University Mohamed Boudiaf - M'sila

FACULTY OF TECHNOLOGY

Department of Electronics

Serial Number……

Registration Number: DUN28012023201536044165

Thesis

Submitted for the Degree of

DOCTORATE LMD

Filière : ***Electronics***

Spécialité : ***Instrumentation***

THEME

Big Data and Artificial Intelligence for Improving the Performance and Efficiency of Large-Scale Grid-Connected PV Power Plant.

Presented by
Amiri Ahmed Faris
Defended on: 17/11/2024

Before the jury composed of:

Name & Surname	Grade	Institution	Title
Mezache Amar	**Professor**	Univ. of **M'sila**	President
Oudira Houcine	**Professor**	Univ. of **M'sila**	Supervisor
Chouder Aissa	**Professor**	Univ. of **M'sila**	Co-Supervisor
Bouzgou Hassen	**Professor**	Univ. of **Batna**	Examiner
Kara Kamel	**Professor**	Univ. of **Blida**	Examiner
Benguesmia Hani	**MCA**	Univ. of **M'sila**	Examiner
Kichou Sofiane	**PhD**	Univ. of **Czech Tech. in Prague**	Guest

Academic Year: 2023/2024

Acknowledgments:

I extend my heartfelt gratitude to my supervisor, Pr. Oudira Houcine and Aissa chouder , whose unwavering support, insightful guidance, and constructive feedback have been invaluable throughout this journey. Your mentorship has not only shaped my research but has also inspired me to strive for excellence in all aspects of academic pursuit. Special thanks to Dr. Sofiane Kichou, with whom I had the pleasure of working. His guidance was instrumental in achieving great results and fostering excellent teamwork.

I sincerely thank the members of my thesis committee for their expertise, encouragement, and valuable critiques. Special thanks to Professor Mezache Amar (*University of M'sila*), President of the jury; Professor Bouzgou Hassen (*University of Batna*); Professor Kara Kamel (*University of Blida*); and Dr. Benguesmia Hani (*University of M'sila*), Examiners, for their insightful feedback and scholarly contributions. Their diverse perspectives have enriched the quality of this work, broadened my understanding, and helped me navigate the complexities of my research.

I would like to express my appreciation to the staff of Faculty of technology, whose dedication to academic excellence has created an inspiring environment for intellectual growth and collaboration. I am particularly thankful to them for their assistance and support at various stages of my doctoral journey.

Special thanks are due to my colleagues and fellow researchers, for their camaraderie, encouragement, and stimulating discussions. Their insights and friendship have made this academic endeavour both rewarding and enjoyable.

I am indebted to my family Specially my father and mother for their unwavering love, encouragement, and understanding throughout the ups and downs of this doctoral pursuit. Their unwavering belief in my abilities has been a constant source of strength and motivation.

To everyone who has contributed, directly or indirectly, to the completion of this thesis, I offer my sincere thanks. Your support and encouragement have been instrumental in this achievement.

Abstract:

This thesis presents reliable methods for fault detection and diagnosis in Photovoltaic (PV) systems.

The first method proposes a two-step approach for developing a reliable PV model and constructed a fault detection procedure using Random Forest Classifiers (RFCs). The first step involves identifying the unknown parameters of the One-Diode Model (ODM) using the Modified Grey Wolf Optimization (MGWO) algorithm and simulating the PV array to extract maximum power point (MPP) coordinates. The second step involves developing two RFCs: one for fault detection and another for fault diagnosis.

The second method uses the Sandia Array Performance Model (SAPM) for accurate photovoltaic system simulation. Parameters are extracted with the Artificial Bee Colony (ABC) algorithm to optimize and reduce errors between measured and simulated data. Additionally, deep learning is employed by combining Convolutional Neural Networks (CNN) and Bidirectional Gated Recurrent Units (Bi-GRU) to analyze dynamic system power outputs at the MPP for fault detection and diagnosis with high precision.

The third work develops a predictive modeling method for PV generation using rigorous feature selection, outlier removal, and hyperparameter tuning. The method is implemented in a MATLAB interface to predict key parameters and evaluate system performance. The efficiency of these methods is evaluated using real data from actual PV systems.

Keywords: Photovoltaic, faults, Random Forest Classifiers, Modified Grey Wolf Optimization, Convolutional Neural Network, Bidirectional Gated Recurrent Unit, deep learning.

Résumé :

Cette thèse propose des solutions efficaces et simples pour la détection et le diagnostic des défauts dans les systèmes photovoltaïques (PV). Le travail est organisé autour de trois méthodes principales :

Première méthode repose sur une approche en deux étapes : dans la première étape, un modèle précis des panneaux photovoltaïques est développé en utilisant le modèle à une diode (One-Diode Model - ODM), avec l'identification des paramètres inconnus grâce à l'algorithme amélioré du loup gris (Modified Grey Wolf Optimization - MGWO), et l'extraction des coordonnées du point de puissance maximale (MPP). La deuxième étape consiste à développer deux classificateurs basés sur la méthode des forêts aléatoires (Random Forest Classifiers - RFCs), l'un destiné à détecter les défauts et l'autre à diagnostiquer leur nature.

Deuxième méthode repose sur l'utilisation du modèle de performance des panneaux photovoltaïques Sandia (SAPM) pour simuler avec précision le comportement des panneaux photovoltaïques. Les paramètres du modèle sont extraits à l'aide de l'algorithme de l'abeille artificielle (ABC), afin d'optimiser les paramètres et réduire les erreurs entre les données mesurées et les données simulées. Cette méthode combine également l'apprentissage profond avec un réseau neuronal convolutif (CNN) et une unité récurrente à portes bidirectionnelles (Bi-GRU) pour analyser les sorties dynamiques des systèmes photovoltaïques au point de MPP, permettant ainsi de détecter et de diagnostiquer les défauts avec une grande précision.

Troisième méthode : Elle est axée sur la modélisation prédictive de la production photovoltaïque et l'évaluation des performances du système. Cette méthode inclut la suppression des données aberrantes, la sélection rigoureuse des caractéristiques pertinentes et l'optimisation des hyperparamètres pour améliorer les modèles prédictifs. Elle est intégrée dans une interface MATLAB pour faciliter son application dans des scénarios réels.

L'efficacité de ces méthodes a été évaluée à l'aide de données réelles provenant de systèmes photovoltaïques, prouvant leur capacité à améliorer la performance des systèmes tout en détectant et diagnostiquant les défauts de manière précise et rapide.

Mots-clés : photovoltaïque, fautes, Classificateurs Forest Random, optimisation, réseau neuronal convolutif, unité récurrente à portes bidirectionnelles, deep learning.

الملخص:

هذه الأطروحة تقدم حلولًا فعّالة وبسيطة للكشف عن الأعطال وتشخيصها في أنظمة الطاقة الكهروضوئية .(PV) يتمحور العمل حول ثلاث طرق رئيسية:

تعتمد الطريقة الأولى على نهج من خطوتين: الأولى تتضمن إنشاء نموذج دقيق للألواح الكهروضوئية باستخدام نموذج الصمام الثنائي الواحد (One-Diode Model) ، حيث يتم تحديد المعلمات غير المعروفة للنموذج باستخدام خوارزمية الذئب الرمادي المحسنة (Modified Grey Wolf) واستخراج إحداثيات نقطة الطاقة القصوى (MPP). الثانية تتمثل في تطوير مصنّفين باستخدام خوارزميات الغابة العشوائية(Random Forest Classifiers) ، حيث يكون الأول مخصصًا لتحديد الأعطال والثاني لتشخيص نوع العطل.

الطريقة الثانية تعتمد على استخدام نموذج أداء المصفوفات الكهروضوئية للساندیا (SAPM) لمحاكاة سلوك الألواح الكهروضوئية بدقة. يتم استخراج المعلمات الخاصة بهذا النموذج باستخدام خوارزمية خلية النحل الاصطناعية (ABC) لتقليل الأخطاء بين البيانات الفعلية والنموذجية. كما تعتمد على التعلم العميق باستخدام مزيج من الشبكة العصبية الالتفافية (CNN) ووحدة البوابة التكرارية ثنائية الاتجاه (Bi-GRU) لتحليل المخرجات الديناميكية للكشف عن الأعطال وتشخيصها بدقة.

الطريقة الثالثة تهدف إلى نمذجة وتوقع إنتاج الطاقة الكهروضوئية وتحليل أداء النظام. تتضمن هذه الطريقة إزالة البيانات الشاذة، اختيار الميزات الأكثر تأثيرًا، وضبط المعلمات لتحسين أداء النماذج التنبؤية. تم تنفيذ هذه الطريقة داخل واجهة MATLAB لتسهيل استخدامها في التطبيقات الواقعية.

تم تقييم فعالية هذه الطرق باستخدام بيانات حقيقية من أنظمة طاقة كهروضوئية، مما يثبت كفاءتها في تحسين أداء الأنظمة والكشف عن الأعطال بدقة وسرعة.

الكلمات الرئيسية: الطاقة الشمسية، الأعطال، مصنفات الغابات العشوائية، خوارزمية الذئب الرمادي المحسنة، الشبكة العصبية الالتفافية وحدة البوابة التكرارية ثنائية الاتجاه، التعلم العميق.

Contents :

List of figures:

List of Tables :

list of Abbreviations:

- **A**: Ampere
- **ABC**: Artificial Bee Colony
- **AC**: Alternating Current
- **AI:** Artificial Intelligence
- **ANFIS :** Adaptive Neuro-Fuzzy Inference System
- **ANN :** Artificial Neural Network
- **ARIMA:** Auto-Regressive Integrated Moving Average
- **αI_{mp}**: Normalized Temperature Coefficient for Imp
- **ANFIS**: Adaptive Neuro-Fuzzy Inference Systems
- **ARMA**: AutoRegressive Moving Average
- **Bi-GRU**: Bidirectional Gated Recurrent Unit
- **βV_{oc}**: Temperature Coefficient for Voc
- **C0, C1, C2, C3:** Empirical Coefficients for PV System Modeling or SAPM Model Parameters
- **CatBoost :** Categorical Boosting
- **CLs**: Convolutional Layers
- **CNN**: Convolutional Neural Network
- **Conv1D**: One-Dimensional Convolutional Layers
- **CV**: Cross Validation
- **DBM**: Deep Boltzmann Machine
- **DBN**: Deep Belief Network
- **DC**: Direct Current
- **DL**: Deep Learning
- **DTL**: Deep Transfer Learning
- **Eg:** Bandgap Energy of the Semiconductor
- **ETent:** Extra Trees with Entropy
- **F1_score:** F1 Score for Classifier Performance
- **FCL :** Fully Connected Layer
- **FDD**: Fault Detection and Diagnosis
- **FC**: Fully Connected
- **FN**: False Negatives

- **FP**: False Positives
- **G**: Irradiance / Solar Irradiance
- **GB**: Gradient Boosting
- **GCPV**: Grid-Connected Photovoltaic
- **GRU**: Gated Recurrent Unit
- **GWO**: Grey Wolf Optimization
- **I :** Current
- **Io :** Diode Reverse Saturation Current
- I_{ph}**:** Photocurrent
- **IRT :** Infrared Thermography
- $\mathbf{I_{SC}}$ **:** Short-Circuit Current
- $\mathbf{I_{DC}}$: Current
- $\mathbf{I_{meas}}$: Measured Current
- $\mathbf{I_{mp}}$: the PV module's current Standard Test Conditions
- $\mathbf{I_{mpp}}$: Current at Maximum Power Point
- $\mathbf{I_{ph}}$: Photogenerated Current
- $\mathbf{I_{sc}}$: Short Circuit Current
- $\mathbf{I_{est}}$: Estimated Current
- **I-V:** Current-Voltage characteristic curve.
- **K:** Boltzmann Constant (1.38×10^{-23} J/K)
- **KNN**: k-Nearest Neighbors
- **KW**: Kilowatt
- **kVA**: Kilovolt-Ampere
- **LGBM:** Light Gradient Boosting Machine
- **LSTM**: Long Short-Term Memory
- **LR**: Linear Regressor
- **ML:** Machine Learning
- **MPP :** Maximum Power Point
- **MPPT:** Maximum Power Point Tracking
- **MAE**: Mean Absolute Error
- **MLP**: Multi-layer Perceptron
- **MGWO**: Modified Grey Wolf Optimization

- **MM5**: Fifth-Generation Penn State/NCAR Mesoscale Model
- **MPP**: Maximum Power Point
- **MPPT**: Maximum Power Point Tracking
- **N**: Number
- **NB :** Naïve Bayes
- **NOCT**: Nominal Operating Cell Temperature
- **N_p**: Number of Strings in Parallel
- **NRMSE**: Normalized Root Mean Square Error
- **Ns**: Number of Modules in Series
- **ODM**: One-Diode Model
- **P:** Power
- **Pa** : Pascal
- **PLs**: Pooling Layers
- **P_{DC}**: PV power
- **Pmp :** Maximum Power
- **PNN :** Probabilistic Neural Network
- **PR :** Performance Ratios
- **PSIM**: Power Simulator / Power Simulation
- **PV**: Photovoltaic
- **q :** Electric Charge (1.602×10^{-19} C)
- **QDA :** Quadratic Discriminant Analysis
- **R :** Resistance
- **ResGRU:** Residual Gated Recurrent Unit
- **ResNet :** Deep Residual Network
- **R^2**: R Squared
- **RBM**: Restricted Boltzmann Machine
- **RF**: Random Forest
- **RFCs**: Random Forest Classifiers
- **RNN**: Recurrent Neural Network
- **RMSE**: Root Mean Square Error
- **R_{sh}**: Shunt Resistance
- **R_s**: Series Resistance

- **SAE**: Stacked Auto Encoder
- **SAPM**: Sandia Array Performance Model
- **STC**: Standard Test Conditions
- **SVM**: Support Vector Machines
- **SVR**: Support Vector Regression
- **T**: Temperature(°C or K)
- **T_m**: Module Temperature
- **TN**: True Negatives
- **TP**: True Positives
- **$T_{\_ ref}$:** Reference Temperature (°C or K)
- **Turing :** Turing Complete
- **UV :** Ultraviolet Radiation
- **V**: Voltage
- **V_{OC}**: Open Circuit Voltage
- **V_{DC}**: Voltage Direct Current
- **V_{mp}**: the PV module's voltage under Standard Test Conditions
- **V_{mpp}**: Voltage at Maximum Power Point
- **Vpv**: Photovoltaic Voltage
- **Vt :** Thermal Voltage (V)
- **WRF-NMM**: Weather Research and Forecasting - Nonhydrostatic Mesoscale Model
- **C:** Coefficient vector.
- **D:** Distance.
- **f :** Fitness function.
- **XGBoost :** Extreme Gradient Boosting
- **Wp**: Watt peak

list of Symbols :

- I_0: Current corresponding to Curve 0. (A)
- I_{0i}: Current point on Curve 0. (A)
- I_{1i}: Current point on Curve 1. (A)
- I_{2j}: Current point on Curve 2. (A)
- I_{3i}: Current point on Curve 3. (A)
- I_4: Current corresponding to Curve 4. (A)
- I_{4i}: Current point on Curve 4. (A)
- I_{sc1}: Short-circuit current of Curve 1. (A)
- I_{sc2}: Short-circuit current of Curve 2. (A)
- S: Set of possible solutions.
- V_0: Voltage corresponding to Curve 0. (V)
- V_{0i}: Voltage point on Curve 0. (V)
- V_{1i}: Voltage point on Curve 1. (V)
- V_{2j}: Voltage point on Curve 2. (V)
- $V_{3i:}$ Voltage point on Curve 3. (V)
- V_4: Voltage corresponding to Curve 4. (V)
- V_{4i}: Voltage point on Curve 4. (V)
- V_{ij}: Neighbouring food source.
- x: Vector of parameters.
- X: Position vector.
- x_i: Possible solutions.
- X_{ij}: Solution vector component.

- x_{max}: Maximum value of solution vector.
- x_{min}: Minimum value of solution vector.
- X_p: Position vector of the prey.
- X_α: Position vector of the alpha wolf.
- $X\beta$: Position vector of the beta wolf.
- X_δ: Position vector of the delta wolf.
- α: Parameter to be ascertained.
- ϕ: Random number between [-1, 1].
- ω: Translation parameter.
- **β :** Temperature Coefficient of Open-Circuit Voltage

Published Works:

Articles:

1. Amiri, A. F., Oudira, H., Chouder, A., & Kichou, S. (2024). Faults detection and

diagnosis of PV systems based on machine learning approach using random forest classifier. Energy Conversion and Management, 301, 118076.

2. Amiri, A. F., Kichou, S., Oudira, H., Chouder, A., & Silvestre, S. (2024). Fault detection and diagnosis of a photovoltaic system based on deep learning using the combination of a convolutional neural network (cnn) and bidirectional gated recurrent unit (Bi-GRU). Sustainability, 16(3), 1012.

3. Amiri, A.F.; Chouder, A.; Oudira, H.; Silvestre, S.; Kichou, S. Improving Photovoltaic Power Prediction: Insights through Computational Modeling and Feature Selection. *Energies* 2024, *17*, 3078. https://doi.org/10.3390/en17133078

Conferences:

1. **Amiri, A. F**., Oudira, H., & Chouder, A. (2022, November). Faults detection of PV systems based on extracted parameters using Modified Grey Wolf algorithm. In *2022 International Conference of Advanced Technology in Electronic and Electrical Engineering (ICATEEE)* (pp. 1-6). IEEE.

2. Oudira, H., Chouder, A., & **Amiri, A. F**. (2023, October). Prediction Model of PV Module Based on Artificial Neural Networks for the Energy Production. In *2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)* (pp. 85-88). IEEE.

3. Oudira, H., **Amiri, A. F**, & Chouder, A. (2024, October). Prediction Model of PV Module Based on Artificial Neural Networks for the Energy Production. In *4th IEEE International Multi-Conference on Smart Systems & Green Process(IMC-SSGP'24), October 30 -November 2, 2024, Djerba-Tunisia.*

GENERAL INTRODUCTION

General Introduction outline:

1. Motivation:

In recent years, global energy policies have prioritized reducing carbon emissions and shifting towards more sustainable energy sources. This shift is evident in the growing adoption of renewable energy sources (RES) aimed at achieving a greener future. Solar photovoltaics (PV) has emerged as a crucial energy source among RES, offering solutions to environmental challenges at highly competitive costs [1]. Consequently, global PV capacity exceeded the terawatt mark in early 2022, representing two-thirds of the anticipated increase in global renewable capacity by 2023 [2]. Furthermore, PV remains the most economically viable choice for new electricity generation in many nations, with expected cost reductions in generation by 2024 [3].

2. Problem Statement:

Photovoltaic (PV) systems are planned to function in severe external environments, enduring extreme weather conditions, wind-induced vibrations, and exposure to ultraviolet radiation [4,5]. In these challenging conditions, a multitude of failures and malfunctions can arise, potentially diminishing the lifespan of PV modules, decreasing the totall energy output of the system, jeopardizing system availability, and presenting safety hazards to personnel engaged in their operation and upkeep [6]. PV systems are vulnerable to an array of faults, spanning from momentary disturbances to lasting breakdowns, all of which can profoundly affect their functionality and safety. Hence, timely identification and diagnosis of these faults are imperative to safeguard the enduring reliability and sustainable functioning of PV systems.

Numerous fault detection and diagnosis techniques have been suggested in the literature, these methods vary in terms of speed, complexity, sensor requirements, and their capacity to identify different types of faults

3. Literature review and background:

Over the past decade, a multitude of techniques for detecting and diagnosing faults in PV systems has emerged. Model-based approaches involve simulating the performance of the actual photovoltaic (PV) installation and comparing the simulated output power with the monitored output [7, 8]. **Chouder and Silvestre** presented in [9] a fault detection strategy for PV systems that relies on power loss analysis. They classified recognized faults into faulty strings, faulty modules, and partial shading by conducting a careful analysis of the ratios

between simulated and measured outputs [9]. **Silvestre *et al.,*** suggested an automated method for detecting faults in grid-connected photovoltaic (PV) systems by monitoring current and voltage indicators. This approach sets thresholds based on normal operational behavior; if these thresholds are surpassed, a fault signal is activated. Faults are then identified by examining the ratios of current and voltage [10]. **Drews *et al.,*** employed a fault detection technique that involves establishing a power residual threshold based on irradiance and temperature data obtained from weather satellites, thus eliminating the need for on-site sensors [11]. Although this approach eliminates the necessity for supplementary on-site sensors, there is a potential compromise in accuracy owing to the likelihood of larger margins of error in weather data. In contrast, **Garoudja *et al.,***'s methodology establishes a threshold on the exponentially weighted moving average of power, current, and voltage residuals, incorporating historical data for fault detection instead of relying solely on the most recent observation [12]. In general, the fault detection techniques outlined above are relatively simple to implement. However, the main challenge lies in accurately determining appropriate thresholds to guarantee their reliability.

Unlike conventional model-based fault detection methods [5–9], which involve simulating the performance of PV installations and comparing the simulated output power with the observed data, machine learning (ML) and deep learning (DL) techniques have become increasingly popular. These methods are seen as promising solutions for fault detection and diagnosis in PV systems. Many studies have assessed the efficiency of ML and DL approaches in identifying and diagnosing faults in these systems. Various ML techniques have been developed for fault detection and diagnosis in PV systems [13–16], including Random Forest (RF), support vector machines (SVM), and artificial neural networks (ANN). **Bendary*et al.,*** suggested two adaptive neuro-fuzzy inference system-based controllers (ANFIS) to handle cleaning, tracking, and faults in PV systems [17] by comparing actual current and voltage values with trained historical data while accounting for ambient conditions like irradiation and temperature. **Madeti and Singh** brought a k-nearest neighbors (KNN) algorithm for real-time fault detection, capable of identifying line-to-line, open circuit, and partial shading faults [18], though its accuracy that could be improved. **Eskandari *et al.,*** developed an ensemble learning method combining SVM, Naïve Bayes (NB), and KNN [19], achieving up to 99.5% accuracy for line-to-line fault detection. **Kapucu*et al.,*** used an ensemble approach with quadratic discriminant analysis (QDA), extra trees with entropy (ETent), and decision trees (DT) [20] to detect partial shading and short circuits, achieving 97.67% accuracy post-

optimization. **Adhya *et al.,*** employed light gradient boosting (LGBM), categorical boosting (CatBoost), and extreme gradient boosting (XGBoost) to recognize faults in PV systems [21], reaching around 99% accuracy, though the method that is complex and needs further refinement.

Akram *et al.,* suggested a monitoring method using a Probabilistic Neural Network (PNN) for detecting short- and open-circuit faults, achieving 98.53% accuracy [22]. **Chen *et al,.*** used RF to classify open circuit, partial shading, degradation, and short circuit faults based on high-frequency current and voltage measurements in parallel circuit substrings, but it was limited to certain weather conditions [23]. **Gong *et al.,*** [24] employed classification regression trees to diagnose PV array faults using I-V curves, achieving 97.9% accuracy. **Mellit *et al.,*** proposed an embedded system for remote monitoring and fault diagnosis based on two ANN models for detection and diagnosis, showing good accuracy in a real-time low-cost edge device setup [25].

Despite these promising AI-based fault detection and diagnostic methods, their accuracy is often limited by the quality of training data. Actual measurements alone are insufficient for training AI models, making the development of accurate PV system models essential to simulate various faults and environmental conditions. Additionally, data processing is crucial for deploying ML-based fault diagnosis procedures, as highlighted by **Wang *et al.,*** who emphasized its importance in achieving 81%-99% accuracy for complex fault categorization [26].

Conversely, Deep Learning (DL) has appeared as the next evolutionary stage of machine learning, attracting significant consideration for its proficiency in pattern recognition, knowledge discovery and data mining. Its remarkable benefit lies in its ability to extract high-level abstract features from extensive datasets, making it particularly advantageous for classification tasks [27]. **Liu *et al.,*** proposed a fault diagnosis method for PV arrays using stacked auto-encoder (SAE) and clustering. This method exploits inherent I-V characteristics, facilitating automatic feature extraction and fault diagnosis [28]. Similarly, **Chen *et al.,*** established an innovative deep residual network (ResNet) for intelligent fault detection and diagnosis. This approach leverages output I-V characteristic curves and input ambient condition data to accurately identify and diagnose faults in PV systems [29]. Additionally, **Gao and Wai** introduced a fault classification method for PV arrays that employs a model combining convolutional neural networks (CNN) and residual gated recurrent units

(ResGRU). This method distinguishes differences in I-V curves under different fault conditions, attaining a classification accuracy of 98.61% [30]. **Eldeghad*et al.,*** proposed a deep learning approach tuned with a particle swarm optimization (PSO) for fault diagnosis in PV plants. This algorithm showed promising outcomes in fault detection, potentially enhancing system reliability, efficiency and safety [31]. **Appiah *et al.,*** used long short-term memory (LSTM) networks to identify fault features, which were then fed into a SoftMax regression classifier for fault detection and identification [32]. Another approach, presented in [33], integrates deep learning with Infrared Thermography (IRT) for fault classification in PV systems. The results indicate that the IRT-DL method outperforms other IRT-ML methods in terms of accuracy and classification. However, the use of IRT for fault detection in PV systems faces challenges such as limitations related to surface defects, susceptibility to dynamic system conditions, augmented equipment costs, and restraints in detecting specific fault types.

4. Contribution:

The purpose of this thesis is highlighted by presenting the summary of our contributions that are deal with the topic cited into motivation section.

The first study outlines a two-step methodology for developing a trusted PV model essential for monitoring and fault detection. Additionally, it establishes a fault detection process using Random Forest Classifiers, optimized via a grid-search algorithm for hyperparameter tuning.

- **First Step**: The focus is on precisely estimating the unknown parameters of the One-Diode Model (ODM) of the PV array working under outside environment. This is achieved using a novel translation technique that corrects randomly measured current-voltage (I-V) curves to reference standard test conditions (STC). Analytical formulations are employed to derive these parameters across various operating conditions, accounting for variations in temperature and irradiance[34]. An optimization algorithm based on the Modified Grey Wolf Optimization (MGWO), introduced by Mirjalili et al. in 2014 [35], is used to determine the five unknown parameters of the ODM at STC. The MGWO algorithm's innovative position updating concept enhances search efficiency and exploitation capabilities while maintaining quick convergence rate. Based on the estimated parameters, the evolution of the

maximum power point (MPP) model is derived and simulated using actual dynamic measurements from a grid-connected PV system in Algeria.

- **Second Step**: The PV array is simulated to extract MPP coordinates and construct its operational databases through PSIM/MATLAB co-simulations. An efficient fault detection and diagnosis process is implemented using the Random Forest Classifier (RFC), involving the development of two RFCs: one for fault detection (binary classifier) and another for fault diagnosis (multiclass classifier). The approach is comprehensively compared with other machine-learning methods for detecting and diagnosing faults in the studied grid-connected PV system. The testing stage includes five operating scenarios: a line-to-line fault, a healthy system, three short-circuited modules in one string, a string disconnected from the array, and shading effects on three panels.

The second study intends to advance fault detection and diagnosis process for PV systems, based on enhancing efficiency, reliability, and safety.

- The novel method integrates Bidirectional Gated Recurrent Units (Bi-GRU) and Convolutional Neural Networks (CNN) within a deep learning structure. This combination of parallel and sequential processing allows the neural network to utilize the power of both convolutional and recurrent layers simultaneously, enabling successful fault detection and diagnosis. Distinct previous works that rely on I-V curves, this approach uses dynamic PV system outputs at maximum power points (MPP), which are easier to get for most PV plants, thus overcoming problems related with precise I-V curve measurements. An accurate PV model is used to create reliable databases representing PV system operation in both healthy and faulty states.

The third work brings several key contributions to the field, the advancement of predictive modeling in the renewable energy sector and offers valuable insights for optimizing for PV systems and management. This work will be structured around two key points

- **Firstly,** it employs rigorous feature selection techniques like **Pearson and Spearman** correlation analysis to identify the most pertinent environmental variables. This ensures that only influential factors are considered, boosting both the interpretability and performance of the predictive model. Secondly, by integrating Isolation Forest for outlier detection during data preprocessing, the method effectively removes

anomalies, enhancing the model's ability to generalize to unseen data. Additionally, the use of RandomizedSearchCV streamlines hyperparameter tuning, with Random Forest emerging as the optimal model choice. Random Forest's ensemble nature and capacity to capture non-linear relationships make it well suited for modeling the intricate dynamics of PV system generation.

- **Finally,** the integration of Python-trained models into a MATLAB interface marks a significant advancement in accurately predicting key parameters crucial for photovoltaic (PV) systems, such as PV generation, P_{DC} (power at the direct current), V_{DC} (voltage at the direct current), and IDC (current at the direct current). Moreover, this interface goes beyond mere prediction by incorporating calculations metrics to evaluate yield, losses, and performance ratios (PR). This comprehensive assessment capability enables users to thoroughly analyze system performance and health, providing valuable insights for optimizing efficiency and addressing potential issues.

5. Thesis organization:

The thesis comprises four chapters, with chapters two and three featuring original research already published.

- **Chapter One** provides a comprehensive overview of PV systems, encompassing various types of faults and methods for their detection and diagnosis. Additionally, it includes a thorough literature review of existing PV generation prediction methods.
- **Chapter Two** surveys the first original work, focusing on fault detection and diagnosis utilizing machine-learning techniques. It introduces a two-step methodology for creating a reliable model of photovoltaic (PV) arrays and implementing a fault detection strategy utilizing Random Forest Classifiers (RFCs). First, we identified the five unknown parameters of the one-diode model (ODM) by incorporating the current-voltage translation technique to predict the reference curve and employing the MGWO algorithm. Next, we simulated the PV array to obtain MPP coordinates and built operational databases through PSIM/MATLAB co-simulations. Following this, we applied two Random Forest Classifiers (RFCs): one for fault detection as a binary classifier, and another for fault diagnosis as a multiclass classifier.

- **Chapter Three** is devoted to the second original work, presenting an innovative use of deep learning for detecting and identifying faults in PV systems through a three-step methodology. Initially, a trusted PV model is proposed and tuned using a heuristic optimization approach. Next, a comprehensive database is constructed, integrating PV model data with monitored module temperature and solar irradiance for both normal and faulty operational states. Finally, fault classification is achieved using features derived from a combination of Bidirectional Gated Recurrent Unit (Bi-GRU) and Convolutional Neural Network (CNN).
- **Chapter Four**, is devoted to the third original work, in which the emphasis is on the identifying the most effective machine learning techniques and supervised learning models to estimate power output from Photovoltaic (PV) plants precisely. The chapter objective also is the implementation of the considered model into a MATLAB application for real-time predictions for enhancing usability and accessibility for a wide range of applications in renewable energy.

The thesis concludes with reflections on the findings and outlines future research directions.

Chapter outline:

I.1. Introduction:

In today's world, electricity plays a crucial role in economic development globally. Its significance continues to grow alongside technological advancements, industrialization, and the increasing demand for modern comforts. Increasing electricity production means enhancing quality of life and generating wealth. However, the majority of electricity production currently relies on fossil fuels. Faced with the depletion of these resources, environmental concerns, and a significant rise in energy demand, the search for new energy sources has become a priority for many countries. Renewable energies offer an environmentally friendly alternative to fossil fuels. Their exploitation could provide electricity nationwide, particularly in remote areas, thereby avoiding the need for new power lines. A better understanding of the importance of renewable energies, including photovoltaic energy, is essential [1, 3].

However, PV systems are susceptible to various failures and anomalies during operation, which can significantly reduce their efficiency, degrade performance, and even shorten their lifespan. Therefore, it is crucial to develop efficient strategies for early detection and diagnosis of faults in PV systems.

This chapter provides a brief overview of PV systems and their main components. It then outlines the different types of failures that can occur in PV systems and their primary causes. After that, it presents a review of the latest strategies proposed for fault detection and diagnosis in PV systems. Finally, an overview of PV systems prediction is presented.

I.2. PV System:

Solar energy presents an alternative to conventional fossil fuels. Available in huge amount and distributed across the entire surface of the Earth, it allows for the capture of up to 1000 W/m^2 in temperate zones. Whether in urban environments or in isolated locations, this energy can be captured and utilized in both thermal and electrical forms.

I.2.1. photovoltaic effect:

The photovoltaic effect is the phenomenon by which photovoltaic cells, or solar cells, directly convert sunlight into electricity. This process can be explained as follows: when a photon from solar radiation strikes a PV cell, its energy can be transferred to an electron within the semiconductor material of the cell. With this additional energy, the electron can

then move from its normal position within the atom (from the valence band to the conduction band, crossing the band gap), creating a gap known as a hole, which contributes to the flow of current in an electrical circuit. This electron-hole pair is referred to as the photovoltaic effect.

I.2.2. PV solar cell:

Photovoltaic (PV) Cell: The fundamental unit of the PV generator is the PV cell. Typically, commercial solar cells produce around 0.6 volts of voltage, with the generated current influenced by sunlight intensity and cell surface area [36].

I.2.3. Equivalent circuit of PV solar cell:

Even though a variety of models has been developed in the literature, as shown in Figure I.1, the ODM is commonly used to describe the solar cells performance [37].

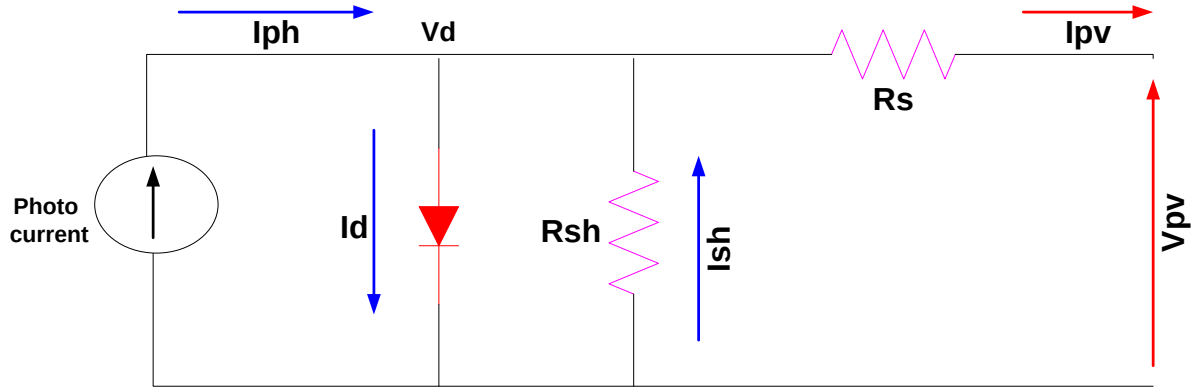

Figure I.1. Equivalent circuit model of ODM.

The relation between the output current and voltage is given by [38]:

$$I = I_{ph} - I_0 \left[\exp\left(\frac{q(V+R_S I)}{nkT}\right) - 1\right] - \frac{V+R_S I}{R_{sh}} \qquad \text{(I.1)}$$

The five parameters namely $I_{ph}, I_0, R_s,\ R_{sh}$ and n of the characteristic equation are generally specified by manufacturers of the system.

Where I_o denotes the diode saturation current *(A)*, I_{ph} signifies the photocurrent *(A)*, and n represents the diode ideality factor. The Boltzmann constant ($1.38 \times 10^{-23} JK^{-1}$) is denoted by *k*, while *T* represents the cell temperature in Kelvin *(K)*. The parameter *q* stands for the electrical charge ($1.602 \times 10^{-19} C$), and *Vt(V)* denotes the thermal voltage, expressed as $V_t = kT/q$. Finally, R_{sh} and R_s refer to shunt and series resistances *(Ω)*.

I.2.3.1. Photovoltaic (PV) Generator: At the heart of the PV system lies the PV generator, responsible for converting solar energy into electrical power through the photovoltaic effect. This phenomenon occurs when sunlight interacts with materials possessing photovoltaic properties, initiating electricity generation.

I.2.3.2. Photovoltaic (PV) Module: While individual solar cells produce some voltage and current, they are insufficient for practical use. Therefore, cells are interconnected in series to increase voltage and in parallel to boost current, forming a PV module. Modern PV modules commonly contain series-connected solar cells, often with a bypass diode for reverse polarization protection.

I.2.3.3. PV String: Multiple PV modules are linked in series to form a PV string, achieving the desired output voltage.

I.2.4. Types of PV Systems:

Grid-connected photovoltaic (GCPV) systems and stand-alone PV setups are the two primary categories in solar energy technology. GCPV systems, also termed grid-tied systems, are engineered to work alongside the local electricity grid, serving as a vital solution for meeting energy demands and seamlessly integrating renewable sources into existing infrastructure. Conversely, stand-alone PV systems operate independently of the grid, often found in remote or off-grid areas lacking centralized power distribution. This study focuses solely on GCPV systems due to their widespread adoption and significance in modern energy landscapes, offering benefits such as surplus energy contribution to grid stability and decreased reliance on traditional energy sources.

In the exploration of GCPV systems, Figure I.2 delineates the core components responsible for capturing solar energy and transforming it into usable electricity. These elements typically encompass PV generators, converters, cables, junction boxes, and protective devices, each playing a pivotal role in ensuring system efficiency and performance. Additionally, Figure I.3 presents an electrical schematic of a GCPV system, illustrating its configuration and interconnections, providing a comprehensive overview of electricity flow from PV modules to grid interfaces.

Despite variations in system types, PV systems share fundamental components and principles. A thorough understanding of these components and their interactions enables

researchers and engineers to devise advanced fault detection and diagnosis techniques, thereby bolstering the reliability and efficiency of GCPV systems.

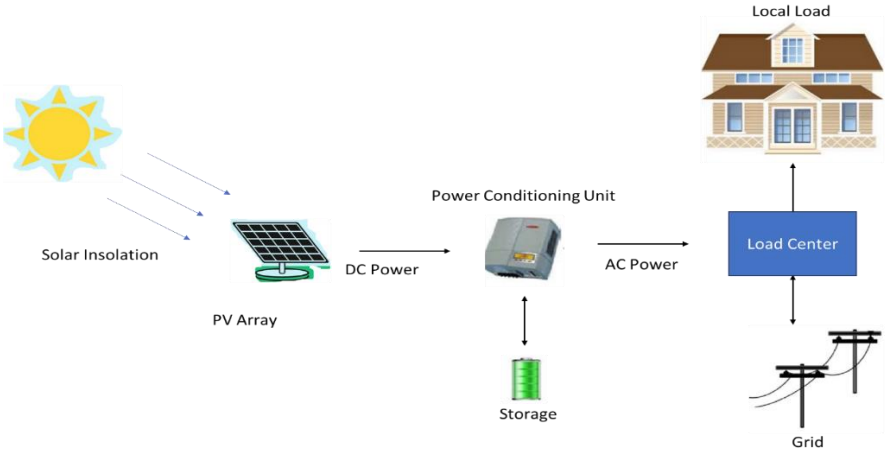

Figure I.2. Grid-connected photovoltaic systems (GCPVS).

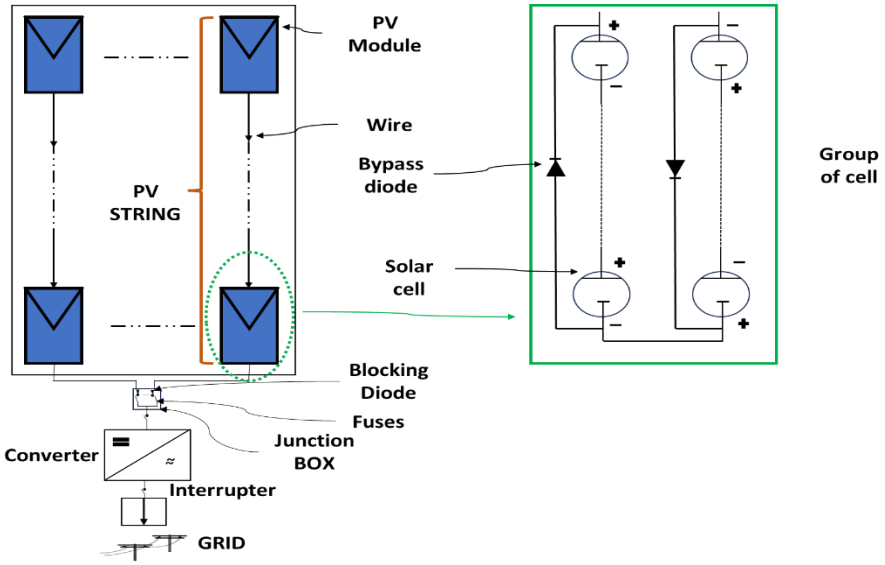

Figure I.3. Electrical schematic of a GCPV system.

I.2.4.1. PV Array: To generate the necessary power within an acceptable voltage range, PV modules are interconnected in a series-parallel arrangement to create a PV array.

I.2.4.2. Converters: PV systems predominantly utilize two types of converters: DC/DC and DC/AC converters. These converters are responsible for extracting maximum power from the PV array (DC) and converting it into alternating current (AC) for grid injection.

I.2.4.3. DC/DC Converters: When connecting a PV generator to a load, the operating point is determined by the intersection of the PV generator's current-voltage characteristic curve and the load's curve. To extract maximum power, DC/DC converters with maximum power point tracking (MPPT) algorithms are utilized to maintain the PV generator at its optimal

operating point. In Figure I.4, the operating point is defined as the intersection of the (I-V) characteristic curve of the PVG and the (I-V) characteristic curve of the load R.

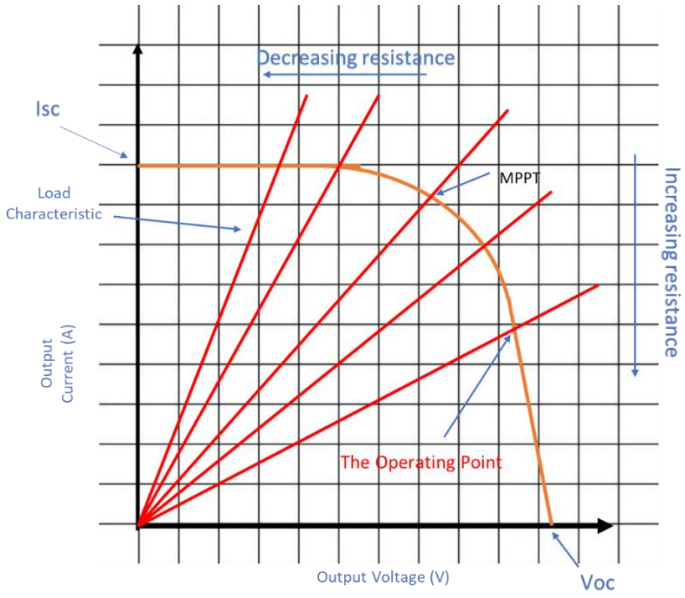

Figure I.4. Operating point variations for different values of resistive load.

