved very low accuracy (25.5% and 67.6%, respectively) , as summarized in Table 5.6 and illustrated in Figure 5.6. This demonstrates the superiority of DGWO 's socially inspired search mechanism in discovering parsimonious and highly effective decision rules compared to traditional greedy or heuristic-based methods.

Overall, DGWO achieves a strong balance between predictive accuracy, simplicity and transparency, establishing it as a robust and industrially viable solution for decision-making environments where reliability and interpretability are essential.

Table 5.6. Comparative Analysis with Other Classifiers

Model	CA (%)	F1-Score (%)	Precision (%)	Recall (%)	Specificity (%)	Generation the Rules (If-Then)
DGWO	**99.48**	**98.96**	**100**	**97.96**	**100**	**YES**
Adaboost [125]	95	94.9	95.8	95	98.3	NO
Random Forest [125]	81	80.6	83.4	81	93.7	NO
Lasso-Regression [125]	76	75.3	79.9	76	92	NO
Ridge-Regression [125]	71	69.4	73.3	71	90.3	NO
Naives Bayes [125]	60	58.9	58.4	60	86.7	NO
CN2 rule induction [125]	50	48.5	51.8	50	83.3	YES
zenor (WEKA Platform)	25.5	10.4	6.5	25.5	74.5	NO
Oner (WEKA Platform)	67.6	42.9	38.7	51.0	84.1	NO
Decision Table (WEKA Platform)	0.735	66.1	85.6	73.5	90.9	YES

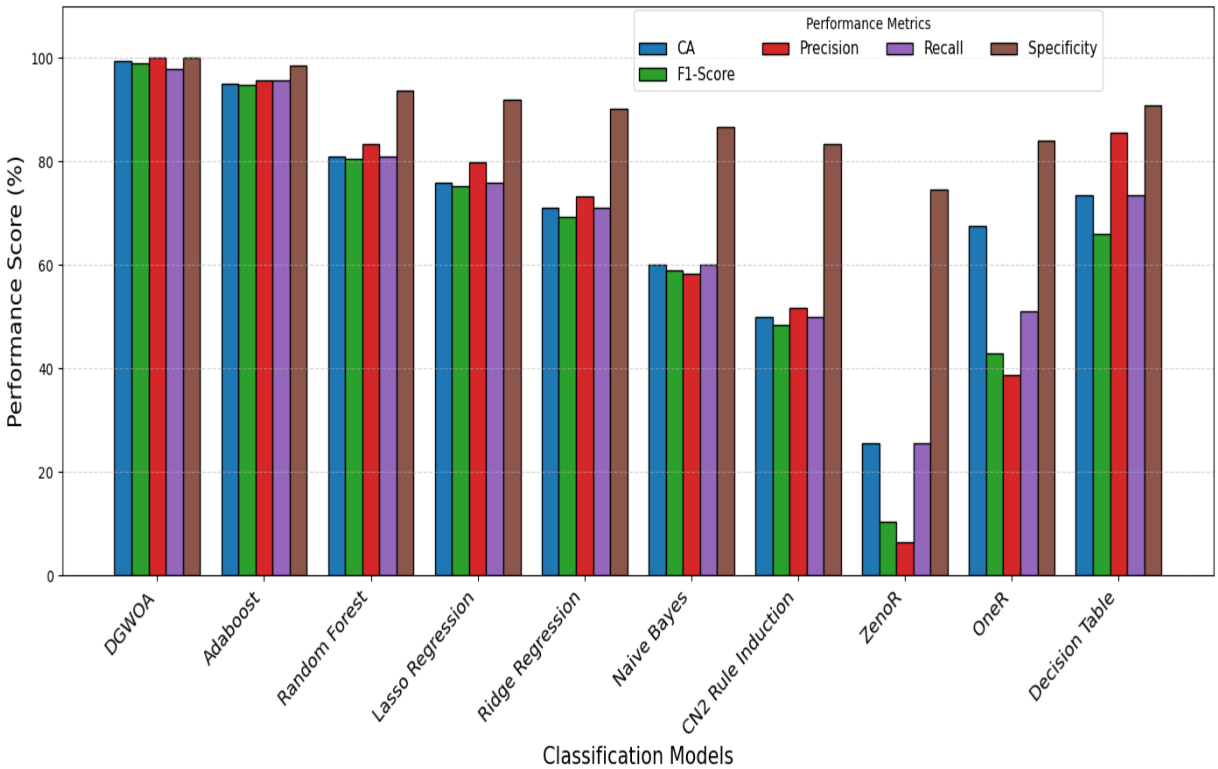


Figure.5.6 Comparing Algorithms Using Different Criteria

5.6 Satistical Robustness and Stability Analysis of the DGWO Algorithm

A. Numerical Stability and Convergence Reliability

As a stochastic population-based metaheuristic, the reliability of the DGWO algorithm must be validated through repeated experimentation. To this end, 30 independent runs were conducted using a population of 110 wolves and 10 internal iterations per run in order to assess numerical stability and convergence behavior.

The results demonstrate that the algorithm achieved the global maximum fitness value of 0.0025641 in 28 out of 30 executions, corresponding to a success rate of 93.3%, which highlights the strong convergence reliability of the proposed method. The arithmetic mean fitness value was 0.00256109, while the median value coincided exactly with the maximum fitness, indicating a highly concentrated distribution around the optimal solution with negligible deviations. Furthermore, the observed standard deviation of 1.128×10^{-5} confirms the extremely low dispersion of the fitness values, underscoring the algorithm's robustness and its reduced sensitivity to random initialization. This behavior indicates that DGWO does not rely on favorable stochastic conditions to achieve high-quality solutions, but instead exhibits consistent and repeatable performance across multiple trials, as illustrated in Figure 5.7.

The minimum fitness value recorded was 0.00251889, occurring in only two runs. Although slightly lower than the optimal value, this marginal deviation can be attributed to temporary entrapment in local optima or reduced population diversity during early exploration phases phenomena commonly observed in metaheuristic optimization algorithms [126], [127]. Nevertheless, the rarity and limited magnitude of these deviations further confirm the numerical stability and overall reliability of the DGWO algorithm.

B. Robustness to Data Noise and Anomaly Resilience

In addition to numerical stability, the robustness of the DGWO engine was evaluated under imperfect data conditions. After the removal of anomalous observations (rows 14, 71 and 148), the algorithm preserved its structural integrity and convergence behavior. Notably, DGWO consistently identified the most resilient feature ranges (Range1 to Range4), demonstrating that the extracted rules are not overfitted to isolated data points.

This resilience indicates that the generated decision rules accurately capture the underlying physical behavior of the PV system rather than reflecting noise-induced artifacts.

Consequently, the DGWO engine proves to be both numerically stable and structurally robust, making it a reliable optimization framework for real-world PV performance modeling.

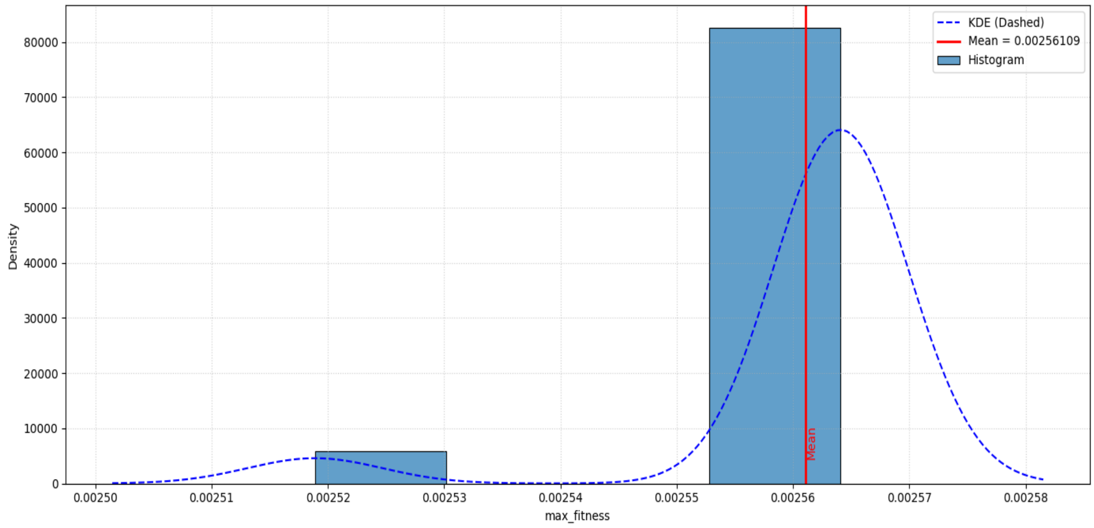


Figure.5.7 Statistical Stability Analysis of DGWO Based on Max Fitness Performance

5.7 Conclusion

This research has shown the efficacy of an upgraded DGWO , in addressing fault detection and diagnostic issues in industrial systems. The proposed method was validated using real-world data from a PV system, wherein continuous variables were discretized to generate interpretable classification rules. The DGWO successfully produced four classification rules, one representing normal operating conditions and three corresponding to fault states (F1, F2 and F3), achieving an overall classification accuracy of 99.48% and a precision of 100% .For a comprehensive assessment the robustness and reliability of the proposed approach, two independent comparative evaluations were conducted. The first involved a set of well-established classification algorithms referenced in prior studies [125], while the second focused on rule-based classifiers available within the Weka platform. In both cases, the DGWOdemonstrated superior performance, highlighting its capacity to generate accurate and generalizable classification rules in complex diagnostic scenarios. Moreover, a statistical stability analysis based on the maximum fitness values was performed, further confirming the algorithm's numerical consistency and resilience to stochastic variation, which is often inherent in population-based metaheuristics. These findings underscore the effectiveness of DGWO not only in achieving high classification accuracy but also in maintaining stable behavior across multiple runs , indicating the statistical stability of the algorithm and its ability to mitigate sensitivity to initial conditions and inherent stochasticity in evolutionary search mechanisms, thereby enhancing its reliability and efficiency in fault detection and diagnostic tasks within complex industrial systems.

