Detection and diagnosis of faults and performance losses in high-power photovoltaic power plants

Edgar Hernando Sepúlveda-Oviedo

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ FÉDÉRALE TOULOUSE MIDI-PYRÉNÉES

Délivré par :
l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue le *16/02/2023* **par :**
Edgar Hernando SEPÚLVEDA OVIEDO

Detection and diagnosis of faults and performance losses in high-power photovoltaic power plants

---

JURY

Nadia Steiner	Professeur d'Universités	Présidente du Jury
Ghaleb Hoblos	Enseignant-Chercheur	Rapporteur
Claude Delpha	Maitre de conférences HDR	Rapporteur
Audine Subias	Professeur d'Universités	Co-encadrante
Victor Grisales-Palacio	Associate Professor	Examinateur
Louise Travé-massuyès	Directrice de recherche	Co-directrice de thèse
Corinne Alonso	Professeur d'Universités	Directrice de thèse
Marko Pavlov	Docteur	Co-encadrante de thèse du monde socio-économique

---

École doctorale et spécialité :
GEET : Génie Electrique
Unité de Recherche :
Laboratoire d'analyse et d'architecture des systèmes
Directeur(s) de Thèse :
Corinne Alonso et *Louise Travé-massuyès*
Rapporteurs :
Ghaleb HOBLOS et *Claude DELPHA*

Contents

List of Figures **x**

List of Tables **xii**

Glossary **xiii**

Résumé **1**

Introduction **29**

1 Background and Project Motivation **33**
1.1 Background . . . 33
1.1.1 Project Motivation . . . 34
1.1.2 Problem statement . . . 36
1.1.3 Aim and objectives . . . 37
1.1.4 Studied fault cases using machine learning . . . 38
1.1.5 Academic products of the thesis . . . 39

2 Fault Diagnosis in Photovoltaic Systems **41**
2.1 Photovoltaic industry . . . 42
2.2 PV system components . . . 45
2.2.1 PV generator . . . 45
2.2.2 Wiring and Junction Box . . . 53
2.2.3 Inverter . . . 53
2.2.4 Protection system . . . 54
2.3 Formal fault dictionary . . . 55
2.3.1 Main causes of faults . . . 56
2.3.2 Multilevel fault classification . . . 57
2.3.3 Frequency of occurrence of faults in the PV system . . . 73
2.3.4 Impact of faults in terms of power loss and human safety . . . 77
2.4 Conventional Fault Detection Methods . . . 82
2.4.1 Visual methods . . . 83
2.4.2 Image-based methods . . . 83
2.4.3 Electrical detection methods . . . 86
2.4.4 Protection Device Based Technique . . . 89
2.4.5 ARC Fault Detector Techniques . . . 89
2.5 Discussion and Conclusions . . . 89

3 Fault Diagnosis in Photovoltaic Systems using Artificial Intelligence 91
3.1 Description of Methodology (Smart B2TE) . . . 92
3.1.1 Document recovery . . . 93
3.1.2 Bibliometric analysis . . . 95
3.1.3 Topic Modeling . . . 98
3.1.4 T-distributed stochastic neighbor embedding . . . 103
3.1.5 Expert qualitative content analysis . . . 106
3.1.6 Summary of the relevant information retrieved. . . . 109
3.2 Methods based on artificial intelligence techniques . . . 110
3.2.1 Supervised Learning . . . 112
3.2.2 Semi-Supervised Learning . . . 120
3.2.3 Reinforcement Learning . . . 122
3.2.4 Unsupervised Learning . . . 123
3.3 Promising Research Topics and Challenges . . . 125
3.4 Discussion and conclusions . . . 128

4 Conventional Data Acquisition in PV plants 131
4.1 Characteristics of a data acquisition system . . . 131
4.1.1 Measured parameters . . . 133
4.1.2 Data acquisition system . . . 137
4.1.3 Wired communication . . . 137
4.1.4 Wireless communication . . . 138
4.1.5 Controller . . . 139
4.1.6 Sample rate . . . 139
4.1.7 Data preprocessing . . . 140
4.1.8 Signal treatment methods . . . 142
4.1.9 Existing data acquisition systems . . . 144
4.2 Tigo industrial and commercial data acquisition system . . . 148
4.2.1 Components and Connection Scheme . . . 148
4.2.2 Instrumented PV plant . . . 149
4.3 Discussion and Conclusions . . . 152

5 Diagnosis-oriented Data Acquisition (Solar Vitality) 155
5.1 Motivation . . . 156
5.2 Characteristics of Solar Vitality . . . 157
5.2.1 Portable data acquisition requirements . . . 157
5.2.2 Measured parameters . . . 158
5.2.3 Hardware . . . 158
5.2.4 Software . . . 164
5.2.5 Electrical power supply . . . 165
5.2.6 Assembled prototype and electrical connection diagram to the PV system. . . . 166
5.2.7 Product evolution cycle and test scenarios . . . 167

5.3 Discussion and Conclusions 177

6 Feature Engineering for Fault Diagnosis 179
6.1 Motivation 180
6.2 Pre-processing 181
6.3 Feature extraction 182
6.3.1 Multi-resolution signal decomposition 183
6.3.2 Features based on signal characterization 186
6.4 Feature Selection 188
6.4.1 Correlation based feature selection 188
6.4.2 Variance based feature selection 189
6.5 Feature Transformation 191
6.5.1 Principal Component Analysis 192
6.5.2 Isometric mapping 194
6.6 Discussion and Conclusions 195

7 An Ensemble Based Diagnosis Algorithm (EB-diag) 197
7.1 Approach description 198
7.2 Dataset 199
7.3 Selected features for fault detection 200
7.4 EB-diag composition 202
7.4.1 k-Nearest-Neighbor 202
7.4.2 Support Vector Machines 203
7.4.3 Decision Trees 205
7.4.4 Majority voting 207
7.5 Discussion and Conclusions 208

8 A Serial Diagnosis Algorithm (Serial-diag) 215
8.1 Approach description 216
8.2 Dataset 217
8.3 DTW Hierarchical clustering 218
8.3.1 Dynamic Time Warping 218
8.3.2 Hierarchical Clustering 219
8.4 Selected features for fault detection 221
8.5 Diagnosis of PV panels 221
8.5.1 PLS Regression model 221
8.5.2 Diagnosis of the health status of PV panels 222
8.5.3 PLS-LDA classification method 224
8.6 Discussion and Conclusions 227

9 An adaptive Diagnosis algorithm (Adaptive-diag) 229
9.0.1 General Scheme of operation 230
9.0.2 Normalization 233
9.0.3 Model based on knowledge 238

9.0.4 Maintenance priority 241
9.1 Discussion and Conclusions 241

Conclusions and perspectives **245**

A Annexes **253**
A.1 Examples of signals captured with the monitoring platform 253
A.2 Test of conventional machine learning algorithms. 257
A.2.1 K-means Clustering 257
A.2.2 Random Forest (RF) 259
A.2.3 Discussion and Conclusions 262

Bibliography **265**

List of Figures

1 Capacité installée solaire photovoltaïque totale 2000-2021 et parts totales installées des 10 principaux marchés solaires photovoltaïques 2021 [SPE 2022]. 1
2 Scénarios annuels mondiaux du marché solaire photovoltaïque 2022-2026 [SPE 2022]. 2
3 Défauts analysés dans les systèmes PV. a) Traces d'escargot; b) Verre cassé; c) Ombrage 6
4 Schéma général de la plateforme de diagnosis proposée. 24

1.1 Faults analyzed in PV systems. a) Snail Trails; b) Broken Glass; c) Shading 37

2.1 Solar electricity generation cost in comparison with other power sources 2009-2021 [SPE 2022]. 42
2.2 Solar electricity generation cost in comparison with conventional power sources 2021 [SPE 2022]. 43
2.3 Annual solar PV installed capacity 2000-2021 and top 10 countries solar share 2021 [SPE 2022]. 43
2.4 Total solar pv installed capacity 2000-2021 and Top 10 solar pv markets total installed shares 2021 [SPE 2022]. 44
2.5 World annual solar pv market scenarios 2022 - 2026 [SPE 2022]. . . . 44
2.6 Current-voltage characteristic curve I(V). 45
2.7 Historical evolution of technology market share and future trends [EPIA 2011]. 46
2.8 Typical, normalized spectral response data for single junction PV technologies, as used for calculation of spectral mismatch factors. [Dirnberger 2015]. 48
2.9 Operation of a PV cell [Jenkins 2017]. 49
2.10 Representation of a solar cell. a) Simple equivalent circuit of an ideal solar cell. b) Symbolic representation 49
2.11 Equivalent circuit of a solar cell with series and shunt resistance. . . 50
2.12 Details of a PV module constituted of PV cells. Cells are connected in series to form PV modules and are associated with secure elements named By-pass diodes. 51
2.13 string PV 52
2.14 Array PV 52
2.15 Examples of configuration of a PV array. a) Series-Parallel (SP) configuration; b) Total-Cross-Tied (TCT) configuration; c) Bridge-Linked (BL) configuration [Andrianajaina 2017] 53
2.16 Aging mechanisms leading to PV module degradation [Dross 2017]. . 58
2.17 Proposed multilevel classification of faults in PV systems 59

2.18 Example of Cell crack 60
2.19 Example of Discoloration 61
2.20 Example of Snail track or Snail Trails 62
2.21 Example of Delamination [Omazic 2019] 62
2.22 Example of Light Induced Energy Degradation (LID) 63
2.23 Example of Frame breakage [Köntges 2014b] 63
2.24 Example of Bubbles [Kim 2021] 64
2.25 Example of Back sheet adhesion loss (BSAL) [de Oliveira 2018] . . . 64
2.26 Example of Burn mark [Omazic 2019] 65
2.27 Example of Shunt hot spot [Aghaei 2015] 66
2.28 Example of Dust and Soiling 66
2.29 Example of Shading 67
2.30 Example of Short circuit (SC) / Open circuit (OC) 68
2.31 Example of PID [Köntges 2014b] 68
2.32 Example of Ground fault (GF) 69
2.33 Example of Line to line fault (LLF) [Alam 2015a] 70
2.34 Example of Arc fault (AF) [Alam 2015a] 71
2.35 Example of Diode fault [Chang 2015, Köntges 2018] 72
2.36 Example of Balance of system (BOS) [Flicker 2016] 72
2.37 Example of Junction box fault [Köntges 2014b] 73
2.38 Example of Ribbon and Solder Bonds Degradation and Broken Interconnect (R and SB) [Rajput 2019] 73
2.39 Percentages of objects studied in [IEC 2016b]. a) Objects studied according to climate zones. b) Objects studied according to PV technologies. 74
2.40 Occurrence distribution of degradation faults 75
2.41 Occurrence distribution of degradation faults that cause measurable energy loss 76
2.42 Occurrence distribution of sudden faults 76
2.43 Occurrence distribution of sudden faults that cause measurable energy loss 77

3.1 5-stage methodology. a) Document recovery, ; b) Bibliometrics analysis; c) Topic modeling; d) t-SNE and e) Expert qualitative content analysis used for construction of the global review on Fault Diagnosis in Photovoltaic Systems using Artificial Intelligence. 93
3.2 Steps of the document recovery stage for the construction of the data corpus: i) To build the search equation using keywords and logical operators; ii) To use the search equation to retrieve the documents in the Scopus and WoS databases; iii) To filter documents by selection criteria; iv) To extend the number of retrieved and filtered documents, adding filtered documents from other sources; and v) Construction of the data corpus. 94

3.3 Statistics of the retrieved documents. Distribution of retrieved documents by type. Number of leaked documents published per year between 2010 and 2022. 95
3.4 Example of a graphical representation Bibliometric network. 96
3.5 Topic coherence analysis based on the coherence score using the metric (C_V). The optimal number of topics is 4. 102
3.6 Visualization Map of co-occurring keywords identifying 6 main clusters. Each cluster corresponds to a type of approach. 107
3.7 Distribution of Document Word Counts by Dominant Topic. In the center the 2D visualization T-SNE of the 4 topics is presented. On the sides, word clouds for each topic are presented. Each of the average values of silhouette coefficients $\bar{\varphi}_g$, $g = 1, \ldots, 4$ is presented above each topic. 109
3.8 Summary scheme of the information found with the Smart B2TE methodology. a) Typical faults studied in the literature; b) typical variables used for fault diagnosis; c) types of machine learning found with bibliometric analysis; and e) types of machine learning found with topical modeling and T-SNE. 110

4.1 A typical data acquisition system aimed at the detection and classification of faults in PV plants. PV plant composed of PV strings, junction box, and inverter. Detection and classification system fed by directly preprocessed data or preprocessed data stored in a database. 132
4.2 Diagram of the most relevant parameters to be measured in real time, according to the IEC 61724 standard [IEC 1998]. 135
4.3 Tigo data acquisition system connection diagram 149
4.4 Tigo platform connection diagram 149
4.5 Tigo supervision website 150
4.6 Geographical location in France 150
4.7 Experimental platform in the LAAS - CNRS, Toulouse - France . . . 150
4.8 Distribution of the panels of the string PV used (experimental platform) with different health statuses: healthy (yellow), broken glass (blue), and big snail trails (red). 151
4.9 Physical location of the panels on the Tigo supervision website. . . . 152
4.10 Example of problems found in the database generated by the Tigo data acquisition system. 152

5.1 Recommended locations to properly measure the temperature of a PV module according to the ISO 16077 standard [ISO 2013] 159
5.2 Portable and adjustable weather station 161
5.3 Electrical diagram of Solar Vitality 166
5.4 Electrical connection diagram. This schematic represents the PV system, the junction box, the PV inverter, the weather station and the external power supply system. 167

5.5 Evolution diagram of the proposed product for Solar Vitality. 168
5.6 Installation of the first version of Solar Vitality in the LAAS-CNRS, Toulouse, France 170
5.7 Installation of the second version of Solar Vitality in Upie, department of Drôme, France 171
5.8 Installation of the third version of Solar Vitality in the LAAS-CNRS 172
5.9 Installation of the fourth version of Solar Vitality in the Delegation 14 of the CNRS, Toulouse, France. 173
5.10 Installation of the last version of Solar Vitality in the LAAS-CNRS . 175
5.11 The last version of Solar Vitality on the terrace of the LAAS-CNRS. Selected PV strings, power supply panel and weather station. 175
5.12 Example of measurements obtained with the weather station. 176
5.13 Example of current and voltage measurements taken with the last version of Solar Vitality. 176

6.1 Example of snail trail/track fault 180
6.2 Example of current of three PV modules in status of health, Healthy (yellow), other fault (blue) and snail trail (red) 180
6.3 Behavior of the current over one day for different health statuses: healthy (yellow), broken glass (blue), and snail trails (red) for a period of 13 hours every minute. 182
6.4 Decomposition into 3 levels of the current signal for a panel with big snail trails (red) and a healthy panel (yellow). The approximation and detail coefficients resulting from the decomposition are presented on the left and right of the figure, respectively. 185
6.5 Correlation matrices: (**a**) Pearson correlation matrix of the η_b initial features of $\mathbb{F}_{morning}$, i.e., before correlation based feature selection; (**b**) Pearson correlation matrix of the η_c uncorrelated features of $\mathbf{F_{morning}}$, i.e., after correlation based feature selection. 190
6.6 Parallel coordinates plot on the matrix $\mathbf{F_{morning}}$. The normalized values of the uncorrelated features η_c are plotted on the vertical axis. The horizontal axis represents the uncorrelated features η_c. 191