I.2.4.4. DC/AC Converters: Essential components of GCPV systems, DC/AC converters (inverters) facilitate the conversion of PV generator energy from DC to AC, enabling grid injection and powering AC devices [36].

I.3. PV FAULTS:

PV arrays and cells are highly sensitive devices that must be installed in open environments to maximize exposure to solar radiation. However, being in such surroundings subjects them to significant environmental and physical stress throughout the year. This stress can lead to physical damages such as corrosion, cracks, and delamination, which in turn reduce their efficiency. PV cells rely on solar radiation to generate current; without it, they cannot produce any current. Shading on parts of the array can cause significant mismatches in the IV characteristics, leading to increased temperatures and potentially severe damage to the cells.

In addition to environmental and physical factors, electrical faults are also common in PV systems. These faults are often due to improper or loose connections of conductors or poor soldering at joints. Such faults, along with the various classes of faults occurring in PV systems, reduce efficiency and consequently lower the resultant power output. If left unaddressed, these errors can lead to power wastage and decreased system performance. Therefore, it is crucial to identify and rectify these faults to avoid dangerous situations and

ensure a high-quality output. Predictive maintenance and proactive fault recognition and rectification are essential for the efficient operation of the system. The different types of faults that can occur in a solar PV system are illustrated in Figure I.5.

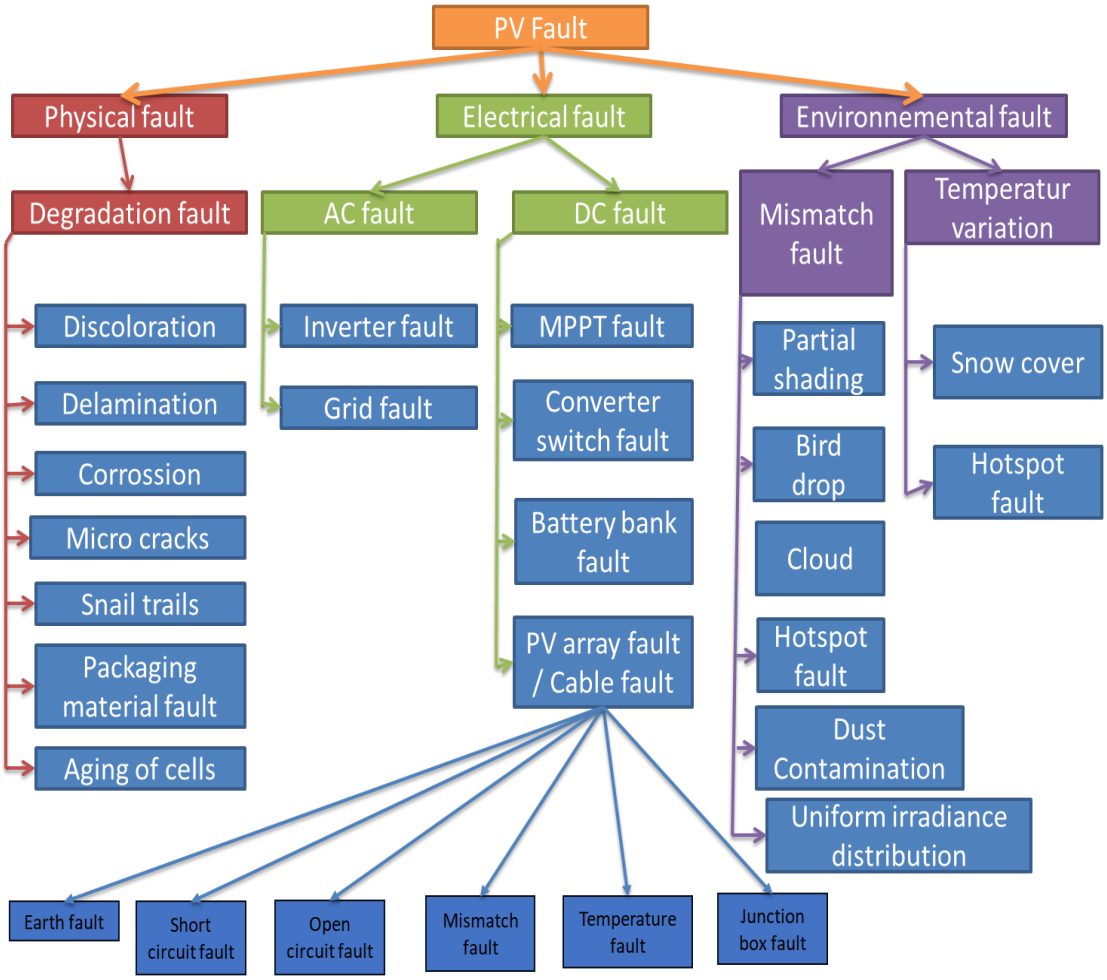

Figure I.5. Different types of faults.

I.3.1. Physical defects:

The majority of physical defects found in PV arrays, such as degradation, encapsulation issues, cracks, and corrosion, typically arise from either mechanical stress or the materials used in manufacturing. Utilizing robust and corrosion-resistant materials can extend the lifespan of PV arrays and decrease the likelihood of mechanical faults occurring.

I.3.1.1. Degradation Issue: Degradation of solar cells can lead to increased series resistance, decreased parallel resistance, and degradation of the anti-reflection coating. These factors may result in a visible brightening of the cell's color and a reduction in photogenerated current,

ultimately reducing power output by nearly half. This section will address various faults associated with degradation [39].

I.3.1.2. Encapsulation Failure: Encapsulation failure occurs due to delamination or discoloration of PV modules, often caused by salt accumulation, moisture penetration, and other environmental factors [40].

- **Discoloration:** Discoloration of PV cells, observable to the naked eye, involves a shift from white to yellow or brown, diminishing light intensity and potentially increasing temperature, affecting PV system performance. Thermal stress, UV exposure, gas or acid accumulation, and corrosion of metallic contacts are common causes of discoloration [41, 42].
- **Delamination:** Delamination, characterized by gaps or detachment between module layers, is typically due to poor adhesion. It results in increased light reflection, moisture or gas infiltration, and can lead to subsequent defects [43].
- **Corrosion:** Encapsulation delamination or cracks can allow moisture ingress, leading to corrosion of PV materials, Figure I.6 show the Corrosion in the solar PV array. Metal contacts and silver fingers are prone to corrosion when exposed to oxygen, sulfur, carbon dioxide, and other corrosive gases. Corrosion can cause leakage current and increased wire resistance, resulting in significant power loss [44].

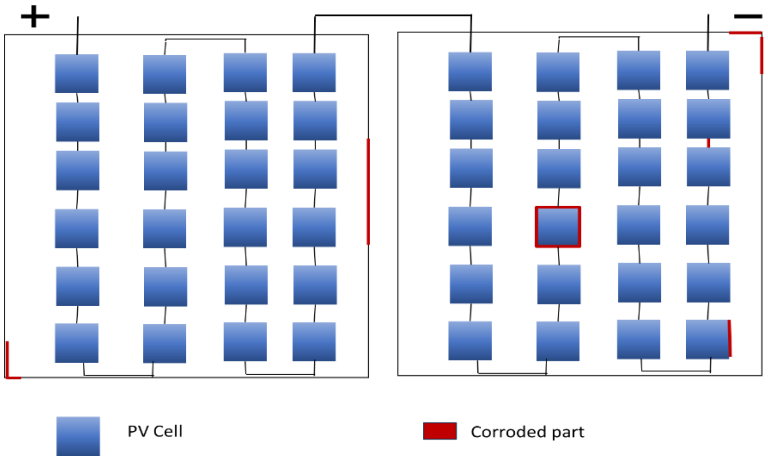

Figure I.6. Corrosion in the solar PV array.

a) **Cell Cracks**: Cracks that develop in the silicon substrate of a PV array are referred to as cell cracks, as illustrated in Figure I.7. Cracks can also happen in various layers of

the cell's lamination. Sometimes these cracks are so small (referred to as micro-cracks) that they are not visible to the naked eye and require Electroluminescence imaging for detection. The main causes of these cracks, as identified in sources [42-46], include mechanical stress during manufacturing, transportation, or installation, shocks during transportation, manufacturing defects, mishandling during packaging, cell aging, high temperatures, and adverse weather conditions such as hailstorms, wind, snow cover, and rain…etc.

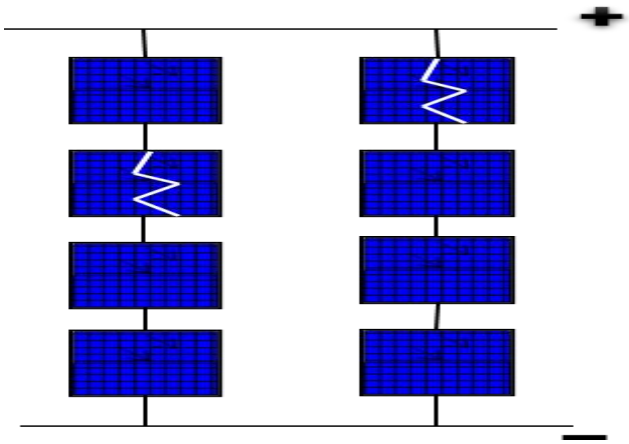

Figure I.7. Crack in solar PV panel.

b) **Silver finger discoloration:** Snail trails denote the discoloration of silver fingers caused by the formation of silver carbonate nanoparticles (Ag_2CO_3). This phenomenon occurs because of the reaction between silver, carbon dioxide, and moisture. These gases can permeate through cracks, making snail trails predominantly visible near cell cracks and edges [46].

c) Packaging pressure fault: Excessive pressure applied during the packaging of PV cells can lead to cell cracking. Any flaws occurring during manufacturing or packaging processes have the potential to diminish the power output of PV cells.

d) **Cell aging:** The typical lifespan of a PV system is approximately 30 years. As cells age, various defects such as discoloration, delamination, and cracks may develop, contributing to the formation of hotspots and ultimately reducing cell efficiency.

I.3.2. Environmental faults:

Numerous environmental factors, including weather conditions, solar radiation, shading, and temperature, influence the operation of PV arrays. These factors can inflict significant

damage on cells, some of which may or may not be reversible. The following subsections delineate the various types of environmental faults:

I.3.2.1. Electrical mismatch: Substantial alterations in electrical characteristics, such as differing IV traits among cells, can lead to discrepancies in PV modules. These faults, stemming from partial shading and temperature fluctuations, can cause permanent damage, such as cracks, hotspots, or soldering faults, resulting in a considerable reduction in overall power output [45, 46].

I.3.2.2. Partial shading: Partial shading occurs when a portion of the module is partially shaded, diminishing the system's power output. Shaded cells act as power dissipaters rather than sources, potentially increasing cell temperature and creating resistance, thereby generating hotspots that can detrimentally affect the PV array. Major contributors to partial shading include bird droppings, shading from buildings and trees, leaves, cloud cover, snow, dust contamination, and uneven irradiance distribution. Figure I.8 illustrate instances of partial shading in solar PV systems.

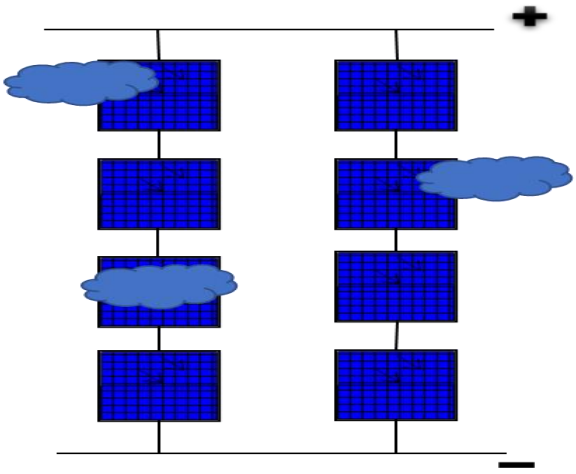

Figure I. 8. Partial shading.

- **Snow coverage:** In winter, layers of snow blanket solar cells often. Failure to promptly remove this snow can result in prolonged partial shading of the cells, potentially leading to mismatch faults. When the IV characteristics of cells mismatch, their temperature rises, facilitating hotspot formation. Additionally, the deposition of snow significantly alters the surface characteristics of PV modules [47, 48].
- **Hotspot fault:** Hotspot faults represent a form of temperature anomaly, generating localized regions of elevated temperatures within the panel compared to its overall temperature profile. These regions exhibit distinct IV characteristics from the rest of

the panel, often stemming from defective cell structures or partial shading. When parameters are not seamlessly transferred from the input to the output side, a mismatch occurs, leading to localized heating on the panel's surface. Hotspot heating occurs when the current value during operation exceeds the specified current value during a short circuit in faulty cells. The severity of hotspot faults correlates with the level and duration of mismatch. Various factors, such as aging, accumulation of dust, soil, snow, and other environmental agents on the panel's surface, contribute to the occurrence of hotspot faults [49–52].

I.3.3. Electrical Faults: Electrical faults encompass issues arising in connections between conductors, short circuits, circuit openings, and malfunctions in electrical appliances and measuring instruments.

I.3.3.1. AC Faults: AC faults occur on the side of the PV system where the alternating current circuit operates, including the grid and the inverter. The following are discussions on grid and inverter faults:

- **Grid fault:** In a grid-connected PV system, the PV array output connects in parallel with the power distribution system or grid via an inverter. During a utility grid power failure, the PV output must disconnect from the grid to prevent power flow from the PV system. Grid faults encompass issues such as PowerStation faults, loose connections, transmission line damage, blackouts, and overloading [45].
- **Inverter fault:** The DC current produced by the PV array is converted into AC by the inverter. Common causes of inverter faults include improper installation, uncontrolled voltage/current, and excessive load power [53]. Grid-tied solar inverters play a vital role in disconnecting the PV system from the grid during grid faults for safety reasons. Inverter failure to perform this action poses a risk to grid workers.

I.3.3.2. DC Faults: DC faults include various issues such as maximum power point tracker (MPPT) faults, battery bank faults, and PV array faults.

- **MPPT fault:** The MPPT maximizes power fed to the inverter from the PV array by adjusting its operation based on certain conditions. Malfunctioning charge regulators can affect MPPT functionality, reducing output power and voltage [45].

- **Battery bank fault:** Battery banks are used to ensure continuous load supply, even in the absence of solar energy. Sometimes, faults occur in these batteries due to abnormal charging conditions.

- **PV array fault:** Various types of PV array faults are discussed, including Earth fault, Line-to-Line fault, Bridging fault, Open circuit fault, Arc fault, Bypass diode fault, and Junction box fault.

Earth fault: An earth or ground fault is a type of short circuit fault occurring when a circuit creates an unintended path to the ground [45], Figure I.9 sows the short circuit between the PV module and the ground. To prevent electrical hazards, non-conductive metals are grounded. When these metals encounter current-carrying conductors, a significant current flow through them and into the ground. In the PV array, a ground fault arises due to a substantial increase in current through non-current-carrying conductors, known as mismatch faults. Ground faults can be lower or upper earth faults.

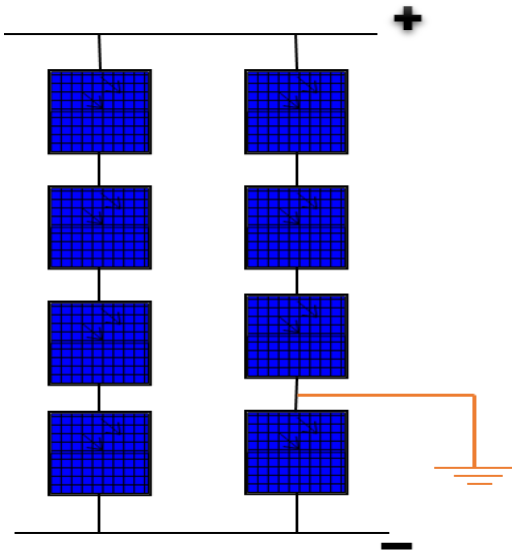

Figure I.9. Short circuit between the PV module and the ground.

Fault between PV array lines: Within a PV array, a line-to-line fault, depicted in Figure I.10, emerges from inadvertent short circuits between disparate potentials or sometimes between array cables. This fault manifests in two forms based on its location: an intra-string fault occurs within the same string, while a cross-string fault spans neighboring strings. A modification in the VI plot of a PV system ensues due to a decrease in voltage [49].

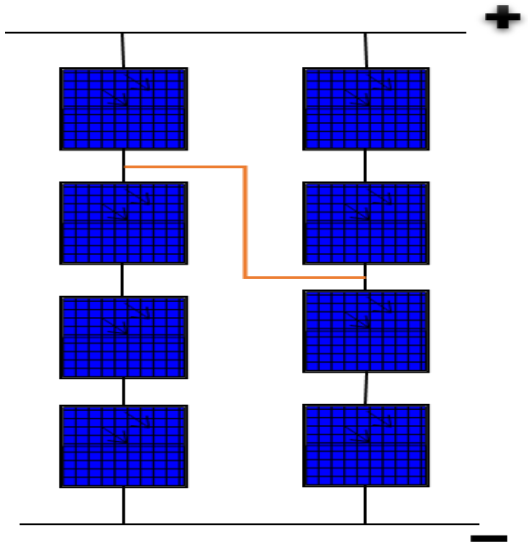

Figure I.10. Line to line fault.

Connection fault: Bridging faults materialize when low-resistance connections form between PV modules, typically stemming from physical damage, corrosion, or cable insulation failure. These faults induce voltage drops and fluctuations in current. The line-to-line fault is often labeled as a bridging fault when it exhibits zero fault impedance [45, 54].

Disruption in circuit: An open circuit fault arises when the line carrying current, series-connected with the load, is severed from the circuit. This fault commonly results from improper or loose connections among system components. Figure I.11 illustrates an open-circuit fault occurring in a PV panel. Consequently, the short-circuit current decreases, and the maximum power output diminishes with each additional disconnected string, while the voltage remains nearly normal [55, 41].

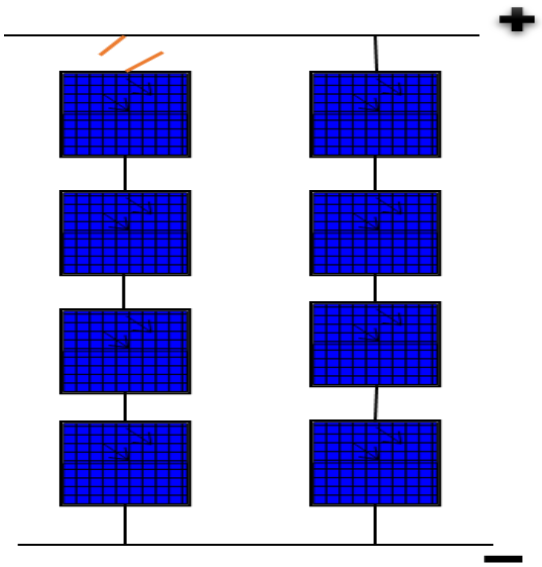

Figure I.11. Open circuit fault.

Short circuit fault: Can happen in DC side in the PV modules, BCD (Blocking diode), and BPD (bypass diode). Caused by bad connection in PV cells or from manufacturing, Figure I.1211 illustrates a short-circuit fault occurring in a PV panel [56-57].

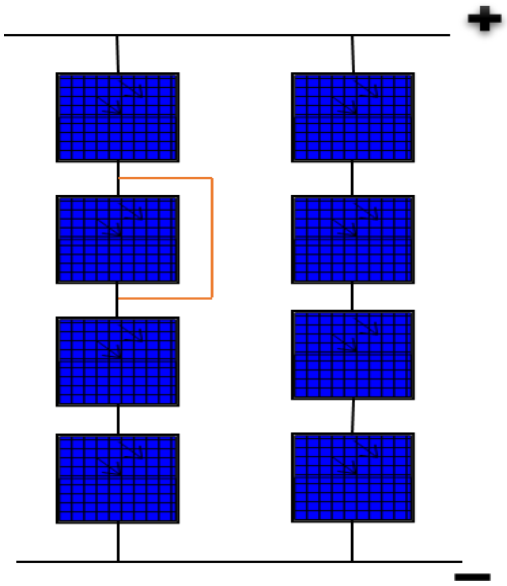

Figure I.12. Short circuit fault.

Arcing fault: Arc faults stem from intricate combinations of connecting structures within circuitry. Faulty soldered joints, loose connections between components, or insulator failures precipitate arc faults. The intense heat generated by arc faults poses combustion risks to susceptible materials. Improper handling may lead to fire accidents as ionization of air by the arc initiates plasma discharge. Arc faults are categorized as series or parallel, with parallel arcs drawing higher currents due to significant potential differences [58–60].

Malfunctioning bypass diode: Bypass diodes establish a shunt connection between select cells in a PV panel. Under normal conditions, no current flows through the diode. However, during partial shading, shaded cells act as power dissipaters, potentially raising cell temperatures. Bypass diodes prevent shaded cells from behaving like resistors by providing a parallel connection. Nonetheless, faulty bypass diodes can cause hotspots, leading to overheating and potentially fire hazards. Employing defective bypass diodes may exacerbate issues, emphasizing the importance of proper component selection [41-61].

Failure in junction box: Junction box failures often result in energy losses from the system. Burnt bypass diodes and improper connections, are primary causes of junction box faults [62].

I.4. Categories of PV fault detection and classification techniques:

The efficient operation of photovoltaic (PV) systems relies heavily on the ability to detect and classify various types of faults promptly. With the increasing deployment of solar energy technology, the need for robust fault detection and classification techniques becomes paramount. In this section, we delve into the diverse categories of methods employed for identifying and categorizing faults in PV systems. These techniques play a crucial role in ensuring the reliability, safety, and optimal performance of solar energy installations. From physical faults arising from environmental stressors to electrical anomalies within the system, a comprehensive understanding of fault detection and classification methodologies is essential for mitigating risks and maximizing the output of PV arrays. Let's explore the multifaceted approaches utilized in the detection and classification of PV faults like illustrating in Figure I.13.

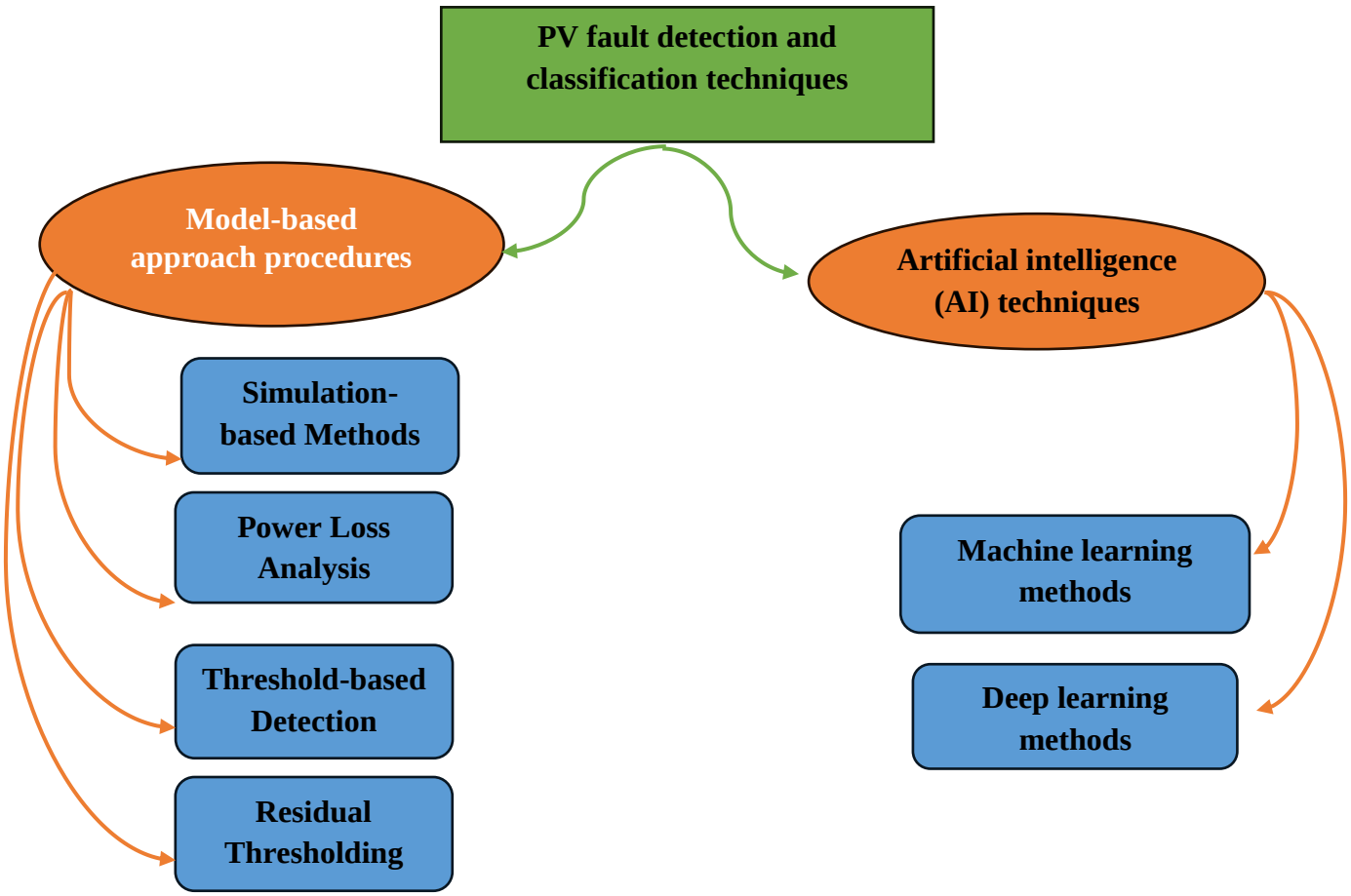

Figure I.13. Approaches utilized in the detection and classification of PV faults.

I.4.1. Model-based approach procedures: Multiple methods for detecting and diagnosing faults in PV systems have emerged over the last decade.

I.4.1.1. Simulation-based Methods: Simulation-based methods involve creating mathematical models or simulations of the PV system under normal and fault conditions. These models are used to predict the behavior of the system and compare it with the actual performance to detect faults. By simulating various fault scenarios, these methods can identify abnormalities in the system's operation [7,8].

I.4.1.2. Power Loss Analysis: Power loss analysis involves monitoring and analyzing the power output of the PV system to detect faults. This method relies on comparing the expected power output based on the system's specifications with the actual power output measured in real-time. Deviations from the expected power output can indicate the presence of faults such as shading, module degradation, or electrical faults. Chouder and Silvester proposed a fault detection strategy for PV systems based on power loss analysis, categorizing identified faults into a faulty string, faulty module, and partial shading through detailed analysis of simulated and measured output ratios [9].

I.4.1.3. Threshold-based Detection: Threshold-based detection involves setting predefined thresholds for specific parameters such as voltage, current, or temperature. When the measured values exceed or fall below these thresholds, it indicates the occurrence of a fault. Threshold-based detection is relatively simple and effective for detecting common faults such as short circuits, open circuits, and overvoltage conditions. Silvestre et al. introduced an automated fault detection procedure for grid-connected PV systems that focuses on current and voltage indicators [10]. The method sets thresholds based on typical operational behavior, triggering a fault signal when these thresholds are exceeded. Faults are identified by analyzing the ratios of current and voltage.

I.4.1.4. Residual Thresholding: Residual thresholding is a technique used in fault detection where residuals, which are the differences between measured and expected values, are analyzed. By setting thresholds on these residuals, deviations beyond certain limits can signal the presence of faults. This method is often used in conjunction with mathematical models or statistical techniques to detect and classify faults in PV systems. Residual thresholding can

provide insights into both known and unknown fault conditions. Drews et al. utilized a fault detection scheme by setting a power residual threshold based on weather satellite data for irradiance and temperature, eliminating the need for additional on-site sensors [11]. However, this approach may compromise precision due to the potentially larger margins of error in weather data. In contrast, Garoudja et al.'s method sets a threshold on the exponentially weighted moving average of voltage, current, and power residuals, incorporating historical data for fault detection rather than relying solely on the most recent observations [12].

Overall, the fault detection methods cited above are uncomplicated to implement. however, the primary challenge lies in accurately selecting appropriate thresholds to guarantee their reliability.

I.4.2. Artificial intelligence (AI) techniques :

In latest years, unlike traditional model-based fault detection methods that simulate the performance of PV installations and compare the simulated output with the actual monitored data [7–12], machine learning (ML) and deep learning (DL) methods have become increasingly popular. These techniques are seen as promising solutions for detecting and diagnosing faults in PV systems. Many studies have assessed the effectiveness of ML and DL approaches in this context. Various artificial intelligence (AI) techniques, including ML and DL, have been integrated as primary methods for PV fault detection and diagnosis because of their superior abilities in feature extraction and classification.

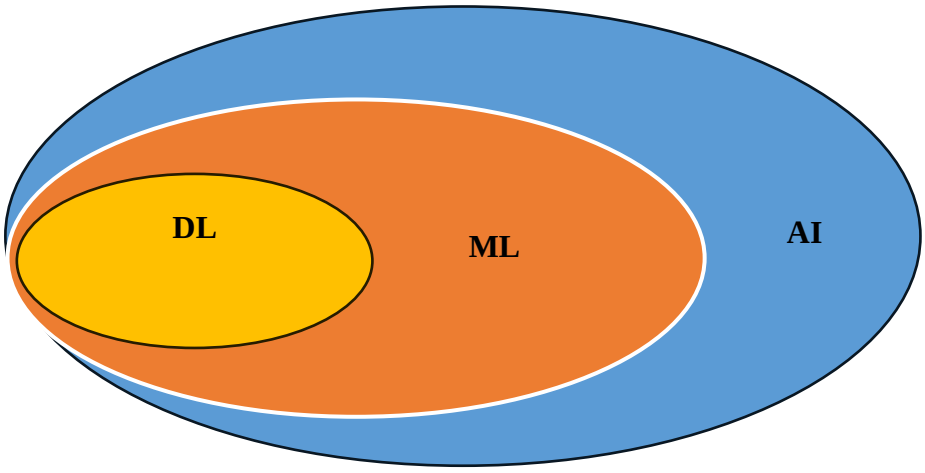

Figure I.14. Interconnected Layers of Intelligence: AI, ML, DL.

I.4.2.1. Machine learning methods:

- **Supervised Learning:** In supervised learning, predefined models are utilized, and systems learn from provided input-output pairs, where the input data and desired output are labeled. Through this approach, the machine can effectively connect the dots with sufficient data knowledge, leveraging labeled sample data and correct output to refine its understanding [64].
- **Unsupervised Learning:** In unsupervised learning, AI operates without predefined target values, necessitating the model to discern patterns within unlabeled input data [64]. Key to this method is learning and refining through trial and error. Unlike supervised learning, unlabeled data is utilized, and the correct output is not provided to the machine. Various algorithms are employed to enable the machine to deduce patterns and relationships through data study and observation. Unsupervised learning often unveils patterns or classifications that may elude human perception due to its capacity to explore data comprehensively and without bias.
- **Semi-Supervised Learning:** Semi-supervised learning amalgamates the benefits of both supervised and unsupervised learning [64]. Initially, training commences with a limited dataset to acquaint the machine with the data. Subsequently, the machine engages in studying and observing the data to augment its knowledge base through inductive reasoning. Another facet of semi-supervised learning involves transductive reasoning, which aids in refining unlabeled data using existing knowledge acquired from labeled data. Despite its advantages, semi-supervised learning remains relatively uncommon in machine learning applications.
- **Reinforcement Learning:** In this scenario, the model is endowed with autonomy to interact within a dynamic environment, receiving feedback through rewards and punishments [64]. Through positive and negative interactions, the model is taught, a departure from the methodologies of the other three learning types. Iteratively, the machine refines its outcomes with each iteration, progressively approaching high-quality output.
- **Multitask Learning Multitask:** learning facilitates the sharing of experiences among multiple algorithms, enabling them to learn concurrently instead of individually [65].
- **Ensemble Learning:** Ensemble learning combines two or more algorithms into a single algorithm, where it's noted that a collection of algorithms typically outperforms individual algorithms when performing specific tasks [66, 67].

- **Instance-Based Learning:** In instance-based learning, the algorithm acquires knowledge of specific patterns and subsequently applies it to new data [66]. As the dataset expands, this learning method evolves in sophistication [68].
- **Evolutionary Computation:** Evolutionary computation stands as a distinct branch within artificial intelligence, drawing inspiration from natural processes [69, 70]. Smart methods leveraging evolutionary algorithms are directed towards resolving diverse real-world challenges; mirroring natural processes involving living entities [70]. Rooted in random processes, data regeneration, and data replacement, evolutionary computation operates within systems such as personal computers or data centers. A spectrum of evolutionary computation approaches finds application across various domains; including image processing, cloud computing, and grid computing [70].

Neural network learning, also known as artificial neural networks (ANN), draws inspiration from the biological structure of brain cells called neurons [66]. Understanding ANN necessitates familiarity with how neurons function [66]. ANNs operate across three layers—input, hidden, and output—paralleling the functioning of brain neurons across four components: dendrites, nucleus, soma, and axon [66]. Data is received by the input layer and processed through the hidden layer before being transmitted as calculated output to the output layer [71].

Conversely, the advent of deep learning (DL) algorithms marks a significant advancement in machine learning, garnering attention for their proficiency in pattern recognition, data mining, and knowledge discovery.

I.4.2.2. Deep learning methods:

- **Convolutional Neural Network Based Fault Diagnosis:** Convolutional Neural Networks (CNNs) consist of three fundamental layers: convolutional layer (CL), pooling layer (PL), and fully connected layer (FCL) [72], [73] (Figure I.15). The CL employs convolution kernels to capture features from the input data or features from preceding layers, utilizing matrix element multiplication within a defined receptive field, and integrating deviations [74]. The dimension of the convolution kernel within the CL regulates the extraction of local spatial correlations within the input data, enhancing specific features while mitigating noise effects [72]. In contrast, the PL aims to shrink the spatial dimensions of the convolved features, reducing computational complexity through dimensionality reduction methods [75].

Additionally, it aids in capturing relevant features invariant to rotation and position, thereby facilitating effective model training [75]. Introducing an FCL serves as a means to learn nonlinear combinations of high-level features extracted by the CL. This layer is tasked with learning potentially nonlinear functions within that feature space [76]. The utilization of CNN-based Fault Detection and Diagnosis (FDD) offers several advantages:

(i) Given the diverse nature of industrial system data [77],[78], CNN can accommodate inputs such as time series [79],[80], spectrograms [81], [82], and images [83],[84], making it well-suited for processing multi-source information [77], [85].

(ii) In the presence of random strong magnetic interference and high temperatures often encountered in complex PV systems, CNN's extracted features exhibit translation invariance [86], [87], thus enhancing the robustness of the diagnostic algorithm and improving CNN's generalization capability.

(iii) Data indicative of faults in PV systems is typically obscured within vast real-time datasets. By learning the probability distribution of real data, CNN-generated countermeasure networks can produce samples, making them applicable even in scenarios with limited sample sizes [88].

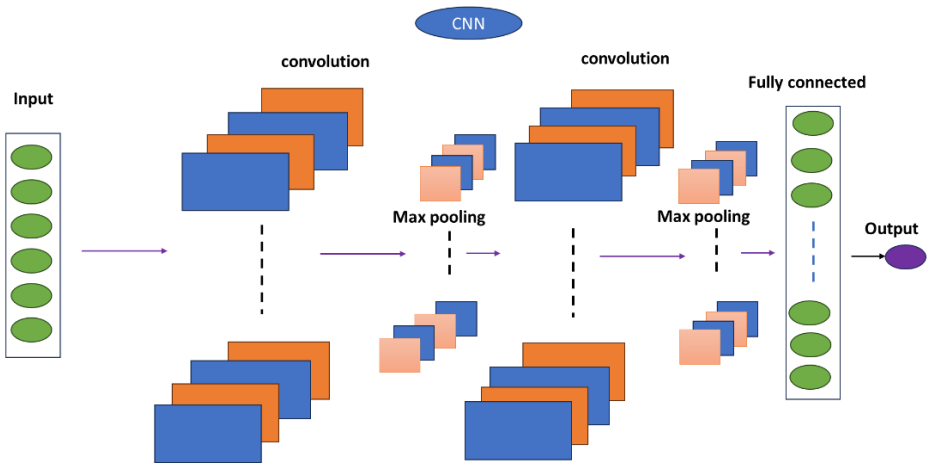

Figure I.15. Architecture of Convolutional Neural Networks (CNNs).

- **Recurrent Neural Network Based Fault Diagnosis:** Recurrent Neural Networks (RNNs) are structured networks where inputs comprise time-series data, with all nodes interconnected sequentially [89]. Unlike multi-layer perceptrons, RNNs possess temporal awareness and memory of prior network states, enabling them to learn

evolving sequences over time [90] (refer to Figure I.16). Long Short-Term Memory networks (LSTMs) and Gated Recurrent Unit (GRU) networks are currently the most widely utilized RNN variants. By incorporating gating mechanisms, each recurrent unit can dynamically capture dependencies across different time scales, mitigating issues such as gradient vanishing and exploding, albeit with inherent length constraints [91] [92]. Consequently, specialized RNN variants like LSTM [90] and GRU [92] have been developed to address long-sequence prediction challenges. The advantages of RNN-based FDD include:

(1) RNN inputs are time-series data, with network depth adapting to input sequence length, ideal for dynamic PV systems monitoring and prediction.

(2) RNNs, being Turing complete, leverage chain connections to extract and represent the dynamic nonlinear characteristics of PV systems.

(3) RNN stability is maintained even with differing lengths between learning and testing sequences—a crucial aspect in PV system control marked by variable-length sequences and irregular sampling.

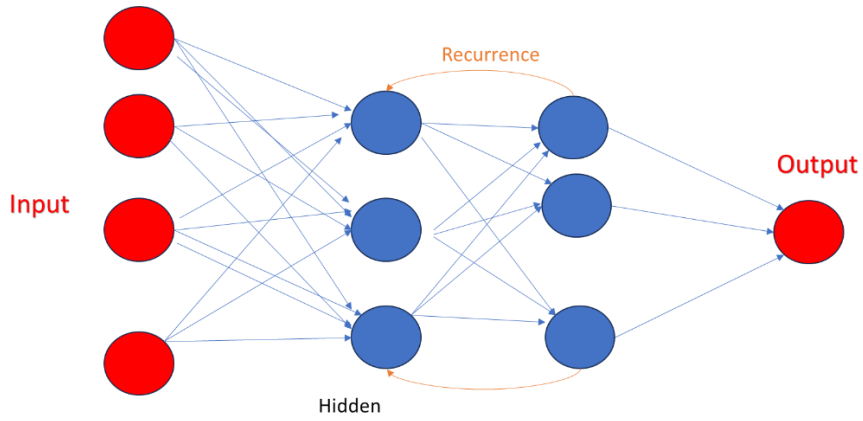

Figure I.16. Recurrent Neural Network.

- **Stacked Auto Encoder Based Fault Diagnosis:** SAE networks represent multi-hidden neural architectures achieved by stacking numerous auto-encoder networks [93], where the output of each layer feeds into the subsequent one [94]. Each auto-encoder comprises two components: an encoder and a decoder, as depicted in Figure I.17. The encoder transforms the network

input into a hidden layer representation, while the decoder reconstructs this representation back to the original input [95].

Typically, faults arise in the high-frequency components linked to the higher-order moments of stochastic processes. Viewed through the lens of Taylor expansion, a function's value near a point can be expressed as an infinite series, incorporating the function's value at that point and the derivatives of each order. Despite the small coefficient of the higher-order term, it differs from the background intricacies. In such scenarios, conventional methodologies struggle to characterize these features effectively. This characterization directly influences the performance of FDD algorithms, particularly in detecting subtle faults. As a multi-layered model, SAE networks employ multiple nonlinear mappings to compute higher-order feature representations, enabling more effective expression of a broader range of functions compared to shallow networks [96].

Utilizing SAE networks for FDD confers several advantages:

(i) The simplicity of SAE network structures aligns well with the predominantly 1D signal nature of data collected from PV systems.
(ii) Given the unlabeled characteristics often present in PV data, SAE networks serve as self-learning mechanisms conducive to unsupervised training in PV applications.
(iii) Given the complexity inherent in PV system data, the layer-by-layer training methodology of SAE networks facilitates the extraction of high-order nonlinear features from data samples while mitigating deep network dispersion.

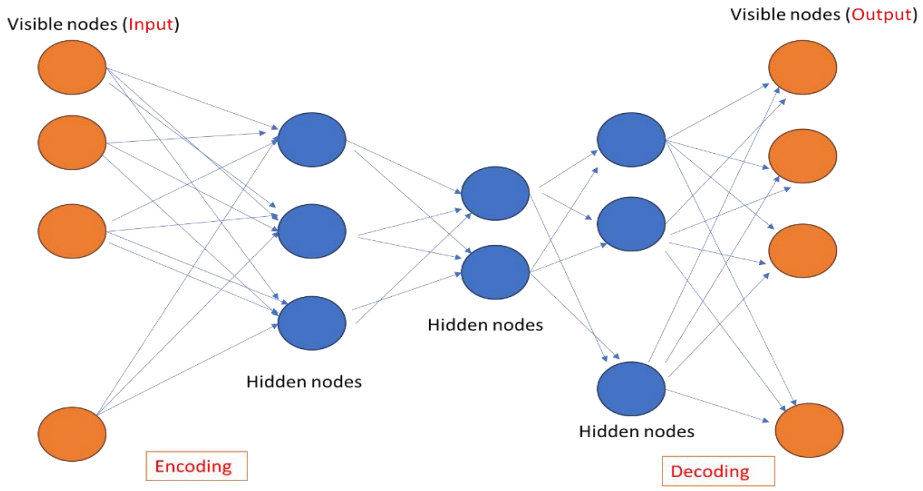

Figure I.17. Components of Autoencoder: Encoder and Decoder.