General Conclusion

and

Future Perspectives

General Conclusion

This thesis has addressed key challenges in PV system modeling and intelligent fault diagnosis by integrating metaheuristic optimization techniques with explainable artificial intelligence. As the deployment of PV systems continues to expand globally, ensuring operational reliability and fostering transparency in diagnostic processes remain critical research priorities. The contributions of this work collectively advance these objectives through a set of complementary methodological innovations.

To enhance signal processing quality for intelligent fault diagnosis, a hybrid filtering strategy was first developed by combining the Moving Average Filter with the GWO Optimizer. This integration improved noise reduction and signal smoothing beyond traditional filtering methods, thereby strengthening the reliability of diagnostic inference independently of the simulation data.

Another significant contribution is the development of the AGWO for feature selection. By dynamically regulating the trade-off between exploration and exploitation, AGWO successfully identified the most discriminative features within high-dimensional datasets. When coupled with state-of-the-art machine learning classifiers such as XGBoost, LightGBM and CatBoost, the selected feature subsets enhanced classification accuracy while reducing computational complexity.

The primary originality of this thesis lies in the introduction of DGWO Algorithm , specifically tailored for rule-based classification. Unlike its continuous counterpart, DGWO operates within discrete search spaces and produces interpretable decision rules for fault detection. Experimental evaluations demonstrated its superior performance in terms of accuracy, robustness and interpretability. The generated classification rules provide actionable insights that support informed decision-making and facilitate transparent diagnostic processes.

Overall, this thesis contributes to the field by:

advancing PV system modeling through simulation and signal enhancement techniques;

improving feature selection using an adaptive metaheuristic framework; and proposing a novel discrete optimization algorithm dedicated to interpretable fault classification.

Together, these contributions establish a comprehensive and coherent framework that effectively addresses both the computational efficiency and explainability requirements of fault diagnosis in PV system.

Perspectives and Future Works

The findings of this thesis are promising and open several avenues for further research. The DGWO framework could be extended to handle multi-objective optimization in rule-based classification, aiming to achieve a better balance among accuracy, interpretability and computational cost. Experimental validation of the proposed algorithms using real PV plant data under diverse operating and fault conditions is also essential to ensure their practical reliability.

Additionally, the AGWO and DGWO frameworks could be adapted to other applications, such as smart grids, renewable energy forecasting and industrial fault detection, demonstrating their generalizability. Further improvements in diagnostic system accuracy and robustness could be achieved by developing hybrid models based on advanced data-driven optimization techniques. Overall, the results confirm that combining metaheuristic optimization with explainable AI offers a powerful pathway toward developing more reliable, interpretable and scalable diagnostic systems, with significant

References

References

[1] T. Region, "MINISTRY OF ENERGY." Accessed: Feb. 04, 2025. [Online]. Available: https://www.energy.gov.dz/?rubrique=energies-nouvelles-renouvelables-et-maitrise-de-lrenergie#

[2] M. Benasla *et al.*, "Algeria's potential to supply Europe with dispatchable solar electricity via HVDC links: Assessment and proposal of scenarios," *Energy Reports*, vol. 11, pp. 39–54, Jun. 2024, doi: 10.1016/j.egyr.2023.11.039.

[3] T. Fares, A. Djeddi and N. Nouioua, "Solar energy in Algeria between exploitation policies and export potential," in *2019 1st International Conference on Sustainable Renewable Energy Systems and Applications (ICSRESA)*, Tebessa, Algeria: IEEE, Dec. 2019, pp. 1–5. doi: 10.1109/ICSRESA49121.2019.9182306.

[4] Athil S. Al-Ezzi and Mohamed Nainar M. Ansari,"Photovoltaic Solar Cells: A Review", Applied System Innovation, vol. 5, no. 4, 2022 , https://doi.org/10.3390/asi5040067

[5] F. Maoulida and M. A. Kassim, "Modélisation d'un module photovoltaïque de type BP SX 150 pour la génération d'énergie électrique," ResearchGate, 2020 https://www.researchgate.net/publication/342048390_Modelisation_d'un_module_photovoltaique_de_type_bp_sx_150_pour_la_generation_d'energie_electrique

[6] M. N. I. Sarkar, "Effect of various model parameters on solar PV cell simulation: a SPICE analysis," Renewables, vol. 3, no. 1, pp. 1–9, Dec. 2016, doi: 10.1186/s40807-016-0035-3.

[7] Narayan Lal Panwar, Fatiha Yettou, and Amor Gama, "A comprehensive review of different types of solar photovoltaic cells and their applications," International Journal of Ambient Energy, vol. 42, no. 10, pp. 1200–1217, 2021, https://doi.org/10.1080/01430750.2019.1592774

[8] S. Al-Ali, A. G. Olabi, and M. Mahmoud, "A review of solar photovoltaic technologies: developments, challenges, and future perspectives," Renewable and Sustainable Energy Reviews, vol. 80, pp. 414–433, Dec. 2017, doi: 10.1016/j.rser.2017.10.038.

[9] J. Cubas and S. Pindado, "Explicit expressions for solar panel equivalent circuit parameters based on analytical formulation and the Lambert W-function," Energies, vol. 7, no. 7, pp. 4098–4115, Jul. 2014. [Online]. Available: https://www.mdpi.com/1996-1073/7/7/4098.

[10] G. MBWAMA BINAMEN, “Etude et dimensionnement d’une centrale photovoltaique pour l’alimentation d’un quartier residentiel,” Memoire Online. [Online]. Available: https://www.memoireonline.com/03/22/12766/Etude-et-dimensionnement-dune-centrale-photovoltaique-pour-lalimentation-dun-quartier-resid.html#_Toc19575766.

[11] L. Bun, “Détection et localisation de défauts pour un système PV,” Thèse de doctorat, Université de Grenoble, 2011. [Online]. Available: https://theses.hal.science/tel-00647189v2

[12] S. Fadhel, “Efficacité énergétique et surveillance d’un microgrid à courant continu alimenté par des panneaux photovoltaïques,” Thèse de doctorat, Université Paris-Saclay (GeePs-L2S), 2020. [Online]. Available: https://theses.hal.science/tel-03096435v1

[13] A. Amir, A. Amir, H. S. Che, A. Elkhateb and N. A. Rahim, “Comparative analysis of high voltage gain DC-DC converter topologies for PV systems,” *Renewable Energy*, vol. 136, pp. 1147–1163, Jun. 2019, doi: 10.1016/j.renene.2018.09.089.

[14] S. Arab Ansari, J. S. Moghani and M. Mohammadi, “Analysis and implementation of a new zero current switching flyback inverter,” *Circuit Theory & Apps*, vol. 47, no. 1, pp. 103–132, Jan. 2019, doi: 10.1002/cta.2577.

[15] H. Abouobaida and E. B. Said, “Practical Performance Evaluation of Maximum Power Point Tracking Algorithms in a PV System,” *IJPEDS*, vol. 8, no. 4, p. 1744, Dec. 2017, doi: 10.11591/ijpeds.v8.i4.pp1744-1755.

[16] G. R. Tapsoba, “Maximisation du transfert de l’énergie d’un champ photovoltaïque tenant compte du phénomène d’ombrage partiel : connexion réseau électrique,” Mémoire de maîtrise, École de technologie supérieure, Montréal, Canada, 2019. [Online]. Available: https://espace.etsmtl.ca/id/eprint/2322/

[17] T. Esram and P. L. Chapman, “Comparison of PV Array Maximum Power Point Tracking Techniques,” *IEEE Trans. On Energy Conversion*, vol. 22, no. 2, pp. 439–449, Jun. 2007, doi: 10.1109/TEC.2006.874230.

[18] A. Mohapatra, B. Nayak, P. Das and K. B. Mohanty, “A review on MPPT techniques of PV system under partial shading condition,” *Renewable and Sustainable Energy Reviews*, vol. 80, pp. 854–867, Dec. 2017, doi: 10.1016/j.rser.2017.05.083.

[19] A.Belaout Abdesslam, Étude et diagnostic des défauts fréquents aux systèmes photovoltaïques par emploi de la caractéristique courant–tension, Magister thesis, Univ. of Sétif 1, Sétif, Algeria, 2014.Available: https://www.researchgate.net/publication/320494826.