7.1 Description of stages of the proposed approach. a) Data acquisition and preprocessing. b) Feature extraction. c) Feature selection. d) Fault detection and classification based on Ensemble Learning. . . . 198
7.2 Electric current signals from 8 photovoltaic modules used in the training of the proposed methodology. The signals are captured during a full day in the 4 seasons of the year. for different health states: healthy (orange) and snail trails (red). The data is captured with a frequency of one minute. The 4 time slices proposed [Sepúlveda Oviedo 2022] and adopted in this chapter are represented using dotted lines. 201

7.3 Example of classifying a new sample using K-Nearest-Neighbor (kNN). 203
7.4 Example of classification hyperplane representation of SVM algorithm. H_1 and H_2 correspond to the hyperplanes of classes 1 and 2 respectively. H corresponds to the optimal hyperplane. 204
7.5 Structure of the Decision Tree (DT) classifier. The DT classifier is composed of two types of nodes. The first type corresponds to the decision node and the second to the leaf node. 206
7.6 Example of fault classification of a PV panel with snail trail using EL based on Majority voting. 208
7.7 Electric current signals from 8 photovoltaic modules used in the testing of the proposed methodology. The signals were captured during a full day in the 4 seasons of the year. The data is captured with a frequency of one minute. The 4 time slices proposed [Sepúlveda Oviedo 2022] and adopted in this chapter are represented using dotted lines. 209
7.8 Confusion matrix of the results of the classification algorithms for each season of the year, after dimensionality reduction using PCA. The class 0 corresponds to healthy panels and the class 1 corresponds to panels with a snail trail. 212

8.1 The five stages of the proposed approach. *i)* Data acquisition and preprocessing; *ii)* DTW Hierarchical clustering; *iii)* Feature extraction; *iv)* Feature selection; and *v)* Health status diagnosis. 216
8.2 Behavior of the current over one day for different health statuses: healthy (yellow), broken glass (blue), and big snail trails (red) for a period of 13 hours every minute. 217
8.3 Example of warping path in the distance matrix $\mathbf{D}$. Each entry $d(i, j)$ represents a local distance between the time series S and T given by the euclidean distance between each point s_i, and t_j. 219
8.4 Dendrogram of the agglomerative hierarchical clustering of current signals. Cluster A in green groups the severely affected panels (broken glass). Cluster B of color B groups the healthy panels and those with big snail trails. 220
8.5 Diagnosis of the PV panels of cluster B with PLS. Diagnosis accuracy with R^2 and $RMSE$ metrics for the four time slices *morning*, *midday*, $afternoon$, and *evening* with PLS. 224
8.6 diagnosis of the PV panels of cluster B with $PLS-LDA$. Diagnosis accuracy with $F-Value$ metric for the four time slices for the time slices *morning*, *midday*, $afternoon$, and *evening* with $PLS-LDA$. 226

9.1 General Scheme of operation. 230
9.2 Offline operation scheme (server). 231
9.3 Offline operation scheme (server). 233

A.1 External connection of the sensors: Irradiance (yellow), Temperature (blue), Current (red) and Electrical power supply (white arrow) . . . 253
A.2 Irradiation comparison. 254
A.3 Current comparison. 254
A.4 Irradiation comparison. 254
A.5 Current comparison. 255
A.6 Data acquisition platform 255
A.7 Positioning of sensors on the panel 256
A.8 Acquisition every second. 256
A.9 Acquisition every 100 milliseconds. 256
A.10 Test of measurements made every 55 milliseconds. 257
A.11 K-means clustering applied to the time series of the current $I_{i\{1:n_I\}}$ of each PV panel PV_i, $i = 1, \ldots, n$. a) the original signals. b) the centroids found using the k-means algorithm. c) An overlay of the original signals and the centroids 259
A.12 Example of classifying a new sample using Random forest (RF) model 261
A.13 Confusion matrix of the results of the RF model. The class 0 corresponds to healthy panels and the class 1 corresponds to panels with a Snail Trail. 262

List of Tables

2.1 Confirmed single-junction terrestrial cell and submodule efficiencies measured under the global AM1.5 spectrum (1000 W/m^2) at 25°C . 47

2.2 Report on the rate of faults observed in PV modules according to the climate [Jordan 2017]. 74

2.3 Definition of potential risks. The risks are divided into human safety risks and system power loss risks. 79

2.4 Summary of the impact of common PV faults. Classification based on the element, the main consequence, the human safety risk and the loss of power. 79

3.1 20 examples of representative words for the 4 topics in the *Optimal LDA Model*, as well as the number of documents in each topic 103

4.1 Parameters to be measured according to BS IEC 61724 [IEC 1998]. . 134

4.2 Conventional PV data acquisition system developed 146

4.3 PV module specifications at STC. 151

5.1 Comparative table of characteristics of different Arduino models . . 163

5.2 Energy consumption and uncertainty of the elements of Solar Vitality. 165

7.1 Fault detection and classification results (F_{value}) for signals captured in Summer. In Without Approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed. 210

7.2 Fault detection and classification results (F_{value}) for signals captured in Fall. In Without Approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed. 211

7.3 Fault detection and classification results (F_{value}) for signals captured in Winter. In Without Approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed. 213

7.4 Fault detection and classification results (F_{value}) for signals captured in Spring. In Without Approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed. 214

8.1 Diagnosis accuracy for the four time slices with the PLS and $PLS-LDA$ methods. 226

9.1 Typical values of the coefficients of Equation (9.15). 241

A.1 Fault detection and classification results with Random forest (RF) model. 261

Glossary

D_w	Wind direction
G	Solar radiation
G_I	In-plane irradiance
I_0	Saturation current of the diode
I_A	Output current
I_{BU}	Output current
I_{FS}	Current from storage
I_{FU}	Current from utility grid
I_L	Load current
I_{ph}	Photocurrent
I_{rs}	Current to storage
I_{ru}	Current to utility grid
Isc	Short-circuit current
K	Boltzmann's constant
n	Ideality factor
P_A	Output power
P_{BU}	Output power
P_{FS}	Power from storage
P_{FU}	Power from utility grid
P_L	Load power
P_{rs}	Power to storage
P_{ru}	Power to utility grid
q	electron charge
R_{sh}	Shunt resistor
Rs	Series resistor

S_w Wind speed

T Absolute temperature

T_{am} Ambient temperature

T_{am} Load voltage

T_m Module temperature

V_A Output voltage

V_{BU} Output voltage

V_{cell} Solar cell terminal voltage

V_{oc} Open circuit voltage

V_S Operating voltage

V_T Thermal Voltage

V_U Utility voltage

DFT Discrete Fourier Transform

DT Decision Trees

DWT Discrete Wavelet Decomposition

EL Ensemble Learning

FT Fourier Transform

$Isomap$ Isometric Mapping

kNN K-Nearest Neighbor

LLE Local Linear Embedding

$MBDM$ Model-Based Difference Measurement

MSD Multiresolution Signal Decomposition

MT Majority voting

PCA Principal Component Analysis

PSD Power Spectral Density

SVM Support Vector Machine

WPT Wavelet Packet Transform

WT Wavelet Transform

Résumé

L'énergie photovoltaïque a pris une place vraiment importante parmi les énergies renouvelables, atteignant une capacité installée mondiale cumulée d'environ 75 GW en 2016 [energy agency 2016] et selon le rapport NREL [Feldman 2022], environ 171 GW de PV seront seront installés dans le monde en 2021, et ils prévoient que dans les années 2022 et 2023, 209 GW et 231 GW seront installés, respectivement. Même des rapports tels que GlobalData estiment que la capacité photovoltaïque installée dans le monde dépassera 1 500 GW en 2030 [Data 2019]. Dans ces installations photovoltaïques, le diagnosis de l'état de santé des composants et des systèmes est essentiel pour garantir la production d'énergie, prolonger la durée de vie utile et prévenir les événements imprévus dans les systèmes solaires photovoltaïques. Dans une étude, Solar Power Europe study [SPE 2022] analyse la capacité solaire photovoltaïque installée cumulée dans le monde, elle a augmenté de 22% pour atteindre 940,0 GW fin 2021, contre 772,2 GW en 2020, comme on peut le voir sur la Figure 2.4. Cela signifie que l'énergie solaire totale a été multipliée par plus de 500 depuis le début du millénaire, lorsque l'ère solaire connectée au réseau a commencé avec le lancement de la loi allemande sur les tarifs de rachat [SPE 2022].

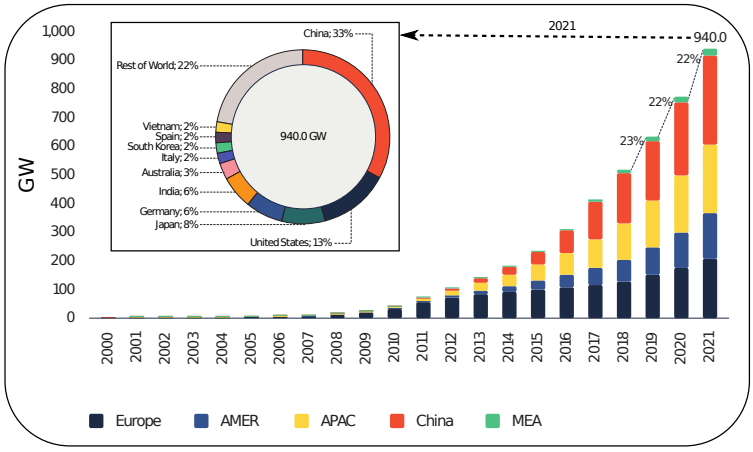

Figure 1: Capacité installée solaire photovoltaïque totale 2000-2021 et parts totales installées des 10 principaux marchés solaires photovoltaïques 2021 [SPE 2022].

En comparant les valeurs de la capacité solaire photovoltaïque installée accumulée pour l'année 2021 et 2010, on peut observer une augmentation de 41,3 GW à 940,0 GW (environ 1 TW), ce qui représente une augmentation impressionnante d'environ 2176,0 %. Dans la même Figure 2.4, une comparaison de pays individuels peut être observée, que la Chine est suivie par les États-Unis, le Japon, l'Allemagne, l'Inde et l'Australie. Enfin, il est intéressant de savoir quelles sont les perspectives de la filière photovoltaïque pour les années à venir. Dans [SPE 2022] une analyse prédictive de l'industrie photovoltaïque jusqu'en 2026 est réalisée. Dans cette analyse, 3 scénarios de marché sont proposés tels que présentés dans la Figure 2.5.

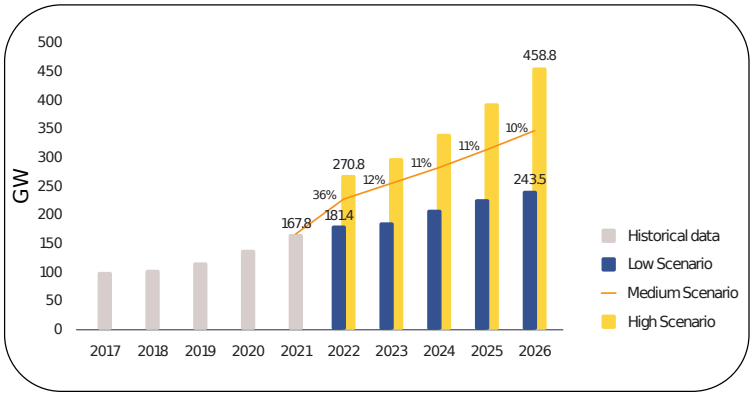

Figure 2: Scénarios annuels mondiaux du marché solaire photovoltaïque 2022-2026 [SPE 2022].

Comme le montre la figure [SPE 2022], dans le scénario moyen, d'ici 2022, les nouvelles capacités installées devraient atteindre 228,5 GW d'ici la fin de 2022, ce qui représente un taux de croissance de 36 % sur la 167,8 GW installés en 2021. Le scénario bas estime une baisse de la demande à 181,4 GW d'ici fin 2022, ce qui, comme mentionné [SPE 2022], est vraiment improbable compte tenu de la forte demande d'énergie solaire ces dernières années. Enfin, le scénario haut prévoit jusqu'à 270,8 GW d'ajouts solaires en 2022. De plus, comme on peut le voir sur la même Figure [SPE 2022], il est prévu qu'en 2026 les capacités installées seront comprises entre 243,5 GW , dans le pire des cas, et 458,8 GW dans le meilleur des cas, soit 1,7 la capacité installée dans le meilleur des cas en 2022.

Dans les grandes installations photovoltaïques représentant plus de 100 kWc et un hectare de surface, on comprend qu'il devient rapidement plus difficile de détecter ou d'identifier l'état physique d'un défaut et même sa nature. Cet objectif peut être difficile à atteindre dans certains cas où la valeur par défaut est cachée dans toutes les caractéristiques de l'installation PV jusqu'à ce qu'il y ait un réel impact sur la production du système PV. Dans le monde entier, certains des aspects qui rendent difficile la détection des défauts sont : i) l'occurrence dans des scénarios de faible irradiation [Yi 2017c]; ii) les défauts qui se produisent en moins d'une seconde [Wang 2013]; iii) présence du dispositif MPPT qui optimise la puissance de sortie d'un champ photovoltaïque [Zhao 2013a]; ou iv) des défauts avec un comportement électrique similaire à celui d'un panneau sain [Hariharan 2016a, Sepúlveda Oviedo 2022]. Si ces défauts ne sont pas détectés, non seulement l'état de la centrale peut se détériorer, mais cela représente également un danger pour la sécurité humaine [Strobl 2010, Wang 2014a, Rabla 2013], générant même de grands incendies [Brooks 2011, Ministry of Housing 2017].

En tant que solution pour évaluer les performances du système PV et calculer la perte d'énergie sur de longues périodes des installations PV, de nombreuses installations PV disposent aujourd'hui de données PV massives (à la fois instantanées et historiques) provenant de sources telles que les stations météorologiques, les onduleurs PV et le réseau public [Zhao 2015a]. Cependant, ces plates-formes de

surveillance ne sont pas orientées vers un diagnosis précis des défauts et ne prêtent donc pas attention à la qualité des données ou n'ont pas de soumissions de données à fréquence d'échantillonnage élevée.

Comme recommandé dans [Dhimish 2018a], pour augmenter la précision du diagnosis, il est nécessaire d'augmenter le temps d'échantillonnage même à des vitesses de quelques microsecondes. Cependant, ces types de systèmes qui capturent des données à ces vitesses sont vraiment un défi lié aux capacités des microprocesseurs utilisés dans les appareils. De plus, l'utilisation de ce type de plate-forme d'instrumentation dans les centrales photovoltaïques au sol est vraiment un défi car l'autonomie énergétique doit être garantie en raison de l'absence de prises de courant ou de systèmes d'alimentation, des conditions d'humidité, entre autres.