- **Deep Belief Network Based Fault Diagnosis:** A Restricted Boltzmann Machine (RBM) is a neural network comprising two layers: a visible layer and a hidden layer, which model high-order interactions between variables using an energy function. The term 'Restricted' signifies that each edge in the bipartite graph must connect one visible unit and one hidden unit [97]. RBM operates under the assumption that, given input data, the activation states of each hidden unit are independent. Conversely, when the hidden unit states are known, the activation states of the visible layer units are independent [98]. RBM can serve as a component of larger architectures like Deep Belief Networks (DBNs) and Deep Boltzmann Machines (DBMs) , Figure I.18 show the Illustration of RBM as a Sub-block in Deep Belief Networks (DBN) and Deep Boltzmann Machines (DBM). DBN, a probabilistic generative model with multiple hidden layers, employs multiple RBMs along with an output layer, typically for classification. Training occurs layer by layer, establishing the joint distribution between observation data and labels [99]. In contrast to the directed connections in DBNs, DBMs feature multiple hidden layers where information flows bidirectionally, with feedback adjustment from top to bottom [99]. Roux and Bengio theoretically demonstrated that given a sufficient number of hidden units, RBM can fit any discrete distribution [98]. FDD based on DBN/DBM offers several advantages:

(i) RBM's generative learning predicts sample probability distributions without restrictive assumptions, which is suitable for handling the random uncertainty inherent in PV systems.

(ii) RBM's unsupervised learning approach expresses data as a probability model, facilitating sample generation expansion, particularly useful in PV systems with limited data samples.

(iii) DBN's creation of activation value sets through feature grouping sequences is well-suited for simulating and controlling complex multivariable nonlinear systems, a common characteristic of unstructured PV system control.

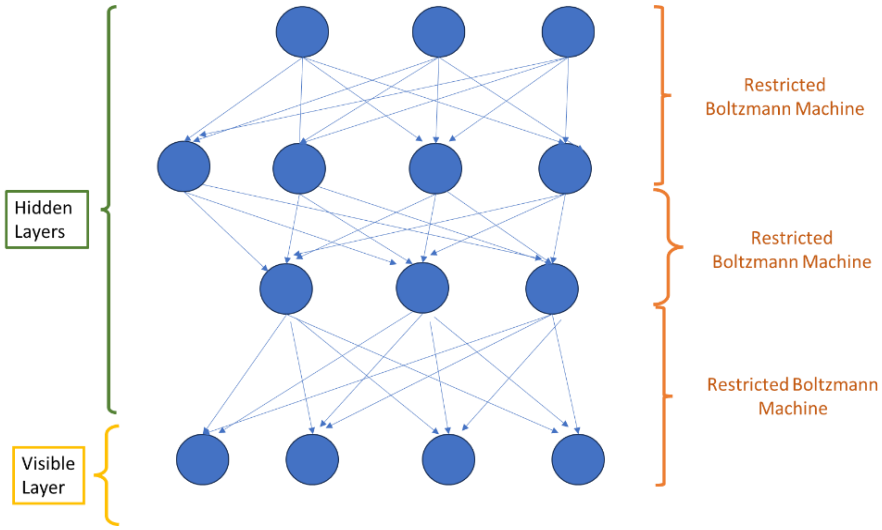

Figure I.18. Illustration of RBM as a Sub-block in Deep Belief Networks (DBN) and Deep Boltzmann Machines (DBM).

- **Deep Transfer Learning Based Fault Diagnosis:** The efficacy of Deep Learning (DL)-based Fault Detection and Diagnosis (FDD) is intricately linked to the quantity of available data. High performance necessitates a large volume of data samples from the same domain for training models [100],[101]. Consequently, DL-based FDD models with complex structures excel when trained on extensive datasets, while their accuracy and reliability diminish with fewer training samples [102]. Moreover, deep models with numerous hidden layers may impact FDD performance [103]. Ensuring that training and testing datasets possess identical feature spaces and distributions is crucial for deep models [104]. However, rebuilding statistical models from scratch due to distribution changes incurs substantial costs in PV systems [105]. Deep Transfer Learning (DTL) emerges as a promising solution to these challenges [106], leveraging existing knowledge to address problems in related fields, thereby mitigating data feature requirements [107].

DTL tools expedite training and enhance classification accuracy by leveraging data from various operating conditions, particularly in scenarios with limited target data [108],[109]. Widely adopted in FDD, DTL yields precise results in complex situations by employing transfer strategies to craft universal diagnostic models [107]. Its primary objective involves applying learned knowledge and skills from data-rich source domains to related target domains with limited data [108], [110], [111]. Figure I.19 illustrates the disparity between DL and DTL-based FDD. DL-based FDD segregates normal and faulty data into training and testing datasets, where the model is trained on the former and evaluated on the latter.

However, the smaller size of the faulty dataset relative to the normal dataset can compromise classification performance [102]. Conversely, DTL-based FDD leverages two sets of data from disparate domains: the source and target domains. Knowledge extraction occurs in the source domain, while the extracted knowledge aids FDD in the target domain. Notably, the faulty data in the source domain tends to be more abundant than in the target domain, facilitating FDD in the latter. DTL facilitates feature extraction and knowledge transfer between target and source domains, categorized into Instances-based, Feature-based, Network-based, and Adversarial-based DTL techniques [108], [112],[113]. Adversarial-based DTL, renowned for revealing a common latent space between target and source domains, has garnered significant attention in transfer learning [114],[115]. DTL has gained prominence across various applications, including image and text recognition, software defect-recognition, and FDD [116], [117].

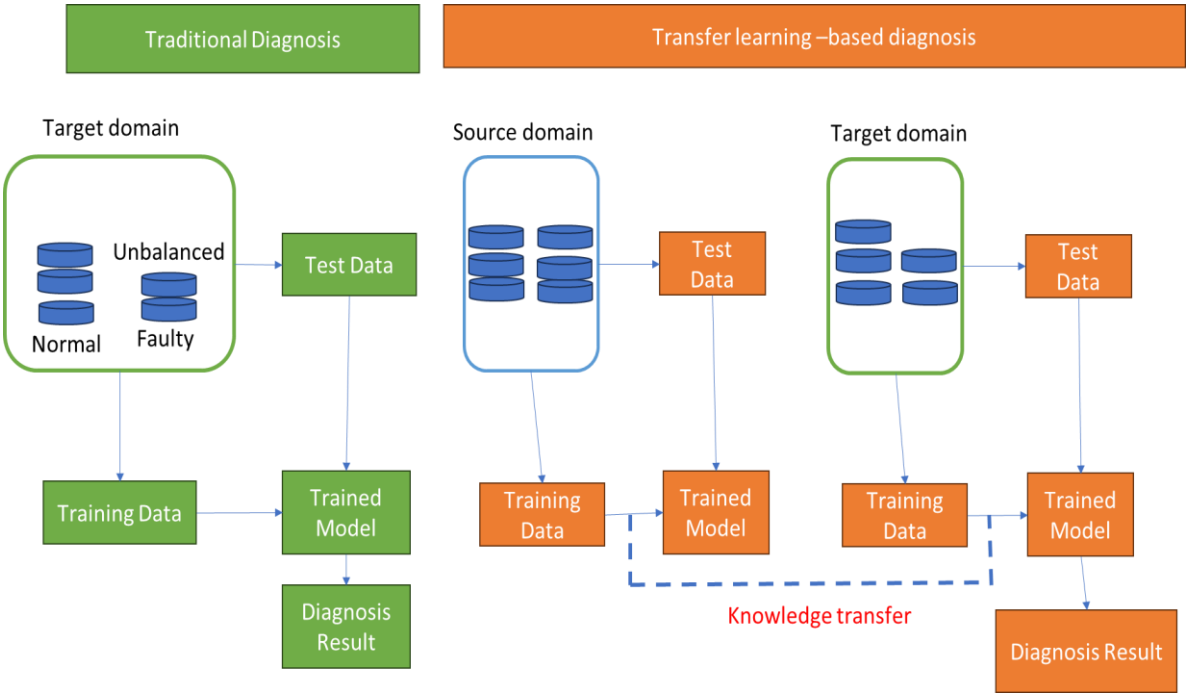

Figure I.19. Difference between DL and DTL-based FDD.

I.5. PV Power Prediction:

Before 2010, there was minimal research and development on photovoltaic generation prediction models, primarily focusing on forecasting incident radiation on solar parks, with power output calculated accordingly. These models relied on data provided by PV panel

manufacturers or empirical equations. However, over the past decade, the proliferation of PV systems globally has spurred a surge in new prediction models, accompanied by studies on solar energy characteristics. Predictive models typically leverage statistical production data and long-term meteorological data to anticipate system behavior using various methods. Interest lies in forecasting energy production in multi-source systems, assessing the power output of each component, facilitating the estimation of energy generation under diverse climatic and operational conditions [118]. Various methodologies for predicting photovoltaic energy systems exist, with some studies employing neural networks for energy generation prediction [119-121]. Today, diverse models for prediction have emerged, allowing for classifications based on different criteria [122], such as linearity or mathematical approach. These classifications delineate models into linear and nonlinear categories or into those based on Artificial Intelligence techniques versus regressive models [123]. Figure I.20 illustrates a classification of PV prediction models, emphasizing two main approaches: those based on past values and atmospheric models.

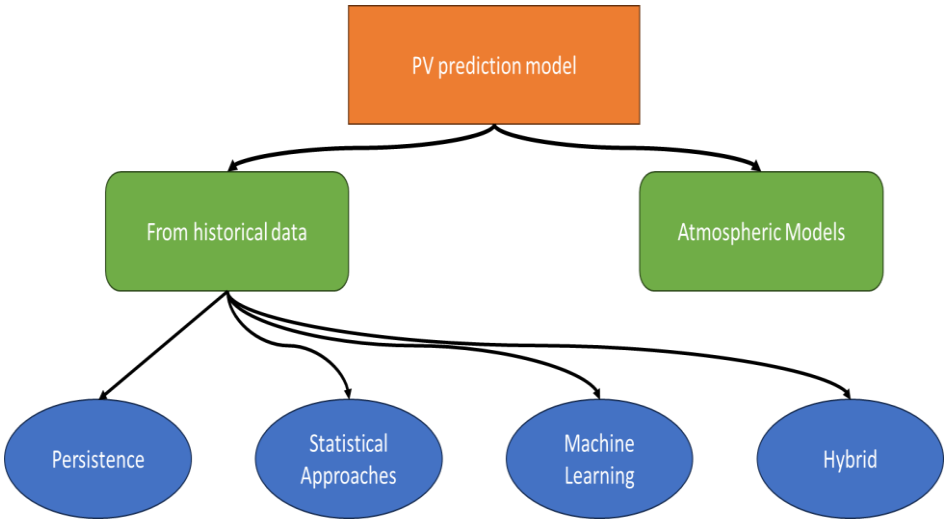

Figure I.20. Classification of PV prediction models.

I.5.1. Models Based on Past Values:

These models rely solely on historical data as input, which may consist of the variable to be predicted alone or supplemented with other variables that could affect it. These variables can include not only those specific to the time they occurred but also locally measured meteorological variables from those past moments. As illustrated in Figure 2, these models can be generally classified as outlined in the following subsections.

I.5.1.1. Persistence Models: These models predict the energy production of a PV system based solely on historical data, using the power production recorded around the same time on previous days of operation. This prediction technique is primarily employed for comparing or benchmarking the performance of other modeling approaches [123].

I.5.1.2. Statistical Approaches: In PV prediction, these methods utilize time series analysis to comprehend the behavior of observed data series or forecast future values. They are particularly effective for short-term estimation of PV power production. Below are some of the techniques commonly used in statistical approaches.

• **Regression models:** In these models, the PV power output is treated as the dependent variable, influenced by meteorological variables [124]. They typically involve mathematical modeling and the inclusion of explanatory variables.

• **Auto-regressive models:** ARMA (Auto Regressive Moving Average) and ARIMA (Auto-Regressive Integrated Moving Average) are widely used techniques for PV prediction using time series data. These methods operate under the assumption that past values of the series, known as its history, influence future values through a combination of Auto-Regressive (AR) and Moving Average (MA) components. In a pure auto-regressive process, future values depend solely on past values. In contrast, in the moving average process, future values of the series depend on independent random variables modeled as white noise [125].

I.5.1.3. Machine Learning Techniques: These models utilize Artificial Intelligence methods and often demand a substantial amount of data to provide accurate estimates of PV energy production. Below are some of the techniques commonly employed in Machine Learning approaches:

• **Artificial Neural Networks (ANN):** Artificial neural networks are mathematical models inspired by the biological nervous system. Most studies focus on Multilayer Perceptron (MLP) networks, which can approximate nonlinear relationships between input and output data. ANN-based approaches have garnered significant interest for solar power prediction [119].

• **Support Vector Machines (SVM):** SVMs are supervised learning algorithms utilized for both classification and regression tasks. They are applied in PV power estimation through time series analysis, and there is increasing interest in these methods [122].

I.5.1.4. Hybrid Models: These models integrate both physical and statistical approaches to improve accuracy in PV power estimation. One example is neuro-fuzzy systems, which combine the supervised learning capability of neural networks with the knowledge representation of fuzzy inference systems. A well-known instance of such a system is Adaptive Neuro-Fuzzy Inference Systems (ANFIS), applied in PV power estimation [126]. Other forms of hybrid models include neural networks optimized using genetic algorithms, the combination of ARMA models with neural networks, the integration of multiple types of neural networks, and the fusion of atmospheric models like MM5 for radiation prediction with fuzzy logic or neural networks for power estimation [127].

I.5.2. Atmospheric Models: Atmospheric Models: These models utilize predictions of meteorological variables obtained from numerical forecasting programs available at various meteorological institutes. These inputs may also be supplemented by data from other sources mentioned earlier. The most commonly used models in this category include MM5 (developed by Pennsylvania University and the National Center for Atmospheric Research) and WRF-NMM (developed by the National Oceanic and Atmospheric Administration/National Centers for Environmental Prediction) [128].

I.6. Conclusion:

This chapter has provided an overview of photovoltaic (PV) systems, their common faults, and the methodologies used for fault detection and diagnosis (FDD). We've explored how PV systems operate, the types of issues they may encounter, and the various approaches employed to detect and address these problems. From traditional model-based techniques to advanced artificial intelligence (AI) methods, there is a diverse toolkit available for monitoring and maintaining PV systems. These approaches range from simulating system performance to analysing data with machine learning algorithms, all aimed at ensuring optimal system efficiency and reliability. Ultimately, the insights gained from this chapter contribute to our collective understanding of PV system maintenance and lay the groundwork for continued advancements in renewable energy technology. As we strive towards a more sustainable future, the effective management of PV system faults will play a crucial role in maximizing energy production and minimizing environmental impact, and this is the aim of the following chapters.

Chapter outline:

II.1. Introduction:

Accurate fault detection procedures play a pivotal role in optimizing the performance of photovoltaic (PV) systems, ensuring their reliability and longevity. Establishing a dependable PV array model serves as the cornerstone for effective monitoring and diagnosing of PV systems. This chapter introduces a comprehensive two-step approach designed to create a reliable PV array model and implement a robust fault detection procedure utilizing Random Forest Classifiers (RFCs).

The first step of the proposed methodology involves the extraction of the five unknown parameters of ODM. This process integrates the current-voltage translation method for predicting the reference curve and employs the modified grey wolf optimization algorithm. Following parameter extraction, the second step focuses on simulating the PV array to obtain maximum power point coordinates and constructing operational databases through co-simulations in PSIM/MATLAB. Subsequently, two RFCs are developed: one for fault detection, serving as a binary classifier, and another for fault diagnosis, functioning as a multiclass classifier.

The chapter provides an overview of the research methodology, highlighting the significance of accurate PV array modelling and fault detection in enhancing the performance and reliability of PV systems.

II.2. Description of the experimental arrangement:

Data from a grid-connected PV system in Algiers, Algeria, were used to fully assess the efficacy of the proposed fault detection methodology and the new approach for figuring out the unknown parameters of the PV model. This PV system, which is located at 36°43'N latitude and 3°15'E longitude, is divided into three sub-arrays, each of which has a capacity of 3.18 kW. The overall capacity of the system is 9.54 kW. Thirty Isofoton 106-12 panels are organized into two parallel strings of fifteen modules each in series to form each sub-array. An IG30 Fronius 2.5 kW single-phase inverter is connected to these PV modules.

The PV plant's inclined and flat solar radiation levels are observed with a *Kipp&Zonen*CM11 thermoelectric pyranometer. Temperature readings of the PV panels are also taken using K-type thermocouples. Meteorological and electrical factors are consistently logged using a data acquisition device (Agilent 34970). Data, encompassing meteorological

conditions (solar irradiance *(G)*, temperature *(T)*, and PV output *(I_{mpp}, V_{mpp}, P_{mpp})* metrics at the Maximum Power Point (MPP)), were gathered at intervals of 1 minute.

The primary specifications of the chosen PV array utilized in this study are outlined in Table II.2, with additional information about the entire PV setup available in [129].

Table II.1. Main specifications of the selected PV array.

Parameter	**Description**
Module technology	Mono-crystalline (mc-Si)
PV array nominal power	3.18 kWp
Inverter type and size	IG30 Fronius single-phase, 2.5 kW
Modules per inverter	30
Modules in series (Ns)	15
Strings in parallel (Np)	2
Tilt - Azimuth	35º - 10º West

Table II.2 presents an overview of the essential electrical characteristics for the Isofoton 106-12 PV module at Standard Test Conditions (STC), which entail a temperature of 25°C and an irradiance level of 1000 W/m².

Table II.2. Electrical characteristics of the considered PV module.

Parameter	**Value**
$P_{mp}(W)$	106
$I_{SC}(A)$	6.54
$V_{OC}(V)$	21.6
$I_{mp}(A)$	6.10
$V_{mp}(V)$	17.4
$\beta V_{OC}(\%/°C)$	-0.36
$\alpha I_{SC}(\%/°C)$	0.06

II.3. Suggestion approach for photovoltaic systems modelling:

Our approach to photovoltaic (PV) modelling is founded on the commonly used one-diode model [37]. This model is widely embraced in PV module modelling across different technologies, including crystalline and thin-film configurations. Its popularity arises from its

ability to balance model complexity with predictive precision. The solar cell's current-voltage *(I-V)* behavior is articulated by the implicit and nonlinear equation provided in Eq. (I.1). For a comprehensive understanding of this model, please refer to the detailed description provided in [38].

PV module manufacturers generally not directly supply the five parameters *(I_{ph}, I_o, n, R_{sh},* and *R_s)*. Research has shown that the actual values of these parameters, when extracted, frequently differ from those calculated using nominal data provided in the datasheet under Standard Test Conditions (STC) [130]. Therefore, achieving a precise correlation between the outputs of the PV model defined by Eq. (I.1) and real-world monitored data is crucial for accurate simulation and fault detection. Consequently, the importance of employing an effective parameter identification process cannot be overstated.

II.3.1. Converting Current-Voltage Characteristics to Standard Conditions:

It is inaccurate to assume the constancy of PV cell parameters due to their significant influence of weather conditions. In addition, the mathematical expressions used in these models depend on having access to reference parameters. However, recreating the standard test conditions is difficult in typical outdoor environments. To tackle this issue, we propose an effective translation approach inspired by a method introduced in [34], initially used for examining the deterioration of modules based on amorphous silicon. This technique converts three different *I-V* curves, acquired under different temperature and irradiance settings, into a standard reference curve.

It is worth mentioning that numerous translation techniques in existing literature often require prior understanding of supplementary parameters. In contrast, our novel approach does not demand any prior knowledge concerning temperature coefficients or internal parameters. Instead, it exclusively depends on data collected from three-measured *I-V* curves (Curves 1, 2, and 3) defined as:

- **Curve 1:** $(V_1[i], I_1[i])$ where $i=1,\dots,n_1$, measured at an irradiance $G_1$and a cell temperature T_1
- **Curve 2:** $(V_2[j], I_2[j])$ where $j=1,\dots,n_2$, measured at an irradiance $G_2$and a cell temperature T_2
- **Curve 3:** $(V_3[k], I_3[k])$ where $k=1,\dots,n_3$, measured at an irradiance $G_3$and a cell temperature T_3

The proposed approach begins with the derivation of a new Curve 0, labelled as $(V_0[i], I_0[i])$, to match the desired conditions G_0 and T_0 at standard test conditions (STC). This involves introducing an intermediary curve, denoted as Curve 4, to facilitate an interpolation process under operating conditions G_4 and T_4. Initially, Curve 4 is generated from Curve 1 and Curve 2. Then, Curve 3 and Curve 4 are utilized to achieve the target Curve 0. The interpolation procedure starts within the irradiance/temperature plane, as depicted in Figure II.1, and is subsequently executed in the voltage/current domain, utilizing identical parameters as detailed below.

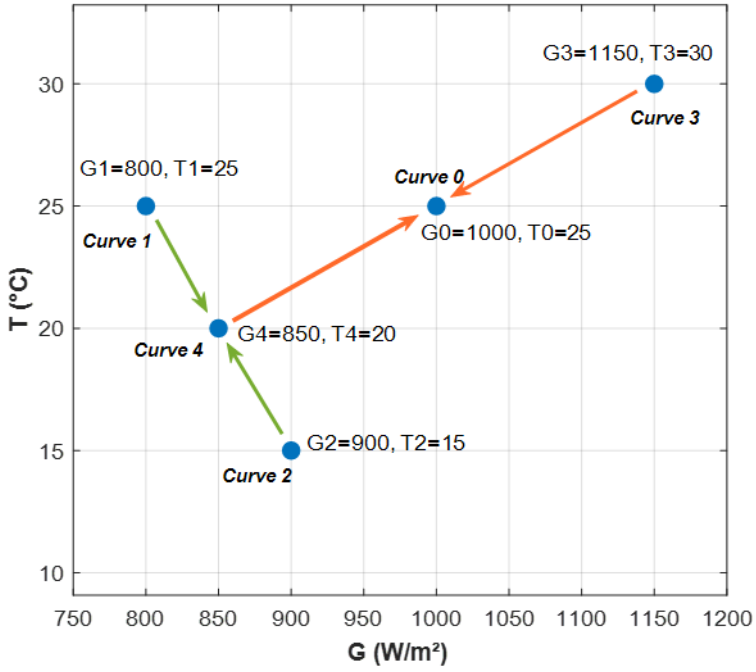

Figure II.1. The operating conditions of curves 1, 2, and 3 are interpolated to obtain the operating conditions of Curves 4 and 0.

The values of G_4 and T_4 are determined through combinations of G_1 and G_2, and T_1 and T_2, respectively, as illustrated in Eqs. (II.1) and (II.2), where the parameter α is to be ascertained. Furthermore, as shown in Eqs. (II.3) and (II.4), the desired irradiance G_0 and temperature T_0 are derived from G_3 and G_4, and T_3 and T_4, respectively, incorporating another unknown parameter $\emptyset$. This setup results in a system of four equations and four unknowns (G_4, T_4, α, and $\emptyset$).

$$G_4 = G_1 + \propto (G_2 - G_1) \quad \text{(II.1)}$$

$$T_4 = T_1 + \propto (T_2 - T_1) \quad \text{(II.2)}$$

$$G_0 = G_3 + \emptyset(G_4 - G_3) \quad \text{(II.3)}$$

$$T_0 = T_3 + \emptyset(T_4 + T_3) \quad \text{(II.4)}$$

The equations under standard conditions have been simplified by introducing a new translation parameter, labeled as ω, defined as the product of $\emptyset$ and α. Moreover, the values of G_4 and T_4 have been incorporated into Eqs. (II.3) and (II.4), leading to the creation of Eqs. (II.5) and (II.6), which can be easily calculated.

$$G_0 - G_3 = (G_1 - G_3).\emptyset + (G_2 - G_1).\omega \quad \text{(II.5)}$$

$$T_0 - T_3 = (T_1 - T_3).\emptyset + (T_2 - T_1).\omega \quad \text{(II.6)}$$

The subsequent step aims to determine the *I-V* curves. It is assumed that I_{SC1} and I_{SC2} represent the short-circuit currents of Curve 1 and Curve 2, respectively. For every point of Curve 1$(V_1[i], I_1[i])$, a corresponding partner $(V_2[j], I_2[j])$ from Curve 2 is sought, ensuring that the following condition is fulfilled: $I_2[j] - I_1[i] = I_{sc2} - I_{sc1}$. Subsequently, a new point $(V_4[i], I_4[i])$ on Curve 4 is derived using Eqs. (II.7) and (II.8). Similarly, for each point of Curve 3 $(V_3[i], I_3[i])$, the most suitable matching point $(V_4[j], I_4[j])$ on Curve 4 is selected, satisfying $I_4[j] - I_3[i] = I_{sc4} - I_{sc3}$, and the point $(V_0[i], I_0[i])$ on Curve 0 is generated based on Eqs. (II.9) and (II.10).

$$V_4[i] = V_4[i] + \propto (V_1[i] - V_2[j]) \quad \text{(II.7)}$$

$$I_4[i] = V_4[i] + \propto (I_1[i] - I_2[j]) \quad \text{(II.8)}$$

$$V_0[i] = V_3[i] + \emptyset(V_3[i] - V_4[j]) \quad \text{(II.9)}$$

$$I_0[i] = V_3[i] + \emptyset(I_3[i] - I_4[j]) \quad \text{(II.10)}$$

II.3.2. Parameter Extraction using Modified Grey Wolf Optimization:

In this section, we present a technique for offline parameter identification through optimization. We opt for this method due to the implicit nature of the characteristic equations in Eq. (I.1), posing challenges for direct parameter determination. Viewing the parameter identification process as an optimization problem, we employ the Modified Grey Wolf Optimization (MGWO) algorithm to tackle this challenge. This algorithm effectively optimizes the unknown parameters to align with the implicit characteristic equations, enabling precise extraction of desired values based on actual measurement data. Our focus lies in quantifying the disparity between the outputs derived from Eq. (I.1) and the data obtained from the current-voltage translation method discussed earlier (section 3.1). We employ the root mean square error as the primary metric for evaluating this difference. For each set of experimental values *(I, V)*, the RMSE is calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N}(f(V,I,x))^2\right)} \quad \text{(II.11)}$$

$$f(V,I,x) = I - \left(I_{ph} - I_o\left[exp\left(\frac{q(V+R_sI)}{nkT}\right) - 1\right] - \frac{V+R_sI}{R_{sh}}\right) \quad \text{(II.12)}$$

In this context, x represents a vector comprising

$x = \left[I_{ph,ref}, I_{o,ref}, R_{sh,ref}, R_{s,ref}, n_{ref}\right]$ while *N* indicates the quantity of data points.

II.3.3. MGWO method:

In 2014, ***Mirjalili et al.*** presented the GWO algorithm that follows the mathematical social behavior model [35].

$$\vec{D} = \left|\vec{C}.\overrightarrow{X_p}(t) - \vec{X}(t)\right| \quad \text{(II.13)}$$

$$\vec{X}(t+1) = \overrightarrow{X_p}(t) - \vec{A}.(\vec{D}) \quad \text{(II.14)}$$

The current iteration is denoted by *t*, where $\overrightarrow{X_p}$represents the position vector of the prey, $\vec{X}$ denotes the position vector of the alpha wolf, and $\vec{A}$and $\vec{C}$ are coefficient vectors calculated as follows:

$$\begin{cases} \vec{A} = 2\vec{a}.\overrightarrow{r_1} - \vec{a} \\ \vec{C} = 2.\overrightarrow{r_2} \end{cases} \quad \text{(II.15)}$$

The elements of vector $\vec{a}$ gradually decrease from 2 to 0 throughout the iterations, while $(\overrightarrow{r_1}, \overrightarrow{r_2})$ represent random numbers within the range of [0,1]. The formula for updating the position is depicted below:

$$\begin{cases} \overrightarrow{D_\alpha} = \left|\overrightarrow{C_1}.\overrightarrow{X_\alpha} - \vec{X}\right| \\ \overrightarrow{D_\beta} = \left|\overrightarrow{C_2}.\overrightarrow{X_\beta} - \vec{X}\right| \\ \overrightarrow{D_\delta} = \left|\overrightarrow{C_3}.\overrightarrow{X_\delta} - \vec{X}\right| \end{cases} \quad \text{(II.16)}$$

$$\begin{cases} \overrightarrow{X_1} = \overrightarrow{X_\alpha} - \overrightarrow{A_1}.(\overrightarrow{D_\alpha}) \\ \overrightarrow{X_2} = \overrightarrow{X_\beta} - \overrightarrow{A_2}.(\overrightarrow{D_\beta}) \\ \overrightarrow{X_3} = \overrightarrow{X_\delta} - \overrightarrow{A_3}.(\overrightarrow{D_\delta}) \end{cases} \quad \text{(II.17)}$$

The iteration process's top three solutions, $\overrightarrow{X_1}, \overrightarrow{X_2},$ and$\overrightarrow{X_3}$, are the positions that each pack wolf updates their position to.

$$\vec{X}(t+1) = \frac{\overrightarrow{X_1}+\overrightarrow{X_2}+\overrightarrow{X_3}}{3} \quad \text{(II.18)}$$

This chapter introduces a flexible strategy that utilizes the grey Wolf Optimizer (GWO) calculation, including a slight adjustment within the determination stage. Outlined in Figure II.2, the chart depicts the steps of the proposed modified GWO (MGWO) strategy, which closely takes after an approach utilized in a past consider [131]. This strategy recognizes *alpha*, *beta*, and *delta* individuals by surveying the wellness work for person positions, particularly focusing on five obscure parameters. Other operators subsequently adjust their positions in a similar fashion.

An inventive methodology for position upgrading is implanted inside GWO, reinforcing both investigation and abuse capabilities while guaranteeing quick merging. This novel concept draws inspiration from the competitive exclusion method found in genetic algorithms [132]. In this approach, only positions from the current iteration of search agents (wolves) that illustrating predominant fitness compared to positions from previous iterations are replaced. Only the top positions are considered during the final phase for selecting new alpha, beta, and delta members. The process iterates to update search agent positions based on these selections, repeating as necessary to reach the maximum number of iterations [133]. The MGWO, supplemented with an extra stage, can seek after ideal results without depending on parameters like conventional methods would.

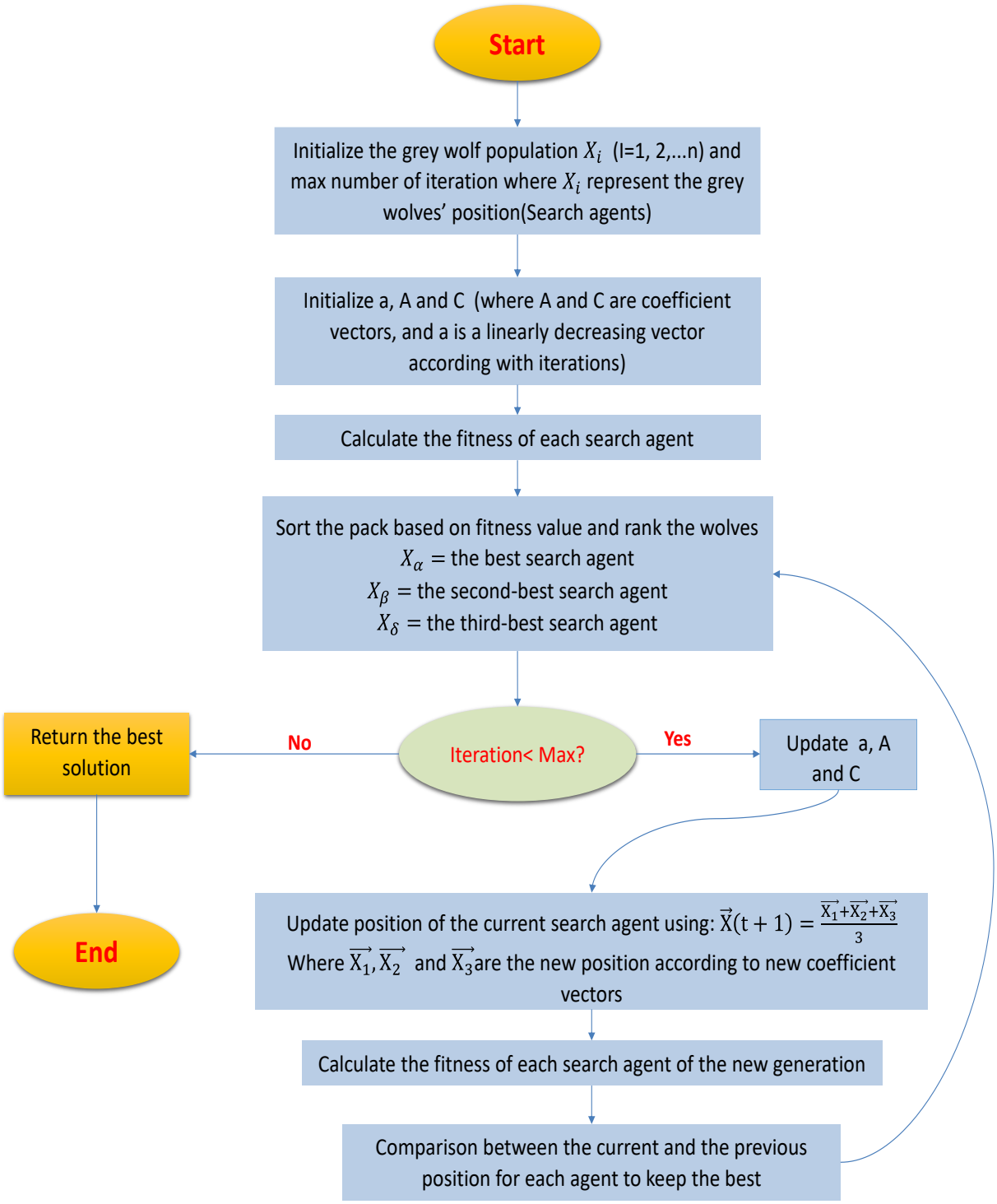

Figure II.2. MGWO algorithm flowchart.

II.3.4. Prediction the photovoltaic (PV) outputs in real outdoor settings:

The subsequent essential stage involves determining the values of the unknown parameters within real operating condition using fully analytical formulas and reference parameters derived from the MGWO algorithm. Equations (II.19) to (II.24) sum up the analytical

formulations that facilitate the computation of the five parameters under consideration as functions of temperature and irradiance [134–136].

$$n(T) = n_{ref}\left({}^{T}/_{T_{ref}}\right) \tag{II.19}$$

$$I_{ph}(G,T) = \frac{G}{G_{ref}}\left[I_{ph,ref} + \alpha(T - T_{ref})\right] \tag{II.20}$$

$$R_{sh}(G) = R_{sh,ref}\left({}^{G_{ref}}/_{G}\right) \tag{II.21}$$

$$E_g = E_{g,ref}\left[1 - 0.0002677(T - T_{ref})\right] \tag{II.22}$$

$$R_s(G,T) = R_{s,ref}\left(\frac{T}{T_{ref}}\right)\left[1 - \beta ln\left({}^{G}/_{G_{ref}}\right)\right] \tag{II.23}$$

$$I_o(G,T) = I_{o,ref}\left(\frac{T}{T_{ref}}\right)^3 e^{\left(\frac{q}{nK_B}\left(\frac{E_{g,ref}}{T_{ref}} - \frac{E_g}{T}\right)\right)} \tag{II.24}$$

Eg represents the band gap energy of the semiconductor, while $E_{g,ref}$ denotes the band gap energy under reference conditions. I_{ph}, I_o, n, R_s, and R_{sh} stand for the five parameters under actual operating conditions. Conversely, $Iph_{,ref}$, $I_{o,ref}$, n_{ref}, $R_{s,ref}$, and $R_{sh,ref}$ refer to the five unknown parameters at the reference conditions determined through the application of the extraction algorithm.

II.4. Faults detection and diagnosis strategy:

Operating a PV system during specific types of malfunctions can result in significant insecurity, severe damages, and safety hazards. This study aims to establish a robust and dependable procedure for detecting faults by employing Random Forest Classifiers to identify anomalies within a PV system and identify their origins. Achieving this goal necessitates the creation of a high-quality database that clearly defines the characteristics of each fault class. Therefore, employing a reliable simulation model that accurately represents the behavior of a PV system under both normal and faulty conditions is essential for addressing this scenario effectively. Figure II.3 offers a detailed flowchart illustrating the steps involved in developing the proposed strategy.

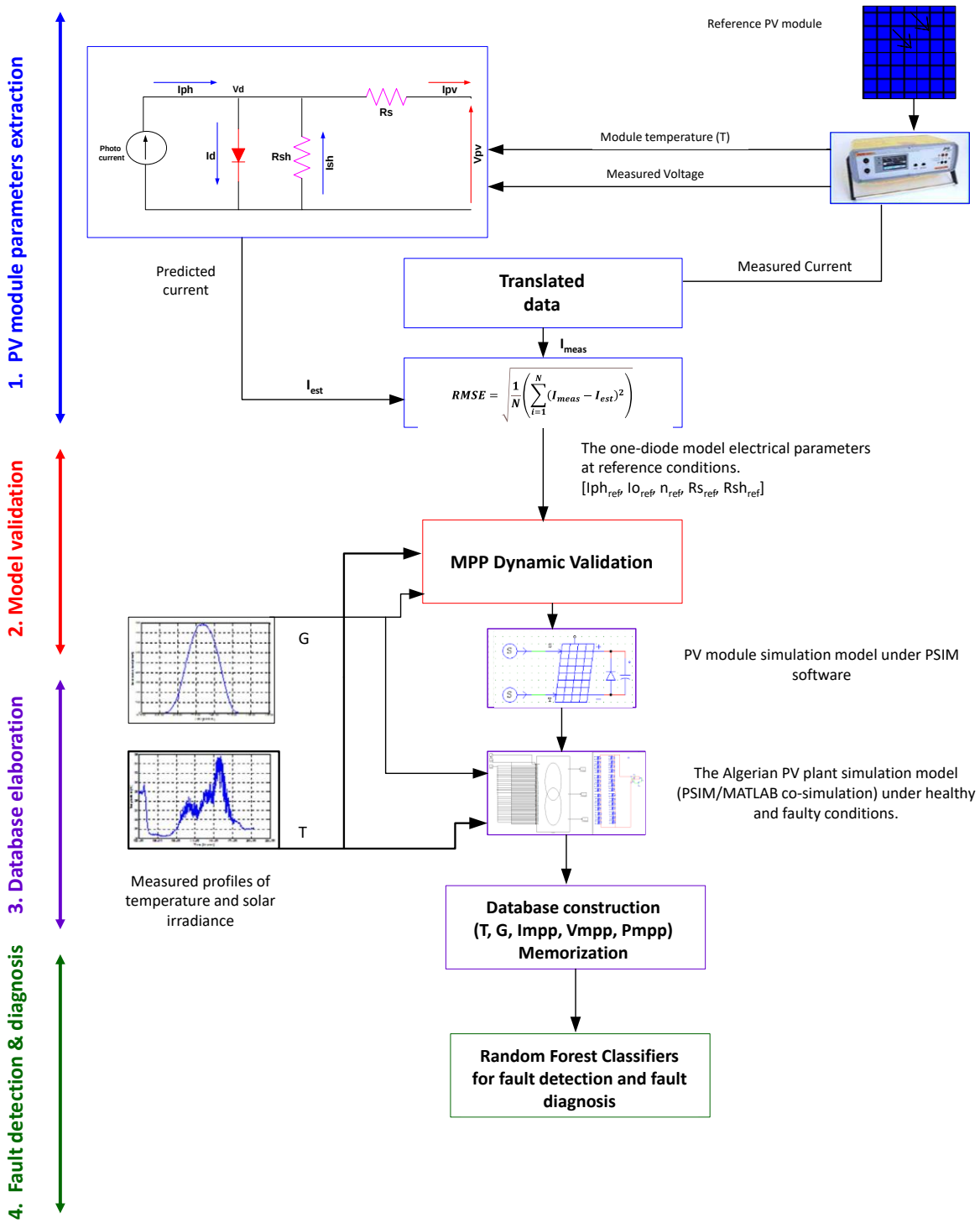

Figure II.3. Flowchart of the proposed fault detection and diagnosis strategy.

The validated PV system model, detailed in the preceding section, forms the basis for generating databases that capture the performance of the PV system under real outdoor conditions. Utilizing daily solar irradiance and module temperature profiles, this PV model is employed to create datasets comprising both optimal operation and intentionally simulated faults. The physical model of the grid-connected PV system is implemented in the PSIM™

software platform. Subsequently, the values of the unknown parameters obtained under reference conditions are integrated into the physical PV array model.

In this study, various simulated scenarios—representing common issues in grid-connected PV systems—are outlined below and depicted in Figure II.4:

a) Healthy system: Reflects normal operation without any anomalies.
b) Three short-circuited modules: Represents a scenario where one string in the PV system has fewer PV panels in operation.
c) Open circuit faults: Simulates a scenario where one string within the PV system becomes non-functional.
d) Line-to-line fault: Represents a short-circuit between two PV strings.
e) Three PV modules shaded: Replicates the effects of partial shading experienced by PV systems due to factors such as cloud movement or nearby objects for a specific duration.

The resulting databases comprise five main attributes—Irradiance, Temperature, and the output Current, Voltage, and Power at Maximum Power Point—extracted from each simulated operational scenario. Figure II.5 illustrates the simulated faults and their impact on the output power of the grid-connected PV system, based on weather data for a typical clear-sky day.

The final stage of the proposed fault detection strategy involves deploying two Random Forest Classifiers (RFCs). The first RFC is dedicated to identifying anomalies within the PV system, while the second RFC is responsible for diagnosing the specific faults detected.

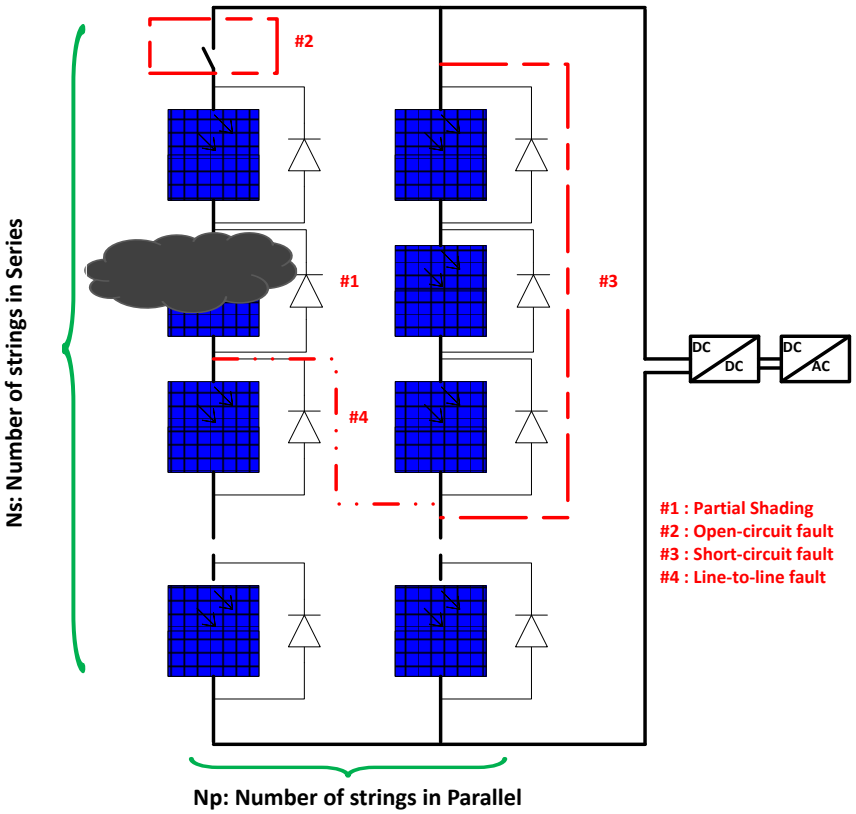

Figure II.4. Failure types considered in the proposed methodology (#1 partial shading, open-circuit fault# 2, #3 short-circuit fault and #4 Line-to-Line fault).