[20] L. Hassaine, “Onduleurs photovoltaïques connectés au réseau électrique,” Bulletin des Énergies Renouvelables, no. 20, CDER, Algérie, 2011. [Online]. Available: https://www.cder.dz/vlib/bulletin/pdf/bulletin_020_05.pdf

[21] L. Bridier, “Modélisation et optimisation d’un système de stockage couplé à une production électrique renouvelable intermittente,” Ph.D. dissertation, Université de La Réunion, La Réunion, France, 2016. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01525696

[22] R. G. Vieira, F. M. U. de Araújo, M. Dhimish and M. I. S. Guerra, “A comprehensive review on bypass diode application on photovoltaic modules,” *Energies*, vol. 13, no. 10, Art. 2472, 2020, doi:10.3390/en13102472.

[23] W. Storr, “Bypass Diodes in Solar Panels and Arrays,” Basic Electronics Tutorials. Accessed: Mar. 07, 2025. [Online]. Available: https://www.electronics-tutorials.ws/diode/bypass-diodes.html

[24] F. A. Syam and O. M. Arafa, “Selection process of PV standalone pumping systems,” *Sustainable Energy res.*, vol. 10, no. 1, p. 22, Dec. 2023, doi: 10.1186/s40807-023-00094-9.

[25] W. Ali, H. Farooq, A. U. Rehman, Q. Awais, M. Jamil and A. Noman, “Design Considerations of Stand-Alone Solar PV Systems,” in *2018 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, Quetta: IEEE, Nov. 2018, pp. 1–6. doi: 10.1109/ICECUBE.2018.8610970.

[26] V. Karthikeyan, S. Rajasekar, V. Das, P. Karuppanan and A. K. Singh, “Grid-Connected and Off-Grid Solar PV System,” in *Smart Energy Grid Design for Island Countries*, F. M. R. Islam, K. A. Mamun and M. T. O. Amanullah, Eds., in Green Energy and Technology. , Cham: Springer International Publishing, 2017, pp. 125–157. doi: 10.1007/978-3-319-50197-0_5.

[27] A. Mahesh and K. S. Sandhu, “Hybrid wind/PV energy system developments: Critical review and findings,” *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 1135–1147, Dec. 2015, doi: 10.1016/j.rser.2015.08.008.

[28] Lucy Meng, “A Complete Guide To Hybrid Solar Systems,’ GYCX Solar Smart Energy Solutions, May 14, 2024. [Online]. Available: https://gycxsolar.com/a-complete-guide-to-hybrid-solar-systems2024/. Accessed: Mar. 13, 2025.”

[29] L. A. García-Gutiérrez, “Development of an active-fault tolerant control applied to PV systems,” Ph.D. dissertation, Université Toulouse III – Paul Sabatier, Toulouse, France, Mar. 2019. [Online]. Available: https://www.theses.fr/2019TOU30071.

[30] B. Yang *et al.*, “Recent Advances in Fault Diagnosis Techniques for PV Systems: A Critical Review,” *Prot. Control Mod. Power Syst.*, vol. 9, no. 3, pp. 36–59, May 2024, doi: 10.23919/PCMP.2023.000583.

[31] A. Javaid, I. Shafi, I. U. Khalil, S. Ahmad, M. Safran, S. Alfarhood and I. Ashraf, “Enhancing PV systems using Gaussian process regression for parameter identification and fault detection,” Energy Reports, vol. 11, pp. 4485–4499, 2024, doi: 10.1016/j.egyr.2024.04.026.

[32] J. Svarc, “Solar Panel Shading Problems & Solutions,” Clean Energy Reviews, Accessed: Mar. 12, 2025. [Online]. Available: https://www.cleanenergyreviews.info/blog/solar-panel-shading-problems-bypass-diodes-optimisers .

[33] G. Schirripa Spagnolo, P. Del Vecchio, G. Makary, D. Papalillo and A. Martocchia, “A review of IR thermography applied to PV systems,” in *2012 11th International Conference on Environment and Electrical Engineering*, Venice, Italy: IEEE, May 2012, pp. 879–884. doi: 10.1109/EEEIC.2012.6221500.

[34] E. H. Sepúlveda-Oviedo, Detection and diagnosis of faults and performance losses in high-power photovoltaic power plants, Ph.D. dissertation, Université Paul Sabatier - Toulouse III, 2023. Available: https://laas.hal.science/tel-04888632v1/document

[35] D. Carrasco-González, P. Horrillo-Quintero, P. García-Triviño, R. Sarrias-Mena, C. Andrés García-Vázquez and L. M. Fernández-Ramírez, “Control of PV power plants with quasi-Z-source cascaded H-bridge multilevel inverters under failure,” *International Journal of Electrical Power & Energy Systems*, vol. 157, p. 109803, Jun. 2024, doi: 10.1016/j.ijepes.2024.109803.

[36] K. Sakthivel and S. A. Alexander, “An extensive critique on machine learning techniques for fault tolerance and power quality improvement in multilevel inverters,” *Energy Reports*, vol. 12, pp. 5814–5833, Dec. 2024, doi: 10.1016/j.egyr.2024.11.016.

[37] G. Xu, S. Nie and Z. Xiong, “Research on PV array reconstruction and Full-cycle maximum power point tracking technology of space solar power station,” *Space Solar Power and Wireless Transmission*, vol. 1, no. 2, pp. 115–128, Sep. 2024, doi: 10.1016/j.sspwt.2024.09.004.

[38] R. G. Mohamed, H. M. Hasanien and M. A. Ebrahim, “Global MPPT controllers for enhancing dynamic performance of PV systems under partial shading condition,” *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 9, p. 100638, Sep. 2024, doi: 10.1016/j.prime.2024.100638.

[39] K. Tidriri, N. Chatti, S. Verron and T. Tiplica, “Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges,” *Annual Reviews in Control*, vol. 42, pp. 63–81, 2016, doi: 10.1016/j.arcontrol.2016.09.008.

[40] G. Ellis, “Introduction to Observers in Control Systems,” in *Control System Design Guide*, Elsevier, 2012, pp. 185–212, doi:10.1016/B978-0-12-385920-4.00010-2.

[41] T. Fares, A. Djeddi, et N. Nouioua, « Solar energy in Algeria between exploitation policies and export potential », in *2019 1st International Conference on Sustainable Renewable Energy Systems and Applications (ICSRESA)*, Tebessa, Algeria: IEEE, déc. 2019, p. 1-5. doi: 10.1109/ICSRESA49121.2019.9182306.

[42] Q. Wang, F. Zheng, W. Qian, et D. Ding, « A practical filter error method for aerodynamic parameter estimation of aircraft in turbulence », *Chin. J. Aeronaut.*, vol. 36, nº 2, p. 17-28, févr. 2023, doi: 10.1016/j.cja.2022.05.008.

[43] M. Gitterman, Qualitative Analysis of Physical Problems. Elsevier, 2012. Available: https://shop.elsevier.com/books/qualitative-analysis-of-physical-problems/gitterman/978-0-12-285150-6

[44] P. Arrighi et G. Dowek, « Causal graph dynamics », *Inf. Comput.*, vol. 223, p. 78-93, févr. 2013, doi: 10.1016/j.ic.2012.10.019.

[45] D. Gonzalez-Jimenez, J. del-Olmo, J. Poza, F. Garramiola, et P. Madina, « Data-Driven Fault Diagnosis for Electric Drives: A Review », *Sensors*, vol. 21, nº 12, p. 4024, janv. 2021, doi: 10.3390/s21124024

[46] Q. Yu, Q. Li, L. Li and Y. Wang, "A survey of data-driven fault-diagnosis methods for large-scale industrial production processes," Chinese Journal of Engineering, vol. 47, no. 4, pp. 780-793, 2025, doi: 10.13374/j.issn2095-9389.2024.05.24.002.

[47] N. E. Breslow and N. E. Day, *Statistical Methods in Cancer Research, Volume I: The Analysis of Case-Control Studies*, IARC Scientific Publications no. 32, Lyon, France: International Agency for Research on Cancer, 1980.

[48] B. Mahesh, « Machine Learning Algorithms - A Review », *Int. J. Sci. Res. IJSR*, vol. 9, nº 1, p. 381-386, janv. 2020, doi: 10.21275/ART20203995.

[49] A.Singh, N. Thakur, et A. Sharma, « A review of supervised machine learning algorithms », in 2016 3rd International Conference on Computing for Sustainable Global Development *(INDIACom)*, mars 2016, p. 1310-1315. https://ieeexplore.ieee.org/abstract/document/7724478.

[50] S. Naeem, A. Ali, S. Anam, et M. M. Ahmed, « An Unsupervised Machine Learning Algorithms: Comprehensive Review », *Int. J. Comput. Digit. Syst.*, vol. 13, nº 1, p. 911-921, avr. 2023, doi: 10.12785/ijcds/130172.

[51] M. S. A. Abid, A. N. Qureshi, and S. A. Khan, "An overview of machine learning classification techniques," in BIO Web of Conferences, vol. 97, no. 4, Art. no. 00133, Apr. 2024, doi: 10.1051/bioconf/20249700133.

[52] P. Sarang, « Naive Bayes », in *Thinking Data Science: A Data Science Practitioner's Guide*, P. Sarang, Éd., Cham: Springer International Publishing, 2023, p. 143-152. doi: 10.1007/978-3-031-02363-7_7.