Malgré la complexité de ces systèmes, l'utilisation croissante du photovoltaïque et la baisse du coût des panneaux solaires suscitent de plus en plus l'intérêt des chercheurs, tant dans le milieu universitaire qu'industriel. L'objectif de cette thèse est le développement de méthodes de diagnosis de défauts embarquées dans des systèmes de surveillance d'installations photovoltaïques de forte puissance, respectant les contraintes industrielles et prenant en compte le rapport coût/bénéfice en productivité ou temps de fonctionnement. La détection précoce des défauts permet de définir efficacement les actions à mener en termes d'utilisation de la centrale photovoltaïque mais aussi en termes de maintenance corrective ou préventive, en tenant compte de l'état de santé de la centrale photovoltaïque ou encore en prenant en compte les prévisions de l'évolution des dégradations.

Motivation du projet

Des centrales photovoltaïques (PV) de grande puissance sont déployées dans le monde entier. Sa durée de vie et son utilisation doivent dépasser 25 ans pour garantir le retour sur investissement des infrastructures. Pour cela, il est préférable d'effectuer la maintenance périodiquement, ou lorsqu'un défaut grave survient et qu'une perte définitive de productivité est détectée. En effet, cette situation peut conduire à l'arrêt total ou partiel de la centrale. Chaque arrêt de production, même lié à un nombre réduit de panneaux et même temporaire, entraîne une perte économique importante dans les centrales de plus de 250 kWc, ce qui justifie la nécessité de diagnostiquer l'état de la centrale et d'anticiper les interventions en fonction des délais.

La productivité des installations photovoltaïques est fortement affectée par des aspects tels que la disponibilité et les performances. La disponibilité fait référence au rapport entre la durée de continuité de la production d'énergie, même sans performance optimale, et la durée totale observée [Díaz 2007]. Le rendement, d'après Díaz et al. [Díaz 2007], fait référence à l'efficacité globale de la chaîne de conversion de puissance. Cette performance est couramment mesurée à l'aide de "l'indice de performance" [CEC 1997, IEC 1998]. Selon des études réalisées par IEA PVPS, le taux de disponibilité annuel d'une installation photovoltaïque bien supervisée peut

atteindre 97 % [Janh 2000]. Comme mentionné dans [Bun 2011b], la productivité d'un système PV peut être améliorée en réduisant le taux de temps d'arrêt et en faisant fonctionner le système à des performances optimales.

Pour réduire les temps d'arrêt ou le temps de production (non optimal), il est nécessaire de réduire le nombre de défauts de composants et le temps de réparation préventive et corrective. C'est là qu'un système de diagnosis intégré dans une plateforme de surveillance robuste est vital pour identifier le défaut le plus rapidement possible. Il est vrai qu'un suivi ou une supervision classique des données historiques de la centrale permet d'identifier (généralement de manière imprécise) la présence d'anomalies dans la production d'énergie. Cependant, la surveillance classique ne permet pas de les détecter immédiatement ou dans des délais plus courts, de sorte que la centrale photovoltaïque continue de fonctionner dans un état sous-optimal. A ce jour, plusieurs sociétés proposent des produits d'aide à la gestion des centrales photovoltaïques, comme S4E avec le produit EnergySoft [S4E 2022] ou Circutor avec son produit scada dédié aux centrales photovoltaïques [Circutor 2022]. Certains fabricants dans le domaine vendent également des onduleurs et des panneaux en tant que SMA et proposent des produits pour aider à estimer les performances de la centrale [SMA 2022]. Cependant, les produits actuellement sur le marché ne sont souvent pas adaptés pour analyser les raisons des pertes de production et se limitent souvent à une visualisation des données, sans autre analyse. La société Feedgy Solar a fait un travail plus poussé avec son outil d'analyse Feedgy pour surveiller et diagnostiquer les systèmes photovoltaïques en ligne, sur la base des données historiques de [Feedgy 2022].

Cependant, malgré ces efforts, les systèmes de surveillance classiques n'ont pas la portabilité pour être couplés à différentes topologies d'installations photovoltaïques, c'est-à-dire qu'ils ne sont pas capables d'être couplés à la fois à des installations résidentielles et à des installations d'autres tailles et topologies telles que les grandes centrales PV au sol. Ces derniers représentent l'un des plus grands défis car ils ne disposent pas de prises de courant ou de sources d'énergie pouvant alimenter le système de détection et de surveillance des défauts, en plus des fortes conditions météorologiques qui affectent ces centrales.

Il est nécessaire de préciser que l'installation d'un système de surveillance avancé ne garantit pas une détection précoce des défauts de l'installation photovoltaïque. Ce système de surveillance doit avoir un système de diagnosis intégré capable de comparer le comportement de différentes chaînes d'une même centrale, en tenant compte : des conditions climatiques, de la technologie, de la topologie, de la dégradation, etc. En d'autres termes, il doit s'agir d'un système de diagnosis soutenu par des experts avec connaissance de la centrale photovoltaïque et mesures de son comportement électrique et météorologique. Un système aussi robuste garantirait que le système photovoltaïque ne fonctionnerait pas pendant des semaines ou des mois dans un état sous-optimal. Ce système de diagnosis de défaut doit intégrer de nouvelles techniques de pointe qui ne nécessitent pas une grande quantité de données pour détecter les phénomènes de défaut. Cette condition est vitale, car lors de l'installation du système dans la centrale, ses données historiques ne seront pas

connues, mais on souhaite également effectuer un diagnosis précoce des défauts pendant les premières heures de fonctionnement du dispositif de diagnosis. Pour cela, il est obligatoire d'augmenter et de standardiser les connaissances existantes sur les causes des défauts et leurs probabilités. Réaliser ce diagnosis précoce des défauts permet de réduire les interventions humaines et leur programmation uniquement basée sur les signaux de défaut.

Toutes ces raisons ont été à la base de cette recherche et révèlent la nécessité de construire un système de diagnosis plus sophistiqué pour détecter et diagnostiquer les défauts afin d'améliorer la productivité de l'installation photovoltaïque. L'énoncé formel du problème de recherche est présenté ci-dessous.

Déclaration de problème

Cette thèse aborde le problème de la perte de performance de l'ensemble du système photovoltaïque et de la réduction de la puissance de sortie générée par l'apparition de défauts. Cette thèse décrira quelques symptômes ou signatures permettant d'identifier les principaux défauts des systèmes photovoltaïques. En outre, il contribuera à l'amélioration de la surveillance classique, en proposant et en construisant une nouvelle plate-forme polyvalente, portable et autonome, au niveau de l'alimentation en énergie, d'acquisition de données et de détection de défaut associée à une station météo proposée et construite qui surveille la vitesse du vent, la température ambiante et irradiation. Cette plateforme dispose de deux systèmes intégrés qui fonctionnent en collaboration. Le premier système intégré s'est concentré sur la collecte et le prétraitement des données axées sur le diagnosis. Le deuxième système aborde le problème du diagnosis des défauts dans les systèmes photovoltaïques avec peu de données. Ce deuxième système embarqué effectue la détection des défauts pendant de petits intervalles de temps et est composé d' algorithmes d'apprentissage automatique qui pourraient être utilisés pour détecter les défauts dans les centrales photovoltaïques de différentes configurations et technologies.

Les défauts trouvés dans cette thèse incluent la trace d'escargot, l'ombrage et le verre cassé. Le défaut de trace d'escargot est inclus pour tester le niveau de diagnosis fin du système embarqué. Ce type de défaut est très difficile à détecter du fait de sa signature électrique très proche de celle d'un panneau sain. Les traces d'escargots extérieures apparaissent sous la forme de lignes brunes décolorées, en particulier autour des bords des cellules et des zones de microfissures. Ce phénomène a été attribué à l'entrée d'humidité et d'oxygène à travers les microfissures, de plus ce défaut peut aggraver les microfissures ou déclencher d'autres défauts plus sévères. Des défauts de type ombrage et verre cassé ont été inclus dans l'intérêt de tester certains défauts courants dont les niveaux d'impact étaient supérieurs à ceux du trace d'escargot. Cette variété de pertes de production nous permet de démontrer que le système proposé est capable de détecter des défauts dans toute la plage d'impact. La Figure 3 montre les trois types de défauts mentionnés.

3

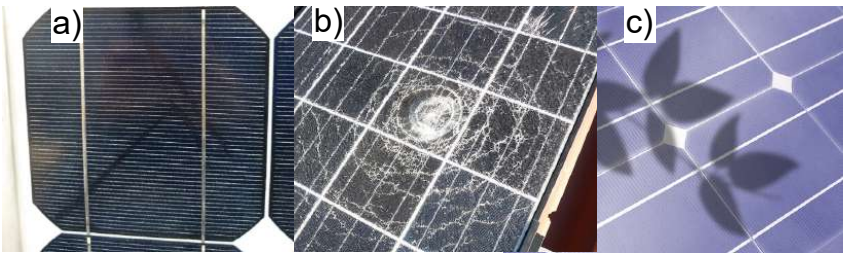

Figure 3: Défauts analysés dans les systèmes PV. a) Traces d'escargot; b) Verre cassé; c) Ombrage

Plusieurs approches ont effectué la détection de défauts dans les systèmes photovoltaïques sur la base de l'analyse de la courbe caractéristique I(V). Ce type de détection est limité pour sa mise en œuvre à grande échelle puisque pour obtenir la courbe caractéristique il faut couper la production du système, rendant cette approche irréalisable du point de vue industriel. Ainsi, cette thèse traite également de la détection de défauts dans les signaux électriques de la centrale en production. Notre système de diagnosis n'a pas besoin de couper ou de suspendre la production de la centrale pour réaliser le diagnosis. Tout le développement de cette thèse est également conçu en respectant les contraintes industrielles et en prenant en compte le gain de coût/productivité ou l'engagement du temps d'exploitation.

Objectifs

L'objectif principal de ce travail de recherche est de concevoir et mettre en œuvre une plate-forme physique d'acquisition de données et de détection de défauts. La surveillance du système photovoltaïque doit avoir une fréquence de capture de données inférieure à une seconde et le système embarqué doit être une approche basée sur l'apprentissage automatique capable de détecter plusieurs défauts qui se produisent dans les installations photovoltaïques sans grandes quantités de données historiques. Le dispositif de détection sera testé sur des centrales photovoltaïques à différentes échelles (petites, moyennes et grandes centrales), emplacements géographiques, conditions météorologiques et technologies. Pour réaliser un diagnosis efficace, il est nécessaire d'avoir une connaissance approfondie des types de défauts, des signatures associées et des méthodes actuellement utilisées. Par conséquent, dans cette thèse, le plus grand nombre d'articles associés à ce sujet sera analysé.

Il y a dix objectifs principaux à cette recherche, qui sont :

1. Proposer deux nouvelles approches pour construire un état de l'art en utilisant des techniques statistiques et d'apprentissage automatique.

2. Construire un dictionnaire de défauts, contenant la description et les principales signatures associées à un large ensemble de défauts.

3. Formuler, concevoir et construire un nouveau système polyvalent d'acquisition de données photovoltaïques. Cette plateforme doit être portable, autonome en énergie et capable de recevoir le signal analogique des capteurs de comportement électrique (tension et courant) et des données météorologiques.

4. Concevoir et construire une station météorologique pouvant être facilement couplée à une centrale photovoltaïque et compatible avec la plate-forme d'acquisition de données. Cette station météo doit surveiller la température ambiante, la vitesse du vent, l'irradiation et envoyer le signal analogique au système d'acquisition de données.

5. Proposer un ensemble de techniques d'extraction et de sélection de caractéristiques, sur des séries temporelles, visant à détecter les défauts dans les systèmes PV.

6. Proposer une approche d'apprentissage automatique pour la détection de défauts fins tels que la traces d'escargot dans les modules PV.

7. Proposer un algorithme hybride univarié d'apprentissage automatique combinant un apprentissage supervisé et non supervisé axé sur la détection de défauts non évidents tels que les défauts de type traces d'escargot et les défauts conventionnels tels que le verre brisé.

8. Proposer un ensemble d'équations pour la normalisation des données électriques et des variables environnementales permettant de comparer les performances des centrales photovoltaïques avec différents âges de démarrage, nombre de panneaux, technologies, etc.

9. Proposer un modèle mathématique PV capable de prédire la production PV d'une centrale en fonction de variables telles que : la température ambiante, la vitesse du vent, l'irradiation, les caractéristiques dans des conditions standard des panneaux PV d'une centrale PV, la date de l'installation de la centrale photovoltaïque, entre autres aspects. Le comportement du modèle doit être comparé aux données réelles d'une centrale photovoltaïque.

10. Proposer une nouvelle approche d'apprentissage automatique adaptatif qui combine l'apprentissage supervisé et non supervisé, ainsi que l'apprentissage basé sur des modèles et des données. De plus, cette approche doit utiliser plusieurs variables électriques et environnementales pour améliorer la détection des défauts au niveau de la branche PV. Enfin, le système doit être intégré dans le nouveau système d'acquisition de données photovoltaïques, être capable de se mettre à jour au fur et à mesure que de nouvelles données provenant d'autres centrales photovoltaïques sont collectées et être capable d'effectuer une détection de défauts dans des centrales photovoltaïques de différentes technologies et topologies.

Étude de cas de défaut à l'aide de l'apprentissage automatique

Sur la base des informations compilées dans cette thèse sur l'état de l'art des défauts possibles dans les champs photovoltaïques, cette recherche choisit d'étudier principalement les défauts de type Snail Trail et Broken Glass (trace d'escargot et verre cassé). Ce choix est dû au fait que les défauts de type verre cassé sont à l'origine de la plus grande perte de production dans les systèmes photovoltaïques. Au contraire, la défaut de type trace d'escargot n'entraîne pas de réduction significative de la production dans les systèmes photovoltaïques, cependant, ce défaut est à l'origine de multiples défauts sévères pouvant générer la perte totale de production d'énergie voire des incendies. Tenant compte de cela, cette recherche part de l'hypothèse que si les algorithmes proposés parviennent à détecter les défauts qui se situent dans les limites supérieure et inférieure de perte de puissance, ils sont capables de détecter toute la gamme de défauts qui se produisent entre eux. Ces défauts sont étudiés uniquement avec des données réelles provenant d'un système photovoltaïque spécifiquement conçu pour cette étude.

Apports de la thèse et aspects innovants

Pour répondre aux problématiques abordées ci-dessus, cette recherche présente des contributions sur dix aspects :

1. **Une revue de l'état de l'art en matière de détection de défauts dans les systèmes PV** qui comprend des méthodes de détection conventionnelles et des méthodes avancées basées sur l'apprentissage automatique. Dans ce contexte scientifique, cette recherche est menée avec deux nouvelles méthodologies d'analyse computationnelle et systématique de la littérature. Ces nouvelles approches peuvent être facilement extrapolées sur la base de la bibliométrie et de la modélisation thématique et couvrir davantage d'articles pour avoir une idée plus précise de l'état actuel de l'art. En outre, ce type de revue présente non seulement des articles pertinents, mais analyse également des aspects tels que : i) les relations de travail collaboratives existantes entre les pays, les auteurs, les institutions scientifiques et les algorithmes d'apprentissage automatique les plus performants dans le domaine en fonction du type d'apprentissage (supervisé, non supervisé, renforcement, semi-supervisé) et les familles de l'algorithme maître d'apprentissage automatique [Domingos 2015]. Cela permet d'identifier, en fonction des conditions du problème, les algorithmes les plus appropriés pour la détection de défauts. Enfin, cette analyse détermine des sujets de recherche intéressants et des défis liés à la détection de défauts dans ces systèmes.