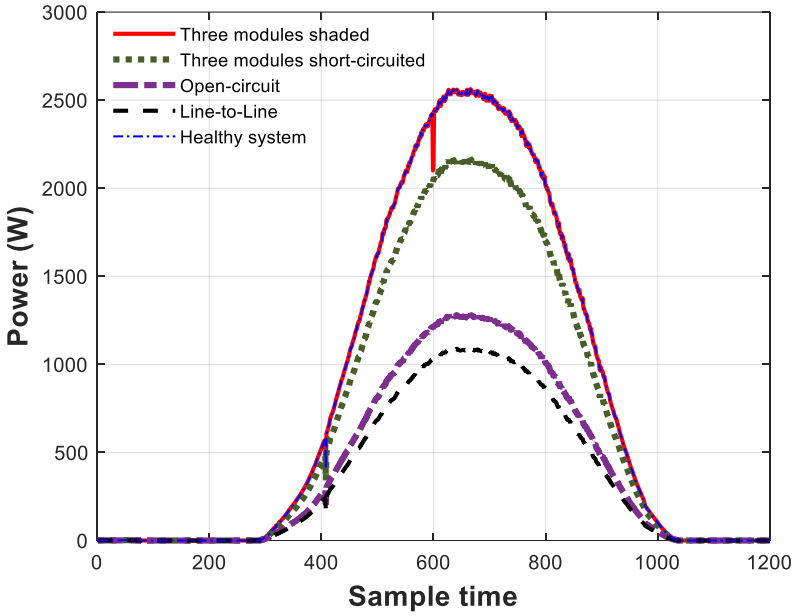

Figure II.5. DC output power of the grid-connected PV system within various fault scenarios.

II.4.1. Implementation of the Random Forest Classifier:

The Random Forest (RF) algorithm stands as a widely used supervised machine-learning approach for both tasks involving classification and regression. Its methodology revolves around ensemble learning, where multiple decision trees are combined across various subsets of input data boost predictive accuracy. The efficacy and problem-solving prowess of Random Forests escalate with the inclusion of more trees, marking it as a cornerstone in machine learning.

In this study, the RF model employed harnesses the advantages of the decision tree algorithm, particularly in terms of speed and accuracy. The model's framework encompasses two pivotal steps: firstly, selecting a sampling method to generate a data subset, and secondly, constructing a decision tree, as described in Figure II.6. Notably, four crucial hyperparameters must be taken into account, comprising the minimum number of samples for leaf nodes, the minimum number of samples for internal node splitting, the maximum number of selections, and the maximum depth of the decision tree. It's important to note that a variety of hyperparameters, including internal node splitting, minimum sample sizes for leaf nodes, and splitting criteria, affect the RFC's performance. This study explores how to best combine these factors for improved outcomes [137].

Following the establishment of the RF model, test set samples are input into the model. Each individual decision tree evaluates the classification outcomes for each sample. Upon completion of this assessment, the class receiving the most votes from all decision trees is designated as the classification for the sample[138]. This is realized through a voting mechanism that consolidates the results of all decision trees. A grid search optimization technique is employed to further refine the RF algorithm's parameters, as depicted in Figure II.7. This methodology aids in pinpointing the most influential parameter combinations for the RF model, thereby enhancing its classification performance. Initially, the Cartesian product is applied to the value set of each hyperparameter to generate the hyperparameter configuration space, encompassing all potential hyperparameter combinations. Subsequently, the grid search algorithm trains a model for each hyperparameter combination in the configuration space. The experiment yielding the best validation set error is then identified as having discovered the optimal hyperparameters [139].

This study employs two RFCs with distinct roles: the first classifier detects signs of faults within the PV system, while the second identifies the specific fault type. Outputs from the

diagnostic model classify various fault scenarios (fault #1, fault #2, fault #3, fault #4) as shown in Figure II.4. Both models share a framework comprising five inputs (T, G, Impp, Vmpp, Pmpp), a data processing module, and employ the Random Forest/Grid search optimization method. However, the detection model differs by offering two outputs (healthy state, faulty state) compared to the diagnosis model. This configuration enables efficient detection and comprehensive diagnosis of faults in the PV system, providing valuable insights into specific fault conditions.

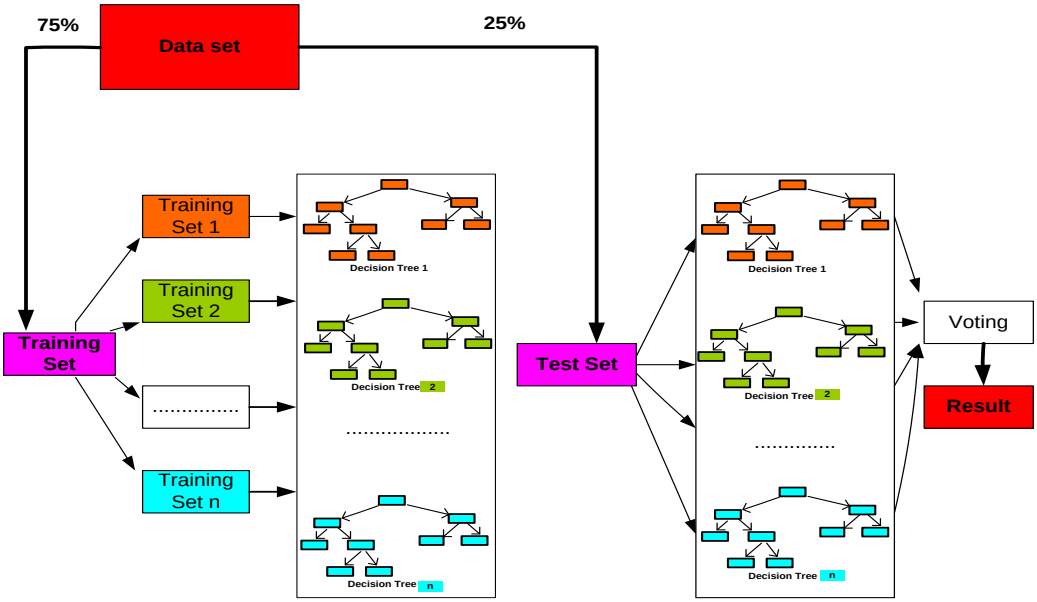

Figure II.6. General structure of the deployed RF model.

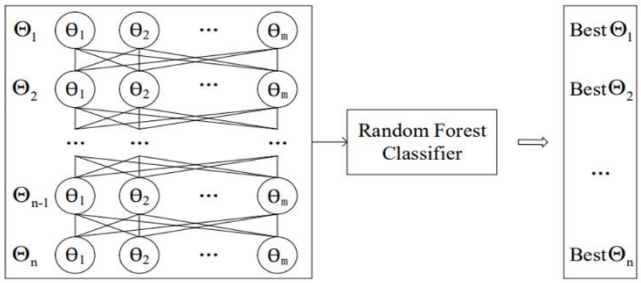

Figure II.7. Grid search algorithm's principle.

II.4.2. Preparing data for the learning and testing stages:

Ensuring the preprocessing of raw data is essential to bolster problem-solving capabilities and achieve higher accuracy. Within this scope, the 'sklearn' library offers a range of functions for managing missing values, enabling effective identification and resolution using the 'isnull' function. To reveal underlying relationships within the data, Pearson's correlation coefficient

is applied. This metric provides values spanning from -1 (reflecting a perfect negative correlation) to +1 (indicating a perfect positive correlation), quantifying the magnitude of linear relationships. Notably, this measure stands apart from correlations among variables [140]. We adopted a normalization procedure grounded in a calibration method to facilitate meaningful comparative analysis of information across attributes in the dataset. This approach centers values around the mean and employs a unit standard deviation. Furthermore, to provide context for the recorded data, appropriate class labels were assigned, streamlining the creation of well-defined data samples. The defined classes alongside their corresponding fault types are detailed in Table II.3.

Table II.3. Defined classes and their corresponding fault type.

Phase	**Class**	**Corresponding fault type**
Detection	0	Healthy
	1	Faulty
Diagnosis	0	#2: Open-circuit fault
	1	#1: Partial Shading
	3	#3: Short-circuit fault
	9	#4: Line-to-line fault

To train and evaluate the two RFCs, Seventy-five percent of the entire data samples were randomly selected for training purposes. Subsequently, the remaining 25% of the data samples were reserved as an independent set of unknown data to evaluate performance in each scenario. Following preprocessing of the original data, the dataset yielded 242,890 data samples designated for detection and 194,400 data samples for diagnosis. As clarified in the preceding section, the classifiers in question were supplied with both training and testing datasets, comprising five key attributes *(T, G, Impp, Vmpp, and Pmpp)*. These attributes serve as input features, with the resulting outputs corresponding to the estimated class labels for each data point. Further details regarding the construction of the detection and diagnosis databases are provided in Table II.4.

Table II.4. Details of the detection and diagnosis database construction.

Phase	Class	Test data set (25%)	Train data set (75%)	Total
Detection	0	12145	36433	242890
	1	48578	145734	
Diagnosis	0	12145	36433	194312
	1	12144	36434	
	3	12145	36433	
	9	12144	36434	

An assessment of classifier performance was undertaken, using the confusion matrix as a key tool for evaluating effectiveness. This matrix offers insights into the accuracy of classifier predictions and identifies areas where errors occurred. Within the matrix, rows signify the actual labels, while columns represent the predicted labels. Diagonal values indicate the frequency of correct predictions, show casing alignment between predicted and actual labels. Values in other cells indicate instances where the classifier assigned incorrect labels to observations, with columns indicating predicted labels and rows representing actual correct labels.

To comprehensively evaluate our proposed system, we utilize metrics such as accuracy, *Precision*, *Recall*, and $F1_{score}$, expressed in the following equations [141]. These metrics are crucial for assessing overall performance and reliability of our models.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(II.24)}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{(II.25)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{(II.26)}$$

$$F1_{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{(II. 27)}$$

In this context, TP (True Positives) represents the samples correctly identified as belonging to class "x." FN (False Negatives) indicates samples that were incorrectly classified into another class but should have been in class "x." TN (True Negatives) denotes samples correctly identified as not belonging to class "x." These were correctly classified differently based on the classifier's decision. Lastly, FP (False Positives) refers to samples incorrectly

labeled as belonging to class "x," despite not meeting the classifier's criteria for this classification.

To improve interpretability in multi-class classification scenarios, we employ averaging techniques to compute macro and weighted averages for Precision, Recall, and F1 score. Macro average (Macro avg) is calculated by taking the unweighted mean, potentially penalizing the model when performance in minority classes is low. In contrast, weighted average (weighted avg) considers the number of true instances in each class, thereby addressing class imbalances and giving more weight to the majority class.

II.5. Results and discussion:

This section emphasizes the validation process for the newly developed PV array modeling approach. Subsequently, we assess the effectiveness of the proposed automatic fault detection system under various weather conditions, covering different patterns of faulty operation in PV arrays. Finally, we evaluate the performance of the fault detection method using the Random Forest Classifier (RFC) by comparing it with other established machine learning techniques. It is noteworthy that these validations rely on data collected from the monitored PV system described earlier.

II.5.1. Approach validation for PV modelling and parameter estimation:

The newly developed method for translating Current-Voltage to Standard Test Conditions (STC) has been validated using three measured curves labeled as Curve 1, Curve 2, and Curve 3, as depicted in Figure II.8. The process involves an intermediate step where Curve 4 is derived based on the operating conditions (T and G) of Curve 1 and Curve 2. Following this, the target curve, representing the operation of the PV array at STC (referred to as Curve 0), is predicted from Curve 4 and Curve 3.

After deriving the reference I-V curve for the PV array, the unknown parameters of the one-diode model were identified using a parameter extraction method that utilizes the Modified Grey Wolf Optimization. This optimization technique has shown remarkable accuracy, achieving an RMSE value of 0.0122, and the specific parameters obtained are outlined in Table II.5.

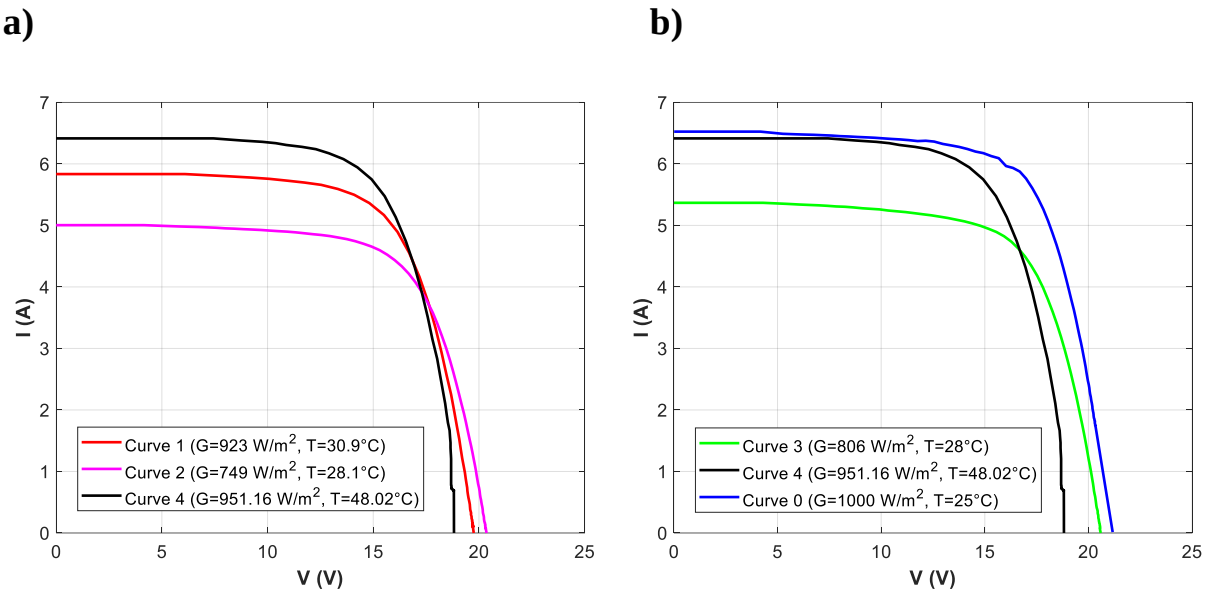

Figure II.8. Predicted I-V curve at STC (Curve 0) using the current-voltage translation method: *(a)* Derivation of Curve 4 from Curves 1 and 2, *(b)* Derivation of Curve 0 from Curves 3 and 4.

Table II.5. Extracted ODM parameters at STC.

Parameter	**Value**
R_p *(Ω)*	42.9633
R_S*(Ω)*	0.2212
I_0 *(A)*	$4.344\ 10^{-7}$
n	45.1606
I_{ph} *(A)*	6.8378
RMSE	0.0122

The methodology proposed for PV array modeling has undergone extensive validation. Extracted parameters were utilized to simulate the PV array under varying irradiance (G) and temperature (T) conditions, as detailed in equations (II.19-II.24). A comparison was made between experimental I-V and P-V curves and simulated data to evaluate model accuracy under steady conditions. The results, depicted in Figure II.9, show strong agreement between measurements and simulated values. This observation is supported by RMSE values of 0.0266 and 0.1024, respectively.

Dynamic validation of the PV array model was conducted using a cosimulation model that integrates MATLAB and PSIM environments. This validation incorporated daily temperature

and irradiation profiles, alongside measured maximum power point (MPP) output profiles from an actual PV system in Algiers, under three distinct weather conditions: a) clear sky, b) semi-cloudy, and c) cloudy day. Figure II.10 illustrates the temporal evolution of the simulated PV array output current. The results indicate significant agreement between measured and estimated values of the MPP current, with RMSE values of 0.1416, 0.216, and 0.2971, respectively. This underscores the efficacy of the identification process and the resilience of the proposed approach.

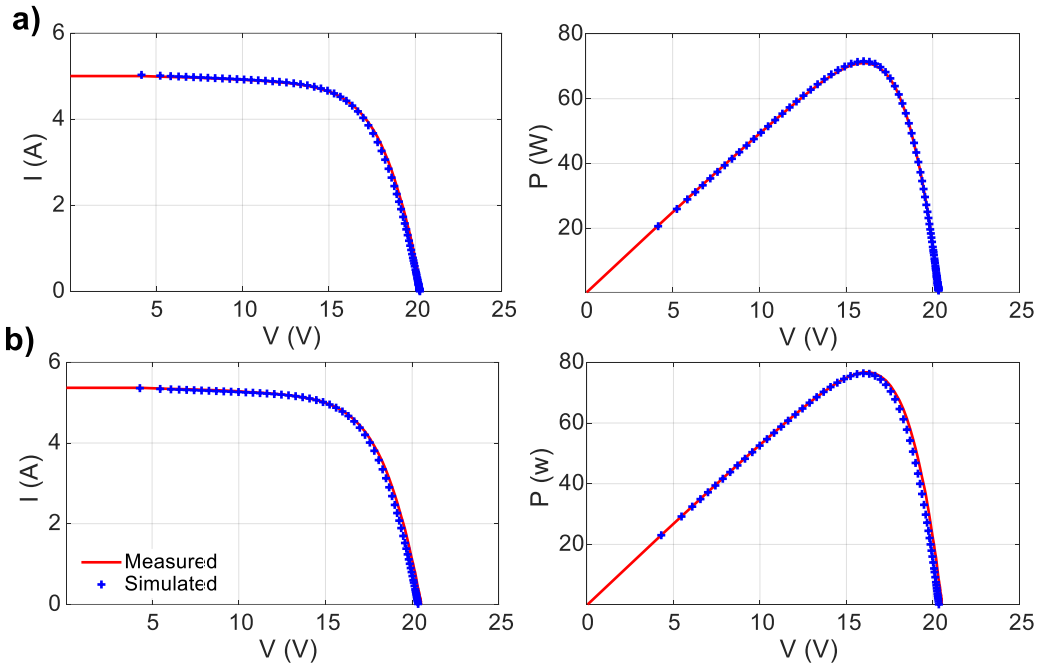

Figure II.9. PV array model validation under a) *T*=28.1, *G*=749, b) T=28.2, *G*=800.

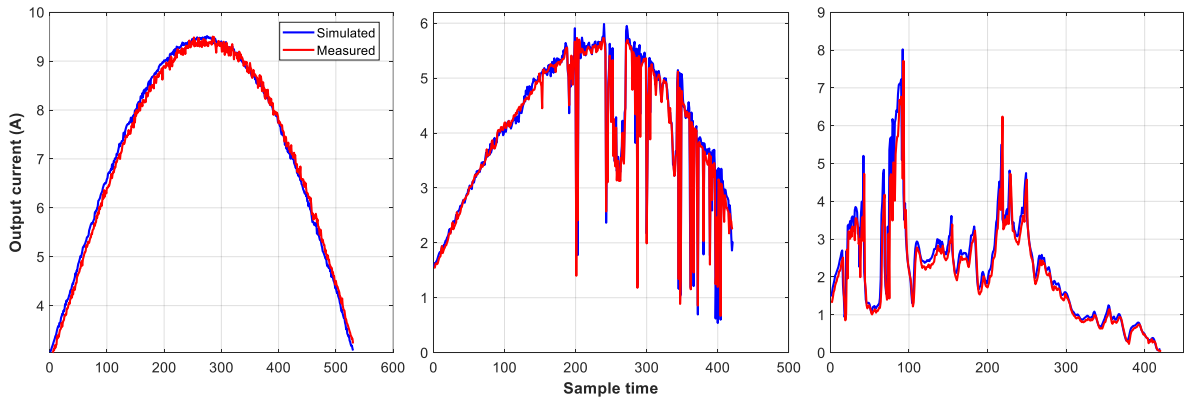

Figure II.10. Dynamic validation of the PV array model under different weather conditions.

II.5.2. Evaluation of the proposed fault detection and diagnosis strategy:

The automated system for detecting faults, developed using Random Forest (RF), is implemented in Python utilizing the scikit-learn library. It employs the "Random Forest Classifier" class from the "ensemble" module. This Python-based approach utilizes essential libraries such as scikit-learn, NumPy, SciPy, seaborn, matplotlib, and the open-source machine learning library dlib [142]. The computational setup for this study involved a personal computer equipped with an Intel Core i7 processor (2.50 GHz), 16 GB of RAM, and a GTX 1060 GPU with 6GB of memory.

As outlined earlier, hyperparameter optimization was conducted using the grid search algorithm. Table II.6 outlines the optimal hyperparameters selected for each RF model.

The summaries of results generated by the reports of the two developed Random Forest Classifiers (RFCs) are detailed in Tables II.7 and II.8. Both classifiers demonstrate exceptional performance in fault detection, achieving an accuracy rate of 99.4%. Precision and Recall (sensitivity) metrics are particularly emphasized due to their effectiveness in handling imbalanced data distributions. However, achieving a balance between these metrics can be challenging, especially evident during the diagnosis phase where the lowest Precision and Recall values were 0.978 and 0.974, respectively. This difficulty stems from the model's struggle to differentiate between faults labeled as #1 (three partially shaded PV modules) and #3 (three short-circuited PV modules), as both types of faults similarly impact PV output power. Despite these challenges, the implemented fault detection system maintains an overall accuracy of 99.4%.

Table II.6. Optimal hyperparameters.

Hyperparameter	**RF Detection model**	**RF Diagnosis model**
max_depth	45	85
n_estimators	65	35
Criterion	gini	entropy
Bootstrap	True	True
Min_samples_leaf	1	1
Min_sample_split	2	2
Max_features	6	6

Table II.7. Classification report of RF detection model.

	Precision	**Recall**	**F1$_{score}$**	**Samples number**
Class0	1.00	0.970	0.985	12145
Class1	0.993	1.000	0.996	48578
Macro avg	0.996	0.985	0.991	60723
Weightedavg	0.994	0.994	0.994	60723
Accuracy (%)	99.4			60723

Table II.8. Classification report of RF diagnosis model.

	Precision	**Recall**	**F1$_{score}$**	**Samples number**
Class0	0.978	1.000	0.989	12145
Class1	1.000	0.974	0.987	12144
Class3	0.999	1.000	1.000	12145
Class9	0.998	1.000	0.999	12144
Macro avg	0.994	0.994	0.994	48578
Weightedavg	0.994	0.994	0.994	48578
Accuracy (%)	99.4			48578

The normalized confusion matrices produced by the two distinct RF models—one specialized in fault detection and the other in fault diagnosis—are depicted in Figures II.11 and II.12, respectively. Analysis of the data given in Table II.7 and Figure II.11 underscores the strong performance of the binary classification model. For the healthy system (Class0), the model demonstrates high precision and minimal false positives, albeit with a slightly lower recall rate, accurately identifying 97% of instances and effectively distinguishing non-risky conditions. Conversely, when handling faulty cases labeled as Class1, the model performs almost flawlessly, exhibiting excellent recall, precision, and an F1 score.

Regarding the diagnostic aspect, as illustrated in Table II.8 and Figure II.12, a detailed breakdown of the rows and columns of the standardized confusion matrix for the RF diagnostic model is provided below:

- In Row 1 (Class0), corresponding to fault 2, the value of 1.00 in the top-left cell indicates precise identification of Class0 instances. Minimal misclassifications into other classes highlight the model's accuracy in detecting open circuit faults.
- Row 2 (Class1), representing fault 1, shows a value of 0.974 in the second cell, indicating the model correctly identifies Class1 instances with a true positive rate of approximately 97.4%. Occasional misclassifications into Class3 and Class0 suggest some confusion but overall effective detection of fault 1.
- Row 3 (Class3), corresponding to fault 3, displays a value of 1.00 in the third cell, indicating accurate recognition of Class3 instances with a true positive rate of 100%. This row shows no misclassifications into other classes, underscoring the model's reliability in identifying fault 3.
- Row 4 (Class9), representing fault 4, exhibits a value of 1.0 in the last cell, indicating a perfect true positive rate and accurate identification of all Class9 instances. No misclassifications into other classes draw attention to how well the model detects line-to-line faults.

the model demonstrates a high true positive rate for Class0 (fault 2), Class3 (fault 3), and Class9 (fault 4), accurately predicting these fault occurrences. It shows precision in identifying these specific defects but also maintains high precision for Class1 (fault 1) with minimal false positives, despite occasional confusion with Class3, as discussed earlier. Overall, the model effectively minimizes misclassifications.

Figure II.11. Normalized Confusion matrix of RF detection model.

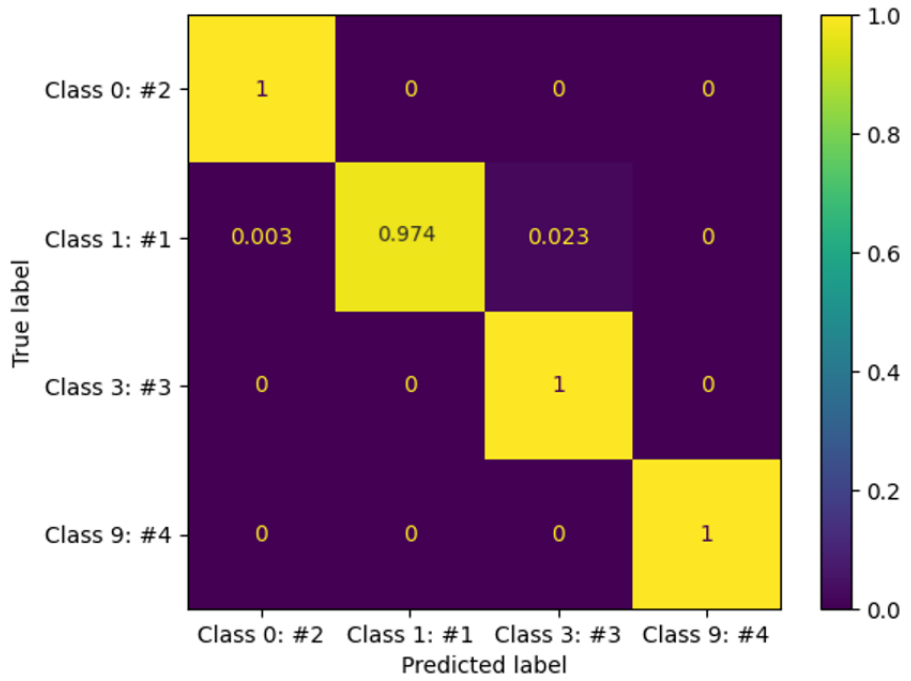

Figure II.12. Normalized Confusion matrix of RF diagnosis model.

To provide a clearer understanding of the classification outcomes, we have generated graphical representations of the confusion matrices for both Random Forest Classifier (RFC) models. These visual summaries are illustrated in Figure II.13, depicting the detection phase, and Figure II.14, illustrating the diagnosis phase. The graphical outputs visually corroborate the data shown in the confusion matrices, ensuring a straightforward interpretation of the model's performance.

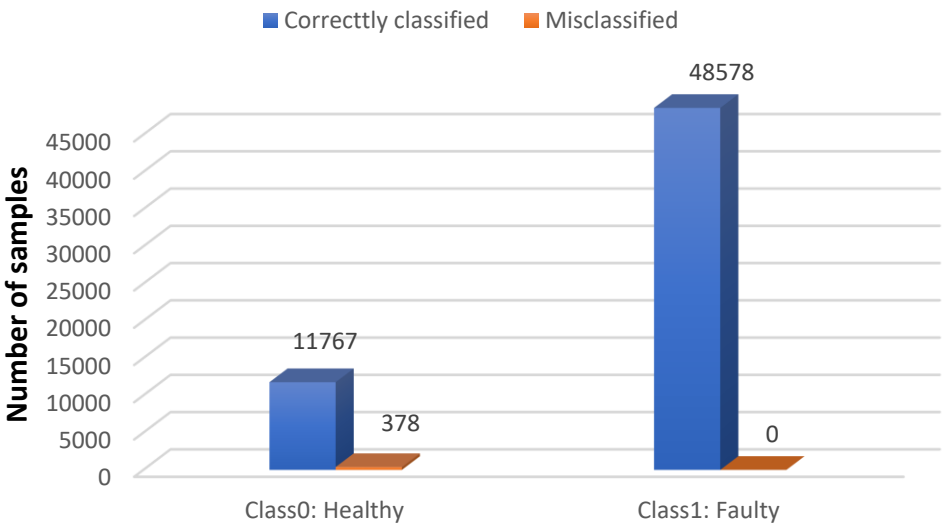

Figure II.13. Fault detection results.

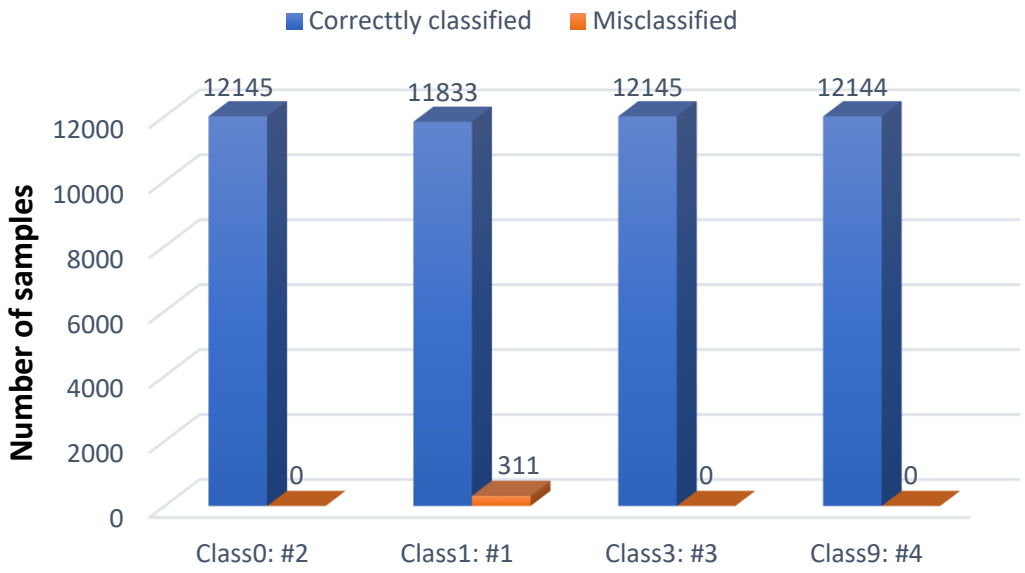

Figure II.14. Fault diagnosis results.

II.5.3. Comparative Analysis:

To emphasize the efficiency of our machine learning-based RFCs in fault detection, we carried out a comparative evaluation with several alternative methodologies, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Neural Networks (MLP Classifier), Decision Trees (DT), and Stochastic Gradient Descent (SGDC). To ensure a fair comparison, we applied consistent methodology as detailed in our study and optimized internal hyperparameters for each algorithm using a grid search approach. The summarized results are outlined in Table II.9.

During the detection phase, SVM achieved an accuracy of 84.5%, while MLP Classifier showed superior performance with an accuracy of 97.3%. SGDC demonstrated the lowest accuracy at 79.6%, whereas both KNN and DT algorithms performed equally well with accuracies of 98.3%. Moving to the diagnosis phase, all algorithms exhibited improved performance. DT achieved the highest accuracy of 98.3%, while SGDC showed notable improvement at 89.8%. Importantly, our proposed RFCs method outperformed all other methods in both detection and diagnosis stages, achieving an overall accuracy of 99.4%. Our RFC model also excelled in precision, recall, and F1 score metrics compared to SVM, KNN, DT, SGDC, and MLP Classifier.

These results are in line with earlier studies. For example, **Eskandari et al.** achieved average accuracies of 96% and 97.5%, respectively, for the detection and classification of Line-to-Line failures (LL) using an SVM-based technique [19]. Even if their accuracy is higher than

ours, it is important to remember that our research includes more than just LL flaws. The classification accuracy of our KNN model was 98.3%, which is in good agreement with findings from **Madeti and Singh** [18], who reported an average accuracy of 98.70% for various short-circuit faults, including those represented by bypass diodes, as well as open-circuit and line-to-line problems. Our model obtained a 98.2% accuracy in the field of Neural Networks, more precisely in the MLP Classifier. Comparatively, **Chine et al.** [143] reported a respectable 90.3% accuracy, maybe as a result of their reduced scope of errors taken into account and inadequate neural network architecture optimization. Furthermore, a fault identification and diagnosis method based on Decision Trees (DT) was provided by **Benkercha and Moulahoum**, with an overall accuracy of almost 99% [144]. Though marginally greater than the DT accuracy of around 98.3% in our analysis, this discrepancy might be attributed to the wider variety of fault types that we examined. **Kapucu and Cubukcu** [20] documented accuracies of 97.46% and 97.67% before and after optimization, respectively, using quadratic discriminant analysis-extra trees-Decision Trees (QDA-ETent-DT) for PV fault detection. Their study primarily addressed partial shading and short-circuit faults, without accounting for fluctuations in weather conditions.

Table II.9. Comparative Evaluation of SVM, KNN, Decision Trees (DT), SGDC, MLP, and RF Utilizing Identical Data Sets.

Phase	**Indicator**	**label**	**SVM**	**MLP Classifier**	**KNN**	**DT**	**SGDC**	**RF**
Detection	Precision	0	0.979	0.913	0.979	0.988	0.000	1.000
		1	0.984	0.990	0.984	0.981	0.796	0.993
	Recall	0	0.939	0.960	0.939	0.927	0.000	0.970
		1	0.995	0.977	0.995	0.997	1.000	1.000
	$F1_{score}$	0	0.959	0.936	0.959	0.956	0.000	0.985
		1	0.990	0.983	0.990	0.989	0.886	0.996
	Accuracy (%)		84.5	97.3	98.3	98.3	79.6	99.4
Diagnosis	Precision	0	0.958	0.992	0.923	0.992	0.851	0.978
		1	1.000	1.000	0.996	0.997	0.876	1.000
		3	0.933	0.974	0.998	0.972	0.986	0.999
		9	0.964	0.964	0.997	0.971	0.908	0.998
	Recall	0	0.928	0.975	0.997	0.972	0.967	1.000
		1	0.951	0.972	0.905	0.975	0.930	0.974
		3	0.968	0.981	0.997	0.985	0.697	1.000
		9	1.000	1.000	1.000	0.997	1.000	1.000
	$F1_{score}$	0	0.943	0.983	0.959	0.982	0.905	0.989
		1	0.975	0.986	0.948	0.986	0.902	0.987
		3	0.951	0.978	0.997	0.979	0.816	1.000
		9	0.981	0.982	0.998	0.984	0.952	0.999
	Accuracy (%)		96.1	98.2	97.5	98.3	89.8	99.4

II.6. Conclusion:

In conclusion, this chapter emphasizes the critical role of fault identification in optimizing photovoltaic (PV) system efficiency, given their susceptibility to various issues such as short circuits and shading. To tackle these challenges, a robust machine learning method employing the Random Forest Classifier (RFC) is introduced for effective fault detection and performance monitoring. The proposed approach relies on an accurate single-diode (SDM) simulation model that faithfully replicates actual PV system performance. Parameters of the SDM are obtained using an innovative blend of the current-voltage transformation technique and the Modified Grey Wolf Optimization (MGWO) algorithm. These parameters are incorporated into a physical model of the PV system, and comprehensive databases representing normal and abnormal operations are constructed through PSIM and MATLAB co-simulations.

Results demonstrate outstanding classification accuracy rates of 99.4% for fault detection and diagnosis, surpassing alternative models such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Neural Networks (MLP Classifier), Decision Trees (DT), and Stochastic Gradient Descent (SGDC). The RFC algorithm proves particularly robust, notably in addressing scenarios involving partial shading.

Chapter outline:

III.1. Introduction:

The transition to sustainable energy sources, particularly photovoltaic (PV) systems, necessitates meticulous monitoring and fault diagnosis to ensure reliability and efficacy. This chapter introduces a novel application of deep learning techniques for fault detection and diagnosis within PV systems, presenting a comprehensive three-step approach.

Initially, a trusted PV model is developed and refined using a heuristic optimization approach, ensuring its accuracy in representing real-world PV system behavior. Subsequently, a detailed database is constructed, integrating PV model data with monitored module temperature and solar irradiance across various operational conditions, including both healthy and faulty states.

The main objectif of the proposed methodology is the use of a combination of Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (Bi-GRU) architectures for fault classification. This combination of parallel and sequential processing allows the neural network to harness the strengths of both convolutional and recurrent layers concurrently, enabling effective fault detection and diagnosis.

III.2. Materials & Methods:

III.2.1. Experimental setup:

The PV system under investigation in this study comprises frameless Glass-Glass Cadmium Telluride (CdTe) thin-film PV panels installed on an experimental residence located in Buštěhrad, Czechia. These panels seamlessly replace the conventional tiles on both the East and West sides of the pitched roof (as shown in Figure III.1). With a total capacity of 3.84 kW, the PV system is connected to four AP systems YC1000^{-3} three-phase micro-inverters. Each side of the roof hosts 24 panels connected to two micro-inverters, with each micro-inverter accommodating four channels linking one string composed of three PV modules. Monitoring of the PV system's outputs, including current *(I_{mpp})*, voltage *(V_{mpp})*, and *power (P_{mpp})* at the maximum power point, has been carried out for each channel at five-minute intervals since October 2018. Additionally, two Si-RS485TC-2T-MB irradiance sensors, one for each side (East and West), equipped with external temperature sensors affixed to the back side of PV modules, are employed to track solar irradiance *(G)* and module temperature *(T_m)* at one-minute resolution. Validation of the proposed fault detection and diagnosis methodology is performed using one year of recorded data from the east-facing

roof. Tables III.1 and III.2 offer an overview of the characteristics of the selected PV generator and the PV panel data, respectively.

Figure III.1. PV system used to validate the proposed fault detection procedure.

Table III.1. Summary of the characteristics of the selected PV generator.

Main parameters	**PV system (East roof)**
PV size	1.92 kW
Inverter nominal power	2 x Microinverter 1kW
Num. modules per inverter	12
Num. modules in series (Ns)	1
Num. strings in parallel (Np)	3 x 4
Tilt – Azimuth	30º – 9º East

Table III.2. PV module electrical data.

Parameter	$\mathbf{P_{mp}}$ (W)	$\mathbf{I_{sc}}$ (A)	$\mathbf{V_{oc}}$(V)	$\mathbf{I_{mp}}$(A)	$\mathbf{V_{mp}}$(V)	$\boldsymbol{\beta_{Voc}}$(%/°K)	$\boldsymbol{\alpha_{Isc}}$(%/°K)
Value	80	2.38	59.4	2.03	43.2	-0.21	0.03

III.2.2. Databases and PV modelling:

Developing a precise database that defines the function of a PV system is crucial for effective fault detection and diagnosis. Therefore, having a reliable simulation model that accurately represents the system's behavior under normal and faulty conditions is essential. In this study, we employed the Sandia Array Performance Model (SAPM), an empirical model developed by Sandia National Laboratories [145], chosen for its simplicity and proven reliability in accurately characterizing and simulating PV array performance. The model has

demonstrated significant accuracy through extensive validation across modules of various technologies, as highlighted in [146]. Estimating the PV array's output current (*Impp*), voltage (V_{mpp}), and power (P_{mpp}) at the maximum power point is directly facilitated by the following equations:

$$Impp = N_p \left[Imp(C_0 Ee + C_1 Ee^2) \left(1 + \alpha_{Imp}(T_m - T_m^*)\right)\right] \quad \text{(III.1)}$$

$$Vmpp = N_s\big[Vmp + C_2 N_{sc}\delta(T_m)ln(Ee) + C_3 N_{sc}(\delta(T_m)ln(Ee))^2 + \beta_{Vmp} Ee(T_m - T_m^*)\big] \quad \text{(III.2)}$$

$$\delta(T_m) = nk(T_m + 273.15)/q \quad \text{(III.3)}$$

$$Ee = \frac{G}{G^*} \quad \text{(III.4)}$$

$$Pmpp = Impp \times Vmpp \quad \text{(III.5)}$$

In this context, I_{mp} and V_{mp} symbolize the PV module's current and voltage under Standard Test Conditions (STC). The coefficients C_0 and C_1, determined empirically and dimensionless, establish the relationship between Imp and the effective irradiance. αI_{mp} ($°C^{-1}$) represents the normalized temperature coefficient for Imp, while C_2 (dimensionless) and C_3 (V^{-1}) are empirical coefficients that link V_{mp} to the effective irradiance (*Ee*). Furthermore, *Nsc* denotes the number of cells in a PV module, $\delta(Tm)$ signifies the thermal voltage per cell at temperature Tm, q represents the elementary charge (1.60218×10^{-19} coulomb), n stands for the ideality factor, k denotes Boltzmann's constant (1.38066×10^{-23} J/K), and β_{Vmp}(V/°C) indicates the temperature coefficient for module *Vmp* at STC.