[53] S. Salcedo-Sanz, "Support vector machines in engineering: An overview," Wiley Interdiscip. Rev.: Data Mining Knowl. Discov., vol. 4, no. 3, pp. 234–267, 2014, doi:10.1002/widm.1125.

[54] B. Charbuty et A. Abdulazeez, « Classification Based on Decision Tree Algorithm for Machine Learning », *J. Appl. Sci. Technol. Trends*, vol. 2, nº 01, p. 20-28, mars 2021, doi: 10.38094/jastt20165.

[55] S. K. Sharma, A. Gaur, V. Saddikuti, et A. Rastogi, « Structural equation model (SEM)-neural network (NN) model for predicting quality determinants of e-learning management systems », *Behav. Inf. Technol.*, vol. 36, nº 10, p. 1053-1066, oct. 2017, doi: 10.1080/0144929X.2017.1340973.

[56] S. Kotsiantis, D. Kanellopoulos and P. Pintelas, "A new approach to classification based on association rule mining," Decis. Support Syst., vol. 42, no. 2, pp. 674–689, Nov. 2006, doi:10.1016/j.dss.2005.03.005.

[57] C.Thanajiranthorn et P. Songram, « Efficient Rule Generation for Associative Classification », *Algorithms*, vol. 13, nº 11, p. 299, nov. 2020, doi: 10.3390/a13110299.

[58] Z. Zhang, « Introduction to machine learning: k-nearest neighbors », *Ann. Transl. Med.*, vol. 4, nº 11, p. 218, juin 2016, doi: 10.21037/atm.2016.03.37.

[59] J. L. Castle, J. A. Doornik, et D. F. Hendry, « Robust Discovery of Regression Models », *Econom. Stat.*, vol. 26, p. 31-51, avr. 2023, doi: 10.1016/j.ecosta.2021.05.004.

[60] F. L. Huang, “Alternatives to logistic regression models in experimental studies,” J. Exp. Educ., vol. 90, no. 1, pp. 213–228, 2022, doi:10.1080/00220973.2019.1699769.

[61] N. Roustaei, « Application and interpretation of linear-regression analysis », *Med. Hypothesis Discov. Innov. Ophthalmol.*, vol. 13, nº 3, p. 151-159, oct. 2024, doi: 10.51329/mehdiophthal1506.

[62] I. C. Gormley, T. B. Murphy, et A. E. Raftery, « Model-Based Clustering », *Annu. Rev. Stat. Its Appl.*, vol. 10, nº Volume 10, 2023, p. 573-595, mars 2023, doi: 10.1146/annurev-statistics-033121-115326.

[63] R. C. Ripan *et al.*, « A Data-Driven Heart Disease Prediction Model Through K-Means Clustering-Based Anomaly Detection », *SN Comput. Sci.*, vol. 2, nº 2, p. 112, févr. 2021, doi: 10.1007/s42979-021-00518-7.

[64] A. Sobrinho Campolina Martins, L. Ramos de Araujo, et D. Rosana Ribeiro Penido, « K-Medoids clustering applications for high-dimensionality multiphase probabilistic power flow », *Int. J. Electr. Power Energy Syst.*, vol. 157, p. 109861, juin 2024, doi: 10.1016/j.ijepes.2024.109861.

[65] Z. Zhao, Z. Jian, G. S. Gaba, R. Alroobaea, M. Masud, et S. Rubaiee, « An improved association rule mining algorithm for large data », *J. Intell. Syst.*, vol. 30, nº 1, p. 750-762, janv. 2021, doi: 10.1515/jisys-2020-0121.

[66] Y. Chi, Y. Dong, Z. J. Wang, F. R. Yu and V. C. M. Leung, “Knowledge-Based Fault Diagnosis in Industrial Internet of Things: A Survey,” *IEEE Internet of Things Journal*, vol. 9, no. 15, pp. 12886–12900, Aug. 2022, doi: 10.1109/JIOT.2022.3163606.

[67] Y. Zakur et L. Flaih, « Apriori Algorithm and Hybrid Apriori Algorithm in the Data Mining: A Comprehensive Review », *E3S Web Conf.*, vol. 448, p. 02021, 2023, doi: 10.1051/e3sconf/202344802021.

[68] A.Sabzaliyev, “Knowledge Representation in Expert Systems: Structure, Classification, and Applications,” Luminis Applied Science and Engineering, vol. 1, no. 2, pp. 1–15, Dec. 2024, doi: 10.69760/lumin.202400005.

[69] I. S. Ahmad, S. Abubakar, F. L. Gambo, and M. S. Gadanya, “A Rule-Based Expert System for Automobile Fault Diagnosis,” International Journal on Perceptive and Cognitive Computing,vol.7,no.1,pp.2025,2021.Available:https://www.researchgate.net/publication/353660280_A_RuleBased_Expert_System_for_Automobile_Fault_Diagnosis

[70] K. Rogulj, J. K. Pamuković, N. Jajac, K. Rogulj, J. K. Pamuković and N. Jajac, “Knowledge-Based Fuzzy Expert System to the Condition Assessment of Historic Road Bridges,” *Applied Sciences*, vol. 11, no. 3, Jan. 2021, doi: 10.3390/app11031021.

[71] A. Yan, Z. Cheng, “A Review of the Development and Future Challenges of Case-Based Reasoning,” Applied Sciences, vol. 14, no. 16, Art. no. 7130, Aug. 2024. doi: 10.3390/app14167130.

[72] T. M. L. Diallo, S. Henry and Y. Ouzrout, “Bayesian Network Building for Diagnosis in Industrial Domain based on Expert Knowledge and Unitary Traceability Data,” *IFAC-PapersOnLine*, vol. 48, no. 3, pp. 2411–2416, Jan. 2015, doi: 10.1016/j.ifacol.2015.06.449.

[73] Y. Wilhelm, P. Reimann, W. Gauchel, et B. Mitschang, « Overview on hybrid approaches to fault detection and diagnosis: Combining data-driven, physics-based and knowledge-based models », *Procedia CIRP*, vol. 99, p. 278-283, janv. 2021, doi: 10.1016/j.procir.2021.03.041.

[74] S. M. Almufti, A. Ahmad Shaban, Z. Arif Ali, R. Ismael Ali, J. A. Dela Fuente, et R. Rajab Asaad, « Overview of Metaheuristic Algorithms », *Polaris Glob. J. Sch. Res. Trends*, vol. 2, nº 2, p. 10-32, avr. 2023, doi: 10.58429/pgjsrt.v2n2a144.

[75] A.Seyyedabbasi, “A comprehensive survey: Evolutionary-based algorithms,” in Decision-Making Models, T. Allahviranloo, W. Pedrycz and A. Seyyedabbasi, Eds., Academic Press, 2024, pp. 77–84, doi:10.1016/B978-0-443-16147-6.00031-1.

[76] S. Katoch, S. S. Chauhan, et V. Kumar, « A review on genetic algorithm: past, present and future », *Multimed. Tools Appl.*, vol. 80, nº 5, p. 8091-8126, févr. 2021, doi: 10.1007/s11042-020-10139-6.

[77] M. A. Abido et A. Elazouni, « Modified multi-objective evolutionary programming algorithm for solving project scheduling problems », *Expert Syst. Appl.*, vol. 183, p. 115338, nov. 2021, doi: 10.1016/j.eswa.2021.115338.

[78] M. C. Rathod and S. R. Vyas, “Optimal Scheduling of Thermal-Solar Power Considering Carbon Emission Based on Evolutionary Programming Algorithm,” in *2024 4th International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)*, Nov. 2024, pp. 1–6. doi: 10.1109/ICEFEET64463.2024.10866479.

[79] M. T. Emmerich, O. M. Shir and H. Wang, “Evolution strategies,” in *Handbook of Heuristics*, A. Ghosh and S. Dutta, Eds., Springer, 2025, pp. 89–123, doi:10.1007/978-3-032-00385-0_13.

[80] B. Chen, H. Ouyang, S. Li and D. Zou, “Differential evolution algorithm with a complementary mutation strategy and data Fusion-Based parameter adaptation,” *Information Sciences*, vol. 668, p. 120522, May 2024, doi: 10.1016/j.ins.2024.120522.

[81] U. Can and B. Alatas, “Physics-based metaheuristic algorithms for global optimization,” American Journal of Information Science and Computer Engineering, vol. 1, no. 3, pp. 94–106, 2015. Available: http://www.aiscience.org/journal/ajisce

[82] Z. Yang, Y. Cai, et G. Li, « Improved Gravitational Search Algorithm Based on Adaptive Strategies », *Entropy*, vol. 24, nº 12, p. 1826, déc. 2022, doi: 10.3390/e24121826.

[83] H. Abedinpourshotorban, S. Mariyam Shamsuddin, Z. Beheshti and D. N. A. Jawawi, “Electromagnetic field optimization: A physics-inspired metaheuristic optimization algorithm,” *Swarm and Evolutionary Computation*, vol. 26, pp. 8–22, Feb. 2016, doi: 10.1016/j.swevo.2015.07.002.

[84] G. Zhang, “Quantum-inspired evolutionary algorithms: A survey and empirical study,” *J. Heuristics*, vol. 17, pp. 303–351, Jun. 2011, doi: 10.1007/s10732-010-9136-0.