2. **Dictionnaire formel des défauts** qui contient quatre types de sources de défauts identifiés : causes externes, interaction matérielle, vieillissement des

composants ou causés par d'autres défauts (cercle cause-effet). À son tour, au sein de ce dictionnaire, une nouvelle classification à plusieurs niveaux des défauts du système est proposée selon le type de défaut, le composant où il se produit (cellule, module, vieillissement, système de protection ou boîte de jonction), qu'il soit structurel, électrique , causé par des augmentations anormales de température (point chaud), de mauvaises connexions ou d'ombre (à cause d'obstacles ou de saleté). Chaque défaut est exposé avec son explication et son illustration graphique. Ce dictionnaire inclut également les aspects de fréquence d'occurrence et d'impact en termes de sécurité humaine et de perte d'énergie.

3. **Une nouvelle plate-forme d'acquisition de données et de surveillance nommée Solar-Vitality** dans les systèmes PV visant à diagnostiquer les défauts dans les chaînes PV. Cette plateforme contient deux systèmes embarqués en charge de : i) un nouveau système de surveillance photovoltaïque polyvalent qui capte le courant et la tension au niveau de la chaîne PV; et ii) détection automatique des défauts.

4. **Une station météo portable** adaptable à différentes configurations de centrales photovoltaïques. Cette station météo capture des variables climatologiques telles que la température ambiante, la vitesse du vent et l'irradiation.

5. **Une contribution au traitement et à l'analyse du signal pour l'extraction et la sélection des caractéristiques de défaut** qui comprend un ensemble d'opérations de transformation effectuées sur les signaux de défaut comme guide pour augmenter la richesse des signaux analysés par les algorithmes d'apprentissage automatique et ainsi améliorer la capacité de discriminer entre les classes.

6. **Un algorithme d'apprentissage d'ensemble nommé EB-diag** capable de détecter les défauts de trace d'escargot dans les modules PV. EB-diag combine plusieurs modèles d'apprentissage, plutôt que d'utiliser un seul modèle d'apprentissage. De plus, cette approche tire parti des techniques d'extraction et de sélection de caractéristiques du point 5, améliorant considérablement la précision de la détection. Les résultats de cette approche démontrent qu'elle est capable de classer les modules PV sains et ceux avec des traces d'escargot de manière efficace et rentable, car elle utilise uniquement le signal de courant électrique des modules obtenu à partir de systèmes d'acquisition de données PV standard. De plus, l'approche est générique et peut être facilement extrapolée à d'autres problèmes de diagnosis dans d'autres domaines.

7. **Un nouvel algorithme hybride d'apprentissage automatique nommé Serial-diag** pour la détection de défauts dans les systèmes photovoltaïques. Ces approches sont même capables de détecter et de diagnostiquer des défauts comme la traînée d'escargot dont le comportement est similaire à celui d'un

panneau sain. Cet algorithme est également testé pour détecter les panneaux avec du verre brisé, en réussissant à les classer efficacement. De plus, cette approche proposée s'est avérée très rapide en termes de calcul car, grâce à la combinaison proposée d'apprentissage non supervisé et supervisé, le calcul le plus lourd n'est effectué que sur une partie des panneaux défaillants.

8. **Une méthode de normalisation efficace pour les données des centrales photovoltaïques** qui apporte une contribution importante. Ce type d'approche permet non seulement de comparer des chaînes PV avec différents nombres de panneaux PV, mais également avec différentes températures, irradiations, vitesses de vent et même technologies. Cette approche inclut également le facteur de dégradation de la centrale photovoltaïque.

9. **Un modèle de prédiction de puissance PV ajusté aux données réelles** qui utilise les variables suivantes : température ambiante, vitesse du vent, irradiation, puissance STC, nombre de panneaux connectés en série, date d'installation de la centrale et taux de dégradation annuel. Grâce à tous ces paramètres, le modèle proposé est capable d'estimer la production PV au niveau du module PV ou de la chaîne PV. Les résultats d'estimation du module sont comparés aux données réelles d'une centrale photovoltaïque, obtenant un niveau élevé de coïncidence avec les données réelles. Ce modèle est également utilisé pour générer une stratégie d'augmentation de données et de génération de défauts synthétiques, comme solution aux problèmes de quantité insuffisante de données ou de données déséquilibrées.

10. **A new approach to machine learning** integrated into the new versatile data acquisition system Solar Vitality.

11. **Une nouvelle approche d'apprentissage automatique adaptatif nommée Adaptive-diag** intégrée dans le nouveau système d'acquisition de données polyvalent Solar Vitality. Cette approche combine l'apprentissage supervisé et non supervisé, ainsi que l'apprentissage basé sur des modèles et des données. Cette approche utilise les données de la modélisation du point 9, les techniques décrites au point 5, ainsi que la normalisation du point 8, pour détecter, localiser et identifier les défauts dans les systèmes PV. Cette approche est capable d'utiliser la vitesse du vent, la température ambiante, l'irradiation, les informations de la fiche technique et l'âge de l'installation photovoltaïque pour générer automatiquement un groupe de panneaux ou de chaînes de référence sains. Avec ces informations, cette approche est non seulement capable de détecter les défauts du système PV, mais également de générer automatiquement un rapport de priorité de maintenance pour les panneaux défectueux. Ce système est également évolutif, car au fur et à mesure que des erreurs sont détectées et que de nouveaux échantillons sont classés, la base de données interne s'agrandit. Une fois que de nouveaux clusters sont détectés, le système s'entraîne et met à jour le modèle de diagnosis de défauts.

Cette thèse démontre plusieurs importantes contributions à la recherche et avancées scientifiques par rapport aux solutions existantes. Parmi les produits académiques obtenus avec cette thèse figurent:

Conférences internationales

1. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Corinne Alonso and Marko Pavlov. *Hierarchical clustering and dynamic time warping for fault detection in photovoltaic systems.* In X Congreso internacional Ingeniería Mecánica, Mecatrónica y Automatización, Bogotá, Colombia, May 2021. (Accepté et présenté)

2. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov, Corinne Alonso. *DTW k-means clustering for fault detection in photovoltaic.* In XI Congreso internacional Ingeniería Mecánica, Mecatrónica y Automatización, Cartagena, Colombia, May 2023. (Accepté et présenté)

3. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov, Corinne Alonso. *Detection and classification of faults aimed at preventive maintenance of pv systems.* In XI Congreso internacional Ingeniería Mecánica, Mecatrónica y Automatización, Cartagena, Colombia, May 2023. (Accepté et présenté)

Conférences nationales

1. Edgar Hernando Sepúlveda Oviedo. *Extraction de signatures et prédiction de l'état de santé des centrales photovoltaïques.* In Journée annuelle de l'école doctorale Geets, Toulouse, France, April 2022. (Accepté et Présenté)

Workshops

1. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov, Corinne Alonso. *Fault detection and diagnosis for PV systems using machine Learning*, Poster, In 9th NextPV workshop, online edition, November 2020. (Accepté et Présenté)

2. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov, Corinne Alonso. *Acquisition de données, et prédiction de l'état de santé de systèmes photovoltaïques.* Oral presentation. In Workshop DO, Mauvezin, October 2021. (Accepté et Présenté)

3. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov, Corinne Alonso. *Advanced machine learning methods, for the detection of fine faults in PV systems, aimed to preventive maintenance.* Oral presentation. In 10th NextPV workshop, Bordeaux, France, January 2023. (Accepté et Présenté)

Articles de revues scientifiques

1. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Corinne Alonso and Marko Pavlov. *Feature extraction and health status prediction in PV systems.* Advanced Engineering Informatics, vol. 53, page 101696, 2022. (journal Q_1)

2. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov and Corinne Alonso. *Artificial intelligence based fault diagnosis in photovoltaic systems Part I: A Bibliometric survey.* Renewable and Sustainable Energy Reviews, vol. 00, page 00, 2022. (Soumis dans un journal Q_1)

3. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov and Corinne Alonso. *An Ensemble Learning-Based Fault Detection and Diagnosis for PV modules.* Sustainability, vol. 00, page 00, 2022. (Soumis dans un journal Q_1)

4. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov and Corinne Alonso. *Artificial intelligence based fault diagnosis in photovoltaic systems Part II: A topic modeling approach.* Renewable and Sustainable Energy Reviews, vol. 00, page 00, 2022. (Soumis dans un journal Q_1)

Brevet

1. Patent Feedgy/LAAS-CNRS (en procès)

Prix obtenus

- Prix de la meilleure communication orale à la conférence sur le génie électrique et la gestion de l'énergie à *Journée annuelle de l'école doctorale Geets.*

Plan de thèse

Le reste de la thèse est organisé en neuf chapitres comme suit :

Chapter 1:

Ce chapitre aborde l'intérêt académique et industriel sous lequel cette thèse est développée, l'approche formelle du problème, le but et les objectifs et les cas qui sont étudiés dans la thèse. Ainsi que les produits académiques obtenus à partir de cette thèse

Chapitre 2

Trois aspects principaux sont présentés dans ce chapitre. En premier lieu, un état de l'art sur les éléments constitutifs d'un système photovoltaïque classique. Deuxièmement, un dictionnaire formel des défauts est proposé qui contient quatre sources de défauts identifiées : les causes externes, l'interaction matérielle, le vieillissement des composants ou causés par d'autres défauts, dans ce que nous appelons le cercle de cause à effet. Dans ce même dictionnaire, une nouvelle classification multiniveau des défauts du système est proposée selon le type de défaut, le composant où elle se produit (cellule, module, vieillissement, système de protection ou boîte de jonction), qu'elle soit structurelle, électrique, par élévation de température anormale (point chaud), mauvaises connexions ou ombrage (ombrage ou saleté). Enfin, ce dictionnaire, basé sur une revue bibliographique intense, contient également une description de chaque défaut en fonction de sa signification ainsi qu'un support illustratif, la fréquence d'occurrence et l'impact en termes de sécurité humaine et de perte d'énergie sont exposés. Enfin, le troisième aspect abordé dans ce chapitre traite des méthodes classiques de détection de défauts dans les systèmes photovoltaïques, en les divisant en cinq grandes catégories : méthodes visuelles, méthodes basées sur l'image, méthodes de détection électrique, technique basée sur des dispositifs de protection et détecteur. détection (AFD).

Ce chapitre part de l'hypothèse que le large éventail de conditions et de scénarios dans lesquels un défaut peut se produire signifie que le choix de la méthode de détection de défaut dépend des connaissances disponibles sur le système (composants du système), parfois de sa taille (niveau d'impact du défaut), ses caractéristiques (électriques, thermiques...), son origine et le type de défaut à diagnostiquer. Pour cette raison, il est évident que des défis subsistent autour de la détection de défauts dans le système photovoltaïque.

De plus, les conditions de détection des défauts dans les systèmes photovoltaïques deviennent plus difficiles lorsque l'influence des conditions météorologiques [Yi 2017c] et ses changements continus de conditions, les sorties non linéaires du système [Fadhel 2018], la la présence ou non de dispositifs de suivi du point de puissance maximale (MPPT) [Zhao 2013a], l'apparition de plusieurs défauts simultanés ou de défauts primaires générant des défauts secondaires plus graves. En fait, il existe également d'énormes scénarios dans ce domaine où le comportement électrique des panneaux avec défauts est très similaire à celui des panneaux sans défaut [Hariharan 2016a, Sepúlveda Oviedo 2022]. Compte tenu de tout cela, il est évident que les approches présentées dans ce chapitre atteignent leurs propres limites lorsque les conditions mentionnées ci-dessus sont remplies ou lorsque de grandes quantités de données de grande dimension (Big data) sont introduites dans les nouveaux systèmes de surveillance.

Par conséquent, cette recherche considère que les stratégies de détection des défauts dans les systèmes PV peuvent être améliorées pour obtenir des systèmes plus efficaces avec des fonctions de détection prédictive des défauts et être appliquées à une large gamme de centrales PV avec divers systèmes de surveillance. De plus,

cette recherche accorde une importance particulière à la détection des défauts avec des signatures similaires à celles des panels sains, en tenant compte de la classification multiniveaux présentée dans ce chapitre et des relations entre les défauts dans le cercle cause-effet. Ceci est primordial, car la détection d'un défaut à l'origine de plusieurs défauts graves augmenterait considérablement le niveau de précision des systèmes de détection de défauts.

Cette thèse considère que des améliorations peuvent être fortement soutenues par des approches d'intelligence artificielle en raison de la capacité de l'intelligence artificielle (IA) à gérer des données multivariées de grande dimension et à extraire des relations cachées au sein des données dans des environnements complexes et dynamiques [Wuest 2016]. De plus, les nouvelles approches doivent être capables de détecter les défauts non seulement lorsque l'impact est sévère mais même dès le début ou lorsque leurs signatures sont similaires à celles d'un panneau sans défaut, comme dans le cas du Snail Trail. Ces approches sont essentielles pour éviter les conséquences destructrices et les risques pour le personnel qui entre en contact avec le système PV. En tenant compte de tous ces aspects, le chapitre 3 traite d'une analyse détaillée de l'état actuel de l'art sur les questions de détection de défauts dans les systèmes PV utilisant l'intelligence artificielle.

Chapitre 3

Ce chapitre présente une étude approfondie de l'état de l'art sur les techniques d'intelligence artificielle utilisées pour la détection de défauts dans les systèmes photovoltaïques. Dans ce chapitre, deux méthodologies computationnelles innovantes qui combinent l'apprentissage automatique, la bibliométrie et l'analyse d'experts sont proposées et utilisées pour extraire les informations pertinentes qui déterminent les domaines de recherche actuels et les défis dans ces domaines. De plus, cela permet de positionner la recherche présentée dans cette thèse en effectuant une analyse approfondie de l'état de l'art qui réduit la subjectivité existante dans les revues conventionnelles et positionne le lecteur à l'avant-garde dans la compréhension des aspects de la détection efficace des défauts dans les systèmes photovoltaïques.

Dans ce chapitre de revue, plus de 620 articles publiés depuis 2010 sur les méthodes d'intelligence artificielle pour la détection de défauts dans les systèmes photovoltaïques sont analysés. Pour extraire les grandes tendances de la recherche, en particulier pour repérer les algorithmes et les approches les plus prometteurs qui s'affranchissent des temps de calcul excessifs, une revue de littérature conventionnelle aurait été extrêmement difficile à réaliser. C'est pourquoi dans ce travail il est proposé de réaliser une revue avec une approche innovante basée sur une méthode statistique dite Bibliométrique et une Analyse Experte de Contenu Qualitative. L'approche bibliométrique a un caractère générique qui peut être facilement utilisé et extrapolé dans les années futures ou appliqué dans d'autres domaines. Cette méthodologie se compose de trois étapes. Tout d'abord, une collecte de données à partir de bases de données est réalisée avec toutes les précautions pour aboutir à une base de données volumineuse, robuste et de qualité. Deuxièmement, de mul-

tiples indicateurs bibliométriques sont choisis en fonction des objectifs à atteindre et analysés pour évaluer leur impact réel, tels que le nombre et le type de publications, les liens de collaboration entre institutions, auteurs et pays, les articles les plus cités, etc. Enfin, l'analyse qualitative du contenu expert réalisée par des experts identifie les sujets de recherche actuels et émergents ayant le plus grand impact sur la détection de défauts dans les systèmes photovoltaïques utilisant l'intelligence artificielle.