The SAPM model incorporates several coefficients and parameters (C_0, C_1, C_2, C_3, n, α_{Imp}, *and* β_{Vmp}) whose values are typically unavailable from the PV module manufacturer. These parameters are often determined through testing and direct measurements of PV modules or arrays under both static and dynamic conditions. The parameter extraction method utilized in this study follows the procedure outlined in [146], employing the artificial bee colony (ABC) optimization algorithm. This algorithm evaluates the model parameters for PV arrays operating under real-world conditions, utilizing daily profiles of solar irradiance, module temperature, and monitored DC output current and voltage profiles. The optimization process aims to minimize the objective function, defined as the root mean square error in Equation (III.6), where $\theta = f$ (C_0, C_1, C_2, C_3, n, α_{Imp}, and β_{Vmp}), and N represents the length of the

measured data, with *Vi* and Ii denoting the measured voltage and current at data point *i*, in the same order.

$$S(\theta) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[Ii - I(Vi,\theta)]^2} \quad \text{(III.6)}$$

A swarm-based meta-heuristic technique called the artificial bee colony algorithm was developed to address multidimensional and multimodal optimization issues. The method is mostly based on the model for honeybee colony foraging behavior [147–148]. The artificial bees are divided into three categories by the ABC algorithm: scouts, observer bees, and employed bees. An employed bee is one that is actively seeking food or making use of a food supply. An onlooker bee is one that waits in the hive to decide which food source to select. Employed bees have their food sources banned and they become scouts when their food sources cannot be improved after a set number of tries. There are a plenty of food sources. At the initialization phase, the ABC generates a randomly distributed initial population of SN solutions. Each solution is produced within its limits according to the equation below:

$$x_i^j = \mathrm{x}_{min}^j + rand[0,1].\left(\mathrm{x}_{max}^j - \mathrm{x}_{min}^j\right) \quad \mathrm{i} = 1, 2\ldots.\mathrm{SN}, \mathrm{j} = 1, 2\ldots\ldots\mathrm{D} \quad \text{(III.7)}$$

In this context, x_{min}^j and x_{max}^j de note the minimum and maximum values of parameter j, respectively, while D represents the number of optimization parameters. Following initialization, the population of solutions undergoes repeated cycles, denoted as C=1, 2..., MCN, involving the search processes of employed bees, onlooker bees, and scout bees. In each cycle, every employed bee generates a new solution v_{ij} according to Equation (III.8), and subsequently assesses its fitness, denoted as fit_i .

$$\mathrm{v}_{ij} = \mathrm{x}_{ij} + \emptyset_{ij}(\mathrm{x}_{ij} - \mathrm{x}_{kj}) \quad \text{(III.8)}$$

Where k ∈ {1, 2, . . ., SN} and j ∈ {1, 2, . . ., D} are randomly chosen indexes. Although k has to be different from i. ij is a random number between [−1, 1]. After the information is shared by the employed bees, each onlooker finds new solution v_{ij} within the neighborhood of x_i by using Eq. (III.8), based on the probability P_i defined as:

$$P_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_i} \quad \text{(III.9)}$$

Where fit_i is the fitness value of the solution x_i. The value of each generated candidate solution v_{ij} that is not within the boundary of the allowed search space is updated so that it will be within this space.

The fitness of each newly generated candidate solution v_{ij} is compared with that of its previous solution. If the new solution demonstrates equal or superior fitness, it replaces the previous one in memory. Conversely, if the new solution's fitness is worse, the previous solution remains stored. Essentially, a greedy selection mechanism is employed to choose between the old and candidate solutions.

At the conclusion of each search cycle, if a solution's fitness cannot be enhanced and the predetermined number of attempts, referred to as the "limit," is exhausted, the solution is abandoned by the scout bee, and a new solution is randomly sought. The new solution x_i is generated using Equation (III.7).

It's evident, from the above explanation, that there are three control parameters in the ABC algorithm: the number of candidate solutions, which equals the number of employed and onlooker bees (SN), the value of the "limit," and the maximum cycle number (MCN).

To realize the SAPM parameter optimization using the ABC algorithm, each candidate solution is defined as a vector of SAPM parameters $\theta = (C0, C1, C2, C3, n, \alpha Imp, \beta Vmp)$, so the optimization problem has multiple parameters to be optimized (D = 7). Thus, the equations for initialization and solution updates become:

$$\theta_i^j = \theta_{min}^j + rand[0,1].\left(\theta_{max}^j - \theta_{min}^j\right) \quad \text{(III.10)}$$

$$v_{ij} = \theta_{ij} + \emptyset_{ij}(\theta_{jmax} - \theta_{jmin}) \quad \text{(III.11)}$$

The fitness of each solution (parameter set) is chosen as the root mean square error between the measured and modeled PV outputs. Therefore, the probability calculation equation becomes:

$$P_i = \frac{S(\theta_i)}{\sum_{n=1}^{SN} S(\theta_i)} \quad \text{(III.12)}$$

To evaluate the candidate parameter sets. The objective is to minimize the RMSE between the measured power and the power calculated using the SAPM parameters. at the end of each search cycle, the abandoned solution that has not improved its fitness over a predetermined

number of cycles (limit) is replaced by a new randomly chosen solution, Figure III.2 illustrate The basic steps of the ABC algorithm.

The application procedure of the proposed method can be divided into four phases:

- **Initialization Phase**: Set the algorithm parameters, including the number of candidate solutions SN, the maximum cycle number MCN, and other relevant parameters. Then, a randomly distributed initial population of parameter sets is generated using Eq. (III.10) and evaluated.

- **Employed Bees Phase**: The employed bees update their parameter sets using Eq. (III.11) and then evaluate the new parameter sets. The greedy selection process is applied, and the probability value Pi for each parameter set θi is calculated using Eq. (III.12).
- **Onlooker Bees Phase**: The roulette wheel selection method is used to recruit onlooker bees for local searching around the chosen parameter set, depending on the calculated probability Pi. The new parameter sets are evaluated, and the greedy selection process is applied.
- **Scout Bees Phase**: Eq. (III.10) is used to replace an abandoned parameter set that has not improved its fitness over a predetermined number of cycles (limit) with a new randomly chosen parameter set. The fitness of the new parameter set is then evaluated.

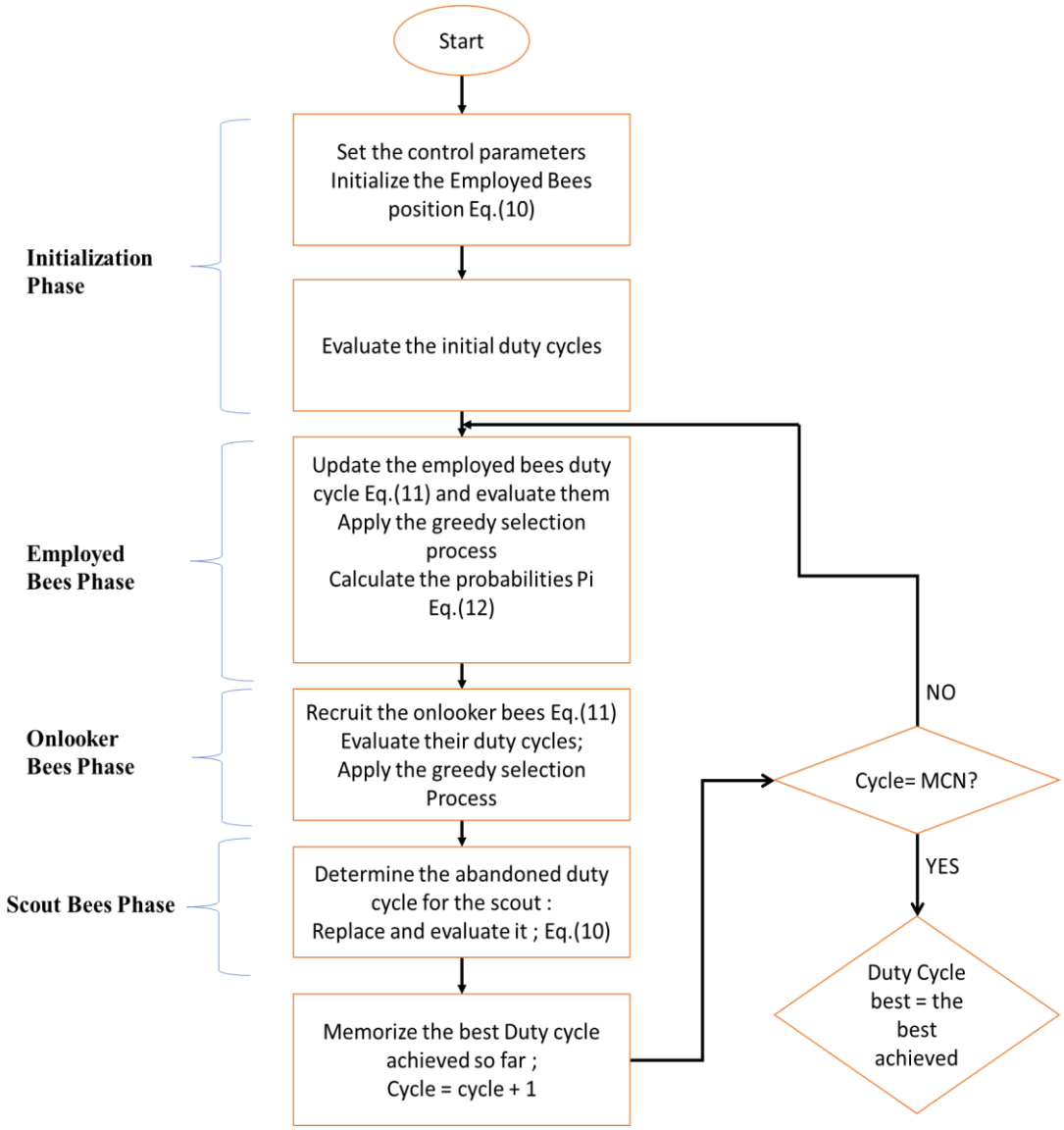

Figure III.2. basic steps of the ABC algorithm.

The completed PV system model serves as the basis for developing databases that thoroughly document the system's performance in outdoor environments. Utilizing yearly profiles of solar irradiance and module temperature, this model is crucial for creating datasets that cover both ideal operation and deliberately simulated faults. The simulated scenarios, representing typical challenges encountered in grid-connected PV systems, are described below and illustrated in Figure III.3:

- ***Healthy system:*** This depicts the standard operation of the PV system without any abnormalities.
- ***Three modules short-circuited:*** This involves the deliberate disconnection of one channel of the micro-inverter.

- ***Six modules short-circuited:*** In this case, two channels from one micro-inverter are intentionally disconnected.
- ***Nine modules short-circuited:*** This scenario entails the disconnection of three channels of one micro-inverter.
- ***Open circuit faults:*** This indicates a situation where one micro-inverter of the PV system malfunctions and becomes inoperative.
- ***Shading faults:*** This scenario simulates the impacts of partial shading experienced by PV systems due to factors such as cloud movement or the presence of nearby objects at specific times. It encompasses various shading patterns occurring on different days and hours throughout the year.

The resulting databases from this procedure capture three essential variables: Irradiance, Temperature, and the power output at the Maximum Power Point derived from every simulated operational scenario.

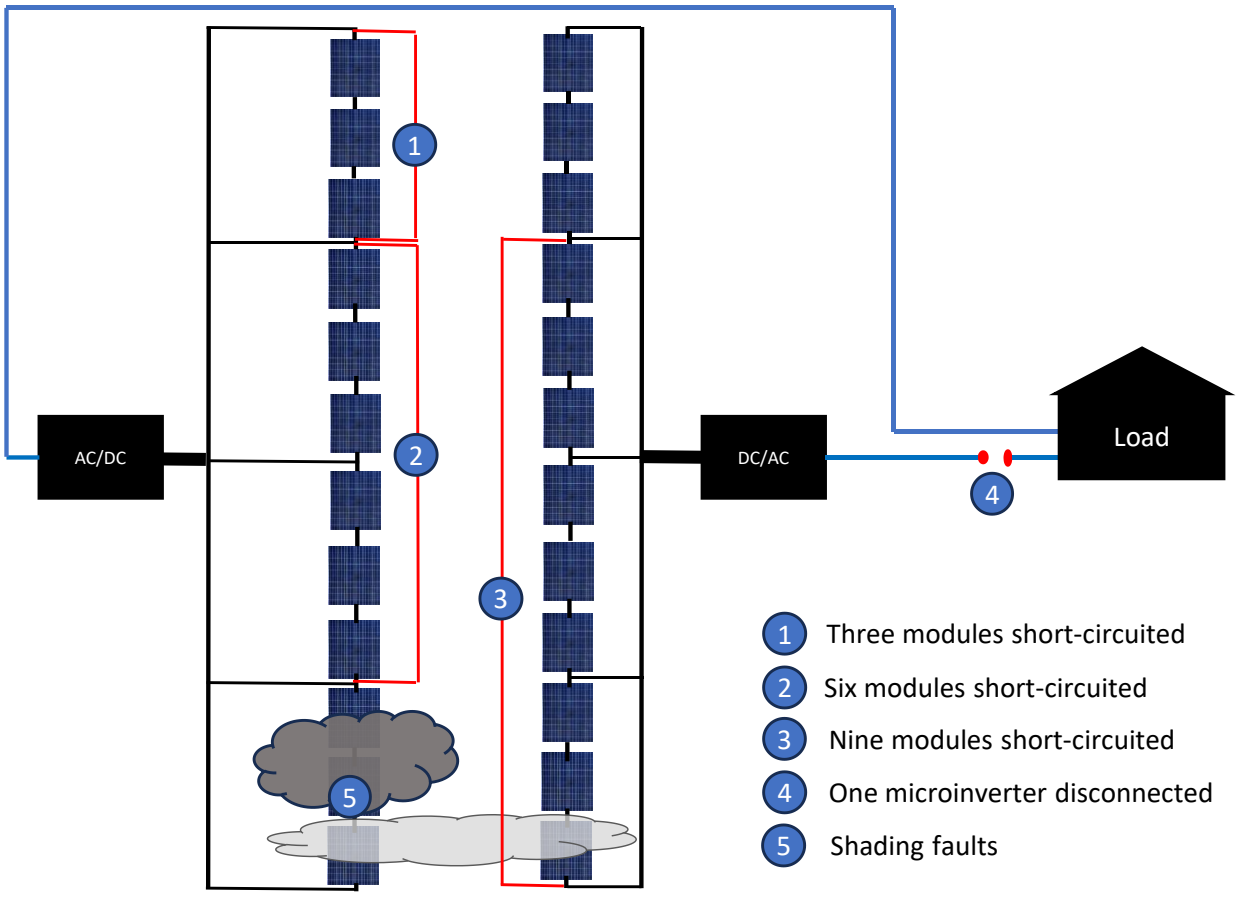

Figure III.3. PV system simple layout with considered faults.

III.2.3. Procedure for Detecting and Diagnosing Faults:

The fundamental goal of this section is to develop a robust and reliable fault identification and diagnosis approach for PV systems using Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (Bi-GRU) deep learning techniques. This subsection begins with a general overview of CNN and Bi-GRU. The proposed hybrid model-based fault detection approach is then described in detail. Finally, it describes the evaluation measures employed to analyze the performance of the suggested method. Figure III.4 offers a flowchart illustrating the steps involved in developing the proposed strategy.

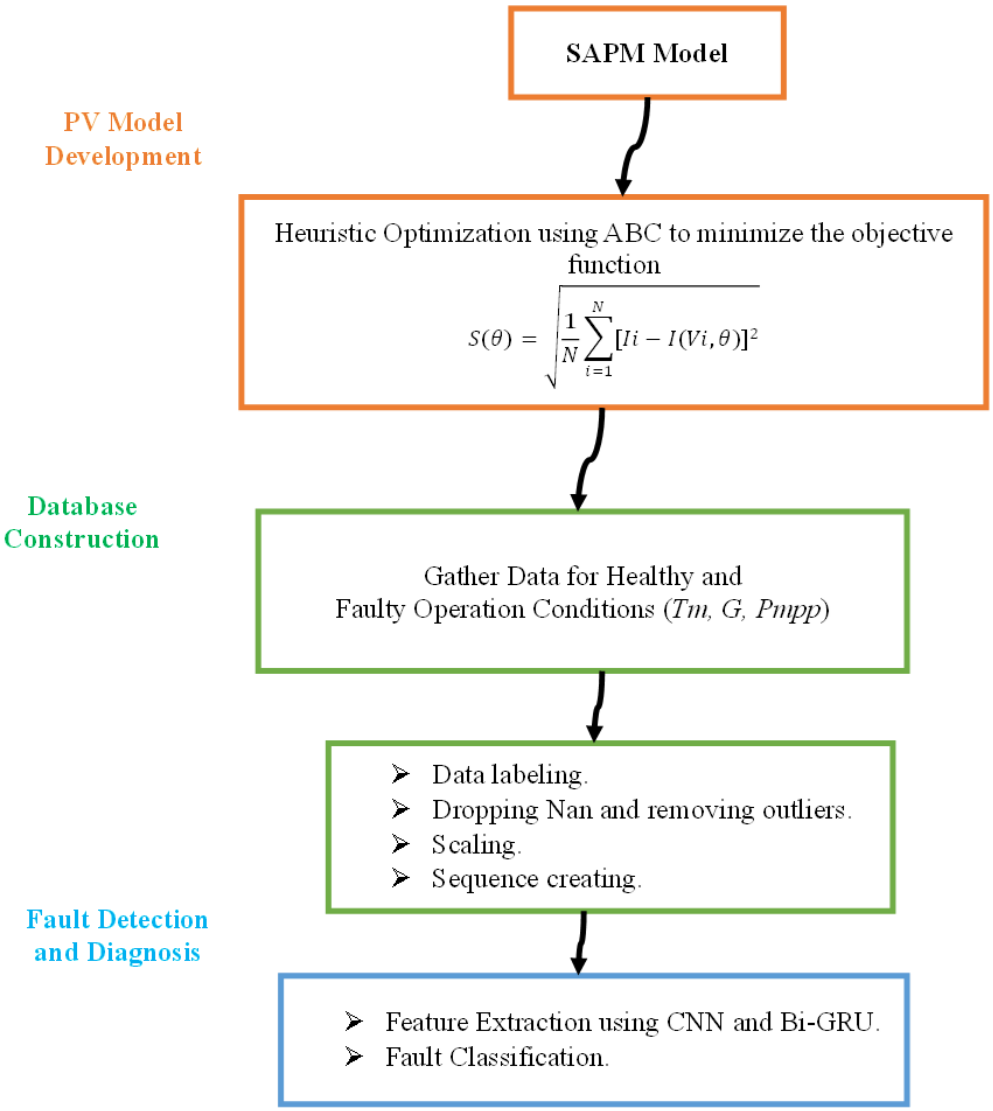

Figure III.4. Flowchart of the proposed CNN and Bi-GRU fault detection and diagnosis strategy.

III.2.3.1. Convolutional Neural Network (CNN):

The CNN belongs to a supervised subset within deep learning algorithms [149], functioning differently from traditional neural networks by employing convolution in its layers instead of matrix multiplication. Its structure consists of two main components: the feature extractor and the classifier. The first one comprises input layers, convolutional layers (CLs), and pooling layers (PLs), organized layer by layer to extract features from the input data matrix. These features, known as feature maps, are generated by passing the input through a series of filters. Subsequently, the convolution maps are flattened and merged into a CNN code feature vector. This CNN code, produced at the output of the convolutional segment, is then linked to the input of the second part, which consists of fully connected layers, forming a multi-layer perceptron. The classifier part specializes in categorization and consists of fully connected layers (FC) along with an output layer. These FC layers accept the features derived from the final pooling layer as input, with the output being the last layer containing one neuron per category [150]. The incorporation of Conv1D layers in our framework underscores CNN's adeptness in feature extraction, particularly suited to datasets exhibiting spatial configurations. CNNs are adept at decoding spatial patterns and intricate networks inherent in data, as exemplified by the Conv1D layers in our setup. This capability is pivotal for a comprehensive understanding of fault progression and manifestation within the PV system, aligning seamlessly with the model's architectural design. The Conv1D Neural Networks General Architecture is illustrated in Figure III.5.

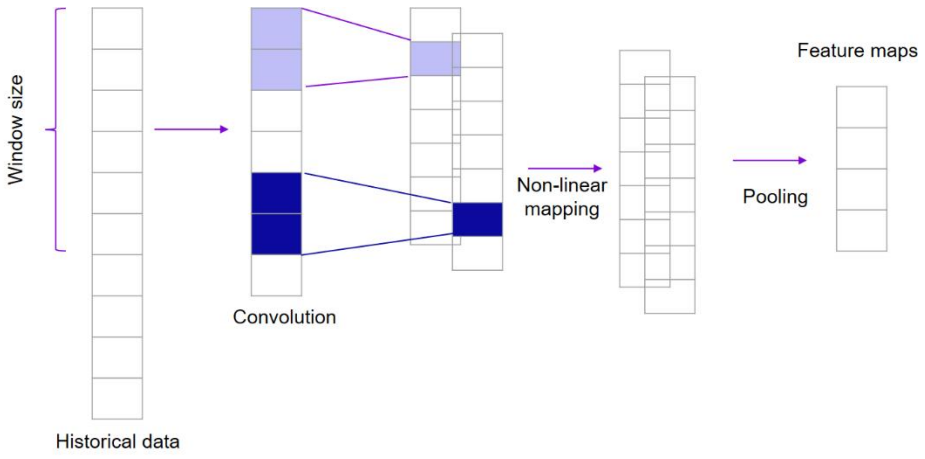

Figure III.5. Conv1D Neural Networks General Architecture.

The utilization of Convolutional Neural Network (CNN) layers within our framework underscores their effectiveness in processing sequential data. Through a systematic approach, we preprocess the data by defining the number of time steps and features, subsequently organizing it into input-output pairs based on defined sequences. This process involves sliding a window across the dataset to extract sequences of input data and their corresponding output labels. The dataset is then split into training, validation, and test sets to facilitate model evaluation. With the model architecture defined, Conv1D layers are incorporated to extract features, followed by batch normalization and max-pooling layers to enhance performance. Dropout layers are included to prevent overfitting by randomly deactivating neurons during training. This approach lays the foundation for a CNN model tailored to process sequential data, with flexibility for further customization based on specific requirements and architectural preferences.

III.2.3.2. Gated recurrent unit (GRU):

The GRU is a distinct variation of Recurrent Neural Networks (RNNs) initially introduced by ***Cho et al.***[151]. It addresses the challenge of gradient vanishing typically found in conventional RNNs by merging the memory capability of Long-Short Term Memory (LSTM) while offering quicker execution owing to fewer parameters during training. [152]. It features two gates, namely the reset gate and the update gate, as depicted in Figure III.6, which control the flow of information.

The update gate controls how much of the previous hidden layer state (ht^{-1}) is kept in the current hidden layer state (*ht*). It uses an activation function to process information from ht-1 and the current moment's input (*xt*), with a smaller activation result indicating higher information retention. The expression for the update gate is as follows:

$$zt = \sigma(Wzxxt + Wzhht - 1 + bz) \quad \text{(III.13)}$$

The reset gate determines the amount of information from the previous time step that is stored in the candidate memory state (ǂt). Like the update gate, it evaluates ht-1 and xt through an activation function, with a higher activation output indicating more information is written to ǂt. The reset gate is expressed as:

$$rt = \sigma(Wrxxt + Wrhht - 1 + br) \quad \text{(III.14)}$$

GRU integrates the reset gate (rt) with ht-1 and xt to generate a candidate memory state (ǂt) according to the following equation:

$$\check{h}t = \tanh[W\check{h}xxt + W\check{h}\check{h}(rt \times ht - 1) + b\check{h}] \quad \text{(III.15)}$$

The current hidden layer state (ht) is derived by integrating the previous hidden layer state (ht−1) with the candidate memory state (ȟt) through the following equation:

$$ht = (1 - zt) \times ht - 1 + zt \times \check{h}t \quad \text{(III.16)}$$

According to the equations provided , *xt* denotes the input at the current time step; ht−1 and ht indicate the hidden layer state at the previous and current time steps, respectively; ȟt represents the candidate memory state; rt and zt signify the reset and update gates, respectively; Wzx, Wrx, and Wȟx refer to the weight matrices associated with xt for the update gate, the reset gate, and the candidate memory state, respectively; Wzh, Wrh, and Wȟȟ refer to the weight matrices associated with ht−1 for the update gate, the reset gate, and the candidate memory state, respectively; bz, br, and bȟ express the corresponding prejudices.

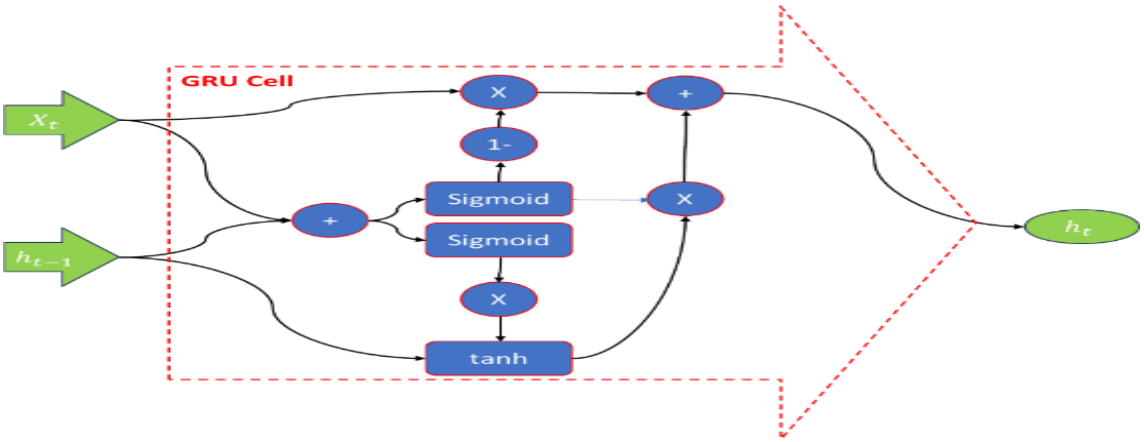

Figure III.6. Structure of GRU.

III.2.3.3. Bidirectional gated recurrent unit (Bi-GRU):

The Bi-GRU unit, originating from the bidirectional RNN [153], consists of two GRU layers with distinct information flow directions, as depicted in Figure III.7. In this Bi-GRU configuration, a backward layer is integrated into the single-layer GRU network to effectively use input data. This design incorporates two hidden layers to capture both past and future data. It acknowledges that excluding one-directional communication could affect the predictive ability of the GRU model. Both hidden layers are linked to the same output layer, and the output of the current state (yt) is formulated as follows:

$$yt = [h^{\gg}, h^{\ll}] \quad \text{(III.17)}$$

In this context, $h^{»}$ and $h^{«}$ represent the outputs of the forward and backward GRU layers, respectively.

In fault detection within PV systems, identifying issues like open circuits, short circuits, and shading faults is crucial to ensuring system integrity. While traditional unidirectional recurrent neural networks focus solely on past-to-future relationships, bidirectional GRU networks excel in capturing bidirectional dependencies by processing information from both past and anticipated future states. By employing two GRU networks moving in opposite directions—interpreting past-to-present and future-to-present data—Bi-GRUs offer a comprehensive insight into a system's temporal dynamics, significantly enhancing fault detection capabilities. This method surpasses mere reliance on historical data by incorporating foresight into potential future scenarios, particularly valuable in predicting shading faults. The proficiency of Bi-GRUs in understanding evolving patterns and relationships aids in identifying subtle fault patterns influenced by future conditions that might otherwise remain undetected.

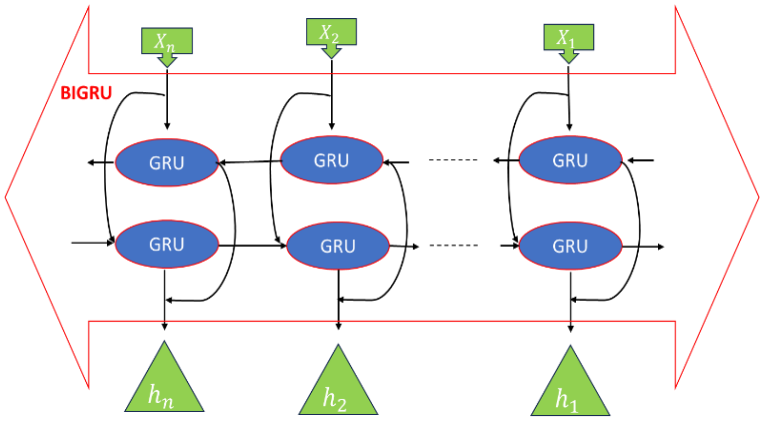

Figure III.7. Structure of Bi-GRU.

III.2.3.4. Hybrid CNN-Bi-GRU Architecture:

The proposed hybrid model, which combines the strengths of CNN and Bi-GRU architectures, aims to improve fault detection accuracy. While CNNs excel in feature extraction through layered processing, they may encounter overfitting issues with high-dimensional data. Conversely, Bi-GRU networks effectively handle high-dimensional and time series data but may overlook specific data features at times. In our model, Conv1D layers serve as feature extractors from CNNs, while Bidirectional GRU (Bi-GRU) layers manage high-dimensional data. The innovative hybrid model's design, as depicted in Figure III.8, seamlessly integrates two Bi-GRU layers into the CNN framework. Placing these Bi-GRU

layers before fully connected layers offers advantages such as effective training on high-dimensional features extracted by the CNN without risking overfitting. Additionally, this integration facilitates the merging of spatial features captured by CNNs and temporal intricacies handled by Bi-GRUs in subsequent layers. Through fully connected layers and skip connections, this combination constructs a comprehensive data representation enriched with fault-specific features. Importantly, the adaptability of both CNNs and Bi-GRUs allows the model to dynamically adjust its feature extraction strategy, effectively aligning with the unique characteristics of the training data. This adaptability ensures that the model can capture and learn from the varied and intricate fault patterns present in the data, thereby enhancing its overall fault detection capabilities.

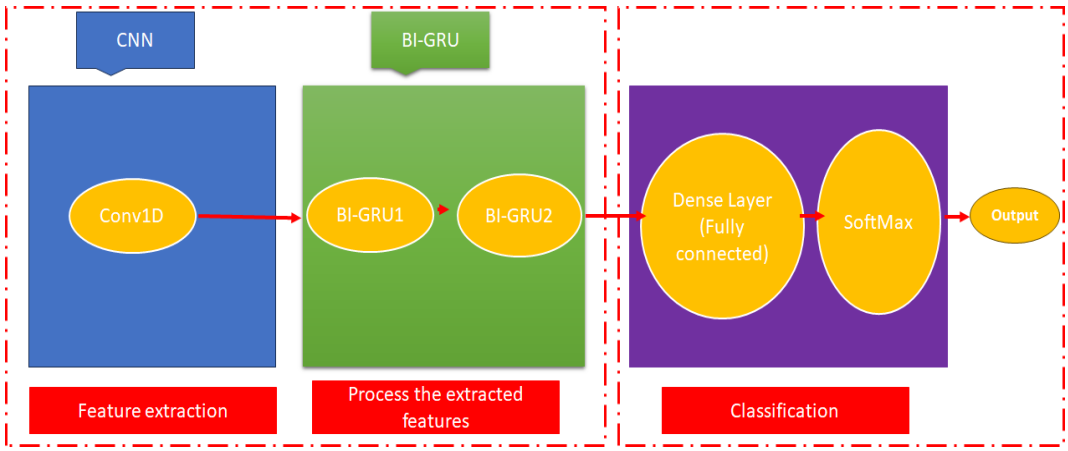

Figure III.8. Adopted hybrid model integrating CNN and Bi-GRU general structure.

The final step of the fault detection approach entails deploying two deep learning models. The initial model aims to detect abnormalities within the PV system, whereas the subsequent model diagnoses the identified faults. Both deep learning models were constructed utilizing Python along with the Tensorflow-Keras and scikit-learn libraries [154–157]. The architecture comprises the following layers:

**Input Layer:* A 3D input layer was established with dimensions (num_time_steps, Variables).

**Convolutional Layers:* Feature extraction was carried out using Conv1D layers, followed by Pooling and Dropout.

**Bidirectional GRU Layers:* Two sets of Bi-GRU layers processed the features obtained from Conv1D. The initial set produced sequences, while the subsequent set yielded only the final output of each sequence. Both GRU layers underwent layer normalization and dropout for regularization.

**Fully Connected Layers:* Dense layers with dropout, kernel_regularizer_l2, kernel_regularizer_l1, and skip connections were incorporated.

**Output Layer:* A dense output layer featuring softmax activation was employed for binary classification in the detection model and multiclass classification in the diagnosis model.

The intricate design of the models for fault detection and diagnosis is outlined in Figure III.9.

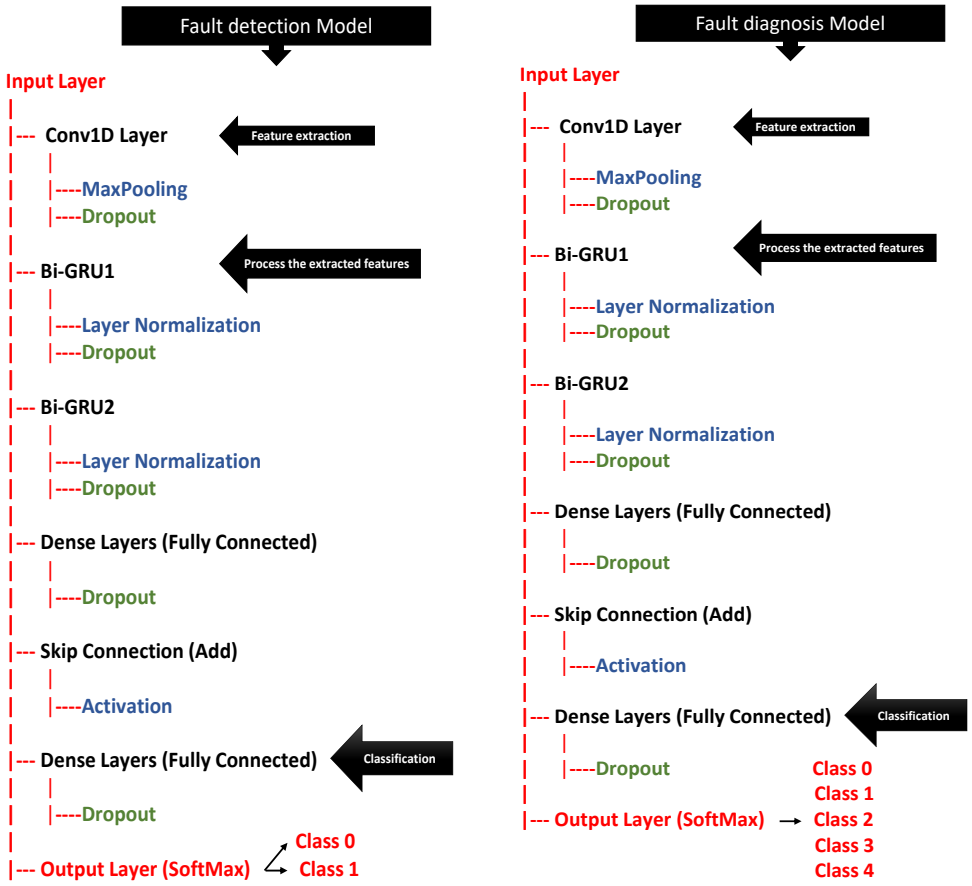

Figure III.9. Designed models for fault detection (left) and fault diagnosis (right).

III.2.4. Process of detecting and diagnosing faults:

To evaluate the effectiveness of the hybrid model proposed in this study, several metrics were utilized, including Categorical Accuracy, Precision, Recall, and F1-Score, as described in Equations (II.24) – (II.27) from chapter II. These metrics incorporate True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [158].

The model was assembled using the Adam optimizer and categorical cross-entropy loss, as described in Equation (III.18). Throughout the training phase, various strategies such as early stopping, model checkpointing, and learning rate reduction callbacks were utilized to enhance

model performance. Subsequently, the model underwent evaluation on the test set, assessing its accuracy through the generated classification report and confusion matrix.

$$Loss = -\sum_{i=1}^{N} \mathrm{Y}_i \times log\hat{\mathrm{Y}}_i \quad \text{(III.18)}$$

III.3. Obtained Results:

This section presents the outcomes obtained from our PV modeling methodology and the prepared datasets, demonstrating the performance of the PV system across different weather conditions and fault scenarios. Following this, we evaluate the effectiveness of our novel fault detection and diagnosis procedure, which combines CNNs and Bi-GRU networks, using various performance metrics.

III.3.1. PV model validation and database construction:

The PV model, based on the empirical SAPM, was applied to replicate the real-world performance of the PV system considered in this study. Daily monitored profiles, encompassing PV output power at MPP (Pmpp), on-plane solar irradiance (G), and module temperature (Tm), were used to deduce the unknown parameters of the model. The derived parameters are outlined in Table III.3, and the validation of the model across multiple days is illustrated in Figure III.10. The outcomes demonstrate a strong agreement between the measured and simulated hourly values of the PV system's output power. The overall Root Mean Square Error (RMSE) value, considering clear sky, semi-cloudy, and overcast days, stands at 2.69%. These results affirm the effectiveness of the parameter identification process and the robustness of the SAPM.The PV system model was then employed to create databases that fully capture the system’s behavior in real outdoor environments. These databases include datasets representing instances of normal operation and deliberately simulated faults, utilizing yearly monitored solar irradiance (G) and module temperature (Tm) profiles. A portion of the simulated PV output, incorporating various faults is visually depicted in Figure III.11. Shading faults are randomly introduced throughout the year. The figure demonstrates the effect of different faults on the PV output, highlighting the significance of accounting for transient defects related to shading.

Table III.3. SAPM PV model extracted unknown parameters.

Parameter	C_0	C_1	C_2	C_3 (V^{-1})	n	β_{Vmp}(V/°C)	α_{Imp} ($°C^{-1}$)
Value	0.915	-0.0446	$1.88\ 10^{-16}$	-7.98	1.31	-0.143	$7.14\ 10^{-4}$

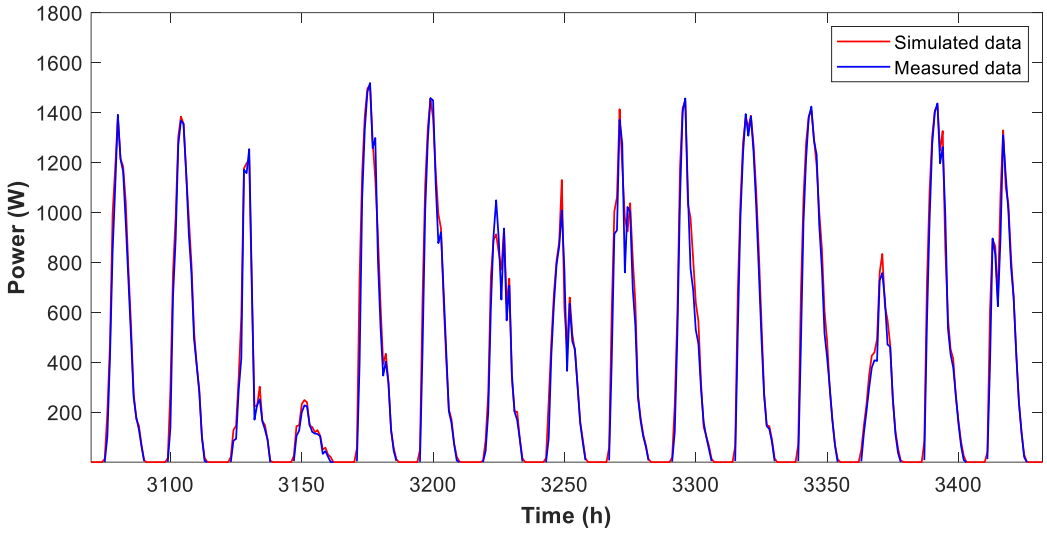

Figure III.10. Comparison between measured and simulated PV output power using SAPM.

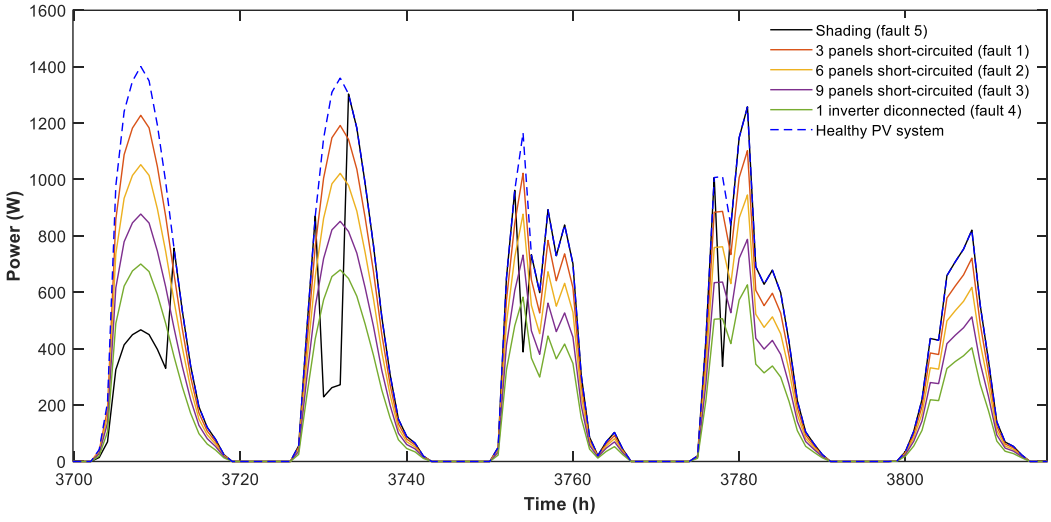

Figure III.11. Representative data of the DC output from the PV system operating under various faults.

III.3.2. Assessing the efficacy of the proposed fault detection method:

During the data generation phase, recorded data is systematically labeled with relevant class annotations. Specifically, out of 52,385 raw data samples, 43,624 are designated for

diagnosis while the remaining are allocated for detection. The initial focus of the preprocessing pipeline prioritizes data quality and reliability. This involves eliminating rows with missing values to enhance data integrity. Additionally, the identification and potential removal of duplicate rows are performed to mitigate redundant information. To ensure the dataset's robustness, efforts are made to address outliers in numerical columns. The Interquartile Range (IQR) method is employed to establish lower and upper bounds for each numerical column, with values beyond these bounds replaced by the corresponding boundary values. This approach enhances the accuracy of data distribution representation and diminishes the influence of extreme values on subsequent analyses [159].

This fundamental stage enhances convergence and boosts performance in deep learning models by ensuring uniform feature scaling. Subsequently, the sequential data length (num_time_steps) and the number of columns (features) in the scaled dataset are determined. This process involves iterative steps to generate sequences and corresponding labels based on the specified time steps. Sequences are formed by considering a data window with a length of num_time_steps. These sequences, along with their corresponding labels, are appended to separate arrays through the loop. For the detection dataset, 23,575 sequences are obtained, leaving 19,589 sequences for the diagnosis dataset. The dataset is then divided into three sets—training, validation, and test—utilizing the train_test_split function from scikit-learn. The training set constitutes 70% of the data, while the remaining 30% is evenly distributed between the validation (15%) and test (15%) sets. Details regarding the construction of the detection and diagnosis datasets are provided in Table III.4.

Ultimately, an input layer is established for the model, personalized to accommodate the designated number of time steps and columns (features) within the sequences *(Tm, G, Pmpp,* equating to 3 features per time step).