[85] G. Klau, N. Lesh, J. Marks and M. Mitzenmacher, “Human-guided search,” *Journal of Heuristics*, vol. 16, pp. 289–310, Jun. 2010, doi: 10.1007/s10732-009-9107-5.

[86] H. Takagi, “Interactive Evolutionary Computation for Analyzing Human Awareness Mechanisms,” *Applied Computational Intelligence and Soft Computing*, vol. 2012, Jun. 2012, doi: 10.1155/2012/694836.

[87] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán and Á. Fernández-Leal, “Human-in-the-loop machine learning: a state of the art,” *Artif. Intell. Rev.*, vol. 56, no. 4, pp. 3005–3054, 2023, doi:10.1007/s10462-022-10246-w.

[88] S. Li, Y. Wei, X. Liu, H. Zhu and Z. Yu, “A New Fast Ant Colony Optimization Algorithm: The Saltatory Evolution Ant Colony Optimization Algorithm,” *Mathematics*, vol. 10, no. 6, p. 925, Jan. 2022, doi: 10.3390/math10060925.

[89] Y.-Y. Hong, L.-F. Chen and W. Zhang, “Classification of PV Faults Using PSO-Optimized Compact Convolutional Transformer,” *IEEE Access*, vol. 11, pp. 140752–140762, 2023, doi: 10.1109/ACCESS.2023.3341889.

[90] D.Teodorović, “Bee colony optimization (BCO),” in *Studies in Computational Intelligence*, vol. 248, 2009, pp. 39–60. doi: 10.1007/978-3-642-04225-6_3.

[91] D. Karaboga and B. Basturk, *Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problems*, vol. 4529. 2007, p. 798. doi: 10.1007/978-3-540-72950-1_77.

[92] N. Johari, A. Zain, N. Mustaffa and A. Udin, “Firefly Algorithm for Optimization Problem,” *Applied Mechanics and Materials*, vol. 421, pp. 512–517, Sept. 2013, doi: 10.4028/www.scientific.net/AMM.421.512.

[93] S. Mirjalili, S. M. Mirjalili and A. Lewis, “Grey Wolf Optimizer,” *Advances in Engineering Software*, vol. 69, pp. 46–61, Mar. 2014, doi: 10.1016/j.advengsoft.2013.12.007.

[94] H. Rezaei, O. Bozorg-Haddad and X. Chu, “Grey Wolf Optimization Algorithm,” in *Advanced Optimization by Nature-Inspired Algorithms*, vol. 720, O. Bozorg-Haddad, Ed., in Studies in Computational Intelligence, vol. 720. , Singapore: Springer Singapore, 2018, pp. 81–91. doi: 10.1007/978-981-10-5221-7_9.

[95] M. H. Nadimi-Shahraki, H. Zamani, Z. Asghari Varzaneh, A. S. Sadiq and S. Mirjalili, “A systematic review of applying grey wolf optimizer, its variants and its developments in different Internet of Things applications,” *Internet of Things*, vol. 26, p. 101135, Jul. 2024, doi: 10.1016/j.iot.2024.101135.

[96] M. Nasir, A. Sadollah, S. Mirjalili, S. A. Mansouri, M. Safaraliev and A. Rezaee Jordehi, “A Comprehensive Review on Applications of GWO in Energy Systems,” *Arch Computat Methods Eng*, pp. 1–41, Dec. 2024, doi: 10.1007/s11831-024-10214-3.

[97] F.-S. Pai and P.-S. Tseng, “An efficient GWO MPPT for a PV system using impedance information acceleration,” *International Journal of Electronics*, vol. 106, no. 4, pp. 648–661, Apr. 2019, doi: 10.1080/00207217.2018.1545929.

[98] S. M. Almufti, Hawar B. Ahmad, Ridwan B. Marqas and Renas R. Asaad, “Grey wolf optimizer: Overview, modifications and applications,” Aug. 2021, doi: 10.5281/ZENODO.5195644.

[99] R. Zriouile, L. Bouhouch, I. Isknan and A. Elfanaoui, “Performance Evaluation of GWO and P&O MPPT for PV Systems under Shading,” in *2025 5th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, May 2025, pp. 1–7. doi: 10.1109/IRASET64571.2025.11007987.

[100] R. Motamarri, N. Bhookya and B. Chitti Babu, “Modified grey wolf optimization for global maximum power point tracking under partial shading conditions in PV system,” *International Journal of Circuit Theory and Applications*, vol. 49, no. 7, pp. 1884–1901, 2021, doi: 10.1002/cta.3018.

[101] I. S. Millah, P. C. Chang, D. F. Teshome, R. K. Subroto, K. L. Lian and J.-F. Lin, “An Enhanced Grey Wolf Optimization Algorithm for PV Maximum Power Point Tracking Control Under Partial Shading Conditions,” *IEEE Open J. Ind. Electron. Soc.*, vol. 3, pp. 392–408, 2022, doi: 10.1109/OJIES.2022.3179284.

[102] S. A. Celtek, S. Kul, M. K. Singla, J. Gupta, M. Safaraliev and H. Zeinoddini-Meymand, “Grey wolf-based heuristic methods for accurate parameter extraction to optimize the performance of PV modules,” *IET Renewable Power Generation*, vol. 18, no. 14, pp. 2248–2260, Oct. 2024, doi: 10.1049/rpg2.13061.

[103] I. Dagal, A.-W. Ibrahim and A. Harrison, “Leveraging a novel grey wolf algorithm for optimization of PV-battery energy storage system under partial shading conditions,” *Computers and Electrical Engineering*, vol. 122, p. 109991, Mar. 2025, doi: 10.1016/j.compeleceng.2024.109991.

[104] L. Liu, “PV MPPT control and improvement strategies considering environmental factors: based on PID-type sliding mode control and improved grey wolf optimization,” *Measurement and Control*, vol. 58, no. 2, pp. 227–244, Feb. 2025, doi: 10.1177/00202940241258821.

[105] M. G. Yenealem, “Mitigation of High PV Penetration Effects in Electrical Grid Systems Using a Hybrid Particle Swarm and Grey Wolf Optimization Approach,” *International Journal of Energy Research*, vol. 2024, no. 1, p. 1481027, Jan. 2024, doi: 10.1155/2024/1481027.

[106] N. Deghfel, A. E. Badoud, A. A. Al-Ahmadi, M. Bajaj, I. Zaitsev and S. S. M. Ghoneim, “Improving maximum power point tracking efficiency in solar PV systems using super-twisting algorithm and grey wolf optimizer,” *IET Renewable Power Generation*, vol. 18, no. 15, pp. 3329–3354, Nov. 2024, doi: 10.1049/rpg2.13138.

[107] A. G. Abera, T. T. Yetayew and A. B. Alyu, “Optimized solar PV integration for voltage enhancement and loss reduction in the Kombolcha distribution system using hybrid grey wolf-particle swarm optimization,” *Results in Engineering*, vol. 26, p. 105484, Jun. 2025, doi: 10.1016/j.rineng.2025.105484.

[108] D. T. Thinh, N. B. H. Quanو ، N. Maneetien ،"Implementation of Moving Average Filter on STM32F4 for Vibration Sensor Application," in Proceedings of the 2018 4th International Conference on Green Technology and Sustainable Development (GTSD), Ho Chi Minh City, Vietnam, Nov. 2018, pp. 627–631, doi: 10.1109/GTSD.2018.8595630.

[109] R. Kumar, "Data-Driven Innovation: How AI and Big Data Reshape Business Development Models," Int. J. Big Data Intell., vol. 1, no. 2, pp. 11–16, 2024. Driven_Innovation_How_AI_and_Big_Data_Reshape_Business_Development_Models

[110] V. R. Mahboobani, C. Y. Lo, K. Y. Ng, M. H. Kwong and K. M. Ku, "Integration of data visualization and Generative AI to accelerate data driven decision: a case study on research topic trend analysis," May 07, 2025, *Research Square*. doi: 10.21203/rs.3.rs-6343303/v1.

[111] M. M. Jayaraj and S. Somu, "Leveraging Big Data and Advanced Technologies for Enhanced Sustainability in Healthcare: An IPO Model Approach," in *Navigating Cyber-Physical Systems With Cutting-Edge Technologies*, IGI Global Scientific Publishing, 2025, pp. 287–308. doi: 10.4018/979-8-3693-5728-6.ch011.

[112] R. Kumar, S. K. Mishra, A. Flah, A. K. Sahoo and S. K. Swain, "Data driven prediction based reliability assessment of solar energy systems incorporating uncertainties for generation planning," *Sci. Rep.*, vol. 15, Art. no. 9335, 2025, doi:10.1038/s41598-025-94106-x.

[113] J. Polo *et al.*, "Standardization of Provider Notes for Accuracy in Data Reporting," *Transplantation and Cellular Therapy*, vol. 31, no. 2, Supplement, pp. S84–S85, Feb. 2025, doi: 10.1016/j.jtct.2025.01.136.

[114] J. Barrera-García, F. Cisternas-Caneo, B. Crawford, M. Gómez Sánchez and R. Soto, "Feature Selection Problem and Metaheuristics: A Systematic Literature Review about Its Formulation, Evaluation and Applications," *Biomimetics,* vol. 9, no. 1, Art. no. 1, Jan. 2024, doi: 10.3390/biomimetics9010009.