De manière complémentaire, dans ce même chapitre, une revue alternative est proposée en utilisant une méthodologie innovante qui combine deux méthodes d'apprentissage automatique : la modélisation thématique (topic modeling) et l'incorporation de voisins stochastiques distribués en t (t-distributed stochastic neighbor embedding t-SNE). Ensuite, un processus d'analyse qualitative du contenu des sujets guidé par des experts permet d'extraire les informations pertinentes qui déterminent les domaines de recherche actuels et les défis dans ces domaines. Cette méthodologie, qui peut être extrapolée à d'autres domaines, réduit la subjectivité existante dans les revues conventionnelles et positionne le lecteur à l'avant-garde dans la compréhension des aspects de la détection efficace des défauts dans les systèmes photovoltaïques.

Une fois ces deux méthodologies appliquées, les résultats sont croisés pour sélectionner un ensemble d'articles représentatifs de l'état de l'art qui sont analysés en détail. Ces articles sont analysés en fonction des algorithmes qu'ils utilisent, des signaux d'entrée, des défauts détectés, du temps d'échantillonnage des signaux, des variables utilisées, entre autres.

Grâce à l'application de ces méthodologies, il est possible d'extraire les paramètres clés, les défis et les opportunités de recherche qui favorisent et orientent la recherche dans le domaine de la détection des défauts. Dans les systèmes photovoltaïques utilisant l'intelligence artificielle, une liste d'algorithmes représentatifs et des documents organisés par groupes sont fournis pour accroître la compréhension des différents types d'algorithmes à considérer dans les recherches futures. L'une des conclusions est que la recherche sur l'apprentissage automatique hybride et les études comparatives de ces méthodes dans la détection de défauts dans les systèmes photovoltaïques sont recommandées. De même, la recherche sur la prévision des systèmes PV devrait être encouragée pour fournir une maintenance préventive basée sur l'état et réduire les temps de retour sur investissement. Même si les demandes industrielles ont tendance à être directement orientées vers des générateurs photovoltaïques entiers, la surveillance au niveau des panneaux photovoltaïques a encore un long chemin à parcourir et doit être construite en parallèle avec des algorithmes d'apprentissage automatique. De plus, la méthodologie d'examen d'art utilisant ces outils innovants tire parti des dernières avancées en matière de traitement du langage naturel pour fournir des détails sur le comportement de la recherche dans un domaine de la connaissance.

Il est important de mentionner que cette recherche n'a pas pris en compte les articles qui ne présentent que le cadre général du sujet sans recherche expérimentale. L'objectif principal de ce chapitre est de favoriser le développement d'approches et

d'outils pour la détection et la classification des défauts dans les systèmes photovoltaïques. Pour cette raison, cette recherche concentre ses efforts sur la fourniture d'une vision objective, cohérente et méta-analytique des recherches actuelles sur l'intelligence artificielle durable appliquée au domaine de l'énergie. En raison du rôle important de la surveillance des systèmes PV pour la détection des défauts dans les systèmes PV, cette thèse propose une nouvelle plate-forme de surveillance des systèmes PV avec un système de détection des défauts intégré. Pour cette raison, et afin de positionner notre nouvelle plateforme de surveillance, le chapitre 4 vise à connaître les caractéristiques et les limites des systèmes de surveillance actuels, et la description de la nouvelle plate-forme construite et testée dans cette thèse.

Chapitre 4

Ce chapitre présente une revue des systèmes de surveillance actuels dans les systèmes photovoltaïques, leurs limites, avantages, inconvénients et défis de développement. De plus, une plate-forme industrielle et commerciale Tigo utilisée dans cette thèse est présentée pour tester les limites de la détection de défauts dans les systèmes photovoltaïques à l'aide de plates-formes de surveillance largement commercialisées. Enfin, un nouveau système de surveillance photovoltaïque est présenté qui suit les directives de la norme CEI 61724. Cette nouvelle plate-forme utilise la carte de développement électronique open source Arduino pour résoudre le problème actuel de surveillance des systèmes photovoltaïques (PV) avec un Raspberry Pi4 . Cette plate-forme peut être utilisée pour surveiller les défauts des systèmes photovoltaïques, des centrales résidentielles aux centrales photovoltaïques de grande puissance, dans les pays développés et en particulier dans les zones reculées ou les régions des pays en développement.

Dans ce chapitre, une revue complète des systèmes de surveillance photovoltaïques existants rapportés dans la littérature est présentée en termes de capteurs et de systèmes d'acquisition utilisés. De plus, les paramètres les plus utilisés dans la surveillance des systèmes photovoltaïques ont été analysés, parmi lesquels : la tension, le courant, le rayonnement solaire, la température et la vitesse du vent. Dans le domaine des systèmes d'acquisition de données, ce chapitre couvre les contrôleurs utilisés pour le système d'acquisition de données, les types de méthodes de transmission de données, le stockage des données et l'analyse des données. De même, il a été identifié qu'un système de transmission de données efficace est essentiel pour assurer la qualité des données, surtout compte tenu des conditions de travail difficiles auxquelles sont soumis ces systèmes d'acquisition de données. Différents moyens de transmission de données (filaires, sans fil et sur courant porteur, etc.) ont été trouvés.

Le choix du meilleur support de transmission est totalement lié aux conditions de fonctionnement du système d'acquisition de données. Plusieurs paramètres peuvent être évalués pour déterminer le meilleur support de transmission de données. Tout d'abord, la zone de couverture et la longueur de la distance entre les capteurs et le système d'acquisition de données doivent être prises en compte. Par exemple, les

câbles coaxiaux ne peuvent pas fonctionner sur de longues distances par rapport au câble à fibre optique. Cependant, pour de courtes distances, il a été démontré que de nombreuses études préfèrent les câbles coaxiaux qui montrent de grandes performances. D'autre part, le protocole WLAN ne peut couvrir qu'une petite zone d'environ 20 km^2, par rapport au GPRS-GSM, mais peut transmettre des données sur des distances de centaines ou de milliers de kilomètres via Internet. Le protocole Power Line Communication (PLC) transmet des informations sur des centaines ou des milliers de mètres en utilisant l'infrastructure filaire existante sans aucune installation supplémentaire. Cependant, il existe des problèmes de vieillissement du câblage et des causes externes qui peuvent affecter l'envoi de données sur des scénarios de plusieurs centaines de mètres. De plus, lorsque le câblage est étendu et que le signal de sortie du capteur est sous tension, des pertes de signal importantes peuvent se produire. Il est donc recommandé de concevoir le système pour une sortie de capteur de courant lorsque le signal doit parcourir de longues distances via un câble.

Chapitre 5

Comme expliqué au chapitre 4, les systèmes actuels de surveillance des centrales photovoltaïques ne sont pas conçus pour le diagnosis des défauts, encore moins, ils visent à détecter les défauts dont l'occurrence est très rapide, ce qui nécessite des vitesses élevées d'échantillonnage des données. Comme proposition pour résoudre ce problème, ce chapitre 5 présente un nouveau système d'acquisition de données orienté diagnosis nommé Solar Vitality. Solar Vitality est conçu avec un accent particulier sur sa précision ou ses incertitudes selon la norme IEC 61724 [IEC 1998]. Solar Vitality utilise la carte de développement électronique open source Arduino pour résoudre le problème actuel des systèmes photovoltaïques (PV) d'acquisition de données avec un Raspberry PI4. Solar Vitality peut être utilisé pour surveiller les défauts des systèmes photovoltaïques, des centrales photovoltaïques résidentielles aux centrales électriques, dans les pays développés et en particulier dans les zones reculées ou les régions des pays en développement. Solar Vitality répond à toutes les exigences pertinentes en termes de précision incluses dans les normes de la Commission électrotechnique internationale (CEI) pour les systèmes photovoltaïques, avec des mesures toutes les 15 millisecondes, comprenant 11 entrées analogiques pour mesurer jusqu'à 4 chaînes photovoltaïques indépendantes et les paramètres météorologiques d'irradiation, la température ambiante et la vitesse du vent. Solar Vitality est complètement autonome en termes d'alimentation électrique, portable et facilement couplée à différentes topologies de systèmes photovoltaïques. Solar Vitality est testé dans différents scénarios et avec différentes topologies de systèmes photovoltaïques en conditions réelles de production. Solar Vitality est testé en fonctionnement continu pendant plus de 6 mois, présentant un fonctionnement robuste même dans les conditions environnementales difficiles de l'été et de l'hiver en France.

Solar Vitality est capable de capturer le comportement électrique du système

photovoltaïque avec une fréquence d'échantillonnage élevée (toutes les 15 millisecondes) et des variables météorologiques telles que la température ambiante, l'irradiation et la vitesse du vent. La conception et la construction de Solar Vitality visent à répondre à deux objectifs. Le premier est de démontrer l'influence de la fréquence d'échantillonnage dans la détection et la classification des défauts dans les systèmes photovoltaïques. La seconde est de démontrer et de quantifier l'effet des variables météorologiques sur le système photovoltaïque. Ensuite, l'explication de la vitalité solaire est présentée. Les résultats des données recueillies à partir des capteurs météorologiques et électriques indiquent que le nouveau système est fiable et présente des performances comparables à celles des systèmes commerciaux. Solar Vitality présente un intérêt particulier pour la recherche et l'industrie. Enfin, Solar Vitality est facilement personnalisable pour les besoins spécifiques de chaque projet et système photovoltaïque et peut même être étendu à d'autres domaines.

Solar Vitality capture le comportement électrique et météorologique de la centrale photovoltaïque. La capture de ces deux comportements est essentielle pour améliorer la précision et le nombre de défauts différents détectés dans les systèmes PV [Blaesser 1997]. La plupart des systèmes d'acquisition de données se sont concentrés sur la capture du comportement électrique du système, car pour le comportement météorologique, ils utilisent des données compilées par des institutions nationales ou européennes [NAS 2022, Atl 2022, PVG 2022]. Ces informations sont utiles pour connaître de manière générale les conditions météorologiques de fonctionnement d'une région ou d'une zone. Cependant, lorsque l'on pense à la détection de défauts dans les systèmes PV, l'utilisation de ces données satellitaires présente plusieurs inconvénients. Premièrement, ces données ne peuvent se substituer aux données spécifiques relevées sur le site. Deuxièmement, il existe de nombreux endroits où ces bases de données ne sont pas disponibles ou sont en train d'être compilées. Troisièmement, bien qu'un large éventail de bases de données météorologiques soient disponibles, elles sont généralement coûteuses, très sophistiquées et difficiles à gérer [Fuentes 2014].

En conséquence, un développement plus poussé des systèmes d'acquisition de données est nécessaire pour collecter et traiter les données électriques et les données météorologiques en fonctionnement, dans le but d'obtenir des valeurs mesurées à l'aide de données précises et faciles à manipuler. Tenant compte de ces aspects, cette recherche présente un nouveau système d'acquisition de données orienté vers le diagnosis (Solar Vitality) qui peut être utilisé pour instrumenter des installations photovoltaïques de toute taille, au niveau d'un panneau PV, d'une chaîne PV ou d'un générateur PV. Il peut également être utilisé pour capturer des données provenant d'installations photovoltaïques avec différentes configurations. Ces caractéristiques sont essentielles pour assurer et faciliter un développement rapide et continu. De plus, Solar Vitality a une grande flexibilité et peut être adapté à chaque cas spécifique (à la fois pour la recherche et les applications industrielles dans les régions développées et en développement). Solar Vitality permettra à la communauté photovoltaïque d'avancer plus rapidement dans certains des domaines de recherche qui ont nécessité une acquisition complète de données photovoltaïques,

mais qui sont limités par des problèmes de coût et de technologie. Il est nécessaire de mentionner cela non seulement parce qu'une acquisition de données et un taux d'échantillonnage de haute qualité peuvent assurer une détection précise des défauts. Il est également nécessaire de mener des recherches approfondies sur le traitement du signal pour extraire les attributs appropriés qui permettent d'identifier et de séparer les différents états de santé de la centrale photovoltaïque. Cette analyse du traitement des attributs est traitée ci-dessous dans le chapitre 6.

Chapitre 6

Ce chapitre 5 présente un ensemble d'approches pour l'extraction et la sélection de caractéristiques et quelques exemples d'apprentissage automatique conventionnel pour la détection de défauts dans les systèmes photovoltaïques. L'objectif de ce chapitre 5 est de comprendre les limites qui existent avec certains des algorithmes d'apprentissage automatique supervisés et non supervisés les plus populaires pour la détection de défauts, tels que Snail Trail et Broken Glass. Ce chapitre 5 propose 3 méthodes d'extraction de caractéristiques. Tout d'abord, la métrique de similarité Dynamic Time Warping (DTW) est utilisée pour comparer directement les mesures actuelles des panneaux PV. Deuxièmement, une décomposition basée sur les ondelettes pour l'extraction de caractéristiques est proposée. Enfin, une extraction des caractéristiques statistiques qui caractérisent le signal électrique est proposée.

En raison de la haute dimensionnalité des matrices obtenues après extraction des caractéristiques, ce chapitre 5 propose également un ensemble de méthodes de sélection des caractéristiques. Les deux premières méthodes appelées analyse en composantes principales (ACP ou PCA en anglais) et cartographie isométrique (Isomap) sont des approches basées sur la transformation d'espaces de caractéristiques. Ensuite, deux méthodes de réduction de dimensionnalité sont proposées en analysant la corrélation et la variance des caractéristiques. Ces deux méthodes préservent les informations pertinentes pour la classification des défauts. Enfin, dans ce chapitre 5, trois algorithmes d'apprentissage automatique sont utilisés, deux non supervisés et un supervisé. Les deux algorithmes non supervisés (K-means et clustering hiérarchique) sont combinés avec la métrique DTW, tandis que l'algorithme supervisé (Random Forest) est réalisé en utilisant en entrée une combinaison des méthodes de sélection et d'extraction de caractéristiques présentées dans ce chapitre.

Les résultats présentés dans ce chapitre montrent effectivement que certaines des méthodes d'apprentissage automatique conventionnelles et les plus représentatives, à la fois non supervisées et supervisées, ont de sérieux problèmes pour détecter les défauts de Snail Trail. Comme discuté au chapitre 2, le défaut Snail Trail ne diminue pas de manière significative les performances des panneaux solaires (il émule un comportement sain) et est donc difficile à détecter au niveau du signal électrique. Cependant, comme indiqué également au chapitre 2, il s'agit d'un défaut qui peut provenir de microfissures et de fissures sévères jusqu'à la corrosion et aux

points chauds. Par conséquent, sa détection précoce est vitale.