Refining the parameters of the deep learning model is achieved through the grid-search optimization method [158]. This methodical tuning procedure systematically explores different combinations of internal hyperparameters to determine the optimal configuration that enhances the performance of our fault detection and diagnosis models. The final hyperparameters for both fault diagnosis and detection models are outlined in Tables III.5 and III.6.

Table III.4. The specifics of constructing the detection and diagnosis dataset.

Phase	Defined faults	Assigned Class	Train dataset (70%)	Test dataset (15%)	Validation dataset (15%)
Detection	Faulty PV system	0	13836	2966	2986
	Healthy PV system	1	2666	571	550
Diagnosis	3 panel short-circuit ed (fault1)	0	2675	564	548
	6 panel short-circuited (fault2)	1	2771	580	635
	9 panel short-circuited (fault3)	2	2783	601	602
	1 inverter disconnected (fault4)	3	2779	610	598
	Shading (fault5)	4	2704	584	555

Table III.5. Hyperparameters tuning for the fault detection model.

Hyperparameter	Range for search	Selected value
epochs	[50 - 250]	136
filters (Conv1D)	[64 - 256]	128
pool_size (MaxPooling1D)	[2 - 4]	2
dropout (Conv1D)	[0.3 - 0.6]	0.4
GRU units (Bidirectional GRU)	[64 - 256]	128
dropout (Bidirectional GRU)	[0.3 - 0.7]	0.5
return_sequences (Bidirectional GRU)	[True, False]	True/False
num_time_step	[12 - 250]	200
activation (GRU, Dense)	['relu', 'tanh', 'sigmoid']	'tanh'
batch_size	[16 - 64]	32
kernel_regularizer_l1	$[10^{-6} - 10^{-3}]$	10^{-5}
kernel_regularizer_l2	$[10^{-5} - 10^{-3}]$	10^{-4}

Table III.6. Hyperparameters tuning for the fault diagnosis model.

Hyperparameter	**Range for search**	**Selected value**
epochs	[50 - 250]	160
filters (Conv1D)	[60 - 250]	128
pool_size (MaxPooling1D)	[2 - 4]	2
dropout (Conv1D)	[0.3 - 0.7]	0.5
dropout (Bidirectional GRU)	[0.3 - 0.6]	0.3
GRU units (Bidirectional GRU)	[64 - 256]	128
return_sequences (Bidirectional GRU)	[True, False]	True/False
activation (GRU, Dense)	['tanh', 'relu', 'sigmoid']	'tanh'
num_time_step	[12 - 250]	200
batch_size	[16 - 64]	32
kernel_regularizer_l1	$[10^{-5} - 10^{-3}]$	10^{-5}
kernel_regularizer_l2	$[10^{-5} - 10^{-3}]$	10^{-4}

Table III.7 presents the findings from the proposed deep learning (DL) architecture, which incorporates CNN and Bi-GRU layers for the defect detection phase, while Figures III.12 and III.13 visualize the results. The examination of the classification report and loss value highlights the binary classification model's remarkable ability in distinguishing between faulty and healthy systems, as shown in Table III.7. For Class 0, which represents the defective System, the model demonstrates exceptional precision, correctly recognizing 100% of instances predicted as defective while reducing false positives. The recall value of 1.00 is also significant, capturing all occurrences of defective systems while reducing false negatives. The F1-score of 1.00 indicates a balanced and effective performance. Referring to Class 1, denoted as the Healthy System, the model adeptly identifies almost all instances predicted as healthy with a precision score of 0.99, and captures 98% of the actual instances of healthy systems, demonstrating a balanced performance. With an F1-score of 0.99, the model's effectiveness in classification is further underscored. In the broader assessment, the model achieves a notable accuracy of 99.45%, highlighting its capability to provide precise predictions across both classes, as depicted in Figure III.12. Moreover, the macro-average and weighted-average values of 0.99 reinforce the overall strong performance.

The analysis of the results reveals that the model has effectively captured underlying patterns during its training phase, as evidenced by its high precision and recall rates for both faulty and healthy systems. The low loss value of 0.0080, as illustrated in Figure III.13, further highlights the model's robust ability in accurately classifying PV system faults. The reported recall value of 0.98 for "Class 1," representing the healthy system, indicates the precise identification of 98% of actual instances, with the remaining 2% being false negatives. In the context of fault detection in PV systems, misclassifying a small portion of healthy system instances is supposed less critical than the reverse situation. Despite occasional misclassifications of healthy systems, the overall performance of the model remains satisfactory for PV system fault detection, thereby alleviating safety concerns associated with misclassifying faults as healthy instances.

Table III.7. Generated classification report for the fault detection model.

	Precision	**Recall**	**F1-score**
Class 0	1.00	1.00	1.00
Class 1	0.99	0.98	0.99
macro avg	0.99	0.99	0.99
weighted avg	0.99	0.99	0.99
accuracy	0.99		

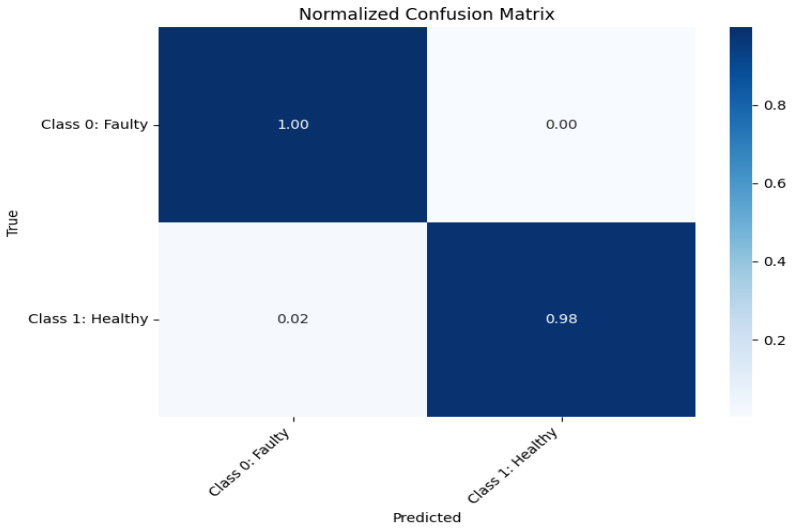

Figure III.12. Normalized confusion matrix of the CNN-Bi-GRU fault detection model.

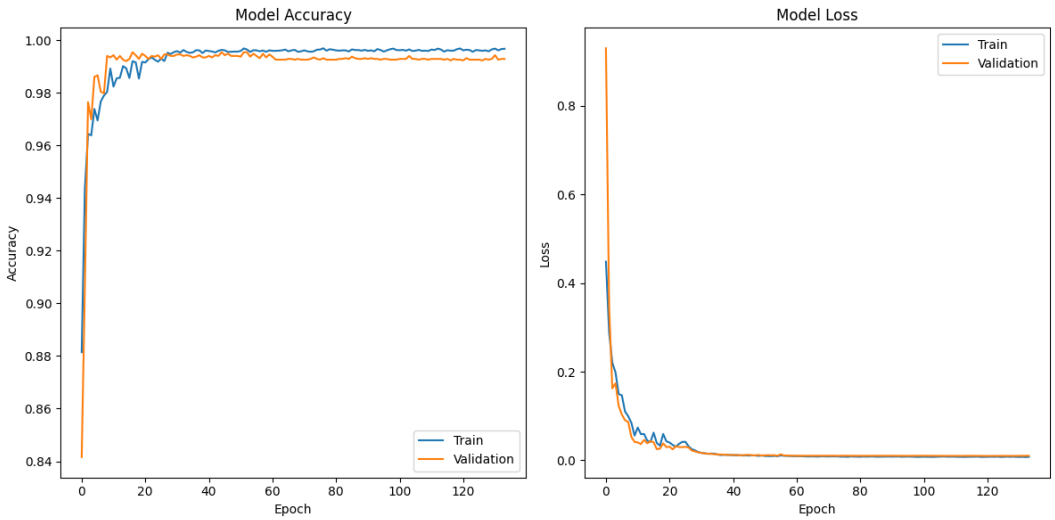

Figure III.13. Accuracy and loss evolution of the CNN-Bi-GRU fault detection model.

The findings of the proposed hybrid model for the diagnostic phase are outlined in Table III.8 and visually depicted in Figures III.14 and III.15. Remarkably, the classification outcomes demonstrate outstanding performance, characterized by a flawless overall accuracy of 100%. The detailed classification report in Table III.8 highlights crucial metrics in the evaluation process. Precision, representing the ratio of correctly predicted positive observations to the total predicted positives, maintains a perfect score of 1.00 across all fault classes, emphasizing the model's accuracy in predicting fault types. Similarly, recall, indicating the ratio of correctly predicted positive observations to all observations in the actual class, also achieves a perfect score of 1.00 for each fault type, confirming the model's effectiveness in capturing all instances of each fault type. The F1-Score, a weighted average of precision and recall, consistently registers a perfect score of 1.00 for all classes, indicating a balanced relationship between precision and recall. Figure III.15 illustrates the attained minimal loss value of 0.0001, approaching zero, highlighting the model's high accuracy in predictions and its ability to learn complex patterns and features crucial for precise classification. Consequently, the model demonstrates exceptional capabilities in fault classification, achieving perfect precision, recall, and F1-Score for each fault type.

Table III.8. Generated classification report for the fault diagnosis model.

	Precision	Recall	F1-score
Class 0	1.00	1.00	1.00
Class 1	1.00	1.00	1.00
Class 2	1.00	1.00	1.00
Class 3	1.00	1.00	1.00
Class 4	1.00	1.00	1.00
macro avg	1.00	1.00	1.00
weighted avg	1.00	1.00	1.00
accuracy	1.00		

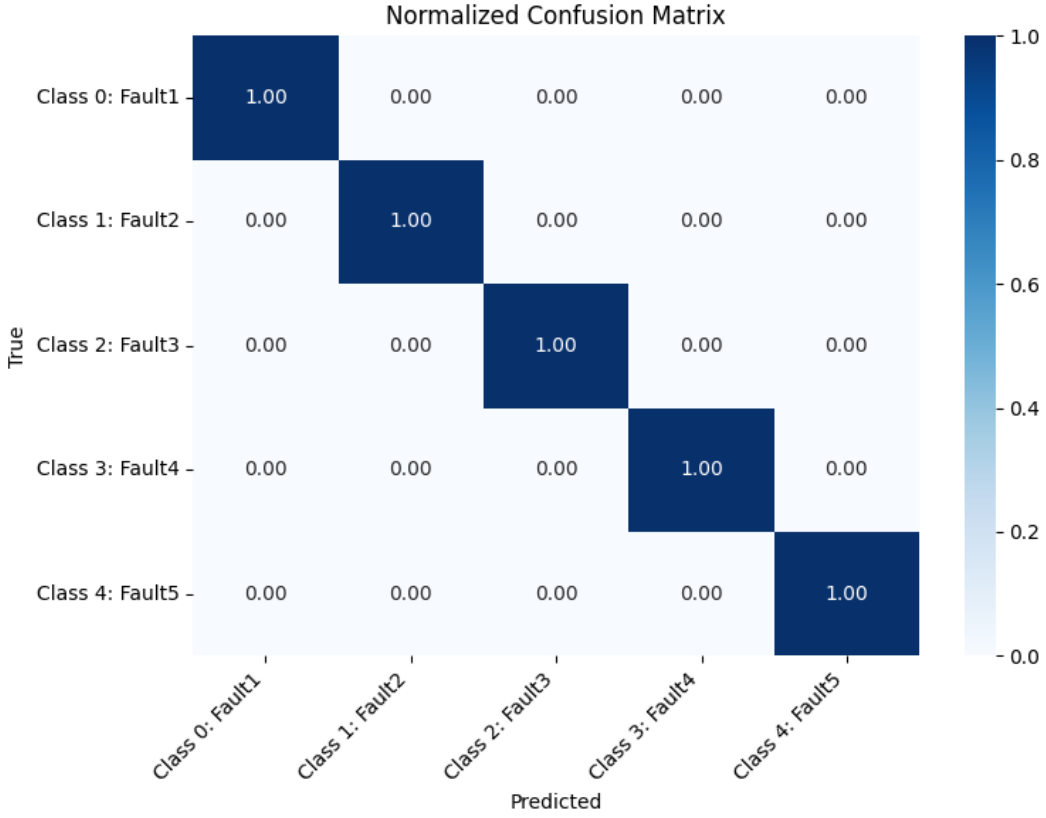

Figure III.14. Normalized confusion matrix of the CNN-Bi-GRU fault diagnosis model.

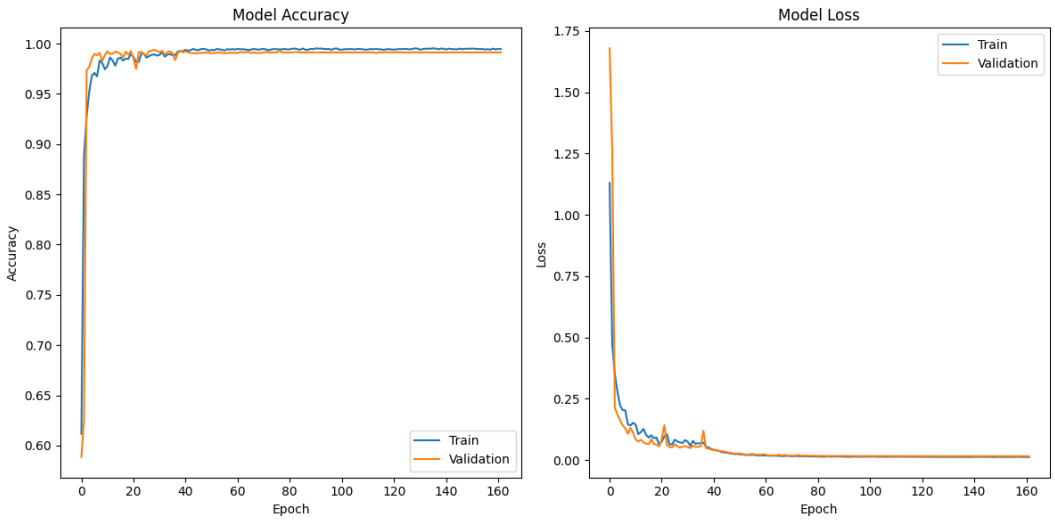

Figure III.15. Accuracy and loss evolution of the CNN-Bi-GRU fault diagnosis model.

III.4. Comparison and Discussion:

In order to underscore the efficacy of our hybrid Deep Learning model architecture, which integrates CNN with Bi-GRU for fault detection and diagnosis, we carried out a comparative assessment against several alternative approaches, including CNN, CNN-LSTM, and CNN-BiLSTM models, all structured similarly to our models. Furthermore, we stocked to the same procedures as in our study to ensure an equitable and thorough comparison. We fine-tuned the internal hyperparameters for each algorithm using the grid search methodology.

The results in Table III.9 indicate that during the detection phase, the CNN-LSTM achieved a precision score of 98.59%, while the CNN-BiLSTM attained the highest overall accuracy at 98.76%. In contrast, the CNN alone had the lowest precision at 96.22%. However, our implemented CNN-Bi-GRU model surpassed all others with an impressive accuracy of 99.46%. Moving to the diagnosis phase, the CNN showed improvement with an accuracy of 97.40%, while both CNN-LSTM and CNN-BiLSTM achieved 100% accuracy. The CNN-Bi-GRU model also reached 100% accuracy, highlighting its superior diagnostic capabilities. These results emphasize the effectiveness of the CNN-Bi-GRU hybrid model in both fault detection and diagnosis. The proposed model excels in accuracy, recall, F1-score, and precision, while also reducing loss in both the detection and diagnosis phases. This improved performance is due to the integrated feature extraction method, leveraging both CNN and Bi-GRU capabilities. Furthermore, the efficiency of the softmax function in classification, similar to artificial neural networks (ANN) known for their effectiveness in fault detection or classification in PV systems, enhances our model's outstanding results.

Examining the suboptimal accuracy of the single CNN-based architecture, we found that it struggles with detecting temporal faults, particularly in distinguishing between healthy and fault states caused by shading, as shown in Figure III.16 (left side). However, by excluding shading faults from the training data and focusing only on permanent faults, the single architecture's performance improved, achieving 100% accuracy in the detection phase, as illustrated in the normalized confusion matrix on the right side of Figure III.16.

Finally, the analyses suggest that a two-phase model structure is more advantageous for handling temporary faults like shading faults. This differentiation is essential because distinguishing between these fault types and the PV system's healthy state is challenging. The complexities of identifying shading faults or similar issues require a more nuanced approach, and a two-phase model provides a more accurate representation of the system's behavior in such scenarios. In contrast, a single-phase model might suffice for permanent faults like open or short circuits, which are generally easier to distinguish from the healthy state. Therefore, selecting the model structure should consider the nature and complexity of the prevalent faults in the PV system, ensuring an optimal balance between accuracy and computational efficiency.

Table III.9. Comparative analysis of the proposed technique with various alternative approaches.

	Metrics	**Class**	**CNN**	**CNN-Lstm**	**CNN-BiLstm**	**CNN-Bi-GRU**
Detection	Precision	0	0.96	0.99	0.99	1.00
		1	0.95	0.97	0.97	0.99
	Recall	0	0.99	0.99	0.99	1.00
		1	0.79	0.94	0.95	0.98
	F1-score	0	0.98	0.99	0.99	1.00
		1	0.87	0.98	0.96	0.99
	Accuracy (%)		96.22	98.59	98.76	99.46
	Loss		0.0680	0.0246	0.0143	0.0080

Diagnosis	Precision	0	0.96	1.00	1.00	1.00
		1	0.97	1.00	1.00	1.00
		2	0.98	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00
		4	0.96	1.00	1.00	1.00
	Recall	0	0.96	1.00	1.00	1.00
		1	0.98	1.00	1.00	1.00
		2	0.98	1.00	1.00	1.00
		3	0.98	1.00	1.00	1.00
		4	0.97	1.00	1.00	1.00
	F1-score	0	0.96	1.00	1.00	1.00
		1	0.97	1.00	1.00	1.00
		2	0.98	1.00	1.00	1.00
		3	0.99	1.00	1.00	1.00
		4	0.97	1.00	1.00	1.00
	Accuracy (%)		97.40	100	100	100
	Loss		0.0948	0.0002	0.0001	0.0001

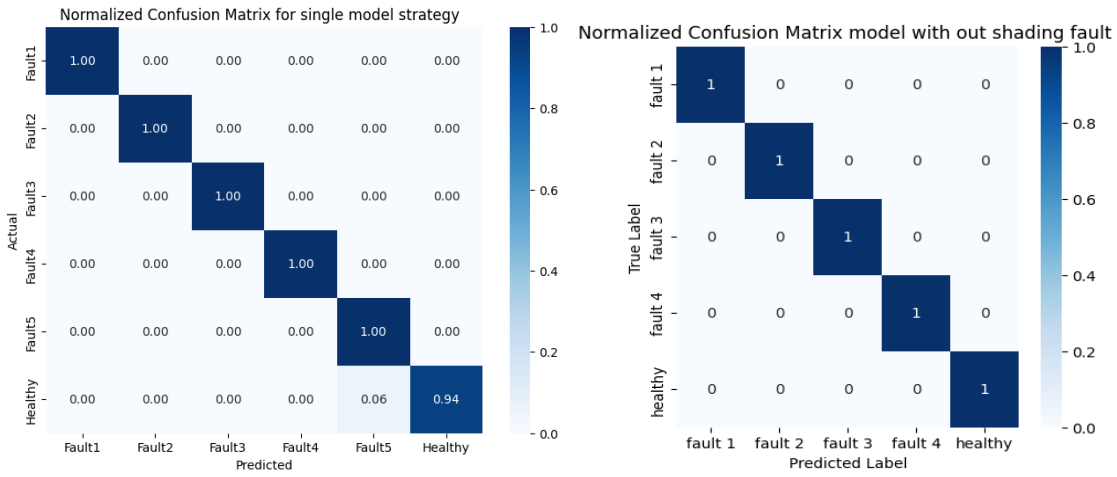

Figure III.16. Single CNN-based architecture model is employed for detecting all types of faults (left side) and specifically for detecting permanent faults (right side).

III.5. Conclusion:

This chapter has been introduced a ground-breaking approach to fault detection and diagnosis in PV systems, that is crucial for ensuring their reliability and efficiency in the transition to sustainable energy sources. Through a comprehensive three-step methodology, including the development of a precise PV model, construction of a detailed database integrating environmental variables, and the use of advanced deep learning techniques for fault classification, significant strides have been made in enhancing fault detection accuracy.

The mixture of CNN and Bi-GRU architectures has proven to be particularly effective in identifying and categorizing various PV fault types, such as open circuits, short circuits, and partial shading. This approach not only improves fault detection accuracy but also lays the groundwork for targeted fault diagnosis and mitigation strategies.

Chapter outline:

IV.1. Introduction:

Considering the numerous strategies that use historical data for predicting PV power, our study delves into identifying the most effective machine learning techniques and supervised learning models for estimating power output from photovoltaic plants situated in Algeria. Through rigorous evaluation, including models like Random Forest, Support Vector Regression (SVR), Multi-layer Perceptron (MLP), Linear Regressor (LR), Gradient Boosting, and k-Nearest Neighbors (KNN), we aim to pinpoint the optimal approach. Our methodology entails particular data preprocessing steps, ensuring dataset integrity by removing missing values and outliers using Isolation Forest. Subsequent feature selection based on correlation threshold aids in identifying relevant features crucial for accurate prediction in PV systems. Leveraging both Pearson and Spearman correlation coefficients, we select features demonstrating significant correlations, ensuring robustness across different correlation methods. Isolation Forest aids in outlier detection, followed by model training and evaluation using key performance metrics. Moreover, integration of the best-performing model into a MATLAB application for real-time predictions enhances usability and accessibility across diverse applications in renewable energy.

IV.2. PV Dataset:

The collected data in this study is from a grid-connected ground-mounted photovoltaic system arranged in Ain El-Melh, Algeria, the site's coordinates are 34°51″ N and 04°11″ E, with an height of 910 meters over ocean level. This photovoltaic control plant is coordinates into Ain El-Melh's medium voltage arrange and is portion of a considerable 400 MW extend supervised by SKTM Company, an auxiliary of Sonelgaz. Sonelgaz, ordered by the Algerian government for renewable vitality advancement, has executed 23 PV power plants over the countries and central locales. The Ain El-Melh plant, gloating a add up to capacity of 20 MWp, ranges over 40 hectares. Key design specifications of this 20 MWp PV facility are detailed in Table IV.1.

Table IV.1. Ain El-Melh PV power plant design parameters (20MWp)

Parameter	**Characteristics**
Type of module	Poly-crystalline silicon
Efficiency of PV module	15%
Tilt and Orientation	33°South
Type of installation	Fixed structure
PV rows distance	5 meters
Inverter nominal power	500 KW
Characteristics of transformers	1250 kVA, 47–52 Hz, 315 V/31.5 kV

The photovoltaic modules are linked to 500 kW inverter cabinets via junction boxes, serving as the primary data source. Data gathering occurred from January 1, 2020, to December 31, 2021, with readings taken every fifteen minutes, resulting in a total of 69,195 data points. This dataset encompasses various parameters such as solar panel temperature, tilt radiation, total radiation, dispersion radiation, direct radiation, wind speed, humidity, pressure, voltage, current, and PV power. Table IV.2 show Summary of Environmental and Electrical Parameters in the Photovoltaic System. Table IV.3 provides the technical specifications of the PV modules utilized within this PV plant.

Table IV.2. PV plant-monitored data.

Feature	**Description**	**Maximum**	**Minimum**	**Average**
Tp	Panel temperature (°C)	74.800	-2.5	27.987833
Gdin	Radiation inclinaison (W/m2)	1651.200	0.0	310.162255
Gtotal	Total radiation (W/m2)	1395.600	0.0	239.705539
Gdisp	Dispersion radiation (W/m2)	686.400	0.0	76.567325
Gdirect	Direct radiation (W/m2)	1365.600	0.0	232.813488
V_V	Wind speed (m/s)	22.200	0.0	3.760438
H	Humidity (%)	71.600	0.0	36.119596
P	Pressure (Pa)	927.000	0.0	912.473873
V_{DC}	Voltage (V)	780.400	0.0	329.776418
I_{DC}	Current (A)	985.400	0.0	183.593662
P_{DC}	PV power (kW)	569.441	0.0	108.289535

Table IV.3. PV module specification (Yingli Solar YL2545-29b).

PV Module	**Specifications**
STC power rating	250 Wp ±5%
Number of cells	60
Vmp	29.8 V
Isc	8.92 A
Imp	8.39 A
Voc	37.6 V
Power temperature coefficient	α %/°C
NOCT (°C)	46±2

IV.3. Data Pre-processing:

IV.3.1. Refining Sensor Data: Procedures for Cleaning and Normalizing:

Data preprocessing is fundamental when working with real data collected from programmed sensors, as they regularly contain errors and inconsistencies. Cleaning and organizing strategies are connected to get ready the data for use with machine learning models. The centre lies on settling minor inconsistencies and evacuating incorrect or missing data from the observing dataset.

One challenge encountered is the presence of empty records, particularly during night-time slots (between 9 p.m. and 4 a.m.), where no measurements are collected. Whereas sun based illumination is intrinsically zero during the night, discuss temperature information may still be missing. In any case, the need of temperature data during the night is regarded unimportant since there's no photovoltaic control generation at that time. Counting night-time data would as it were include excess data, increasing model complexity and calculation time without giving important comes about. To anticipate the negative effect of empty records on learning models, columns containing invalid information are eliminated. The same procedure is applied to remove duplicated values or deficient records.

After these preprocessing steps, the database eventually contains 33,465 points. To optimize the model's execution and guarantee data homogeneity, the min-max normalization strategy is applied. This process scales each information point to a run between 0 and 1. The equation for calculating the normalized value x_{norm} for a given value x is:

$$x_{norm} = \frac{x-\min(x)}{\max(x)-\min(x)} \quad \text{(IV.1)}$$

This normalization method serves different purposes, counting speeding up the optimization process, minimizing aberrations between information values, evacuating dimensional impacts, and reducing computational requirements.

IV.3.2. Correlation Coefficient Examination:

The examination inspected the relationship variables to discover the connections among P_{DC} and person climate components. The correlation coefficient, indicated as r, implies the degree of relationship between two factors, x_i and y_i, is expressed as follow [160]-[162]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x_\iota})(y_i - \overline{y_\iota})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x_\iota})^2 \sum_{i=1}^{n}(y_i - \overline{y_\iota})^2}} \quad \text{(IV.2)}$$

$$\overline{x_\iota} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{(IV.3)}$$

$$\overline{y_\iota} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad \text{(IV.4)}$$

By applying (3) and (4) to (2), the following expression can be driven:

$$r = r_{xiyi}$$

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i{}^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i{}^2 - (\sum_{i=1}^{n} y_i)^2}} \quad \text{(IV.5)}$$

Where $\{\overline{x_\iota}, \overline{y_\iota}\}$ and n are the mean and sample size, respectively, and $\{x_i, y_i\}$ are the individual sample points indexed by i.

Two strategies are utilized for estimating correlation and correlation coefficients between two factors: The Pearson strategy evaluates the straight relationship between factors, showing a corresponding alter between them. Then again, the Spearman strategy assesses a basic (ordinal or rank) relationship, where factors tend to alter together without essentially being proportional.

This think about utilized the Pearson relationship strategy to analyze the correlation between P_{DC} and Metrological factors. Figure IV.1 outlines the results of this relationship investigation, within the heatmap histograms within the inclining plots illustrating the recurrence conveyances of P_{DC} and Metrological factors.

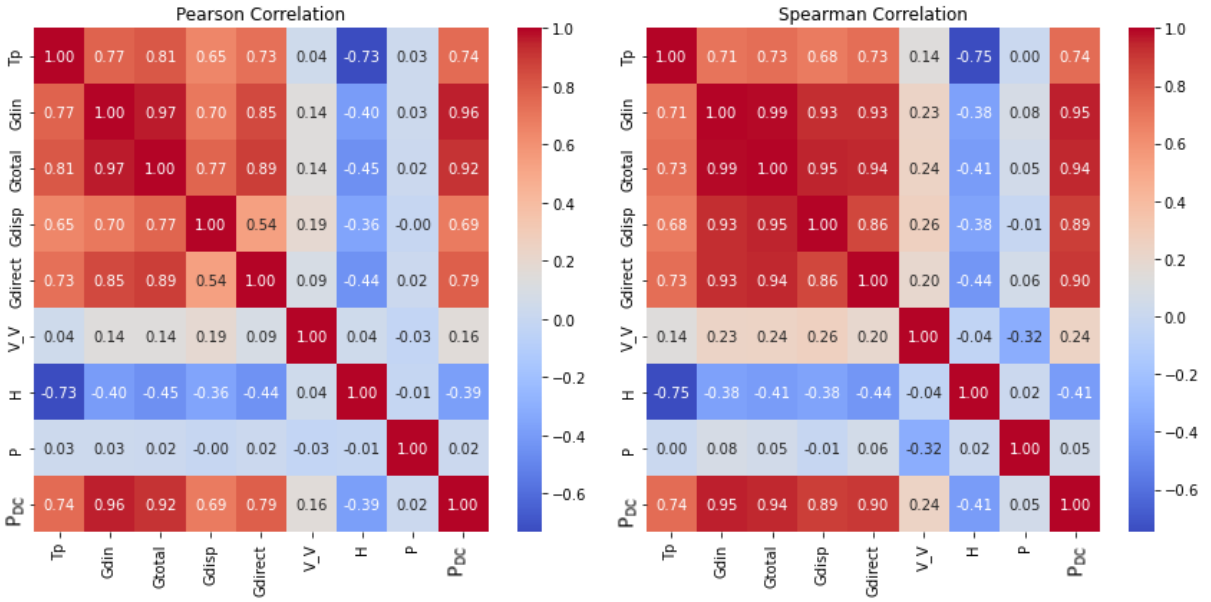

Figure IV.1. Heat map of the outcomes of this correlation analysis.

The correlation matrix provided offers insights into the relations between PV power generation, voltage, or current and various environmental variables. Each cell in the matrix presents the correlation coefficient between two variables, ranging from -1 to 1. The sign of the coefficient indicates the direction of the relationship: "+" denotes a positive correlation, and "-" denotes a negative correlation. A higher absolute value of the correlation coefficient signifies a stronger association between the variables. [163-164]:

Analyzing the correlation matrix, we notice several remarkable patterns. Variables such as incline solar radiation "G_{din}", Total Irradiance "G_{total}" and direct solar radiation "G_{direct}" exhibit strong positive correlations with PV power generation ("P_{DC} "), indicating that higher values of these environmental factors leads to increased PV power generation. Conversely, the variable "H" (representing humidity) demonstrates a notable negative correlation with PV power generation, suggesting that higher humidity levels may lead to decreased PV power output.

Also, a few factors, such as "Tp" (temperature of PV panel) and " G_{disp} " (representing dispersed solar radiation), appear direct positive relationships with PV power generation. These relationships mean that temperature and dispersed solar radiation may also play critical parts in influencing PV power generation, yet to a lesser degree compared to other variables like direct solar radiation ("G_{direct}").

Additionally, certain factors, such as "V_V" (wind speed) and "P" (pressur), display weaker relationships with PV power generation, as shown by their correlation coefficients that is near to zero. Whereas these factors may still have a few impact on PV generation, their affect shows up to be moderately minor compared to other environmental variables.

In general, this relationship examination gives profitable experiences into how different environmental factors relate to PV power generation. Understanding these connections can advise decision-making process related to optimizing the performance of PV system, forecasting energy production, and planning more productive renewable energy system.

After loading the dataset and drop any lines with missing values and removing the outliers. We characterize the target variable P_{DC}. At that point, we compute both Pearson and Spearman correlation coefficients independently with the target variable. We combine theme from both strategies by selecting the maximum absolute value. After that, we channel out highlights whose outright correlation coefficient with target variable is less than or break even with to 0.1. This procedure chooses a subset of the first features that meet the correlation basis. The number of input features remains the same; we do not evacuate any features from the dataset itself but distinguish which features are important based on the relationship threshold. This approach guarantees that we capture significant correlations regardless of the method used. the figure Figure IV.2 demonstrates features choice based on relationship threshold for P_{DC} data ensuring the distinguishing proof of related features vital for precise prediction and analysis in photovoltaic systems.

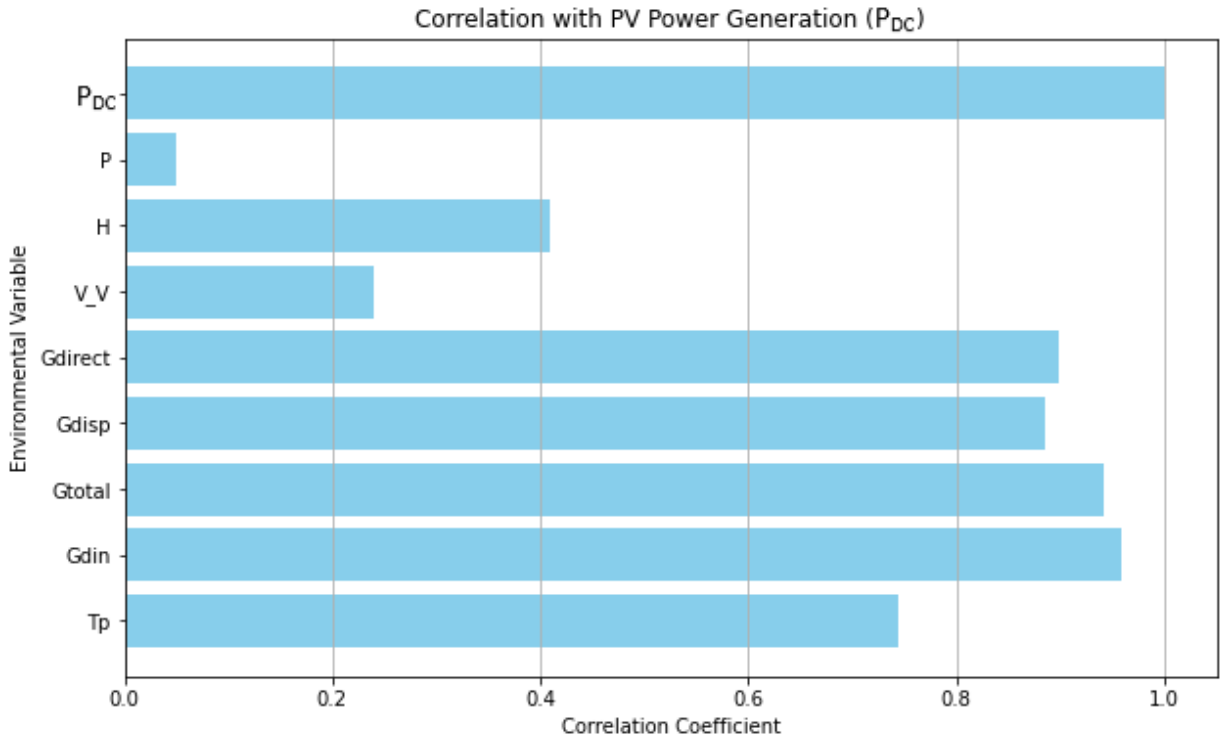

Figure IV.2. Feature Selection based on Correlation Threshold for P_{DC}.

IV.3.3. Enhancing Regression Model Performance: Outlier Detection with Isolation Forest:

Isolation Forest is a popular algorithm used for outlier detection in machine learning. It works by confining anomalies in the dataset instead of modeling the typical data points. This approach is especially compelling for high-dimensional datasets with complex structures. The fundamental rule behind Isolation Forest is that anomalies are regularly uncommon and have properties that make them easy to separate. During the confinement process, inconsistencies are anticipated to be disconnected with less parts compared to typical data points. In this manner, the way length to separate an inconsistency is regularly shorter than that of a ordinary data point. Therefore, the path length to isolate an anomaly is typically shorter than that of a normal data point. By measuring the average path length across multiple isolation trees, Isolation Forest assigns anomaly scores to each data point. Data points with shorter average path lengths are considered more anomalous.

In this work, Isolation Forest is used for outlier detection before training the regression models. Specifically, after loading and preprocessing the dataset, Isolation Forest is used to detect and remove outliers from the dataset using the Isolation Forest class from the sklearn

library. ensemble module. After setting the contamination parameter, representing the anticipated extent of outliers within the dataset. Outlier predictions are then used to filter out the outliers from the original dataset, resulting in a cleaned dataset containing only the corrected data points. In Figures IV.3, we present the distribution of P_{DC}, both before and after the removal of outliers. The X-axis represents PV generation values, while the Y-axis represents the frequency of occurrence. By comparing the two distributions, we gain insights into how the removal of outliers affects the overall distribution of PV generation values.

By removing outliers before training the regression models, Isolation Forest helps improve the strength and performance of the models by reducing the influence of anomalous data points on the training process. This ensures that the models are better able to capture the underlying patterns and relationships in the data, leading to more accurate predictions of PV generation.

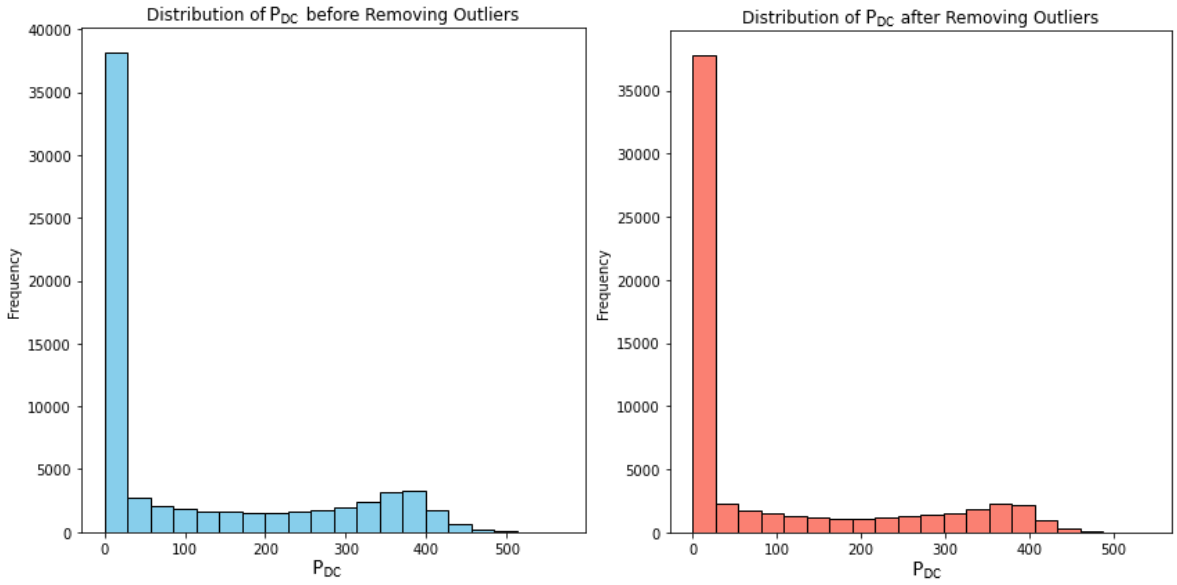

Figure IV.3. Distribution of P_{DC} Before and After Removing Outliers

This visualization enables us to compare the distribution of the target variable 'P_{DC}' before and after outlier removal, offering insights into how outlier removal impacts data distribution. By identifying and eliminating outliers, the Isolation Forest method effectively isolates anomalous data points that could distort the distribution. By excluding these outliers, Isolation Forest ensures that the resulting histograms accurately represent the distribution of normal data points within each bin. This allows for a clearer understanding of the data distribution and the effects of outlier removal on the overall dataset [165].

IV.4. Methodology:

There are numerous strategies based on historical data for predicting PV power. This chapter proposes modeling PV power production using computational methods, rely on historical data from a generation system located in Algeria as described in Section 2. Machine Learning, a vast field within computer science, offers suitable techniques for making predictions. This study aims to explore various Machine Learning techniques and supervised learning models to determine which provides the most accurate estimation of power produced by photovoltaic plants for our dataset. The performance of these methods was evaluated using experimental data. The proposed models hold potential for simulating and implementing similar PV systems in the region, thereby helping to meet energy demand.

The methodology began with thorough data preprocessing to ensure dataset integrity. Missing values were either imputed or removed, and features highly correlated with the target variable were identified using Pearson and Spearman correlation coefficients, then merged. Additionally, outlier detection and removal were performed using the Isolation Forest algorithm to enhance model robustness. The dataset was subsequently split into training and testing sets for model evaluation.

Following data preprocessing, six regression models were selected for evaluation: Random Forest, Support Vector Regression (SVR), Multi-layer Perceptron (MLP), Linear Regressor (LR), Gradient Boosting, and k-Nearest Neighbors (KNN). Each model subjected to hyperparameter optimization using randomized search, which involved tuning various hyperparameters such as the number of estimators, maximum depth, learning rate, kernel type, activation function, and number of neighbors.

Once hyperparameters were tuned, the performance of each model was evaluated using multiple metrics, including Root Mean Squared Error (RMSE), Normalized Root Mean Squared Error (NRMSE), Mean Absolute Error (MAE), and R-squared (R^2). These metrics provided insights into the accuracy, precision, and goodness of fit of the models [166-168].

$$MAE = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i| \tag{IV.5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2} \tag{IV.6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n-1}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n-1}(\bar{y}_i - y_i)^2}, \overline{y} = \sum_{i=0}^{n-1} y_i \quad \text{(IV.7)}$$

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \quad \text{(IV.8)}$$

The results of the model evaluation were then analyzed and compared to identify the best-performing model for predicting PV generation. This analysis helped highlight the strengths and weaknesses of each model and facilitated the selection of the most suitable model for the task at hand.

The methodology employed in this study provided a systematic approach to comparing regression models for predicting PV generation. By following a structured methodology encompassing data preprocessing, model selection, hyperparameter tuning, and model evaluation, valuable insights were gained into the performance of each model, contributing to the advancement of predictive modeling in the renewable energy sector.

After identifying and selecting the best prediction model based on its performance metrics, the next step involves integrating this model into a MATLAB application. This process typically entails exporting the model, along with any necessary preprocessing steps or feature engineering techniques (specially the normalization process), into a format compatible with MATLAB. Once integrated, the model can be deployed within the MATLAB app after converting it into a desktop application using MATLAB App Designer, allowing users to input relevant data and receive predictions or insights based on the model's calculations. This seamless integration facilitates real-time or on-demand predictions within the MATLAB environment, enhancing the usability and accessibility of the predictive model for various applications and users.