[115] Z. Sadeghian, E. Akbari, H. Nematzadeh and H. Motameni, "A review of feature selection methods based on meta-heuristic algorithms," J. Exp. Theor. Artif. Intell., vol. 37, no. 1, pp. 1–51, 2025, doi:10.1080/0952813X.2023.2183267.

[116] X. Li, Q. Fu, Q. Li, W. Ding, F. Lin and Z. Zheng, "Multi-objective binary grey wolf optimization for feature selection based on guided mutation strategy," *Applied Soft Computing*, vol. 145, p. 110558, Sep. 2023, doi: 10.1016/j.asoc.2023.110558.

[117] S. Reka, T. S. Praba, K. K. Manchala and A. Venkateswarlu, “Comprehensive Feature-Driven PCOS Predictor: A Reinforcement Learning-Based Binary Equilibrium Optimization Approach,” *Int J Comput Intell Syst*, vol. 18, no. 1, p. 188, Jul. 2025, doi: 10.1007/s44196-025-00931-3.

[118] Z. Ye *et al.*, “Hybrid rice optimization algorithm inspired GWO for high-dimensional feature selection,” *Sci Rep*, vol. 14, no. 1, p. 30741, Dec. 2024, doi: 10.1038/s41598-024-80648-z.

[119] X. Bai, Y. Zheng, Y. Lu and Y. Shi, “Chain hybrid feature selection algorithm based on improved Grey Wolf Optimization algorithm,” *PLOS ONE*, vol. 19, no. 10, p. e0311602, Oct. 2024, doi: 10.1371/journal.pone.0311602.

[120] W. Long, “GWO based on Nonlinear Adjustment Control Parameter,” presented at the 2016 4th International Conference on Sensors, Mechatronics and Automation (ICSMA 2016), Atlantis Press, Dec. 2016, pp. 643–648. doi: 10.2991/icsma-16.2016.111.

[121] R. Ahmadi, G. Ekbatanifard and P. Bayat, “A Modified GWO Based Data Clustering Algorithm,” *Applied Artificial Intelligence*, Jan. 2021, Accessed: Oct. 24, 2025. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/08839514.2020.1842109

[122] C. Tang *et al.*, “Multi-strategy GWO for Engineering Problems and Sewage Treatment Prediction”, doi: 10.1002/aisy.202300406.

[123] M. O. Ayansiji, N. Okposo and J. S. Apanapudor, “Enhanced GWO With Adaptive Control And Chaotic Initialization For Global Optimization,” *FNAS-JMNS*, vol. 2, no. 3, pp. 15–21, May 2025, doi: 10.63561/jmns.v2i3.861.

[124] S. Singh, J. C. Bansal, H. Nematzadeh and H. Motameni, “Mutation-driven grey wolf optimizer with modified search mechanism,” Expert Syst. Appl., vol. 194, p. 116450, May 2022, doi:10.1016/j.eswa.2021.116450.

[125] S. S. M. Ghoneim, A. E. Rashed and N. I. Elkalashy, “Fault Detection Algorithms for Achieving Service Continuity in PV Farms,” Intelligent Automation & Soft Computing, vol. 29, no. 3, pp. 467–479, 2021, doi: 10.32604/iasc.2021.016681.

[126] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*. 2016, p. 794. doi: 10.1145/2939672.2939785.

[127] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, USA, 2017. Available: https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf

[128] L. Prokhorenkova, A. Gusev, A. Vorobev, A. Dorogush and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada, 2018. Available: https://papers.nips.cc/paper/7898-catboost-unbiased-boosting-with-categorical-features.pdf

[129] A. Elly, A. Robert and A. Okunola, "The Role of Ensemble Learning in Improving Efficiency of Machine Learning Algorithms," ResearchGate, Dec. 2024. [Online]. Available: https://www.researchgate.net/publication/386452234_The_Role_of_Ensemble_Learning_in_Improving_Efficiency_of_Machine_Learning_Algorithms

[130] F. G. Beshaw, T. H. Atyia, M. F. M. Salleh, M. K. Ishak and A. S. Din, “Utilizing Machine Learning and SHAP Values for Improved and Transparent Energy Usage Predictions,” Computers, Materials & Continua, vol. 83, no. 2, pp. 3553–3583, Apr. 2025, doi: 10.32604/cmc.2025.061400.

[131] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[132] V. M. Patro and M. R. Patra, “Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy,” *Transactions on Machine Learning and Artificial Intelligence*, vol. 2, no. 4, pp. 77–91, Aug. 2014, doi: 10.14738/tmlai.24.328.

[133] J. Abdulhasan Salim, B. M. Albaker, M. Shyaa Alwan and M. Hasanuzzaman, “Hybrid MPPT approach using Cuckoo Search and GWO for PV systems under variant operating conditions,” *Global Energy Interconnection*, vol. 5, no. 6, pp. 627–644, Dec. 2022, doi: 10.1016/j.gloei.2022.12.005.

[134] F. Abdelmalek, H. Afghoul, F. Krim, M. Bajaj and V. Blazek, “Experimental validation of novel hybrid Grey Wolf Equilibrium Optimization for MPPT to improve the efficiency of solar PV system,” *Results in Engineering*, vol. 25, p. 103831, Mar. 2025, doi: 10.1016/j.rineng.2024.103831.

[135] I. Dagal, A.-W. Ibrahim and A. Harrison, “Leveraging a novel grey wolf algorithm for optimization of PV-battery energy storage system under partial shading conditions,” *Computers and Electrical Engineering*, vol. 122, p. 109991, Mar. 2025, doi: 10.1016/j.compeleceng.2024.109991.

[136] L. Liu, “PV MPPT control and improvement strategies considering environmental factors: based on PID-type sliding mode control and improved grey wolf optimization,” *Measurement and Control*, vol. 58, no. 2, pp. 227–244, Feb. 2025, doi:10.1177/00202940241258821.

[137] M. G. Yenealem, “Mitigation of high PV penetration effects in electrical grid systems using a hybrid particle swarm and grey wolf optimization approach,” Int. J. Energy Res., vol. 2024, no. 1, p. 1481027, Jan. 2024, doi:10.1155/2024/1481027.

[138] N. Deghfel, A. E. Badoud, A. A. Al-Ahmadi, M. Bajaj, I. Zaitsev and S. S. M. Ghoneim, “Improving maximum power point tracking efficiency in solar PV systems using super-twisting algorithm and grey wolf optimizer,” *IET Renewable Power Generation*, vol. 18, no. 15, pp. 3329–3354, Nov. 2024, doi: 10.1049/rpg2.13138.

[139] A. G. Abera, T. T. Yetayew and A. B. Alyu, “Optimized solar PV integration for voltage enhancement and loss reduction in the Kombolcha distribution system using hybrid grey wolf-particle swarm optimization,” *Results in Engineering*, vol. 26, p. 105484, Jun. 2025, doi: 10.1016/j.rineng.2025.105484.

[140] M. Supriyamenon and P. Rajarajeswari, “A review on association rule mining techniques with respect to their privacy preserving capabilities,” *International Journal of Applied Engineering Research*, vol. 12, no. 24, pp. 15484–15488, 2017.

[141] C.-W. Tsai, C.-F. Lai, M.-C. Chiang and L. T. Yang, “Data mining for Internet of Things: a survey,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 77–97, 2014, doi: 10.1109/SURV.2013.103013.00206.

Appendix

Appendix A: Simulation Framework and Dataset Construction

A.1 Overview

This appendix provides a comprehensive description of the simulated PV system and the methodology adopted for dataset generation. The presented framework ensures the reproducibility of the experimental setup and supports the evaluation of fault detection and classification algorithms under controlled conditions.

The study is based on a grid-connected PV power plant with a nominal capacity of 250 kW, modeled using MATLAB/Simulink. The simulation environment enables the generation of both normal and faulty operating scenarios, which are subsequently used to construct the training and testing datasets.

A.2 Architecture of the Simulated PV System

The overall structure of the simulated system is illustrated in Figure A1. The PV plant is interfaced with the electrical grid through power electronic converters and transmission components.

Figure A1 presents the detailed structure of the simulated PV system:

- Figure A1(a) : General schematic of the grid-connected PV system
- Figure A1(b) : MATLAB/Simulink model of the electrical grid and transmission network

The Figure.A1 provides a comprehensive visualization of the system components involved in dataset generation, including power conversion, transmission, and generation subsystems.