Il est intéressant de noter que pour la détection de défauts dont la signature électrique est différente de celle d'un panneau sain, les algorithmes de détection non supervisés tels que le clustering hiérarchique et les k-means représentent une grande opportunité avec un faible coût de calcul puisqu'ils ne nécessitent pas la fonctionnalité multiple extraction. Un inconvénient des deux méthodes non supervisées est que le nombre de classes souhaitées doit être défini a priori. Par exemple, comme mentionné dans [Nielsen 2016], le résultat final du HC dépend du niveau auquel les grappes sont coupées. Cette caractéristique du HC pourrait être considérée comme un avantage, si l'on souhaite différencier le niveau d'affectation entre panneaux d'un même défaut, puisque, si la coupe -off augmente le niveau, il est possible de déterminer des sous-groupes liés au niveau du défaut. Cet aspect est essentiel pour établir une priorité dans la maintenance préventive. Pour tester la robustesse des algorithmes HC et k-means, les deux ont été testés avec des fenêtres temporelles différentes, obtenant toujours le même résultat même avec des fenêtres de 3 minutes. Cet aspect est important pour établir une priorité dans la maintenance préventive.

En ce qui concerne l'approche Random Forest (RF), il est possible de remarquer comment l'augmentation du nombre de caractéristiques augmente également la précision de la détection et de la classification des défauts. De plus, on peut voir que bien que la RF ait des limites pour la détection des traces d'escargots, elle est capable de détecter 3 panneaux sur 4 avec le défaut de trace d'escargot malgré le petit nombre d'individus (échantillons ou panneaux) et la grande similitude entre les échantillons. chacune des classes. L'un des avantages de toutes les méthodes présentées dans cette section est qu'elles parviennent à regrouper les signaux à l'aide d'une seule variable, qui est le courant du panneau. Cela peut se traduire par une réduction significative du nombre de capteurs de diagnosis de défaut dont la signature électrique est différente de celle d'un panneau sain. De plus, il n'est pas nécessaire de couper la production photovoltaïque pour réaliser le diagnosis et un nombre réduit d'individus de chaque classe (panneaux par classe) est utilisé, ce qui évite d'avoir recours à un grand nombre d'échantillons pour entraîner le système de diagnosis.

Considérant que les défauts Snail Trail sont actuellement détectés par des visites régulières du personnel dans les centrales photovoltaïques, ces approches apportent vraiment une contribution importante à la détection automatique des défauts. En général, le processus de diagnosis proposé ici au moyen de l'extraction et de la sélection de signatures avec des algorithmes d'apprentissage automatique supervisé, bien qu'encore limité en termes de détection de panel complet avec Snail Trail, est une étape vers la compréhension des limitations existantes. Pour les raisons énoncées ci-dessus, un nouvel algorithme de détection de défaut est présenté au chapitre 6. Cet algorithme essaie d'améliorer les performances de détection fine des défauts dans les systèmes PV avec des approches d'IA non conventionnelles.

Chapitre 7

La performance, la sécurité et la fiabilité des centrales photovoltaïques sont fortement liées à la capacité à détecter les pertes anormales de production d'énergie et les défauts dès leur apparition. Pour ces raisons, un objectif majeur dans ce domaine est de développer des paradigmes intelligents de détection et d'isolation des défauts (FDI) qui peuvent grandement bénéficier de l'apprentissage d'ensemble (EL). Jusqu'à présent, ces techniques appliquées aux sources d'énergie durables, telles que les centrales photovoltaïques à haute énergie, s'avéraient trop complexes. Cependant, les avancées récentes de la communauté scientifique rendent ces techniques plus applicables et peuvent donc assurer un fonctionnement performant des systèmes photovoltaïques. La technique proposée combine plusieurs modèles d'apprentissage, à savoir Support Vector Machine (SVM), K-Nearest Neighbor (kNN) et Decision Trees (DT), au lieu d'utiliser un seul modèle d'apprentissage. Le modèle combiné est orienté vers la détection de défauts classiques, mais sa particularité par rapport aux modèles existants est sa capacité à détecter des défauts dont les caractéristiques électriques sont similaires à celles d'un panneau sain. Dans la méthodologie proposée, dans un premier temps, une matrice de prédicteurs est construite en extrayant les caractéristiques temps-fréquence (à l'aide de la décomposition en ondelettes) et les statistiques du signal de courant photovoltaïque des panneaux PV. Ensuite, en raison de la dimension élevée de la matrice de prédicteurs, deux algorithmes de sélection de caractéristiques et de réduction de dimensionnalité (PCA et Isomap) sont utilisés. Enfin, la matrice de prédicteurs réduite est constituée entre les algorithmes d'apprentissage d'ensemble. Cette méthode est validée avec une vraie chaîne photovoltaïque de 8 panneaux (4 sains et 4 avec Snail Trail).

L'approche proposée dans ce chapitre 6 vise à apporter une contribution significative à la maintenance préventive des systèmes photovoltaïques. Une amélioration de la maintenance préventive des installations se traduit par une augmentation de la garantie de production continue de ces systèmes photovoltaïques. Cela devient critique si l'on tient compte du fait que les systèmes photovoltaïques distribuent environ 2% de la consommation totale d'énergie dans le monde [Pillai 2019a] et présentent des pertes de plus ou moins 18,9% par an dues à l'apparition de défauts [S 2021]. De même, ce type de recherche est vital étant donné que la croissance de l'énergie photovoltaïque devrait se poursuivre au cours des prochaines décennies, et on estime même que d'ici 2050 l'énergie photovoltaïque fournira environ 11 % de la production mondiale d'électricité et réduira de 2,3 gigatonnes. (Gt) d'émissions de CO^2 par an [IEA 2007c, IEA 2007b]. De même, il est important de souligner que ce chapitre 6 propose et développe une approche basée sur l'apprentissage automatique qui ne nécessite qu'un ensemble de signaux de courant MPP au fil du temps. L'approche proposée dans cette recherche utilise uniquement le signal de courant des panels, un nombre réduit d'individus, ainsi qu'un nombre réduit de fonctions qui réduisent fortement les coûts de collecte de données, de stockage des données et de temps de calcul. De plus, le processus de diagnosis proposé ici s'est avéré être simple et efficace sur le plan informatique.

Enfin, un autre aspect intéressant est que cette approche est capable de détecter ce type de défaut même dans des conditions de faible irradiation où il est plus difficile de diagnostiquer les défauts. Ces résultats démontrent son fort potentiel pour classer ou discriminer les panneaux présentant des défauts dont la réduction de puissance est faible, mais qui peuvent être à l'origine d'autres défauts sévères, même dans des conditions de faible irradiation. L'analyse par fenêtres temporaires est un autre aspect intéressant de cette approche, puisqu'elle considère que la détection d'un défaut dans un intervalle de temps peut devenir ultérieurement un défaut grave ou tout simplement disparaître, entraînant une légère perte de performance. Par conséquent, cette méthode proposée fournit un outil de surveillance de l'évolution des défauts qui contribue directement à la maintenance préventive et corrective des grandes installations photovoltaïques. Cette approche a permis de détecter avec une grande précision des défauts de type Snail Trail, qui à ce jour ne peuvent être détectés qu'en visitant régulièrement la centrale photovoltaïque, ce qui est extrêmement coûteux.

Chapitre 8

Le diagnosis vise à prédire l'état de santé des composants et des systèmes. Dans les systèmes photovoltaïques (PV), il est vital d'assurer la production d'énergie et de prolonger la durée de vie des centrales photovoltaïques. Plusieurs algorithmes de prédiction et de classification ont été proposés dans la littérature à cette fin. La précision de ces algorithmes dépend directement de la qualité des données avec lesquelles ils sont ajustés ou entraînés, c'est-à-dire des caractéristiques. Dans ce chapitre 7, une approche innovante est proposée pour la prédiction de l'état de santé des systèmes photovoltaïques, qui comprend une étape de sélection des caractéristiques. Cette approche discrimine d'abord les panneaux PV gravement touchés en utilisant les caractéristiques électriques de base. Dans un second temps, il discrimine les autres panels défectueux en utilisant des caractéristiques temps-fréquence plus élaborées et en sélectionnant les caractéristiques les plus pertinentes par corrélation et analyse de variance. Enfin, l'approche prédit l'état de santé des panneaux photovoltaïques à l'aide d'une méthode de régression non linéaire appelée moindres carrés partiels. Ceci est ensuite combiné avec une analyse discriminante linéaire et comparée. L'approche est validée avec des données de courant réel d'une centrale photovoltaïque composée de 12 panneaux photovoltaïques d'une puissance comprise entre 205 et 240 Wp dans trois états de santé (verre cassé, sain, traces d'escargot). Les résultats obtenus montrent que l'approche proposée prédit efficacement les trois états de santé. Détermine le niveau de dégradation des panneaux, indiquant les priorités pour les actions de maintenance correctives et prédictives. De plus, il est rentable car il n'utilise que des mesures électriques déjà disponibles dans les systèmes d'acquisition de données photovoltaïques standard. Surtout, l'approche est générique et peut être facilement extrapolée à d'autres problèmes de diagnosis dans d'autres domaines.

Pour résumer, l'approche utilise, dans un premier temps, un simple clustering

hiérarchique basé sur le Dynamic Time Warping, pour regrouper les panneaux PV en deux groupes A et B, où le groupe A contient les panneaux PV sévèrement affectés et le groupe B contient les plus sévèrement panneaux PV concernés, le reste. À ce stade précoce, la méthode discrimine clairement les types sains et trace d'escargot des verres cassées, ciblant les actions prioritaires de maintenance prédictive et réduisant par conséquent les coûts globaux. Dans un deuxième temps, l'utilisation d'un ensemble de caractéristiques temporelles et fréquentielles détaillées permet une approche plus précise pour détecter des défauts infimes et montre sa capacité à discriminer les panneaux faiblement affectés (trace d'escargot) des panneaux sains.

La seconde étape a été validée en identifiant avantageusement les panneaux photovoltaïques présentant de gros défauts de traces d'escargot malgré la difficulté à les discriminer des panneaux sains. Cela représente une nette contribution par rapport aux travaux antérieurs tels que [Garoudja 2017a] qui ne parvient pas à détecter les défauts dont le comportement est très similaire à celui des panneaux sains. Il est également important de noter que notre méthode a le net avantage de nécessiter un suivi très simple. En fait, seul le courant MPP est nécessaire. Comme dans le cas du chapitre 6, ces algorithmes contribuent à la détection d'un type de défaut aujourd'hui généralement détectable en visitant régulièrement la centrale photovoltaïque. De plus, le processus de diagnosis pour être efficace doit être informatiquement simple et efficace.

Un avantage supplémentaire est que l'approche proposée dans cet article ne nécessite qu'un petit nombre d'individus de chaque classe, réduisant ainsi le coût d'acquisition et de stockage des données. Un autre point intéressant est que, comme dans l'algorithme proposé au chapitre 6, la méthode proposée présente les meilleures performances dans les cas de faible irradiation, comme en début et en fin de journée, où il est plus difficile de diagnostiquer ce type de fins défauts.

Cette méthode et celle proposée au chapitre 6 fournissent des informations sur des moments spécifiques de la journée qui doivent être surveillés. Ainsi, ce diagnosis par fenêtres temporaires permet d'analyser l'impact et l'évolution des défauts dans le temps. Notez que différents intervalles de temps pourraient être utilisés pour augmenter la résolution dans le diagnosis des défauts tels que les défauts d'arc [Wang 2013], l'ombrage partiel [Kumar 2018], les défauts LL [Dadhich 2019] qui se produisent avec de faibles niveaux de irradiation.

En ce qui concerne les aspects temporels, il convient également de noter que la décomposition du signal multi-résolution est extrêmement efficace pour détecter le moment exact où un signal change, ainsi que le type et l'étendue du changement [Misiti 2013]. Cela offre un avantage par rapport à la transformée de Fourier car si le défaut se manifeste plus rapidement que la fenêtre d'échantillonnage de l'analyse de Fourier, comme c'est le cas avec les défauts d'arc, il est très probable qu'il ne soit pas détecté.

Les différents apports mis en évidence ci-dessus font de la méthode proposée une méthode efficace de surveillance des systèmes photovoltaïques et est susceptible de réduire significativement les coûts de maintenance. Pour toutes les raisons

énoncées ci-dessus, il est possible d'affirmer que l'approche développée est très efficace en termes de calcul informatique, de qualité de l'information et de prédiction de l'état de santé, ainsi qu'économique puisqu'elle ne nécessite pas l'installation de capteurs supplémentaires. Il montre également des réalisations de bon augure dans l'extraction de caractéristiques et la réduction de données expérimentales grâce au diagnosis de défauts dans les systèmes photovoltaïques.

Fait intéressant, la méthode proposée est basée sur des algorithmes génériques qui pourraient être appliqués aux défauts de générateurs photovoltaïques qui ne sont pas considérées dans ce chapitre 7, ainsi qu'à d'autres applications dans le secteur de l'énergie. Ceci est considéré dans nos travaux futurs. Enfin, le chapitre 8 propose et implémente un algorithme encore plus complexe embarqué dans le système de monitoring proposé dans cette thèse. Ce système a été testé dans des conditions réelles de fonctionnement.

Chapitre 9

Ce chapitre présente les derniers développements en termes d'intelligence artificielle et de surveillance des systèmes photovoltaïques réalisés dans cette thèse. Ce chapitre propose une approche innovante de machine learning intégrée dans la plateforme de monitoring proposée et construite dans le chapitre 4 de cette thèse. Cette nouvelle approche conçue au niveau String PV a été testée à différents niveaux de configuration photovoltaïque, d'un panneau photovoltaïque à plusieurs centrales photovoltaïques à grande échelle, montrant de hautes performances en termes de précision de détection, d'adaptation à différentes conditions de formation et d'installation physique. Ce système est capable de détecter, d'identifier et de localiser le défaut, ainsi que d'identifier la priorité de maintenance de la chaîne défaillante.

En termes d'efficacité de calcul, l'approche de diagnosis proposée parvient à condenser les informations des caractéristiques initialement extraites en un plus petit nombre de caractéristiques dans un autre espace, en conservant les informations essentielles et en éliminant les informations redondantes ou non pertinentes. Amélioration significative des résultats de diagnosis des défauts. Les grandes lignes de l'approche sont présentées dans la Figure 4.

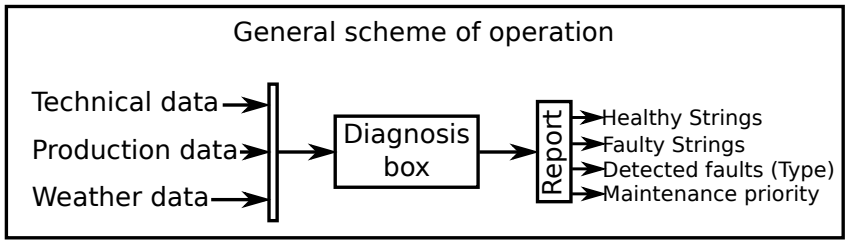

Figure 4: Schéma général de la plateforme de diagnosis proposée.