The methodology framework, illustrated in Figure IV.4, guides our approach. Initially, we engage in data collection and preprocessing, encompassing database exploration, and normalization, followed by the segmentation of data into training and testing data. In the modeling phase, our objective is to train the chosen algorithms using the training data until we obtain a satisfactory model. To achieve this, we employ a randomized search algorithm to identify the best-performing model. Finally, the last stage entails evaluating the models using testing data, calculating estimation errors, then we save the best model along with its scaler. Additionally, we incorporate k-fold cross-validation in the training process with a fold size of 5 to enhance the robustness of our evaluations.

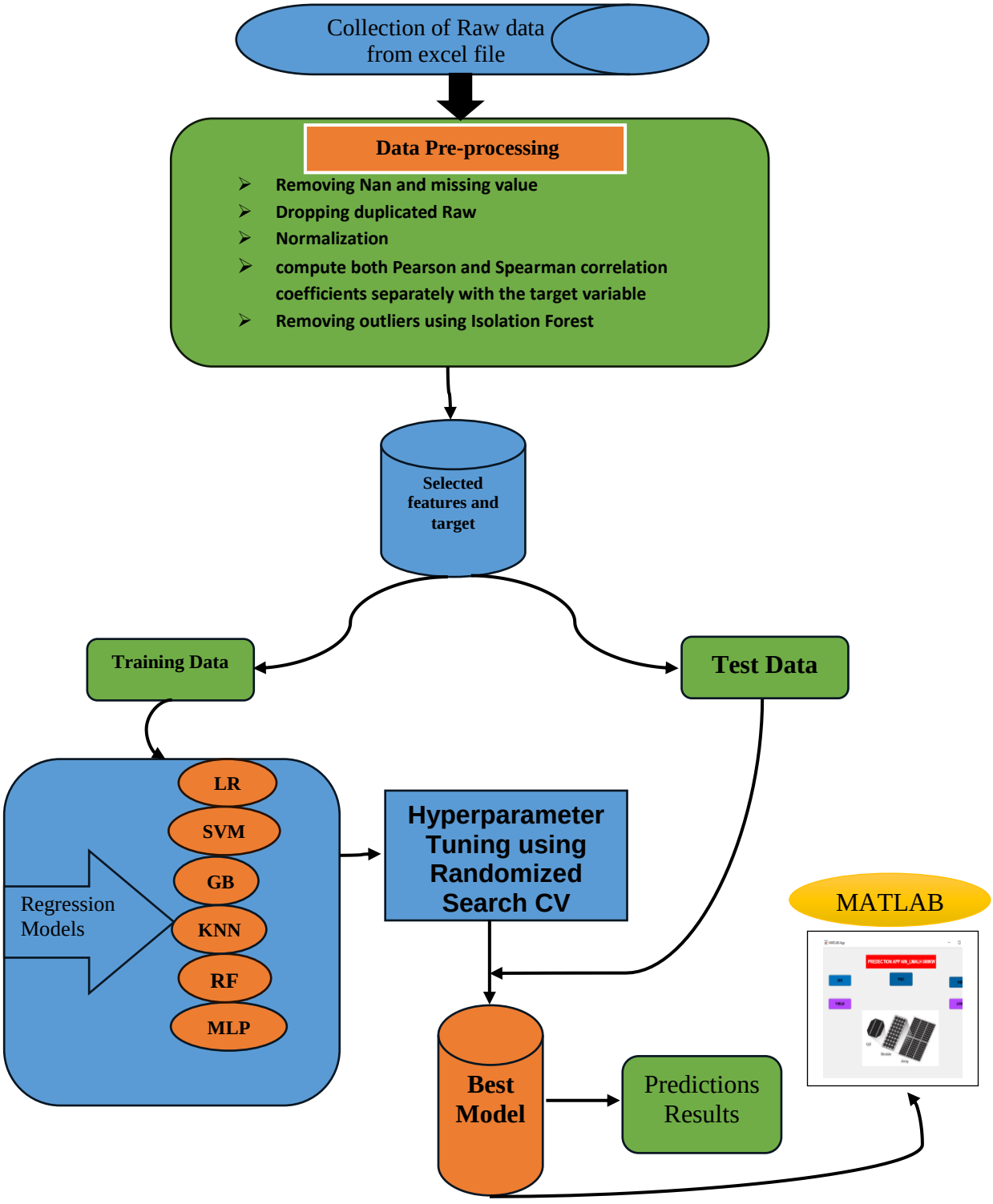

Figure IV.4. Methodology framework.

IV.4.1. Regression Models:

Various regression models have been used to predict power production from the data. The performance of each model is evaluated using MAE, MSE, RMSE, R^2 error, explained variance score, and prediction plots.

In predicting PV power generation, several regression models have been employed to optimize accuracy and efficiency. Among these models: k-Nearest Neighbors (KNN) [169], Support Vector Machines (SVR) [170], Random Forest (RF) [171], Linear Regression (LR) ,Multi-layer Perceptron (MLP) and Gradient Boosting (GB) regression [172]. KNN uses the principle of similarity to predict output values based on the characteristics of neighboring data points. LR establishes a linear relationship between input and output variables, assuming a linear correlation. SVM constructs hyperplanes in a high-dimensional space to classify data points, making it versatile for regression tasks. RF, employing an ensemble of decision trees, provides robust predictions by averaging multiple estimators. GB regression enhances predictive performance through sequential training of weak learners, iteratively minimizing errors.

Each model offers unique advantages, from KNN's simplicity to RF's resilience against overfitting and GB's boosting capabilities. By comparing and contrasting these approaches, researchers can identify the most suitable regression model for PV power prediction tasks, ensuring accurate and reliable results.

IV.4.2. Hyperparameter optimization using Randomized Search CV and Evaluation Metrics

Hyperparameters, which are set before the learning process begins, significantly influence a model's performance. Adjusting these parameters is crucial for optimizing model effectiveness, often requiring the exploration of various combinations. Techniques such as Grid Search and Random Search have been developed for hyperparameter optimization. Grid Search systematically explores a defined subset of the hyperparameter space, evaluating each combination using cross-validation on the training data, as demonstrated in previous work [173-174] from chapter II and chapter III. In contrast, the Random Search Algorithm, also known as the Monte Carlo method or stochastic algorithm [175], operates by iteratively sampling parameter settings from a specified distribution and evaluating the model using cross-validation [176]. Unlike Grid Search, which tests all parameter values, Random Search samples a number of settings. Random Search demonstrates more efficient performance than

Grid Search, as it avoids allocating excessive trials for less important dimensions to optimize the hyperparameters for the all used models [177]. To optimize the hyperparameters for all the models used in this research, hyperparameter tuning is performed using a randomized search algorithm. The Randomized Search CV function from the sci-kit-learn library is implemented for this purpose [178]. Randomized SearchCV randomly selects hyperparameters and evaluates the results through cross-validation, where the data is divided into training and validation subsets. This study employs 5-fold cross-validation to achieve a robust model.

Cross-validation refers to techniques used to assess a predictive model's performance by splitting data into multiple subsets. K-fold cross-validation is the most common form, where the data is randomly partitioned into k equal folds. The model is trained on k-1 folds and tested on the remaining fold, repeating the process k times to ensure each fold serves as the test set once. The results from each fold are averaged to provide an overall performance estimate. Figure IV.5 illustrates the process of the 5-fold cross-validation technique used in this study.

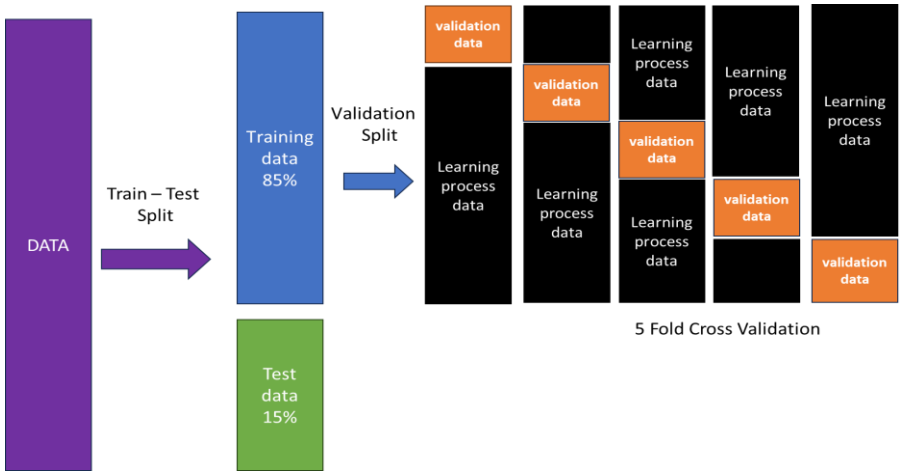

Figure IV.5. Process of the used cross-validation technique with 5-fold cross-validation.

IV.5. Development of a MATLAB App for Power Prediction:

In predictive analytics for power systems, the integration of Python-based machine learning models with MATLAB's versatile application framework (App Designer) signifies a new level of efficiency and accuracy. This work focuses on designing and implementing a user-friendly MATLAB application customized for power prediction tasks. The application interface, created using MATLAB's intuitive App Designer tool, allows for easy interaction

and seamless integration with underlying algorithms. Users can input relevant data, select prediction parameters, and visualize both measured and predicted results in real-time.

Key features of the developed application include:

- **Tab-Based Interface:** The application is organized into tabs corresponding to different prediction tasks, such as predicting power demand, voltage, and current.
- **Interactive Controls:** Users can interact with various components such as buttons, state buttons, and toggle buttons to initiate prediction tasks and customize parameters.
- **Visualization Tools:** Graphical representations, including UIAxes components, facilitate the visualization of measured and predicted data, aiding in the analysis and interpretation of results.
- **Export Functionality:** The application allows users to export prediction results for further analysis or integration with external systems.
- **Streamlined Integration:** The generated Excel file from the PV station can be used directly for real-time prediction without any preprocessing required.

The designed MATLAB app offering a user-friendly interface with distinct tabs catering to various prediction types such as Power PV generation (P_{DC}), Voltage PV generation (V_{DC}), Current PV generation (I_{DC}), Yield, and Loss calculations is shown in Figure IV.6.

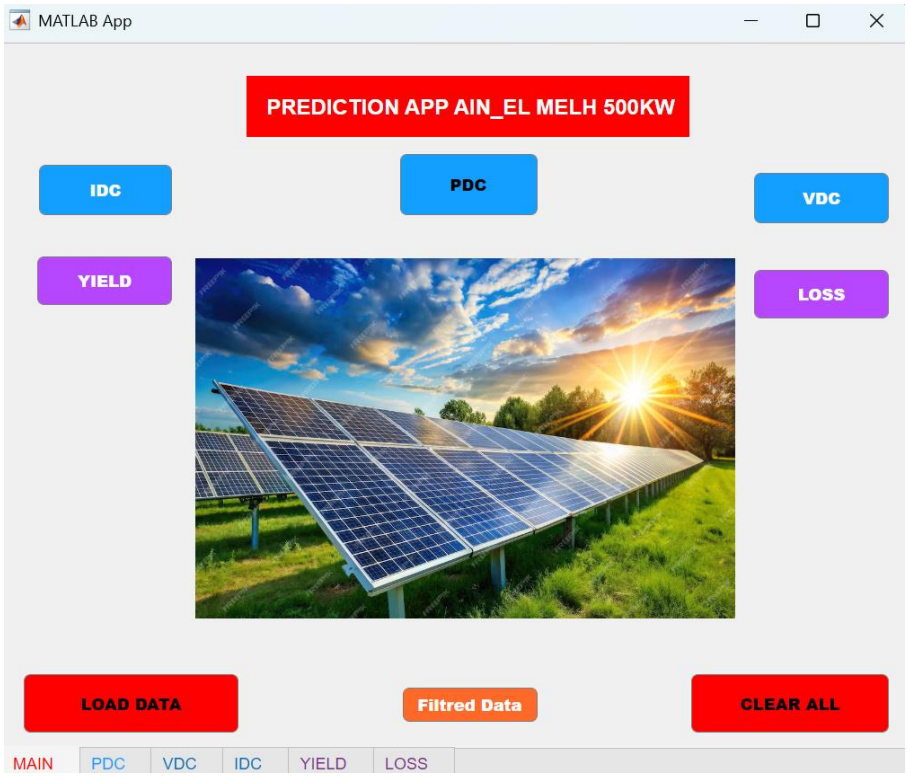

Figure IV.6. Main page of the MATLAB App.

Each tab in the interface is meticulously crafted with intuitive functionality. It features clear visualization through UI Axes and streamlined operations with buttons for tasks like clearing data, triggering predictions, and exporting results.

Additionally, in the yield tab and loss tab, as illustrate in Figure IV.7, we perform the following calculations:

- **Reference Yield (Yr for Measured or actual and YR for predicted):** This is the time that the sun must be shining with G0 = 1kW/m² to radiate the energy Ht to the PV array of the PV module.

$$Reference\ Yield = Ht / G0.$$

- **Array Efficiency (Ya for Measured or actual and YA for predicted):** This is the time that the PV system needs to work at the nominal power of the PV array P0 to produce the output DC energy EDC.

$$Array\ Efficiency = EDC / P0.$$

- **Final Yield (Yf for Measured or actual and YF for predicted):**

This is the time that the PV system needs to operate at the nominal power of the PV array P0 to produce the output AC energy EAC.

$$Final\ Yield = EAC/P0.$$

- **System Losses (Ls for Measured or actual and LS for predicted):**

$$System\ Losses = Array\ Efficiency - Final\ Yield$$

- **Array Capture losses:** (Lc for Measured or actual and LC for predicted)

$$Array\ Capture\ losses = Reference\ Yield - Array\ Efficiency$$

- **Performance ratio (Pr for Measured or actual and PR for predicted):**

Represents the ratio between the effective energy EAC and the energy that would be generated from an ideal, lossless PV installation, assuming a 25ºC solar cell temperature with the same radiation level.

$$Performance\ ratio = (Final\ Yield / Reference\ Yield) \times 100.$$

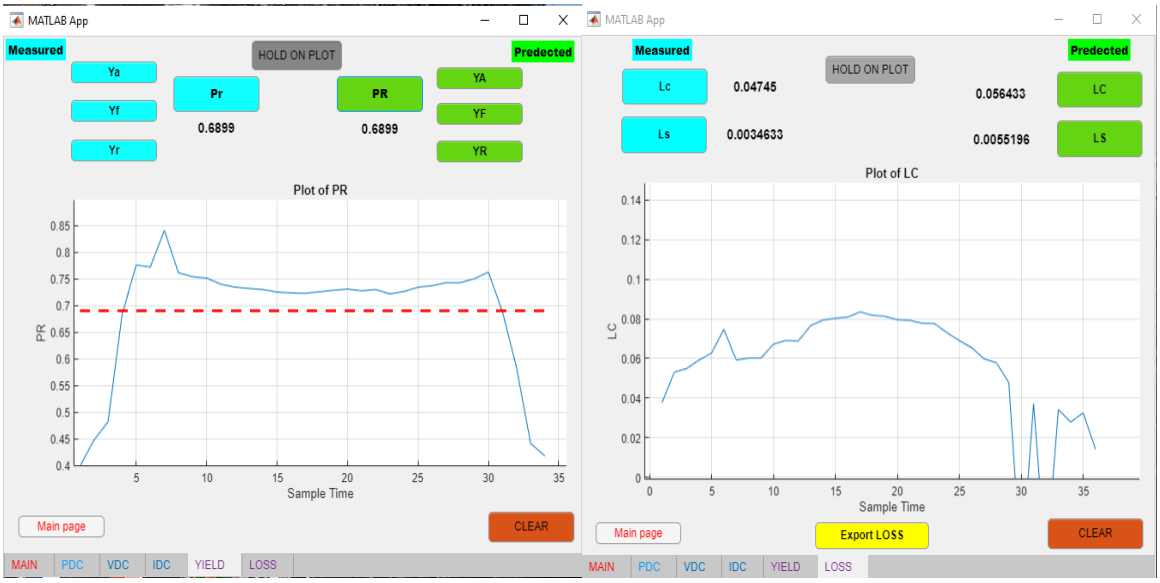

Figure IV.7. Yield and Loss Tab of the designed MATLAB App

The "hold on plot" button enables users to maintain the UI axes, allowing for the simultaneous plotting of two or more graphs for comparison purposes. Every aspect of the design is equipped with buttons to ensure clarity and ease of use for any user, enhancing accessibility. Whether predicting P_{DC} loads, I_{DC} current, analyzing V_{DC} voltages, or calculating losses, this MATLAB app empowers users with seamless integration of machine learning models, paving the way for informed decision-making and optimized performance in power system management. The development and utilization of such an application for power prediction offer a valuable tool for researchers and practitioners in the field of power systems analysis and management.

IV.6. Obtained Results:

This section displays the findings of our technique, as well as the datasets utilized to demonstrate the predicted outcomes of PV system generation under various weather situations. Furthermore, the findings received from the MATLAB program are displayed in different graphics. These findings provide the performance metrics for several regression models using P_{DC} datasets. The picture also includes a full comparison of the performance of several regression models across multiple metrics for the P_{DC} dataset. Figures IV.8 and Figures IV.9 show a comparison of the measured and predicted P_{DC} plots using RF.

P_{DC}, I_{DC} and V_{DC}

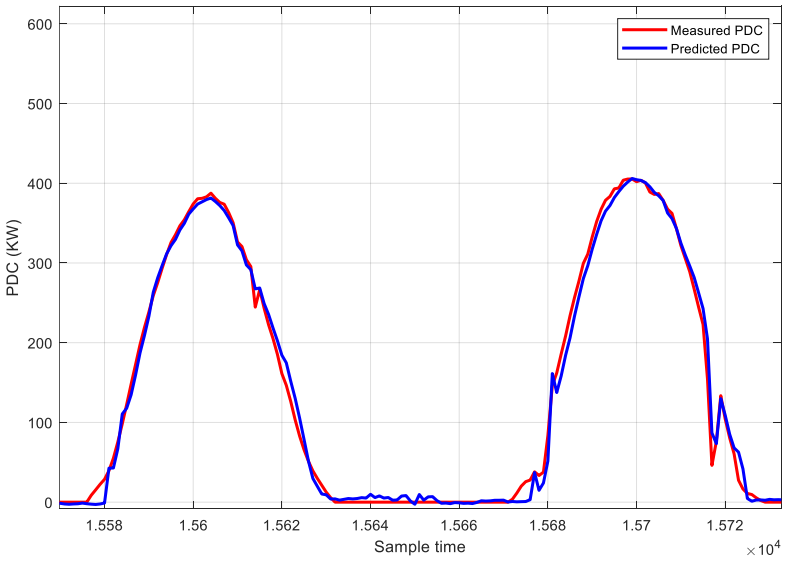

Figure IV.8. Random Forest predictions across the test datasets for P_{DC}.

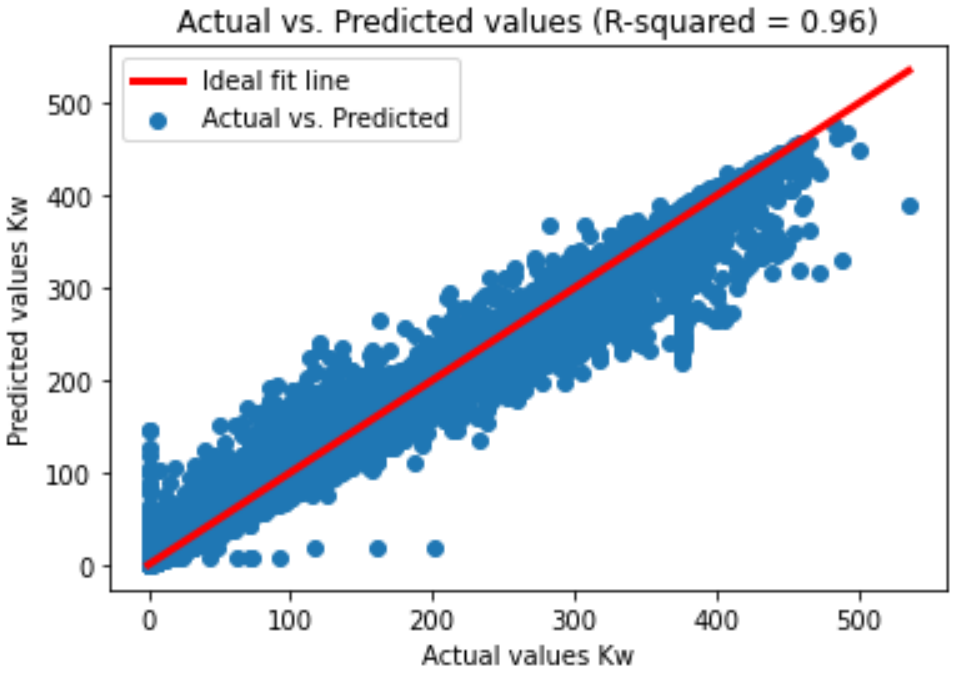

Figure IV.9. Actual and predicted plots using RF for P_{DC}.

Figure IV.10 and Table IV.4 show the comparative analysis of machine learning algorithms for predicting PV power outputs.

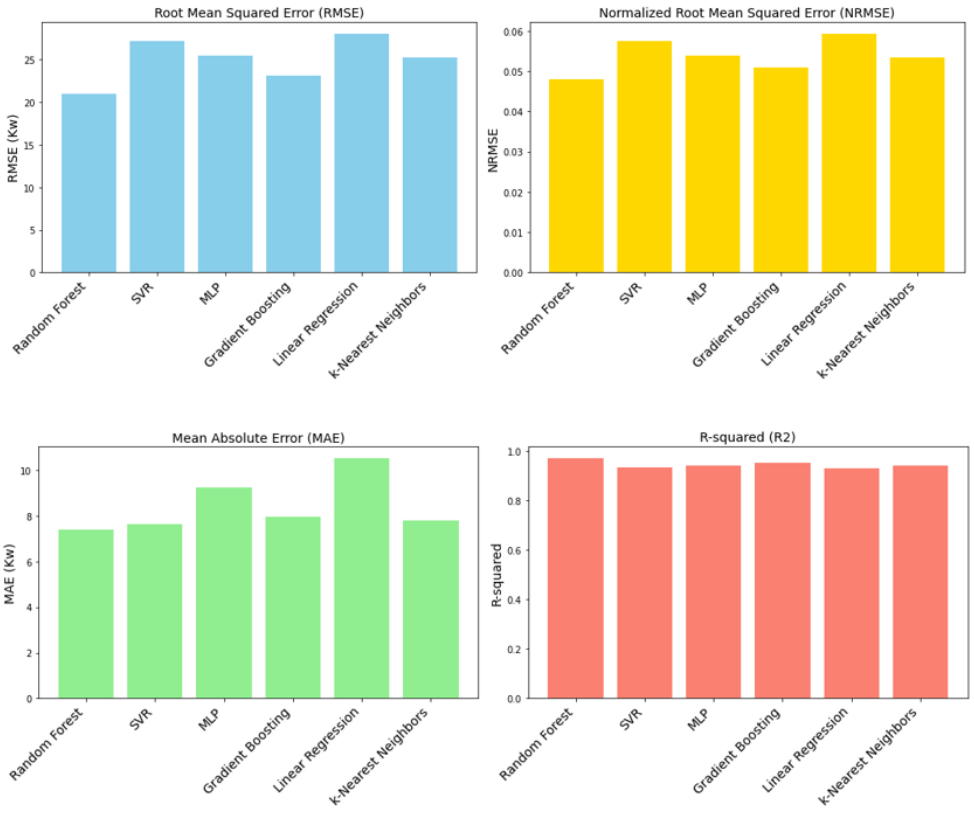

Figure IV.10. Error metrics of PV power outputs for the different machine learning algorithms used.

Table IV.4. Comparative Analysis of Machine Learning Algorithms for Predicting PV Power Outputs.

Models	**RMSE (Kw)**	**NRMSE**	**MAE (Kw)**	**R-squared**
Random Forest	21.02	0.048	7.40	0.968
SVR	27.1202	0.0574	7.6380	0.9319
MLP	25.4615	0.0539	9.2456	0.9400
Gradient Boosting	23.1536	0.0510	7.9418	0.9504
Linear Regression	27.9645	0.0592	10.5077	0.9276
k-Nearest Neighbors	25.2593	0.0535	7.7948	0.9409

From Table IV.4, Figure IV.8, Figure IV.9, and Figure IV.10, the comparative analysis of machine learning algorithms for predicting PV power outputs reveals that Random Forest

emerges as the top-performing model across all metrics. It boasts the lowest RMSE of 21.02, highest R-squared of 0.968, and the smallest MAE and NRMSE values of 7.40 and 0.048, respectively, indicating its superior predictive accuracy and robustness. Gradient Boosting also demonstrates competitive performance, particularly in terms of R-squared (0.9504) and MAE (7.9418). Conversely, Linear Regression and SVR exhibit comparatively poorer performance. Linear Regression yields an RMSE of 27.9645, R-squared of 0.9276, MAE of 10.5077, and NRMSE of 0.0592, while SVR produces an RMSE of 27.1202, R-squared of 0.9319, MAE of 7.6380, and NRMSE of 0.0574. These findings suggest that for accurate predictions of PV power outputs, leveraging Random Forest or Gradient Boosting models would be most beneficial, offering superior predictive capabilities over alternative algorithms.

We performed the same study and used the same methods for the P_{DC} dataset for both I_{DC} and V_{DC}. Additionally, for the V_{DC} dataset, we included a filter to exclude all values where the PV generator was not working for better training performance. Below are the results for P_{DC}, I_{DC}, and V_{DC} datasets, presenting various metrics and relevant parameters in Table IV.5. The results presented in Figure IV.11 and Figure IV.12 pertain to the Random Forest predictions across the test datasets for I_{DC} and V_{DC}, respectively.

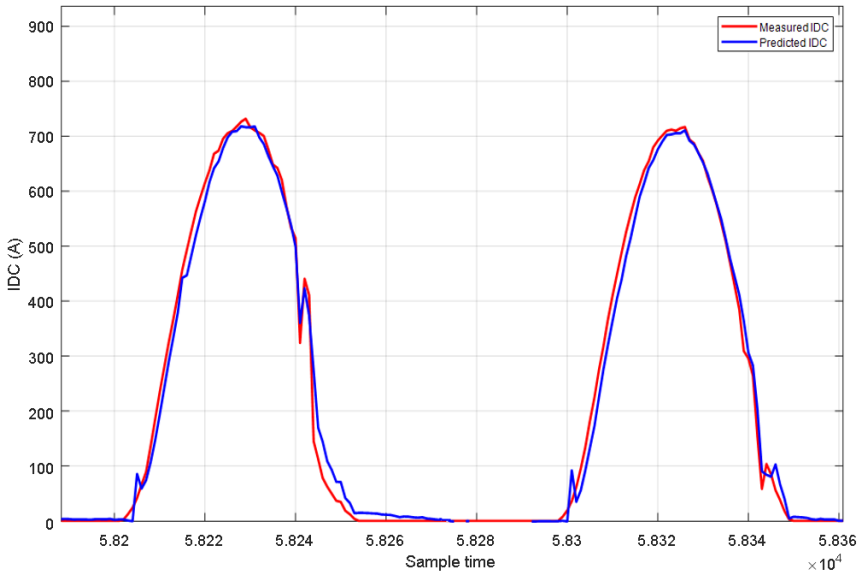

Figure IV.11. Random Forest predictions across the test datasets for I_{DC}.

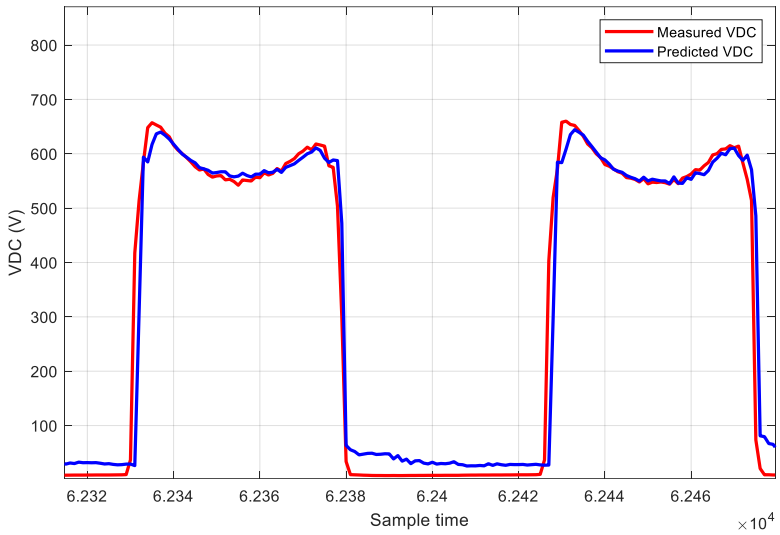

Figure IV.12. Random Forest predictions across the test datasets for V_{DC}.

Table IV.5. Optimization Results and Performance Evaluation of Machine Learning Models for Power Distribution Predictions.

Dataset	**Parameter**	**Value**
P_{DC}	Best Parameters	{'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}
	Best RMSE	19.413
	NRMSE	0.034
	MAE	6.677
	R-squared	0.968
I_{DC}	Best Parameters	{'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}
	Best RMSE	24.499
	NRMSE	0.0476
	MAE	8.089
	R-squared	0.957
V_{DC}	Best Parameters	{'max_depth': 30, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 150}
	RMSE	11.691
	NRMSE	0.060
	MAE	7.424
	R-squared	0.953

Figure IV.13 shows the comparison between actual and predicted values for V_{DC} by using RF. The same information is presented in Figure IV.14 for I_{DC}.

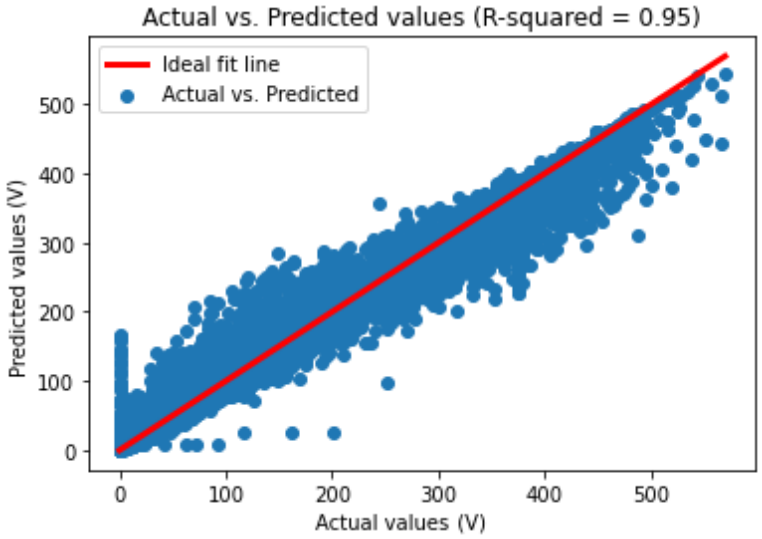

Figure IV.13. Actual and predicted plots using RF for V_{DC}.

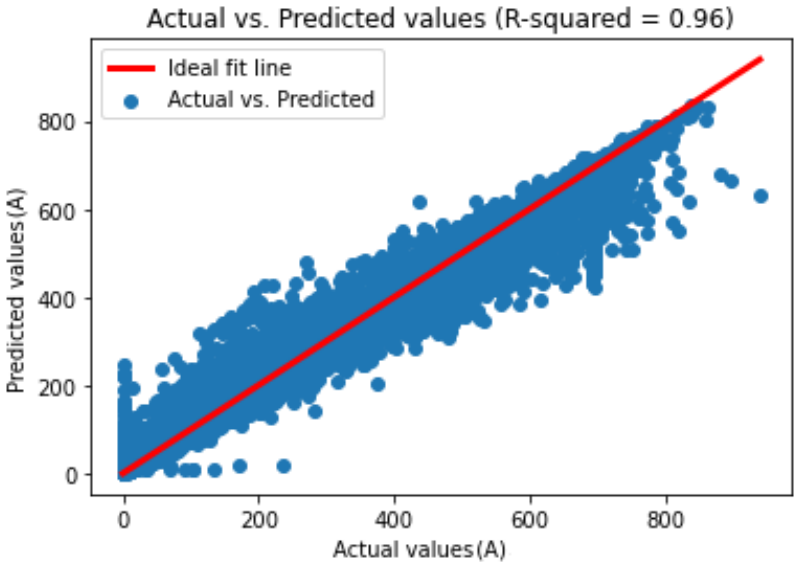

Figure IV.14. Actual and predicted plots using RF for I_{DC}.

The conducted analysis highlights the proficiency of the Random Forest Regressor model in discerning the complex interactions between environmental factors and PV system performance. For the P_{DC} dataset, the Random Forest model demonstrated impressive results with optimal parameters {'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}, achieving a best RMSE of 21.02 kW, NRMSE of 0.048%, MAE of 7.40 kW, and an R-squared value of 0.968. Similarly, in I_{DC} prediction, using the same parameters,

the Random Forest model achieved strong performance with a best RMSE of 24.499 kW, NRMSE of 0.0476, MAE of 8.089 kW, and R^2 of 0.957, effectively capturing I_{DC}'s complexity despite variability due to various influences. For the V_{DC} dataset, the Random Forest model, optimized with {'max_depth': 30, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 150}, showed excellent performance, with an RMSE of 11.691 kW, NRMSE of 0.060%, MAE of 7.424 kW, and R^2 of 0.953.

The analysis illustrated that each regression model's performance varied. The Random Forest model delivered strong results with low RMSE, NRMSE, and MAE, coupled with a high R^2 score. The SVR also yielded promising outcomes, particularly with specific kernel types and regularization settings. The MLP displayed adaptability with various activation functions and hidden layer configurations, though it required careful tuning to avoid overfitting. Both Gradient Boosting and k-Nearest Neighbors exhibited moderate performance, potentially benefiting from further optimization or feature engineering. The outcomes derived from the MATLAB app, employing the random forest regressor trained models to predict under diverse weather conditions for P_{DC}, I_{DC}, and V_{DC} respectively, are illustrated in Figures IV.15, IV.16, and IV.17 for a cloudy day, and in Figures IV.18, IV.19, and IV.20 for a clear day.

- **Interface result Clair day:**

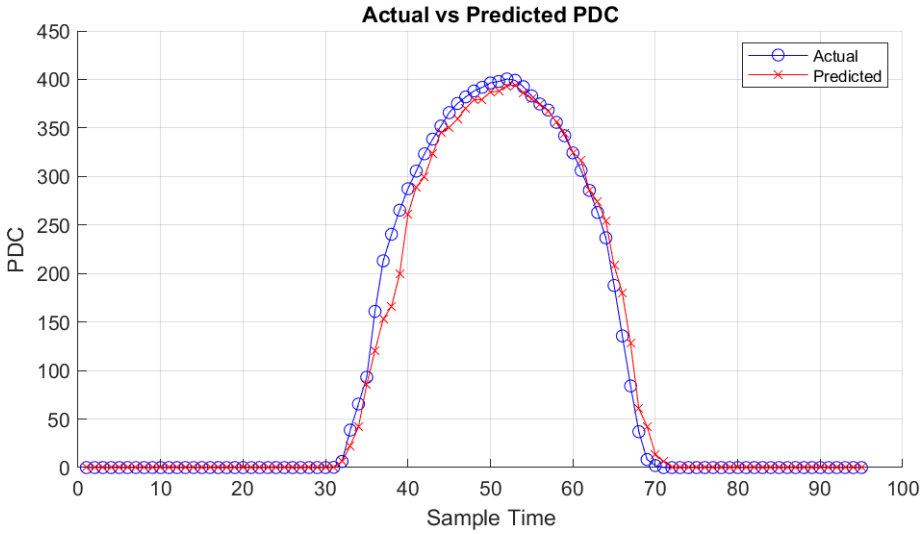

Figure IV.15. P_{DC} prediction results obtained from the MATLAB app for Clair day.

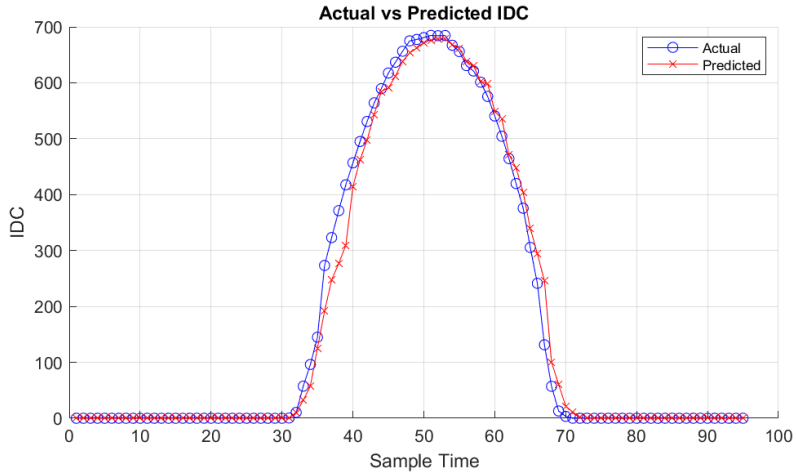

Figure IV.16. I_{DC} prediction results obtained from the MATLAB app for Clair day.

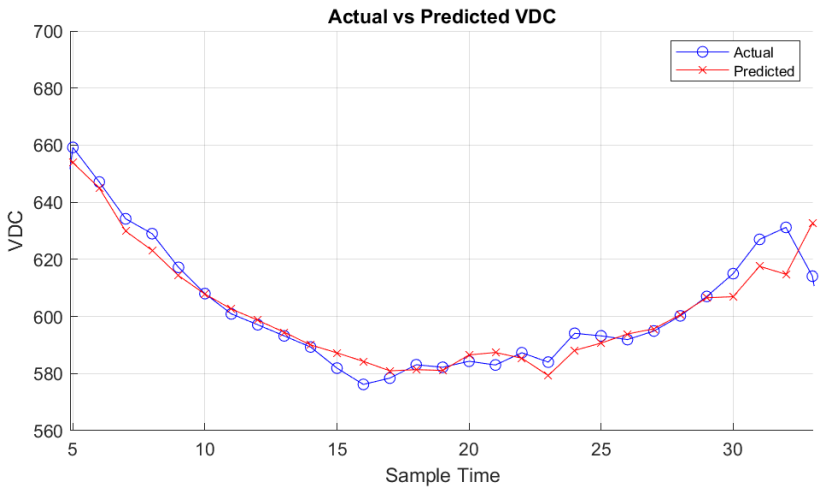

Figure IV.17. V_{DC} prediction results obtained from the MATLAB app for Clair day.

- **Interface result cloudy day:**

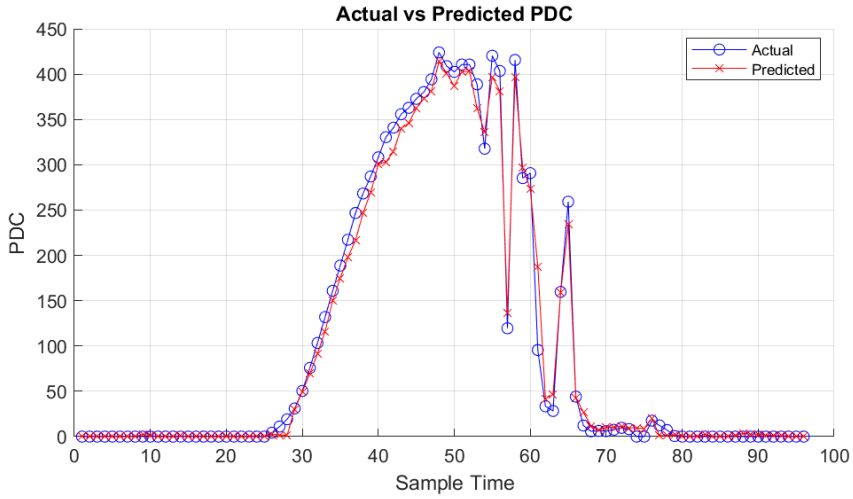

Figure IV.18. P_{DC} prediction results obtained from the MATLAB app for cloudy day.

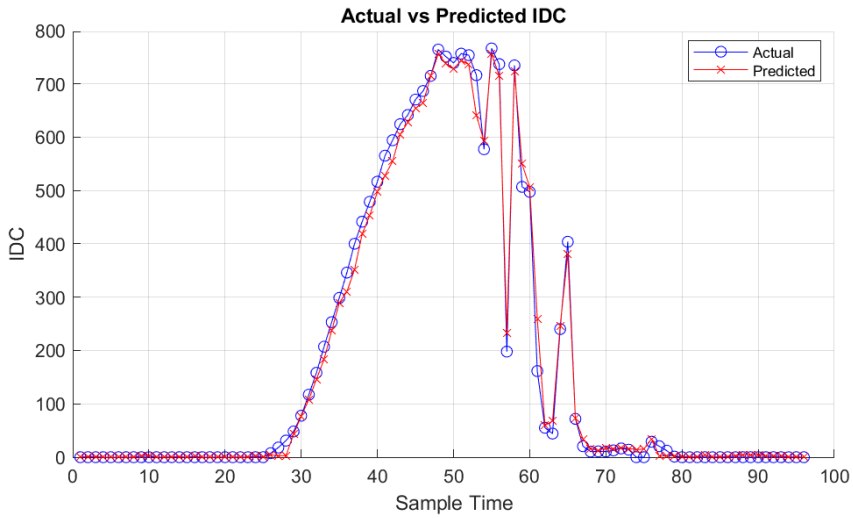

Figure IV.19. I_{DC} prediction results obtained from the MATLAB app for cloudy day.

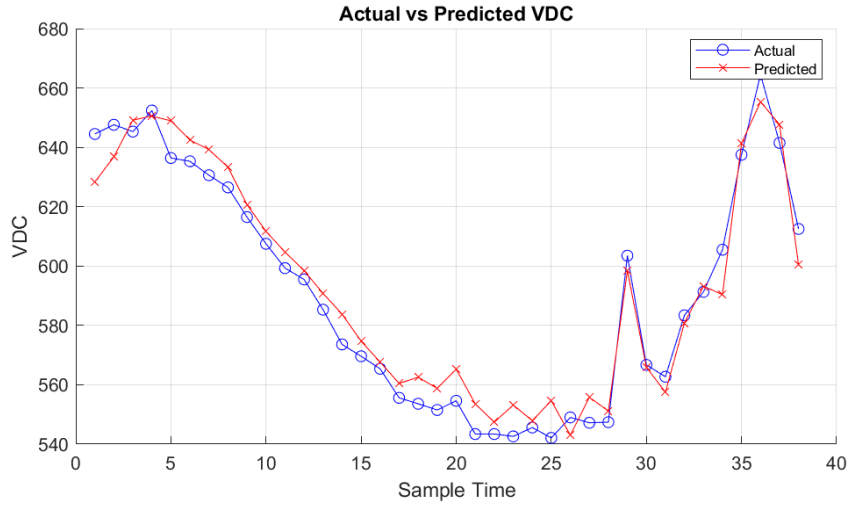

Figure IV.20. V_{DC} prediction results obtained from the MATLAB app for cloudy day.