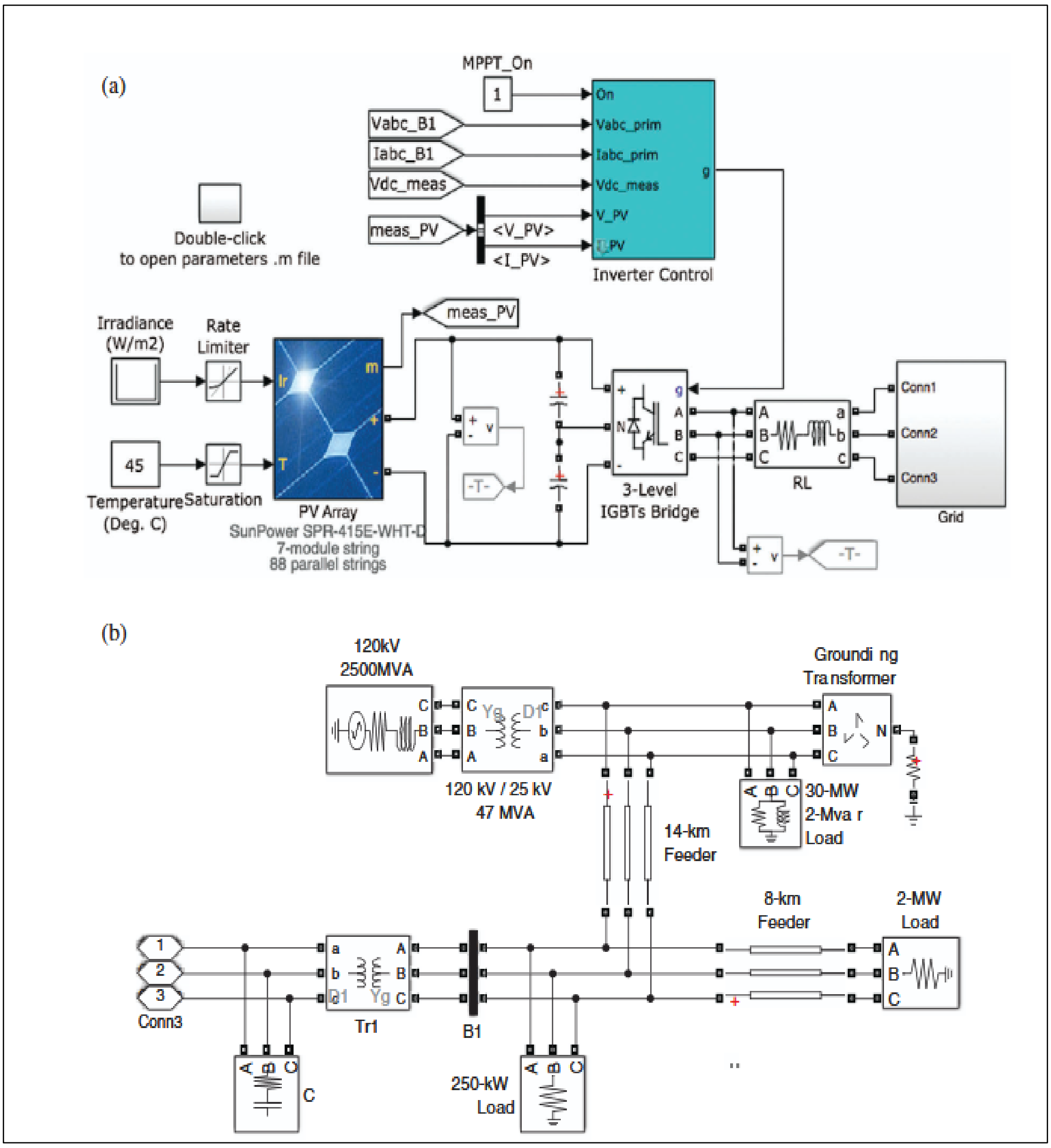


Figure A1. Simulated photovoltaic (PV) power system in MATLAB/Simulink

A.2.1 Power Conversion Stage

The PV system is connected to the grid via a three-level inverter based on insulated gate bipolar transistor (IGBT) technology, controlled using pulse-width modulation (PWM). The inverter control system is designed to ensure optimal energy extraction through a maximum power point tracking algorithm based on the perturb and observe method.

A.2.2 Grid Interface

The grid connection is achieved through the following components:

- A step-up transformer (0.25/250 kV)
- A 120 kV equivalent power grid

Two transmission feeders:

- A 14 km feeder connected to the main grid
- An 8 km feeder supplying a local load

Additionally, the system includes static loads and grounding transformers to emulate realistic grid conditions.

A.3 Photovoltaic Array Configuration

As shown in Figure A2.

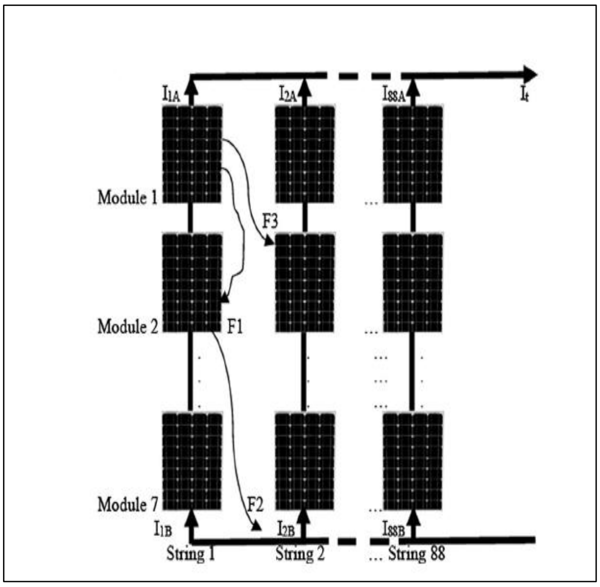


Figure A2. Detailed electrical configuration and internal architecture of the PV farm.

The PV generator is structured as follows:

- 88 parallel strings
- Each string consists of 7 series-connected PV modules

Each module includes:

- 128 solar cells
- Maximum power: 414.8 W
- Voltage at maximum power: 72.9 V

- Current at maximum power: 5.69 A
- Open-circuit voltage: 85.3 V
- Short-circuit current: 6.09 A

This configuration enables accurate modeling of the electrical behavior under varying environmental and fault conditions.

A.4 Dataset Generation Methodology

A.4.1 Simulation-Based Data Acquisition

The dataset was generated through time-domain simulations of the PV system under different operating conditions. The total simulation duration was 0.4 s, with faults introduced at 0.2 s .

- Training data were collected in the interval: 0.2 s – 0.4 s (post-fault steady-state)
- Testing data were collected in the interval: 0.1 s – 0.3 s (including transient behavior)

This approach ensures that the dataset captures both steady-state and dynamic system responses.

A.4.2 Fault Scenarios

Four operating conditions were considered:

- Class 0: Fault-free operation
- Class 1: String fault
- Class 2: String-to-ground fault
- Class 3: String-to-string fault

These fault types represent the most common failure modes in PV systems and significantly impact system performance and safety.

A.4.3 Dataset Composition

The dataset consists of:

- Training set: 600 instances
- Testing set: 100 instances
- Total features: 30 attributes and 1 class label

The class distribution in the training dataset is as follows:

- 16.67%: Normal operation (F0)
- 25.5%: String faults (F1)
- 24.83%: String-to-ground faults (F2)
- 33%: String-to-string faults (F3)

This distribution ensures a representative sampling of all system conditions.

A.4.4 Dataset Representation

To provide a clearer understanding of the dataset structure, a representative sample extracted from the training data is presented in Figure A.3. Each row corresponds to a single observation, while each column represents a specific feature or the associated class label.

The figure illustrates a subset of the dataset in tabular form, highlighting the organization of input features alongside their corresponding class labels under different operating conditions. This representation facilitates the interpretation of the dataset structure and the relationships between variables.

The complete dataset is stored in CSV format and is publicly available online to ensure the reproducibility and transparency of the research work. It can be accessed via the Kaggle platform at the following link:

https://www.kaggle.com/datasets/amrezzeldinrashed/fault-detection-dataset-in-photovoltaic-farms