Le système de diagnosis, présenté à la Figure 4, a besoin de trois types d'informations pour effectuer le diagnosis. Premièrement, il collecte les informations techniques de la centrale PV, ces informations proviennent des rapports de

terrain de la centrale PV et de sa fiche technique. Parmi les aspects compilés figurent l'âge de l'installation, les modifications apportées aux panneaux, la technologie, la topologie, le comportement des panneaux dans des conditions standard, etc.

Deuxièmement, la plate-forme utilise la "Diagnosis Box" pour collecter le comportement électrique de la centrale photovoltaïque. Le comportement de l'installation est représenté uniquement par la capture du courant et de la tension en fonction du temps. Enfin, le comportement météorologique est enregistré en capturant la température ambiante, la vitesse du vent et l'irradiation en utilisant la station météo. À la suite de l'analyse interne de la box, un rapport est obtenu qui contient 4 résultats : *i)* Healthy Strings; *ii)* Faulty Strings; *iii)* Detected Faults (Type); and *iv)* Maintenance priority.

A la connaissance des chercheurs ayant participé à cette recherche, c'est la première fois qu'une plateforme de diagnosis aussi complète que celle-ci est proposée. Plusieurs tests ont été effectués sur différents systèmes PV, montrant des performances élevées dans chacun d'eux. Le diagnosis de la centrale est réalisé en fin de journée après avoir collecté toutes les données électriques et météorologiques de la journée. La plateforme de diagnosis se comporte avec deux processus principaux.

Pour réaliser le processus de diagnosis complet, le boîtier de diagnosis dispose de deux processus qui peuvent être exécutés en fonction des besoins. Le premier processus se produit en ligne et le second hors ligne. Le processus en ligne est responsable du diagnosis des installations photovoltaïques dont l'état n'est pas connu auparavant. Alors que le processus hors ligne est chargé de fournir la caractéristique évolutive et auto-adaptative du modèle d'apprentissage automatique. C'est-à-dire qu'il est chargé de recycler le modèle à l'aide de la base de données augmentée avec les nouveaux échantillons de chaînes ou de panneaux dont l'état (ou les étiquettes de défaut) est connu.

Comme on peut le voir sur la même Figure 4, cette approche est capable de déterminer la priorité de maintenance. Pour déterminer la priorité de maintenance, un calcul résiduel est effectué entre les signaux de la chaîne PV et le centre du cluster de chaînes saines. Le boîtier de diagnosis est également autonome en termes de consommation d'énergie, puisqu'il dispose de l'ensemble du système de génération PV qui garantit le fonctionnement de tous les composants. Ce système n'a pas non plus besoin d'Internet, il peut donc être installé dans des installations photovoltaïques telles que de grandes centrales photovoltaïques où il n'y a pas de réseaux sans fil. De plus, l'approche est suffisamment polyvalente pour pouvoir être connectée à différentes topologies de systèmes photovoltaïques, des installations résidentielles aux installations de grande puissance.

Un autre aspect intéressant de cette approche est qu'elle propose un ensemble d'équations de normalisation de variables. Ces variables nous permettent de mettre toutes les variables dans une échelle correcte pour améliorer la précision des résultats dans la détection des défauts, et d'augmenter les performances de l'algorithme de classification en cas de faible irradiation.

Le système est également facilement reconfigurable en fonction des caractéris-

tiques de l'installation photovoltaïque et du besoin de sortie de données. Autrement dit, si l'analyse doit être effectuée hors ligne sur un serveur, ce système peut être facilement configuré pour que les données soient enregistrées dans un fichier externe sous différents formats (csv, sql, etc.). Sinon, par défaut, le système effectue le stockage sur un serveur local, ce qui réduit la taille du stockage des données significativement.

Perspectives

Cette thèse laisse la porte ouverte à de multiples travaux futurs car elle aborde deux grands domaines : le matériel et le logiciel

Le matériel

Acquisition de données orientée vers le diagnosis (Solar Vitality)

Dans les travaux futurs, les dimensions de Solar Vitality devraient être réduites. Pour cela, une analyse de marché des différents composants qui maintiennent la qualité du signal mais réduit la consommation d'énergie et les dimensions doit être réalisée. Cela peut également améliorer le point faible de Solar Vitality, qui est le coût associé aux capteurs performants. De plus, les microcontrôleurs utilisés dans la dernière version de Solar Vitality pourraient être remplacés par ceux qui consomment moins d'énergie. Un écran tactile pourrait également être ajouté à la plate-forme, afin de réduire sa dépendance à un ordinateur pour configurer le système et le démarrer. L'inclusion de l'écran permet également d'ajouter une nouvelle fonctionnalité qui est la supervision en temps réel.

De plus, cela faciliterait le paramétrage du système avant le démarrage. Les deux cartes d'acquisition de données et les cartes de traitement de données pourraient être remplacées par une seule carte avec le système embarqué afin de réduire la taille du prototype et la consommation d'énergie. Si possible, une carte électronique personnalisée permettant d'intégrer les diviseurs de tension doit être réalisée. Cela augmenterait la robustesse de la plateforme étant donné qu'il s'agit d'un prototype portable. Une autre étude sur les panneaux défaillants à différentes vitesses de capture de données devrait être effectuée pour déterminer si un ADC à haute vitesse est vraiment nécessaire ou pourrait être remplacé par un ADC avec moins de fonctionnalités. Cela réduirait le prix, ainsi que la quantité de données à stocker.

Station météo

D'autres protocoles de communication sans fil devraient être explorés pour éviter le câblage entre la station météo et la nouvelle plateforme de surveillance du système PV. Cela facilitera également le couplage de la station avec le système PV. Il serait intéressant d'examiner la possibilité d'ajouter des capteurs d'humidité liés à la dégradation accélérée des modules PV et à l'apparition d'autres défauts comme démontré au chapitre 2. Des connecteurs spéciaux doivent être mis sur les bornes des

câbles des capteurs. la station météo qui va directement à une boîte de connexion électrique qui accélère le processus de connexion et évite les mauvaises connexions des capteurs.

Le logiciel

Stockage des données

Le système pourrait être complété en passant à des tables relationnelles qui permettent d'avoir des clés primaires et secondaires pour éviter la confusion entre les données. Un système de sauvegarde automatique pourrait être mis en place pour éviter la perte de données. Les données ne pourraient plus être stockées sur un serveur phpmyadmin local et aller à la place sur des plateformes en ligne telles que celles fournies par Amazon ou d'autres.

Pré-traitement des données

Le système d'estimateur de kalman qui élimine le bruit dans les signaux capturés par l'Arduino doit être vérifié avec différents tests sur le terrain. L'un des tests les plus importants est le retrait et l'apparition de la source de mesure pour adapter adéquatement les coefficients aux temps de réponse des capteurs et éviter de masquer les défauts.

Système de diagnosis

Les approches présentées dans les chapitres 7 et 8 pourraient être améliorées en utilisant la normalisation des données présentée dans le chapitre 9. De plus, ils pourraient être modifiés pour effectuer une classification multivariée afin d'inclure des aspects tels que l'humidité, la vitesse du vent, l'irradiation et la température ambiante. Ceci, compte tenu des excellents résultats obtenus dans les chapitres 7 et 8. De plus, tous les paramètres du Chapitre 9 du modèle pourraient être améliorés en incluant un algorithme d'optimisation (heuristique ou Méta-heuristique).

La base de données de défauts pour améliorer la formation doit continuer à être construite sur le terrain, idéalement avec des centrales photovoltaïques qui ont des configurations, des technologies et des types de panneaux différents. Cela augmenterait de façon exponentielle la robustesse du système. De plus, cela montrerait s'il est nécessaire de modifier davantage la normalisation des données proposée dans la thèse.

Une autre idée serait d'explorer la possibilité de coupler le système avec un système de véhicule sans pilote qui capture des images afin de localiser efficacement les défauts de la centrale qui ont une signature thermique au niveau du panneau. Des tests doivent être effectués pour expédier les algorithmes proposés dans les Chapitres 6-8 dans des appareils tels que des onduleurs ou des optimiseurs afin de tester leur utilisabilité et leur précision.

De plus, un nouveau système couplé dédié au suivi de points MPPT basé sur des techniques d'apprentissage automatique pourrait être facilement couplé aux systèmes déjà présentés dans cette thèse.

Système de rapport de priorité de maintenance

Pour le système de rapport de maintenance, un système codé en Visual Basic pourrait générer automatiquement des rapports pour le client, en envoyant une série de recommandations de changement en fonction des défauts détectés.

Introduction

The constant increase in the global demand for electricity, of the price of oil and gas products and of environmental pollution have driven in consequence an important interest in the systematic use of renewable energies [Onar 2008]. Among them, photovoltaic energy is classified as a sustainable electrical energy resource with a constantly decreasing cost [Ray 2018, Romero-Cadaval 2015]. In addition, photovoltaic energy can be implemented in all continents and in various climates, is considered as a clean resource [Shahsavari 2018], can be used in both small installations and large-scale power plants due to its scalability. Among the different solutions in the market [Navid 2021a], it is cataloged as the best way to generate energy from the environment [Madeti 2017a]. It can be also considered as a vital tool to promote social transformation and sustainable economic development [Hariharan 2016b].

Currently, the use and the production of photovoltaic panels has increased substantially [Jean 2015]. This increase in this type of energy has highlighted four major challenges that these systems face: i) the potential occurrence of faults and the response time required to detect and solve them [Araneo 2009, Chen 2018b]; ii) degradation faults [Ndiaye 2013] which can reduce the total energy production by up to 17.5% per year [Dhere 2012] and deteriorate the system at rates of 0.8 % per year or may even cause discontinuity or total system failure [Wohlgemuth 2011, Chamberlin 2011]; iii) reduction in the rate of recovery of initial investments and increase in maintenance costs of photovoltaic systems [**?**] and iv) the need to improve supervision systems of photovoltaic plants due to the appearance of recurring undetectable faults [Parida 2011]. In an effort to counteract production losses, improvements in photovoltaic cell efficiency and maximum power extraction have been developed, which increase the efficiency, stability, reliability and robustness of photovoltaic systems [Seyedmahmoudian 2016]. However, when a fault occurs in the system or is imminent, it is also needed to detected and classified as quickly as possible faults or failures [Huang 2018] in order to trigger appropriate preventive or corrective maintenance of the photovoltaic system in a timely manner [Upadhyay 2014].

Faults in photovoltaic systems can be caused by aspects such as the useful life of components, increases in temperature during operation, external factors (environmental and non-environmental) or interactions between materials [Fadhel 2018]. However, detection is difficult due to the high dependence of the PV system on weather conditions [Yi 2017c], the presence of maximum power point tracking (MPPT) devices [Zhao 2013a], scenarios in which the electrical behavior of panels with degradations is very similar to others panels with normal behaviors [Hariharan 2016a, Sepúlveda Oviedo 2022] or scenarios where the occurrence of the fault is so fast that it does not seem to have occurred [Zhu 2018]. For these reasons, faults can go undetected for hours and not only degrade the state of the

photovoltaic panel, but can also cause it to catch fire and pose a danger to human safety [Strobl 2010, Wang 2014a, Rabla 2013]. Thus, early diagnosis of faults in photovoltaic systems can sometimes be a real challenge [Haque 2019] and also an attractive and developing area of research [Ahmad 2018].

Selection of the correct fault detection technique depends on factors such as the type of fault (line-to-line, short circuit, open circuit, hot spot, partial shade, etc.), severity, and the potential for multiple faults to occur. Multiple review papers have been published as a guide for the selection of the best fault detection approach in PV systems [Madeti 2017b, Alam 2015a]. Some of these reviews range from conventional fault detection methods to more recent artificial intelligence approaches based on Machine Learning algorithms. Although these reviews provide an overview of the research area, they require a long time to achieve an in-depth review, and they hardly eliminate the subjective factors in the literature selection, which leads to some important literature being ignored [Shen 2021].

Thesis Outline

The rest of the thesis is organized into nine chapters as follows:

Chapter 1: This chapter addresses the academic and industrial interest under which this thesis is developed, the formal approach to the problem, the purpose and objectives and the cases that are studied in the thesis. As well as the academic products obtained from this thesis.

Chapter 2: This chapter provides a description of the components of a photovoltaic installation, the faults and methods (that do not use artificial intelligence) for the detection of faults in PV systems. The aim of this chapter is to remember physical aspects and behavior of a PV plant from a PV cell to a complete system of a high power PV plant. It is built to help the reader to more understand the real challenge in this field to improve its reliability and why classical techniques of monitoring are not sufficient.

Chapter 3: This chapter presents an extensive study of the state of the art on artificial intelligence techniques used for fault detection in PV systems. In this chapter, two computational methodologies that combine machine learning, bibliometric and expert analysis are proposed and used to extract the relevant information that determines the current research areas and the challenges in these areas. In addition, it allows positioning the research presented in this thesis by performing an in-depth analysis of the state of the art that reduces the existing subjectivity in conventional reviews and positions the reader at the forefront in understanding aspects of effective fault detection in photovoltaic systems.

Chapter 4: In order to guarantee the correct operation and performance of a photovoltaic system, it is essential to implement a robust, effective, low-cost and sustainable monitoring unit that is capable of monitoring, recording data and analyzing the number of parameters that are measured in a photovoltaic plant of small, medium and large scale. As a tool to help understand the necessary

conditions to correctly monitor a PV system, this chapter deals with two main aspects. First, a comprehensive review of various photovoltaic monitoring systems is presented. That review includes a detailed description of all the main photovoltaic monitoring systems, based on the sensors used and their principles of operation, controllers used in data acquisition systems, data transmission methods and data storage. Finally, it presents a commercial and industrial data acquisition platform.

This chapter presents the new Diagnosis-oriented Data Acquisition named Solar Vitality. This platform is built and tested in this thesis (complying with the IEC 61724 standard). The platform is capable of capturing the current and voltage variables of multiple strings, as well as capturing meteorological variables such as temperature, irradiation, and wind speed at a speed of 15 milliseconds. In addition, the weather station where the weather sensors are coupled is completely versatile and adaptable to different topologies of PV plants.

Chapter 6: This chapter provides an explanation about some feature extraction and selection approaches that are used for fault detection in this thesis. Furthermore, some of these approaches are tested with well-known supervised and unsupervised learning algorithms to determine the limitations of conventional algorithms in our application case.

Chapter 7: This chapter presents the first AI approach proposed in this thesis. This approach combines the features of multiple machine learning and signature extraction and selection algorithms for snail trail fault detection. This proposed approach called Ensemble Learning (EL) combines several learning models, namely Support Vector Machine (SVM), K-Nearest Neighbor (kNN), and Decision Trees (DT), instead of using a single learning model.