The MATLAB application's results show case significant success in predicting PV system generation across diverse weather conditions. Figures IV.15 to IV.20 depict predictive outcomes for both cloudy and clear days. A notable advantage of this approach is the meticulous selection of relevant features, achieved through Pearson and Spearman correlation analyses. By computing these correlation coefficients, the method ensures a comprehensive understanding of the relationships between environmental variables and PV system generation, enhancing model interpretability and performance. Integration of Isolation Forest for outlier detection enables robust data preprocessing, effectively improving model generalization. Furthermore, Randomized SearchCV facilitates efficient hyperparameter tuning, with Random Forest emerging as the best-performing model due to its ensemble nature, ability to handle non-linear relationships, versatility, scalability, and resilience to overfitting. Integration of the trained Random Forest model into the MATLAB app enables efficient management of PV systems, aiding in resource allocation and decision-making.

IV.7. Conclusion:

This Chapter highlights machine learning's efficiency in accurately predicting photovoltaic (PV) power generation, particularly in Algeria. Random Forest emerges as the top model, with meticulous data preprocessing, feature selection, and model evaluation contributing to its selection. Historical data and computational methods yield impressive performance metrics (low RMSE of 19.413, high R^2 of 0.968), emphasizing the importance of feature selection and outlier detection.

Integration of the top model into a MATLAB application for real-time predictions enhances usability and accessibility in renewable energy. This end contribute significantly to advancing predictive modeling techniques for PV systems, offering valuable insights for simulating and implementing similar systems to address energy demands and promote sustainability. Deep learning and hybrid techniques offer avenues for further enhancing prediction accuracy and robustness. Additionally, incorporating more weather data, like cloud cover, could improve predictive capabilities, especially in regions with variable weather patterns.

GENERAL CONCLUSION:

Conclusion:

This thesis has explored innovative methodologies and techniques to enhance the reliability, efficiency, and performance of photovoltaic (PV) systems, crucial for their integration into sustainable energy grids. Through a multi-faceted approach spanning fault detection, diagnosis, and predictive modeling, significant advancements have been made in addressing key challenges faced by PV installations.

- **In Chapter Two**, a robust fault detection and diagnosis framework leveraging machine learning, specifically the Random Forest Classifier (RFC), was introduced. This approach, underpinned by a precise one-diode model (ODM) simulation, showcased exceptional accuracy rates in detecting and categorizing various faults, thereby enabling effective performance monitoring and maintenance strategies.
- **In Chapter Three** delved into the realm of deep learning, presenting a groundbreaking approach to fault detection and diagnosis in PV systems. By integrating Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (Bi-GRU) architectures, this methodology achieved remarkable accuracy in identifying and classifying PV faults, paving the way for targeted mitigation measures and improved system reliability.
- **In Chapter Four**, the focus shifted towards predictive modeling of PV power generation, emphasizing the importance of meticulous data preprocessing and feature selection. Leveraging machine learning techniques, particularly Random Forest, precise predictions of PV output were attained, offering valuable insights into optimizing system performance and addressing energy demands.

The integration of Python-trained models into a MATLAB interface not only enhances prediction accuracy but also facilitates real-time decision-making and system optimization, bridging the gap between research and practical implementation.

Looking forward, opportunities exist for further enhancing prediction accuracy and robustness through the exploration of deep learning and hybrid techniques. Additionally, the incorporation of additional weather data, such as cloud cover, holds promise for improving predictive capabilities, especially in regions with variable weather patterns.

Despite the promising results for fault detection and diagnosis, the strategies carried out in this thesis raise a number of questions and provide some directions for future works. In particular, the following points merit serious consideration:

1. **Integration of Real-world Data:** While the fault detection and diagnosis methodologies demonstrated exceptional accuracy in simulated environments, the integration of real-world data from operational PV systems could provide valuable insights into the performance and scalability of these techniques. Collaborating with industry partners or deploying pilot projects to collect and analyze real-time data can validate the effectiveness of the proposed methodologies in practical scenarios.
2. **Enhancement of Fault Classification:** Further research is warranted to enhance the fault classification capabilities of the developed methodologies. Exploring advanced machine learning algorithms, ensemble techniques, or hybrid approaches could improve the robustness and generalizability of fault detection and diagnosis systems, particularly in complex fault scenarios or dynamic operating conditions.
3. **Adaptation to Regional Variability:** PV systems operate in diverse geographical regions, each characterized by unique environmental conditions and operational challenges. Consideration should be given to adapting the developed methodologies to address regional variability, such as variations in solar irradiance, temperature profiles, or grid infrastructure, ensuring their applicability across different geographical contexts.
4. **Integration of Multi-modal Data:** Explore the integration of multi-modal data sources, including sensor data, satellite imagery, and weather forecasts, to enrich the fault detection and diagnosis process. By leveraging complementary data sources, the accuracy and reliability of fault detection systems can be further improved, enhancing their effectiveness in identifying and mitigating system faults.
5. **Robustness to External Factors:** Investigate the robustness of fault detection and diagnosis methodologies to external factors such as environmental noise, sensor inaccuracies, or system uncertainties. Developing techniques to mitigate the impact of these factors and enhance the resilience of fault detection systems will be crucial for their practical deployment in real-world PV installations.
6. **Scalability and Computational Efficiency:** Considerations should be given to the scalability and computational efficiency of the developed methodologies, particularly in large-scale PV installations or resource-constrained environments. Exploring

techniques for optimizing model performance, reducing computational complexity, and enhancing scalability will be essential for facilitating widespread adoption of fault detection and diagnosis systems.

In conclusion, while significant advancements have been made in fault detection, diagnosis, and predictive modeling for PV systems, there remain important avenues for future research and development. By addressing these challenges and opportunities, we can further enhance the reliability, efficiency, and performance of PV installations, accelerating the transition towards sustainable energy grids. This thesis underscores the significance of continuous innovation and interdisciplinary approaches in advancing PV system reliability and efficiency. By proposing actionable solutions to key challenges, this research contributes to the ongoing transition towards sustainable energy sources, driving progress towards a greener and more resilient future.

REFFERENCES

References:

[1]. Tripling Renewable Power and Doubling Energy Efficiency by 2030: Crucial Steps towards 1.5 °C. Available online: https://www.irena.org/Publications/2023/Oct/Tripling-renewable-power-and-doubling-energy-efficiency-by-2030 (accessed on 27 November 2023).

[2]. Europe, S.P. Global Market Outlook for Solar Power 2023–2027; Technique Report; European Photovoltaic Industry Association: Brussels, Belgium, 2023; Available online: https://www.solarpowereurope.org/insights/market-outlooks/global-market-outlook-for-solar-power-2023-2027-1 (accessed on 27 November 2023).

[3]. IEA—International Energy Agency—IEA. Available online: https://www.iea.org/reports/worldenergy-outlook-2023 (accessed on 27 November 2023).

[4]. Marion, B., Schaefer, R., Caine, H., & Sanchez, G. (2013). Measured and modeled photovoltaic system energy losses from snow for Colorado and Wisconsin locations. Solar Energy, 97, 112-121.

[5]. Potnuru, S. R., Pattabiraman, D., Ganesan, S. I., & Chilakapati, N. (2015). Positioning of PV panels for reduction in line losses and mismatch losses in PV array. Renewable energy, 78, 264-275.

[6]. Pillai, D. S., & Rajasekar, N. (2018). Metaheuristic algorithms for PV parameter identification: A comprehensive review with an application to threshold setting for fault detection in PV systems. Renewable and Sustainable Energy Reviews, 82, 3503-3525.

[7]. Daliento, S., Chouder, A., Guerriero, P., Pavan, A. M., Mellit, A., Moeini, R., & Tricoli, P. (2017). Monitoring, diagnosis, and power forecasting for photovoltaic fields: A review. International Journal of Photoenergy, 2017.

[8]. Hariharan, R., Chakkarapani, M., Ilango, G. S., & Nagamani, C. (2016). A method to detect photovoltaic array faults and partial shading in PV systems. IEEE Journal of Photovoltaics, 6(5), 1278-1285.

[9]. Chouder, A., & Silvestre, S. (2010). Automatic supervision and fault detection of PV systems based on power losses analysis. Energy conversion and Management, 51(10), 1929-1937.

[10]. Silvestre, S., Kichou, S., Chouder, A., Nofuentes, G., & Karatepe, E. (2015). Analysis of current and voltage indicators in grid connected PV (photovoltaic) systems working in faulty and partial shading conditions. Energy, 86, 42-50.

[11]. Drews, A., De Keizer, A. C., Beyer, H. G., Lorenz, E., Betcke, J., Van Sark, W. G. J. H. M., ... & Heinemann, D. (2007). Monitoring and remote failure detection of grid-connected PV systems based on satellite observations. Solar energy, 81(4), 548-564.

[12]. Garoudja, E., Harrou, F., Sun, Y., Kara, K., Chouder, A., & Silvestre, S. (2017). Statistical fault detection in photovoltaic systems. Solar Energy, 150, 485-499.

[13]. Dhimish, M., Holmes, V., Mehrdadi, B., & Dales, M. (2018). Comparing Mamdani Sugeno fuzzy logic and RBF ANN network for PV fault detection. Renewable energy, 117, 257-274.

[14]. Momeni, H., Sadoogi, N., Farrokhifar, M., & Gharibeh, H. F. (2019). Fault diagnosis in photovoltaic arrays using GBSSL method and proposing a fault correction system. IEEE Transactions on Industrial Informatics, 16(8), 5300-5308.

[15]. Yi, Z., & Etemadi, A. H. (2016). Fault detection for photovoltaic systems based on multi-resolution signal decomposition and fuzzy inference systems. IEEE transactions on smart grid, 8(3), 1274-1283.

[16]. Leva, S., Mussetta, M., & Ogliari, E. (2018). PV module fault diagnosis based on microconverters and day-ahead forecast. IEEE Transactions on Industrial Electronics, 66(5), 3928-3937.

[17]. Bendary, A. F., Abdelaziz, A. Y., Ismail, M. M., Mahmoud, K., Lehtonen, M., & Darwish, M. M. (2021). Proposed ANFIS based approach for fault tracking, detection, clearing and rearrangement for photovoltaic system. Sensors, 21(7), 2269.

[18]. Madeti, S. R., & Singh, S. N. (2018). Modeling of PV system based on experimental data for fault detection using kNN method. Solar Energy, 173, 139-151.

[19]. Eskandari, A., Milimonfared, J., & Aghaei, M. (2020). Line-line fault detection and classification for photovoltaic systems using ensemble learning model based on IV characteristics. Solar Energy, 211, 354-365.

[20]. Kapucu, C., & Cubukcu, M. (2021). A supervised ensemble learning method for fault diagnosis in photovoltaic strings. Energy, 227, 120463.

[21]. Adhya, D., Chatterjee, S., & Chakraborty, A. K. (2022). Performance assessment of selective machine learning techniques for improved PV array fault diagnosis. Sustainable Energy, Grids and Networks, 29, 100582.

[22]. Akram, M. N., & Lotfifard, S. (2015). Modeling and health monitoring of DC side of photovoltaic array. IEEE Transactions on Sustainable Energy, 6(4), 1245-1253.

[23]. Chen, Z., Han, F., Wu, L., Yu, J., Cheng, S., Lin, P., & Chen, H. (2018). Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. Energy conversion and management, 178, 250-264.

[24]. Gong, S., Wu, X., & Zhang, Z. (2020, July). Fault diagnosis method of photovoltaic array based on random forest algorithm. In 2020 39th Chinese Control Conference (CCC) (pp. 4249-4254). IEEE.

[25]. Mellit, A., Benghanem, M., Kalogirou, S., & Pavan, A. M. (2023). An embedded system for remote monitoring and fault diagnosis of photovoltaic arrays using machine learning and the internet of things. Renewable Energy, 208, 399-408.

[26]. Wang, M., Xu, X., & Yan, Z. (2023). Online fault diagnosis of PV array considering label errors based on distributionally robust logistic regression. Renewable Energy, 203, 68-80.

[27]. Hong, Y. Y., & Pula, R. A. (2022). Methods of photovoltaic fault detection and classification: A review. Energy Reports, 8, 5898-5929.

[28]. Liu, Y., Ding, K., Zhang, J., Li, Y., Yang, Z., Zheng, W., & Chen, X. (2021). Fault diagnosis approach for photovoltaic array based on the stacked auto-encoder and clustering with IV curves. Energy Conversion and Management, 245, 114603.

[29]. Chen, Z., Chen, Y., Wu, L., Cheng, S., & Lin, P. (2019). Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions. Energy Conversion and Management, 198, 111793.

[30]. Gao, W., & Wai, R. J. (2020). A novel fault identification method for photovoltaic array via convolutional neural network and residual gated recurrent unit. IEEE access, 8, 159493-159510.

[31]. Eldeghady, G. S., Kamal, H. A., & Hassan, M. A. M. (2023). Fault diagnosis for PV system using a deep learning optimized via PSO heuristic combination technique. Electrical Engineering, 105(4), 2287-2301.

[32]. Appiah, A. Y., Zhang, X., Ayawli, B. B. K., & Kyeremeh, F. (2019). Review and performance evaluation of photovoltaic array fault detection and diagnosis techniques. International Journal of Photoenergy, 2019.

[33]. Boubaker, S., Kamel, S., Ghazouani, N., & Mellit, A. (2023). Assessment of machine and deep learning approaches for fault diagnosis in photovoltaic systems using infrared thermography. Remote Sensing, 15(6), 1686.

[34]. Piliougine, M., Sánchez-Friera, P., Petrone, G., Sánchez-Pacheco, F. J., Spagnuolo, G., & Sidrach-de-Cardona, M. (2022). Analysis of the degradation of amorphous silicon-based modules after 11 years of exposure by means of IEC60891: 2021 procedure 3. Progress in Photovoltaics: Research and Applications, 30(10), 1176-1187.

[35]. Mirjalili, S. M. S. M., Mirjalili, S. M., & Lewis, A. (2014). Grey Wolf Optimizer Adv Eng Softw 69: 46–61. ed.

[36]. Bun, L., "Détection et Localisation de Défauts pour un Système PV," Université Grenoble Alpes, 2011.

[37]. Castañer L, Silvestre S. Modelling photovoltaic systems using PSpice 2002:358.

[38]. Khatibi, A., Razi Astaraei, F., & Ahmadi, M. H. (2019). Generation and combination of the solar cells: A current model review. Energy Science & Engineering, 7(2), 305-322.

[39]. Nilsson, D. (2014). Fault detection in photovoltaic systems.

[40]. Triki-Lahiani, A., Abdelghani, A. B. B., & Slama-Belkhodja, I. (2018). Fault detection and monitoring systems for photovoltaic installations: A review. Renewable and Sustainable Energy Reviews, 82, 2680-2692.

[41]. Abdulmawjood, K., Refaat, S. S., & Morsi, W. G. (2018, April). Detection and prediction of faults in photovoltaic arrays: A review. In 2018 IEEE 12th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG 2018) (pp. 1-8). IEEE.

[42]. Sarikh, S., Raoufi, M., Bennouna, A., Benlarabi, A., & Ikken, B. (2019). Photovoltaic discoloration and cracks: experimental impact on the IV curve degradation. In Proceedings of the 1st International Conference on Electronic Engineering and Renewable Energy: ICEERE 2018, 15-17 April 2018, Saidia, Morocco 1 (pp. 609-616). Springer Singapore.

[43]. Abd Allah, S., & Gharabawy, E. (2018). Review on Corrosion in Solar Panels. Int. J. Smart Grid, 2(4).

[44]. Singh, D., & Kathuria, R. S. (2018). Fault prediction and analysis techniques of solar cells and PV modules. Int J Eng Sci Res Technol, 7, 384-99.

[45]. Mahalakshmi, R., Karuppasamypandiyan, M., Bhuvanesh, A., & Ganesh, R. J. (2016). Classification and detection of faults in grid connected photovoltaic system. Int J Sci Eng Res, 7, 149-54.

[46]. Niazi, K. A. K., Yang, Y., & Sera, D. (2019). Review of mismatch mitigation techniques for PV modules. IET Renewable Power Generation, 13(12), 2035-2050.

[47]. Andrews, R. W., Pollard, A., & Pearce, J. M. (2013). The effects of snowfall on solar photovoltaic performance. Solar Energy, 92, 84-97.

[48]. Heidari, N., Gwamuri, J., Townsend, T., & Pearce, J. M. (2015). Impact of snow and ground interference on photovoltaic electric system performance. IEEE Journal of Photovoltaics, 5(6), 1680-1685.

[49]. Appiah, A. Y., Zhang, X., Ayawli, B. B. K., & Kyeremeh, F. (2019). Review and performance evaluation of photovoltaic array fault detection and diagnosis techniques. International Journal of Photoenergy, 2019.

[50]. Bharadwaj, P., Karnataki, K., & John, V. (2018, June). Formation of hotspots on healthy PV modules and their effect on output performance. In 2018 IEEE 7th world conference on photovoltaic energy conversion (WCPEC)(A joint conference of 45th IEEE PVSC, 28th PVSEC & 34th EU PVSEC) (pp. 0676-0680). IEEE.

[51]. Wang, Y., Itako, K., Kudoh, T., Koh, K., & Ge, Q. (2016, October). Voltage-based hot-spot detection method for PV string using projector. In 2016 IEEE International Conference on Power and Renewable Energy (ICPRE) (pp. 570-574). IEEE.

[52]. Wendlandt, S., Drobisch, A., Buseth, T., Krauter, S., & Grunow, P. (2010, September). Hot spot risk analysis on silicon cell modules. In 25th European Photovoltaic Solar Energy Conference and Exhibition (pp. 4002-4006). Valencia, Spain.

[53]. Batzelis, E., Samaras, K., Vokas, G., & Papathanassiou, S. A. (2016, June). Off-grid inverter faults: diagnosis, symptoms and cause of failure. In Materials Science Forum (Vol. 856, pp. 315-321). Trans Tech Publications Ltd.

[54]. Nehme, B., Msirdi, N. K., Namaane, A., & Akiki, T. (2017). Analysis and characterization of faults in PV panels. Energy Procedia, 111, 1020-1029.

[55]. Pei, T., & Hao, X. (2019). A fault detection method for photovoltaic systems based on voltage and current observation and evaluation. Energies, 12(9), 1712.

[56]. Itoh, U., Yoshida, M., Tokuhisa, H., Takeuchi, K., & Takemura, Y. (2014). Solder joint failure modes in the conventional crystalline Si module. Energy Procedia, 55, 464-468.

[57]. Jiang, L. L., & Maskell, D. L. (2015, July). Automatic fault detection and diagnosis for photovoltaic systems using combined artificial neural network and analytical based methods. In 2015 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[58]. Pillai, D. S., & Rajasekar, N. (2018). A comprehensive review on protection challenges and fault diagnosis in PV systems. Renewable and Sustainable Energy Reviews, 91, 18-40.

[59]. Strobl, C., & Meckler, P. (2010, October). Arc faults in photovoltaic systems. In 2010 Proceedings of the 56th IEEE Holm Conference on Electrical Contacts (pp. 1-7). IEEE.

[60]. Miao W, Liu X, Lam KH et al. Arc-faults detection in PV systems by mea-suring pink noise with magnetic sensors. IEEE Trans Magn 2019;55:1–6.

[61]. Artale, G., Caravello, G., Cataliotti, A., Cosentino, V., Cara, D. D., Guaiana, S., ... & Tinè, G. (2020, September). DC series arc faults in PV systems. Detection methods and experimental

characterization. In 24th IMEKO TC4 International Symposium and 22nd International Workshop on ADC and DAC Modelling and Testing, IWADC (pp. 135-40).

[62]. Artale, G., Caravello, G., Cataliotti, A., Cosentino, V., Cara, D. D., Guaiana, S., ... & Tinè, G. (2020, September). DC series arc faults in PV systems. Detection methods and experimental characterization. In 24th IMEKO TC4 International Symposium and 22nd International Workshop on ADC and DAC Modelling and Testing, IWADC (pp. 135-40).

[63]. Vieira, R. G., de Araújo, F. M., Dhimish, M., & Guerra, M. I. (2020). A comprehensive review on bypass diode application on photovoltaic modules. Energies, 13(10), 2472.

[64]. Abubakar, A., Almeida, C. F. M., & Gemignani, M. (2021). Review of artificial intelligence-based failure detection and diagnosis methods for solar photovoltaic systems. Machines, 9(12), 328.

[65]. Caruana, R. (1997). Multitask learning. Machine learning, 28, 41-75.

[66]. Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-38665.

[67]. Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. Journal of artificial intelligence research, 11, 169-198.

[68]. Wikipedia. Instance-Based Learning. Available online: https://en.wikipedia.org/wiki/Instance-based_learning (accessed on 31 March 2023).

[69]. Al-Sahaf, H., Bi, Y., Chen, Q., Lensen, A., Mei, Y., Sun, Y., ... & Zhang, M. (2019). A survey on evolutionary machine learning. Journal of the Royal Society of New Zealand, 49(2), 205-228.

[70]. Yar, M. H., Rahmati, V., & Oskouei, H. R. D. (2016). A survey on evolutionary computation: Methods and their applications in engineering. Mod. Appl. Sci, 10(11), 131139.

[71]. Sharma, V., Rai, S., & Dev, A. (2012). A comprehensive study of artificial neural networks. International Journal of Advanced research in computer science and software engineering, 2(10).

[72]. Wang, X., Zhao, Y., & Pourpanah, F. (2020). Recent advances in deep learning. International Journal of Machine Learning and Cybernetics, 11, 747-750.

[73]. Manohar, M., Koley, E., Ghosh, S., Mohanta, D. K., & Bansal, R. C. (2020). Spatio-temporal information based protection scheme for PV integrated microgrid under solar irradiance intermittency using deep convolutional neural network. International Journal of Electrical Power & Energy Systems, 116, 105576.

[74]. Aziz, F., Haq, A. U., Ahmad, S., Mahmoud, Y., Jalal, M., & Ali, U. (2020). A novel convolutional neural network-based approach for fault classification in photovoltaic arrays. IEEE Access, 8, 41889-41904.

[75]. Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. Artificial intelligence review, 53, 5455-5516.

[76]. Lu, X., Lin, P., Cheng, S., Lin, Y., Chen, Z., Wu, L., & Zheng, Q. (2019). Fault diagnosis for photovoltaic array based on convolutional neural network and electrical time series graph. Energy Conversion and Management, 196, 950-965.

[77]. Khalyasmaa, A. I., Eroshenko, S. A., Tashchilin, V. A., Ramachandran, H., Piepur Chakravarthi, T., & Butusov, D. N. (2020). Industry experience of developing day-ahead photovoltaic plant forecasting system based on machine learning. Remote Sensing, 12(20), 3420.

[78]. Dagnely, P., Ruette, T., Tourwé, T., & Tsiporkova, E. (2018, September). Ontology-driven multilevel sequential pattern mining: mining for gold in event logs of photovoltaic plants. In 2018 International Conference on Intelligent Systems (IS) (pp. 1-7). IEEE.

[79]. Canizo, M., Triguero, I., Conde, A., & Onieva, E. (2019). Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study. Neurocomputing, 363, 246-260.

[80]. Zhao, B., Lu, H., Chen, S., Liu, J., & Wu, D. (2017). Convolutional neural networks for time series classification. Journal of Systems Engineering and Electronics, 28(1), 162-169.

[81]. Kovács, G., Tóth, L., Van Compernolle, D., & Ganapathy, S. (2017). Increasing the robustness of CNN acoustic models using autoregressive moving average spectrogram features and channel dropout. Pattern Recognition Letters, 100, 44-50.

[82]. Jana, G. C., Sharma, R., & Agrawal, A. (2020). A 1D-CNN-spectrogram based approach for seizure detection from EEG signal. Procedia Computer Science, 167, 403-412.

[83]. Lin, Z., Ji, K., Leng, X., & Kuang, G. (2018). Squeeze and excitation rank faster R-CNN for ship detection in SAR images. IEEE Geoscience and Remote Sensing Letters, 16(5), 751-755.

[84]. Beeravolu, A. R., Azam, S., Jonkman, M., Shanmugam, B., Kannoorpatti, K., & Anwar, A. (2021). Preprocessing of breast cancer images to create datasets for deep-CNN. IEEE Access, 9, 33438-33463.

[85]. Shuai, L., Yuanning, L., Xiaodong, Z., Guang, H., Jingwei, C., Qixian, Z., ... & Chaoqun, W. (2020). Multi-source feature fusion and entropy feature lightweight neural network for constrained multi-state heterogeneous iris recognition. IEEE Access, 8, 53321-53345.

[86]. Lin, B., Deng, S., Gao, H., & Yin, J. (2020). A multi-scale activity transition network for data translation in EEG signals decoding. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18(5), 1699-1709.

[87]. Kauderer-Abrams, E. (2017). Quantifying translation-invariance in convolutional neural networks. arXiv preprint arXiv:1801.01450.

[88]. Mohd Razak, S., & Jafarpour, B. (2020). Convolutional neural networks (CNN) for feature-based model calibration under uncertain geologic scenarios. Computational Geosciences, 24(4), 1625-1649.

[89]. Xia, M., Zheng, X., Imran, M., & Shoaib, M. (2020). Data-driven prognosis method using hybrid deep recurrent neural network. Applied Soft Computing, 93, 106351.

[90]. Lee, C. K., Ng, K. K., Chen, C. H., Lau, H. C., Chung, S. Y., & Tsoi, T. (2021). American sign language recognition and training method with recurrent neural network. Expert Systems with Applications, 167, 114403.

[91]. Zhu, R., Tu, X., & Huang, J. X. (2020). Deep learning on information retrieval and its applications. In Deep learning for data analytics (pp. 125-153). Academic Press.

[92]. Li, X., Ma, X., Xiao, F., Wang, F., & Zhang, S. (2020). Application of gated recurrent unit (GRU) neural network for smart batch production prediction. Energies, 13(22), 6121.

[93]. 91Park, P., Marco, P. D., Shin, H., & Bang, J. (2019). Fault detection and diagnosis using combined autoencoder and long short-term memory network. Sensors, 19(21), 4612.

[94]. Yang, Z., Gjorgjevikj, D., Long, J., Zi, Y., Zhang, S., & Li, C. (2021). Sparse autoencoder-based multi-head deep neural networks for machinery fault diagnostics with detection of novelties. Chinese Journal of Mechanical Engineering, 34(1), 54.

[95]. Li, R., Wu, Q., Liu, J., Wu, Q., Li, C., & Zhao, Q. (2020). Monitoring depth of anesthesia based on hybrid features and recurrent neural network. Frontiers in neuroscience, 14, 26.

[96]. Chang, C. H. (2015). Deep and shallow architecture of multilayer neural networks. IEEE transactions on neural networks and learning systems, 26(10), 2477-2486.

[97]. Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing, 115, 213-237.

[98]. Souriau, R., Lerbet, J., Chen, H., & Vigneron, V. (2021). A review on generative Boltzmann networks applied to dynamic systems. Mechanical Systems and Signal Processing, 147, 107072.

[99]. Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. Parallel distributed processing: Explorations in the microstructure of cognition, 1(282-317), 2.

[100]. Malik, A., Haque, A., Satya Bharath, K. V., & Jaffery, Z. A. (2021). Transfer learning-based novel fault classification technique for grid-connected PV inverter. In Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2021 (pp. 217-224). Springer Singapore.

[101]. Xiao, D., Huang, Y., Zhao, L., Qin, C., Shi, H., & Liu, C. (2019). Domain adaptive motor fault diagnosis using deep transfer learning. Ieee Access, 7, 80937-80949.

[102]. Cho, S. H., Kim, S., & Choi, J. H. (2020). Transfer learning-based fault diagnosis under data deficiency. Applied Sciences, 10(21), 7768.

[103]. Shao, S., McAleer, S., Yan, R., & Baldi, P. (2018). Highly accurate machine fault diagnosis using deep transfer learning. IEEE Transactions on Industrial Informatics, 15(4), 2446-2455.

[104]. Wen, L., Gao, L., & Li, X. (2017). A new deep transfer learning based on sparse auto-encoder for fault diagnosis. IEEE Transactions on systems, man, and cybernetics: systems, 49(1), 136-144.

[105]. Li, J., Wang, Y., Zi, Y., & Jiang, S. (2020). A local weighted multi-instance multilabel network for fault diagnosis of rolling bearings using encoder signal. IEEE Transactions on Instrumentation and Measurement, 69(10), 8580-8589.

[106]. Han, J., Miao, S. H., Yin, H. R., Guo, S. Y., Wang, Z. X., Yao, F. X., & Lin, Y. J. (2021, March). Deep-adversarial-transfer learning based fault classification of power lines in smart grid. In IOP conference series: Earth and environmental science (Vol. 701, No. 1, p. 012074). IOP Publishing.

[107]. Wang, P., & Gao, R. X. (2020). Transfer learning for enhanced machine fault diagnosis in manufacturing. CIRP Annals, 69(1), 413-416.

[108]. Zhang, R., Tao, H., Wu, L., & Guan, Y. (2017). Transfer learning with neural networks for bearing fault diagnosis in changing working conditions. Ieee Access, 5, 14347-14357.

[109]. Yang, B., Lei, Y., Jia, F., Li, N., & Du, Z. (2019). A polynomial kernel induced distance metric to improve deep transfer learning for fault diagnosis of machines. IEEE Transactions on Industrial Electronics, 67(11), 9747-9757.

[110]. Wan, Z., Yang, R., Huang, M., Zeng, N., & Liu, X. (2021). A review on transfer learning in EEG signal analysis. Neurocomputing, 421, 1-14.

[111]. Zhang, D., & Zhou, T. (2021). Deep convolutional neural network using transfer learning for fault diagnosis. IEEE Access, 9, 43889-43897.

[112]. Li, Q., Shen, C., Chen, L., & Zhu, Z. (2021). Knowledge mapping-based adversarial domain adaptation: A novel fault diagnosis method with high generalizability under variable working conditions. Mechanical Systems and Signal Processing, 147, 107095.

[113]. Fang, X., Gong, G., Li, G., Chun, L., Li, W., & Peng, P. (2021). A hybrid deep transfer learning strategy for short term cross-building energy prediction. Energy, 215, 119208.

[114]. Deng, Y., Huang, D., Du, S., Li, G., Zhao, C., & Lv, J. (2021). A double-layer attention based adversarial network for partial transfer learning in machinery fault diagnosis. Computers in Industry, 127, 103399.

[115]. Zhang, Z., Li, X., Wen, L., Gao, L., & Gao, Y. (2019, August). Fault diagnosis using unsupervised transfer learning based on adversarial network. In 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE) (pp. 305-310). IEEE.

[116]. Sun, G., Liang, L., Chen, T., Xiao, F., & Lang, F. (2018). Network traffic classification based on transfer learning. Computers & electrical engineering, 69, 920-927.

[117]. Lu, S., He, Q., & Zhao, J. (2018). Bearing fault diagnosis of a permanent magnet synchronous motor via a fast and online order analysis method in an embedded system. Mechanical Systems and Signal Processing, 113, 36-49.

[118]. Kumar, K. P., & Saravanan, B. (2017). Recent techniques to model uncertainties in power generation from renewable energy sources and loads in microgrids–A review. Renewable and Sustainable Energy Reviews, 71, 348-358.

[119]. Bou-Rabee, M., Sulaiman, S. A., Saleh, M. S., & Marafi, S. (2017). Using artificial neural networks to estimate solar radiation in Kuwait. Renewable and Sustainable Energy Reviews, 72, 434-438.

[120]. Abdel-Nasser, M., & Mahmoud, K. (2019). Accurate photovoltaic power forecasting models using deep LSTM-RNN. Neural computing and applications, 31, 2727-2740.

[121]. Wang, K., Qi, X., & Liu, H. (2019). A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network. Applied Energy, 251, 113315.

[122]. Ahmed, R., Sreeram, V., Mishra, Y., & Arif, M. D. (2020). A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. Renewable and Sustainable Energy Reviews, 124, 109792.

[123]. Das, U. K., Tey, K. S., Seyedmahmoudian, M., Mekhilef, S., Idris, M. Y. I., Van Deventer, W., ... & Stojcevski, A. (2018). Forecasting of photovoltaic power generation and model optimization: A review. Renewable and Sustainable Energy Reviews, 81, 912-928.

[124]. Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2018). Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. Energy, 164, 465-474.

[125]. Wang, J., Li, P., Ran, R., Che, Y., & Zhou, Y. (2018). A short-term photovoltaic power prediction model based on the gradient boost decision tree. Applied Sciences, 8(5), 689.

[126]. Tovar, M., Robles, M., & Rashid, F. (2020). PV power prediction, using CNN-LSTM hybrid neural network model. Case of study: Temixco-Morelos, México. Energies, 13(24), 6512.

[127]. Niccolai, A., Dolara, A., & Ogliari, E. (2021). Hybrid PV power forecasting methods: A comparison of different approaches. Energies, 14(2), 451.

[128]. Powers, J. G., Klemp, J. B., Skamarock, W. C., Davis, C. A., Dudhia, J., Gill, D. O., ... & Duda, M. G. (2017). The weather research and forecasting model: Overview, system efforts, and future directions. Bulletin of the American Meteorological Society, 98(8), 1717-1737.

[129]. Garoudja, E., Chouder, A., Kara, K., & Silvestre, S. (2017). An enhanced machine learning based approach for failures detection and diagnosis of PV systems. Energy conversion and management, 151, 496-513.

[130]. Kichou, S., & Wolf, P. (2018). Approach for Simulating outputs of PV module/array of different technologies with high accuracy. EUPVSEC proceedings.

[131]. Amiri, A. F., Oudira, H., & Chouder, A. (2022, November). Faults detection of PV systems based on extracted parameters using Modified Grey Wolf algorithm. In 2022 International Conference of Advanced Technology in Electronic and Electrical Engineering (ICATEEE) (pp. 1-6). IEEE.

[132]. Herrea, F. (1998). Tackling real-coded genetic algorithms: operators and tools for behavioral analysis. Artificial intelligence review, 12(4), 265-319.

[133]. Ali, M., El-Hameed, M. A., & Farahat, M. A. (2017). Effective parameters' identification for polymer electrolyte membrane fuel cell models using grey wolf optimizer. Renewable energy, 111, 455-462.

[134]. Chouder, A., Silvestre, S., Sadaoui, N., & Rahmani, L. (2012). Modeling and simulation of a grid connected PV system based on the evaluation of main PV module parameters. Simulation Modelling Practice and Theory, 20(1), 46-58.

[135]. Elkholy, A., & Abou El-Ela, A. A. (2019). Optimal parameters estimation and modelling of photovoltaic modules using analytical method. Heliyon, 5(7).

[136]. Tossa, A. K., Soro, Y. M., Azoumah, Y., & Yamegueu, D. (2014). A new approach to estimate the performance and energy productivity of photovoltaic modules in real operating conditions. Solar energy, 110, 543-560.

[137]. Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9(3), e1301.

[138]. Resende, P. A. A., & Drummond, A. C. (2018). A survey of random forest based methods for intrusion detection systems. ACM Computing Surveys (CSUR), 51(3), 1-36.

[139]. Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. arXiv preprint arXiv:2106.11342.

[140]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[141]. Mellit, A., & Kalogirou, S. (2022). Assessment of machine learning and ensemble methods for fault diagnosis of photovoltaic systems. Renewable Energy, 184, 1074-1090.

[142]. King, D. E. (2009). Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, 10, 1755-1758.

[143]. Chine, W., Mellit, A., Lughi, V., Malek, A., Sulligoi, G., & Pavan, A. M. (2016). A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. Renewable Energy, 90, 501-512.

[144]. Benkercha, R., & Moulahoum, S. (2018). Fault detection and diagnosis based on C4. 5 decision tree algorithm for grid connected PV system. Solar Energy, 173, 610-634.

[145]. Kratochvil, J. A., Boyson, W. E., & King, D. L. (2004). Photovoltaic array performance model (No. SAND2004-3535). Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA (United States).

[146]. Kichou, S., Silvestre, S., Guglielminotti, L., Mora-López, L., & Muñoz-Cerón, E. (2016). Comparison of two PV array models for the simulation of PV systems using five different algorithms for the parameters identification. Renewable Energy, 99, 270-279.

[147]. D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Technical report-tr06, Erciyes university, engineering faculty, computer engineering department2005.

[148]. W. Gao, S. Liu, and L. Huang, "A global best artificial bee colony algorithm for global optimization," Journal of Computational and Applied Mathematics, vol. 236, pp. 2741-2753, 2012.

[149]. Aziz, F.; Ul Haq, A.; Ahmad, S.; Mahmoud, Y.; Jalal, M.; Ali, U. A Novel Convolutional Neural Network-Based Approach for Fault Classification in Photovoltaic Arrays. IEEE Access 2020, 8, 41889–41904, doi:10.1109/ACCESS.2020.2977116.

[150]. Mansouri, M.; Trabelsi, M.; Nounou, H.; Nounou, M. Deep Learning-Based Fault Diagnosis of Photovoltaic Systems: A Comprehensive Review and Enhancement Prospects. IEEE Access 2021, 9, 126286–126306, doi:10.1109/ACCESS.2021.3110947.

[151]. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[152]. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. Neural Comput. 1997, 9, 1735–1780, doi:10.1162/NECO.1997.9.8.1735.

[153]. Schuster, M.; Paliwal, K.K. Bidirectional Recurrent Neural Networks. IEEE Trans. Signal Process. 1997, 45, 2673–2681, doi:10.1109/78.650093.

[154]. Keras: Deep Learning for Humans Available online: https://keras.io/ (accessed on 27 November 2023).

[155]. Pedregosa FABIANPEDREGOSA, F.; Michel, V.; Grisel OLIVIERGRISEL, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; et al. Scikit-Learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. J. Mach. Learn. Res. 2011, 12, 2825–2830.

[156]. Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. arXiv preprint arXiv:2106.11342.

[157]. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2016.

[158]. Mellit, A., & Kalogirou, S. (2022). Assessment of machine learning and ensemble methods for fault diagnosis of photovoltaic systems. Renewable Energy, 184, 1074-1090.

[159]. Dash, C. S. K., Behera, A. K., Dehuri, S., & Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. Decision Analytics Journal, 6, 100164.

[160]. Spearman, C.; The proof and measurement of association between two things. Amer. J. Psychol. 1904, 15, 1, 72-101.

[161]. I-Kuei Lin, L.; Concordance correlation coefficient to evaluate repro-ducibility. Biometrics 1989, 45, 1, 255-268.

[162]. Best, D.J.;and Roberts, D.E.; Algorithm AS 89: The upper tail proba-bilities of Spearman's ρ. J. Roy. Statist. Ser. C (Appl. Statist.) 1975, 24, 377–379.

[163]. Revelle W. Psych v1.8.4, 2018. Available online: https://www.rdocumentation.org/packages/psych/versions/1.8.4/topics/pairs.panels (accessed on 5 May 2024).

[164]. Weisstein E.W. S.; Rank correlation, coefficient 1999. Available online: https://mathworld.wolfram.com/SpearmanRankCorrelationCoefficient.html (accessed on 15 May 2024).

[165]. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD) 2012, 6.1, 1-39.

[166]. Shah, I.; Iftikhar, H.; Ali, S. Modeling and Forecasting Electricity Demand and Prices: A Comparison of Alternative Ap-proaches. J. Math. 2022, 3581037.

[167]. Shah, I.; Jan, F.; Ali, S. Functional data approach for short-term electricity demand forecasting. Math. Probl. Eng. 2022, 6709779.

[168]. Lisi, F.; Shah, I. Forecasting next-day electricity demand and prices based on functional models. Energy Syst. 2020, 11, 947–979.

[169]. Margoum, S. et al. Prediction of Electrical Power of Ag/Water-Based PVT System Using K-NN Machine Learning Technique. In Proceedings of the International Conference on Digital Technologies and Applications, Fez, Moroco, 27 Jan. 2023.

[170]. Kuriakose, A. M.; Kariyalil, D. P. ; Augusthy, M. ; Sarath, S. ; Jacob, J. ; Antony, N. R. ; Comparison of Artificial Neural Network, Linear Regression and Support Vector Machine for Prediction of Solar PV Power. In Proceedings of the 2020 IEEE Pune Section International Conference (PuneCon), Pune, India, 16 December 2020.

[171]. Khalyasmaaet, A. al. Prediction of Solar Power Generation Based on Random Forest Regressor Model. In Proceedings of the International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Novosibirsk, Russia, 21 October 2019.

[172]. Gupta, R.; Yadav, A.K.; JHA, S.K.; et al. Predicting global horizontal irradiance of north central region of India via machine learning regressor algorithms. Engineering Applications of Artificial Intelligence 2024, 133, 108426.

[173]. Amiri, A. F., Oudira, H., Chouder, A., & Kichou, S. (2024). Faults detection and diagnosis of PV systems based on machine learning approach using random forest classifier. *Energy Conversion and Management*, *301*, 118076.

[174]. Amiri, A. F., Kichou, S., Oudira, H., Chouder, A., & Silvestre, S. (2024). Fault detection and diagnosis of a photovoltaic system based on deep learning using the combination of a convolutional neural network (cnn) and bidirectional gated recurrent unit (Bi-GRU). *Sustainability*, *16*(3), 1012.

[175]. Vapnik, V.N; Statistical learning theory. Wiley, New York, USA,1998.

[176]. Rojas-Dominguez, L.C.; Padierna, J.M.; Carpio Valadez, H.J.; Puga-Soberanes and. Fraire, H.J.; Optimal Hyper-Parameter Tuning of SVM Classifiers with Application to Medical Diagnosis I.EEE Access 2018, 6, 7164–7176.

[177]. Ramaprakoso. Analisis-Sentimen, GitHub. Available online: https://github.com/ramaprakoso/analisis-sentimen/blob/master/kamus/acronym.txt. (accessed on 20 March 2024).

[178]. Ahmad, M.; Aftab, S.; Salman, M.; Hameed, N.; Ali, I. and Nawaz, Z. SVM Optimization for Sentiment Analysis. International Journal of Advanced Computer Science and Applications 2018, 9, 4, 393-398.