1	I1	I2	I1MAX	I1MIN	I1VAR	I2MAX	I2MIN	I2VAR	I3	I4	I3max	I3min	I3var	I4MAX	I4MIN	I5	I6	Itotal1
2	3.464131585	3.464131585	3.77651478	3.433125243	5.37438E-05	3.77651478	3.433125243	5.37438E-05	3.755050256	3.755050256	3.782291689	3.745652791	5.4144E-05	3.782291689	3.745652791	3.755050256	3.755050256	426.8767281
3	2.244765809	2.244765809	2.611713911	2.21097809	0.000350392	2.611713911	2.21097809	0.000350392	2.572121517	2.572121517	2.61461827	2.553621648	0.000381478	2.61461827	2.553621648	2.572121517	2.572121517	426.4354417
4	3.878360313	3.878360313	4.101281713	3.854098385	9.97587E-06	4.101281713	3.854098385	9.97587E-06	4.094576909	4.094576909	4.10679979	4.088955247	7.86493E-06	4.10679979	4.088955247	4.094576909	4.094576909	426.9477648
5	2.816388592	2.816388592	3.285793348	2.75814925	0.000192359	3.285793348	2.75814925	0.000192359	3.247759244	3.247759244	3.289337483	3.233320946	0.000203956	3.289337483	3.233320946	3.247759244	3.247759244	426.7386021
6	1.238798405	1.238798405	5.455666573	0.141519477	0.001153152	5.455666573	0.141519477	0.001153152	5.465640192	5.465640192	5.472441612	5.457335473	7.54072E-06	5.472441612	5.457335473	5.465640192	5.465640192	497.6608565
7	4.967361071	4.967361071	5.114487582	4.947663247	7.98331E-06	5.114487582	4.947663247	7.98331E-06	5.119613436	5.119613436	5.126098833	5.110910315	6.98892E-06	5.126098833	5.110910315	5.119613436	5.119613436	467.5491241
8	3.157037846	3.157037846	4.779538892	2.7537744	0.000170701	4.779538892	2.7537744	0.000170701	4.779906848	4.779906848	4.790567799	4.770081784	1.2192E-05	4.790567799	4.770081784	4.779906848	4.779906848	440.5913292
9	2.257497805	2.257497805	2.781367272	2.183820683	0.000358516	2.781367272	2.183820683	0.000358516	2.741517643	2.741517643	2.78445936	2.7238151	0.000359431	2.78445936	2.7238151	2.741517643	2.741517643	426.5186394
10	3.089721529	3.089721529	4.26569134	2.805382583	8.3981E-05	4.26569134	2.805382583	8.3981E-05	4.265938427	4.265938427	4.271811178	4.261446985	2.77787E-06	4.271811178	4.261446985	4.265938427	4.265938427	426.9819555
11	2.9252459	2.9252459	4.433857584	2.545095954	0.000145401	4.433857584	2.545095954	0.000145401	4.437534556	4.437534556	4.444814235	4.430006251	8.77198E-06	4.444814235	4.430006251	4.437534556	4.437534556	427.8399591
12	4.349454349	4.349454349	4.567401016	4.334717137	1.9529E-05	4.567401016	4.334717137	1.9529E-05	4.554301694	4.554301694	4.572821502	4.547245903	1.80492E-05	4.572821502	4.547245903	4.554301694	4.554301694	430.4926871
13	4.693365669	4.693365669	5.12684107	4.608856753	1.76308E-05	5.12684107	4.608856753	1.76308E-05	5.133616295	5.133616295	5.141864788	5.125496173	8.20991E-06	5.141864788	5.125496173	5.133616295	5.133616295	468.1412751
14	-3.62548279	-3.62548279	1.586466106	-4.46929027	0.001918841	1.586466106	-4.46929027	0.001918841	1.588439372	1.588439372	1.591558046	1.586446115	6.13828E-07	1.591558046	1.586446115	1.588439372	1.588439372	426.9688783
15	2.81784044	2.81784044	3.12588557	2.757478294	6.83789E-06	3.12588557	2.757478294	6.83789E-06	3.130705307	3.130705307	3.134544496	3.126720534	1.86647E-06	3.134544496	3.126720534	3.130705307	3.130705307	426.9379592
16	3.462297954	3.462297954	3.693148883	3.425118836	5.12361E-06	3.693148883	3.425118836	5.12361E-06	3.695431521	3.695431521	3.699889494	3.690744155	2.44886E-06	3.699889494	3.690744155	3.695431521	3.695431521	426.6792982
17	4.447619495	4.447619495	4.690709116	4.408073032	5.92637E-06	4.690709116	4.408073032	5.92637E-06	4.695446141	4.695446141	4.700252886	4.689878224	2.9648E-06	4.700252886	4.689878224	4.695446141	4.695446141	435.7152542
18	1.631619762	1.631619762	1.837313653	1.61781406	0.000202533	1.837313653	1.61781406	0.000202533	1.808981042	1.808981042	1.840038901	1.794300966	0.000225179	1.840038901	1.794300966	1.808981042	1.808981042	426.3479035
19	2.089427371	2.089427371	3.148090794	1.897308686	0.001264834	3.148090794	1.897308686	0.001264834	3.098405671	3.098405671	3.151183171	3.068761952	0.000817501	3.151183171	3.068761952	3.098405671	3.098405671	425.9708714
20	3.938369657	3.938369657	4.220086498	3.887706386	1.10079E-05	4.220086498	3.887706386	1.10079E-05	4.225701417	4.225701417	4.233053097	4.217994423	7.27597E-06	4.233053097	4.217994423	4.225701417	4.225701417	426.4722441
21	0.640430851	0.640430851	0.873409663	0.596542921	6.07727E-06	0.873409663	0.596542921	6.07727E-06	0.871117321	0.871117321	0.876171706	0.86916476	2.82879E-06	0.876171706	0.86916476	0.871117321	0.871117321	426.8767352
22	0.623284161	0.623284161	0.952200364	0.572848567	0.000103381	0.952200364	0.572848567	0.000103381	0.939553467	0.939553467	0.954831506	0.930658351	6.47623E-05	0.954831506	0.930658351	0.939553467	0.939553467	426.0723788
23	4.585800796	4.585800796	4.838017273	4.559888127	0.000838537	4.838017273	4.559888127	0.000838537	4.769850632	4.769850632	4.842253929	4.743090213	0.000897264	4.842253929	4.743090213	4.769850632	4.769850632	438.9168184
24	1.106496634	1.106496634	1.339409214	1.063414418	3.61465E-06	1.339409214	1.063414418	3.61465E-06	1.338162456	1.338162456	1.342182909	1.337105185	6.02809E-07	1.342182909	1.337105185	1.338162456	1.338162456	426.9689059
25	5.322665889	5.322665889	5.536780454	5.286619504	1.35247E-05	5.536780454	5.286619504	1.35247E-05	5.545854762	5.545854762	5.555035277	5.536707812	1.16367E-05	5.555035277	5.536707812	5.545854762	5.545854762	506.9262263
26	2.542369233	2.542369233	2.888739813	2.472998988	8.05511E-06	2.888739813	2.472998988	8.05511E-06	2.892126376	2.892126376	2.895488557	2.888654671	1.61883E-06	2.895488557	2.888654671	2.892126376	2.892126376	426.892619
27	-0.18863295	-0.18863295	0.994098188	-0.45536471	9.24147E-05	0.994098188	-0.45536471	9.24147E-05	0.997915348	0.997915348	0.998959126	0.996469017	2.09464E-07	0.998959126	0.996469017	0.997915348	0.997915348	426.8418717
28	2.398980351	2.398980351	2.626549298	2.376383568	0.000395761	2.626549298	2.376383568	0.000395761	2.5857835	2.5857835	2.629497386	2.561828266	0.000443041	2.629497386	2.561828266	2.5857835	2.5857835	426.347901
29	0.487364179	0.487364179	0.875779531	0.402136413	9.72038E-06	0.875779531	0.402136413	9.72038E-06	0.877318383	0.877318383	0.878821145	0.875939961	4.3556E-07	0.878821145	0.875939961	0.877318383	0.877318383	426.8926175
30	2.47405265	2.47405265	2.727193412	2.453982273	4.76744E-05	2.727193412	2.453982273	4.76744E-05	2.704769238	2.704769238	2.730509874	2.698095381	4.95283E-05	2.730509874	2.698095381	2.704769238	2.704769238	426.905384
31	1.706359569	1.706359569	2.077314969	1.64036145	1.51251E-05	2.077314969	1.64036145	1.51251E-05	2.06734505	2.06734505	2.080371834	2.064434374	7.98232E-06	2.080371834	2.064434374	2.06734505	2.06734505	426.9759579
32	4.622142459	4.622142459	5.692056385	4.444069298	0.001117107	5.692056385	4.444069298	0.001117107	5.616424548	5.616424548	5.697462732	5.584412892	0.001162242	5.697462732	5.584412892	5.616424548	5.616424548	511.5346567
33	1.643976068	1.643976068	2.531988529	1.486654875	0.000823221	2.531988529	1.486654875	0.000823221	2.493106425	2.493106425	2.5348731	2.467795774	0.000507922	2.5348731	2.467795774	2.493106425	2.493106425	426.1655494

Figure A3: Sample extracted from the training dataset illustrating input features and class labels.

A.5 Measured Variables and Extracted Features

The dataset includes electrical and environmental measurements collected from the simulated PV system.

A.5.1 Electrical Measurements

- Currents at the top and bottom of each string: I1A, I1B, I2A, I2B, I3A, I3B
- Total current (Itotal)
- DC voltage (Vdc)
- DC power (Pdc)

A.5.2 Environmental Variables

- Temperature (T): 10°C to 35°C
- Solar irradiance (IR): 100 W/m² to 1000 W/m²

A.5.3 Statistical Features

For each electrical signal, the following statistical descriptors were computed:

- Mean
- Maximum
- Minimum
- Variance

These features enhance the capability of machine learning models to capture system behavior under different conditions.

A.6 Feature Engineering

To improve classification accuracy, additional derived features were introduced:

- Range1 = I1Amax − I1Amin
- Range2 = I2Amax − I2Amin
- Range3 = I1A − I1B
- Range4 = I2A − I2B

These range-based features capture variations in current distribution and are particularly effective in distinguishing between different fault types.

A.7 Data Characteristics and Variability

The dataset incorporates randomness in key parameters to ensure generalization:

- Temperature: uniformly distributed between 10°C and 35°C
- Irradiance: uniformly distributed between 100 W/m² and 1000 W/m²
- Fault resistance: varies between 1 Ω and 2000 Ω

This variability allows the dataset to reflect realistic operating conditions and improves the robustness of the trained models.

A.8 Reproducibility and Reliability

The simulation-based dataset generation approach provides:

- Controlled fault injection
- Repeatable experimental conditions
- High-quality labeled data

This ensures the reliability of the evaluation process and enables fair comparison between different fault detection algorithms.