Chapter 8: In this chapter a detailed analysis focused on reducing computational calculation time while maintaining accuracy in fault detection at the PV panel level is presented. The innovative hybrid approach (unsupervised and supervised learning) of machine learning proposed in this chapter pays special attention to the quality of the data with which it is fitted or trained. The approach presented in this chapter is capable not only of identifying Snail trails and Broken glass faults with high precision and reduced computational time, but in addition to evaluating the evolution of the fault over time, it determines the level of degradation of the panels. This is a very important aspect since it allows indicating priorities for corrective and predictive maintenance actions. Furthermore, this proposed approach is cost-effective as it uses only electrical measurements that are already available in standard photovoltaic data acquisition systems. Above all, the approach is generic and can be easily extrapolated to other diagnosis problems in other domains. Finally, this approach can be extrapolated to PV string configurations or other configurations.

This chapter presents the latest development in terms of artificial intelligence and monitoring of PV systems made in this thesis. This chapter proposes an innovative machine learning approach embedded within the monitoring platform described in Chapter 5 of this thesis. This new approach applied at the PV string level has been tested at different levels of PV configuration, from a PV panel to

multiple large-scale PV plants, demonstrating high performance in terms of detection accuracy, adaptation to different training conditions and physical installation. This system is capable of detecting, identifying and locating the fault, as well as identifying the maintenance priority of the faulty string.

Finally, at the end of the manuscript, we draw a conclusion about this work and present some prospects for future work.

Chapter 1

Background and Project Motivation

Contents

1.1 Background 33
1.1.1 Project Motivation 34
1.1.2 Problem statement 36
1.1.3 Aim and objectives 37
1.1.4 Studied fault cases using machine learning 38
1.1.5 Academic products of the thesis 39

In this chapter, all issues related to the productivity of a photovoltaic installation and how power losses require a fault detection and diagnosis study are addressed. Likewise, the academic and industrial interest under which this thesis is developed, the formal approach to the problem, the purpose and the objectives are described, and finally, the academic products obtained from this thesis.

1.1 Background

Photovoltaic energy has taken a truly important position among renewable energies, reaching a cumulative global installed capacity of approximately 75 GW in 2016 [energy agency 2016] and according to the report of NREL [Feldman 2022], it is estimated that 171 GW of PV will be installed worldwide in 2021, and they project that in the years 2022 and 2023, 209 GW and 231 GW will be installed, respectively. Even reports such as GlobalData estimate that the photovoltaic power installed worldwide will exceed 1,500 GW in 2030 [Data 2019]. In existing and future photovoltaic installations, independently of their nature and sizes, the diagnosis of the health status of the components and systems become vital to guarantee energy production, extend the useful life and prevent unexpected events in photovoltaic solar systems.

In large PV systems representing more than 100 kWp and one-hectare area, it is understandable that it becomes rapidly more difficult to detect or identify a physical fault position and/or its origin. This target might prove difficult to meet in some cases where the fault goes unnoticed until generating a significant negative impact on the PV system production. Globally, some of the aspects that

make fault detection difficult are: i) occurrence under low irradiation scenarios [Yi 2017c]; ii) faults that occur in less than one second [Wang 2013]; iii) presence of the MPPT device that optimizes the output power of a photovoltaic array [Zhao 2013a]; or iv) faults with electrical behavior similar to that of a healthy panel [Hariharan 2016a, Sepúlveda Oviedo 2022]. If these faults are not detected also be qualified as incipient faults, not only can the state of the plant deteriorate, but it also represents a danger to human safety [Strobl 2010, Wang 2014a, Rabla 2013], even generating drastic fault as large fires [Brooks 2011, Ministry of Housing 2017].

As a solution for evaluating photovoltaic system performance and calculating energy loss over long periods of time, many PV installations today have massive PV data (both instantaneous and historical) from sources such as weather stations, photovoltaic inverters and the public network [Zhao 2015a]. However, these supervision platforms are not oriented to fine diagnosis of faults and therefore they do not pay attention to the quality of the data or they have data shipments with low sampling frequencies. As recommended in [Dhimish 2018a], to increase diagnosis accuracy it is necessary to increase the sampling time even at speeds of a few microseconds. However, the types of systems that capture data at these speeds present a challenge linked to the capabilities of the microprocessors used in the devices. In addition, using this type of instrumentation platform in different types of PV plants is really a challenge since energy autonomy must be guaranteed due to the absence of electrical outlets or power supply systems, local humidity and other environmental conditions must be considered to ensure a robust platform operation, among others.

Despite the complexity of these systems, the increasing use of photovoltaic energy and the reduction in the cost of solar panels are increasingly attracting the interest of researchers, both in academia and in the industrial world. The objective of this thesis is the development of fault diagnosis methods embedded in high power data acquisition systems compatible with photovoltaic installations, respecting industrial limitations and taking into account the cost/benefit trade-off in productivity or operating time. The early detection of faults allows to define effectively the actions to be carried out in terms of the use of the photovoltaic plant but also in terms of corrective or conditional maintenance, taking into account the current state of health of the photovoltaic plant or even taking into account forecasts of the evolution of degradation.

1.1.1 Project Motivation

High-power photovoltaic (PV) power plants are being deployed all over the world. Its useful life and use must exceed 25 years to guarantee the return on investment in infrastructure. For this, it is preferable to carry out maintenance either periodically, or when a serious fault occurs and a definitive loss of productivity is detected. In fact, this situation can lead to the shutdown of all or part of the plant. Each production loss, even linked to a reduced number of panels and even temporary, leads to a significant financial loss in plants in particular ones of more than 250

kWp, which justifies the need to diagnose the state of the plant and anticipate maintenance interventions at the time will be exact useful.

The productivity of PV plants is strongly affected by aspects such as availability and performance. Availability refers to the relationship between the duration of continuity in energy production, even without optimal performance, and the total period observed [Díaz 2007]. On the other hand, performance refers to the overall efficiency of the energy conversion chain. Commonly this performance is measured using the "performance index" [CEC 1997, IEC 1998]. According to studies carried out by IEA PVPS, the annual availability rate of a well supervised photovoltaic installation can reach 97 % [Janh 2000]. As mentioned in [Bun 2011b], the productivity of a PV system can be improved by reducing the downtime rate and operating the system at maximum performance.

To reduce downtime or (non-optimal) production time it is necessary to reduce the number of component faults and the time for both preventive and corrective maintenance. It is there, where a diagnosis system embedded in a robust data acquisition platform is vital to identify the fault as quickly as possible. It is true that a classic data acquisition or supervision of the plant's production data allows to identify (imprecisely) the presence of anomalies in energy production. However, classic data acquisition does not allow for an early detection of faults, so the photovoltaic plant continues to operate in a sub-optimal state. To date, several companies offer products to help manage photovoltaic power plants, such as S4E with the EnergySoft product [S4E 2022] or Circutor with its Scada product dedicated to photovoltaic power plants [Circutor 2022]. Some inverter manufacturers offer products and services to help estimate [SMA 2022] plant performance. However, the products currently on the market are often not adapt at analyzing the reasons for production losses and are often limited to a visualization of the data, without further analysis. More advanced work is done by the company Feedgy with its Feedgy Analytics tool to monitor and diagnose PV systems online, based on historical data [Feedgy 2022].

However, despite these efforts, classic data acquisition systems do not have the portability to be coupled to different PV plant topologies. That is signified they are not capable of being coupled to both residential installations and plants with other sizes and topologies such as those large plants on the ground. The latter represent one of the greatest challenges since they do not have electrical outlets or energy sources that can feed the data acquisition and fault detection system, in addition to the strong weather conditions that affect these plants.

It is necessary to clarify that installing an advanced data acquisition system does not guarantee early detection of faults in the PV plant. This data acquisition system must have embedded a diagnosis system capable of comparing the behavior of different strings of the same plant, taking into account several constraints such as weather conditions, technology, topology, degradation, etc. In other words, it must be a diagnosis system supported by expert knowledge of the PV plant and measurements of its electrical and meteorological behavior. Such a robust system would guarantee that the photovoltaic system does not work for weeks or months in

a sub-optimal state. This fault diagnosis system must integrate new cutting-edge techniques that do not require a large amount of data to detect fault phenomena. This condition is vital, because when the system is installed in the plant, its historical data is not known, but it is also desired to carry out an early diagnosis of faults during the first hours of operation of the diagnosis device. For this, it is mandatory to increase and standardize the existing knowledge about the causes of faults and their probabilities. Achieving this early diagnosis of faults allows a reduction in human interventions and their programming only based on signs of fault.

All these reasons are the basis for this research and reveal the need to build a more sophisticated diagnosis system to detect and diagnose faults in order to improve the productivity of the photovoltaic installation. The formal statement of the research problem is presented below.

1.1.2 Problem statement

This thesis addresses the problem of loss of performance of the entire photovoltaic system and reduction of the output power generated due to the appearance of faults. This thesis will describe some symptoms or signatures that allow identifying the main faults in photovoltaic systems. In addition, it will make a contribution to the improvement of classical data acquisition, proposing and building a new versatile, portable and autonomous data acquisition and fault detection platform at the energy supply level associated with a proposed and built weather station that monitors wind speed, ambient temperature and irradiation. This platform has two embedded systems working in collaboration. The first embedded system focused on the collection and previous-treatment of diagnosis-oriented data. The second system addresses the problem of fault diagnosis in PV systems with few levels of data. This second embedded system performs fault detection for small time intervals and is made up of novel machine learning algorithms that could be used to detect faults in photovoltaic plants of different configurations and technologies.

The faults detected in this thesis include snail trail, shading even partial shading and broken glass faults. The snail trail type fault is included to test the fine diagnosis level of the embedded system. This type of fault is very difficult to detect due to its electrical signature highly similar to that of a healthy panel. Snail trails (also known as snail tracks or worm marks) in outdoor conditions appear as brownish discolored contact fingers, especially around cell edges and areas of microcracks [Kim 2016, Köntges 2014b]. This phenomenon is attributed to the entry of moisture and oxygen through microcracks, in addition this fault can worsen the microcracks or trigger other more severe faults. The broken glass and shading type faults are included with the interest of testing some common faults whose impact levels are higher than those of the snail trail. This variety of production losses allows us to demonstrate that the proposed system is capable of detecting faults in the entire impact range. Figure 6.1 shows the three types of faults mentioned.

Multiple approaches have performed fault detection in PV systems based on I(V) characteristic curve analysis. This type of detection is limited for its implementation

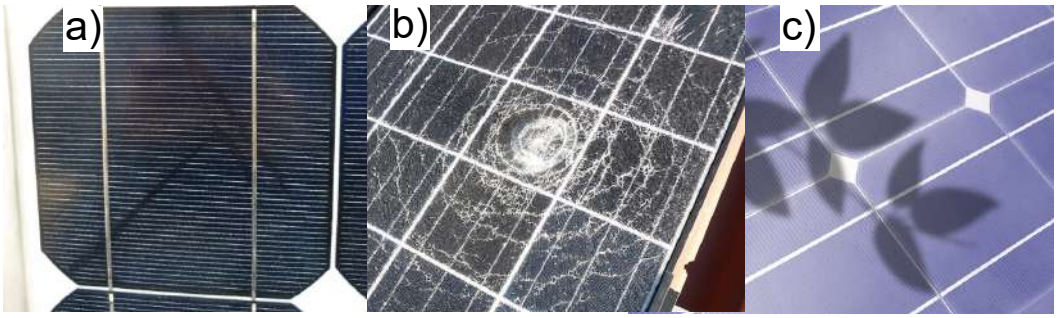

Figure 1.1: Faults analyzed in PV systems. a) Snail Trails; b) Broken Glass; c) Shading

on a large scale since to obtain the characteristic curve it is necessary to cut the production of the system, making this approach infeasible from the industrial point of view. So, this thesis also deals with the detection of faults on the electrical signals of the plant in production. The proposed diagnosis system does not need to cut or suspend the production of the plant to carry out the diagnosis. Furthermore, the system developed in this thesis is designed respecting the industrial objectives and therefore taking into account aspects such as the cost, durability and ease of use, in view of industrializing and commercializing the solution.

1.1.3 Aim and objectives

The main objective of this research work is to design and implement a physical platform for data acquisition and fault detection. The data acquisition of the PV system must have a data capture frequency of less than one second and the embedded system must be capable of detecting multiple faults that occur in photovoltaic installations without large amounts of historical data. The detection device must be tested on PV plants at different scales (small, medium and large plants), geographic locations, weather conditions and technologies. To achieve effective diagnosis, it is necessary to have a broad knowledge of the types of faults, associated signatures and methods currently used. Therefore, in this thesis a large number of articles associated with this topic are analyzed.

There are ten main objectives of this research, which are:

1. To propose two new approaches to build a state of the art using statistical and machine learning techniques.

2. To build a fault dictionary, containing the description and main signatures associated with a wide set of faults.

3. To formulate, design and build a new and versatile photovoltaic data acquisition system. This platform must be portable, autonomous in terms of energy and capable of receiving the analog signal from electrical behavior sensors (voltage and current) and meteorological data.

4. To design and build a weather station that can be easily coupled to a PV plant and is compatible with the data acquisition platform. This weather station must monitor ambient temperature, wind speed, irradiation and send the analog signal to the data acquisition system.

5. To propose a set of extraction and selection techniques of features, on time series, aimed at detecting faults in PV systems.

6. To propose a machine learning approach for the detection of fine faults such as snail trail in PV modules.

7. To propose a hybrid univariate Machine Learning algorithm combining supervised and unsupervised learning focused on the detection of faults non evident as snail trail type faults and conventional faults such as broken glass.

8. To propose a set of equations for the normalization of electrical data and environmental variables that allows comparing the performance of PV plants with different ages of start-up, number of panels, technologies, etc.

9. To propose a PV mathematical model that is capable of predicting the PV production of a plant based on variables such as: ambient temperature, wind speed, irradiation, the characteristics in standard conditions of the PV panels of a PV plant, the date of installation of the PV plant, among other aspects. The behavior of the model must be compared with real data from a PV plant.

10. To propose a new adaptive machine learning approach that combines supervised and unsupervised learning, as well as model- and data-based learning. In addition, this approach should use multiple electrical and environmental variables to improve fault detection at the PV string level. Finally, the system must be embedded in the new photovoltaic data acquisition system, be able to update itself as new data from other PV plants is collected, and be able to perform fault detection in PV plants of different technologies and topologies.

1.1.4 Studied fault cases using machine learning

Based on the information analyzed on the state of the art of possible faults in photovoltaic fields, this research opts to study mainly snail trail and broken glass type faults. This choice is due to the fact that broken glass type faults are the cause of the greatest loss of production in PV systems. On the contrary, the snail trail type fault does not cause a significant reduction in production in PV systems, however, this fault is the cause of multiple severe faults that can generate the total loss of energy production or even fires. Bearing this in mind, this research starts from the hypothesis that if the proposed algorithms manage to detect the faults that are in the upper and lower limits of power loss, they are capable of detecting the entire range of faults that occur between them. These faults are studied only with real data from a PV system specifically designed for this study. The detailed

configuration of these faults and the presentation of the photovoltaic field to be studied are detailed in Chapters 4 and 5 of this thesis.

1.1.5 Academic products of the thesis

The list of academic products of the thesis is listed below.

1.1.5.1 International conferences

1. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Corinne Alonso and Marko Pavlov. *Hierarchical clustering and dynamic time warping for fault detection in photovoltaic systems.* In X Congreso internacional Ingeniería Mecánica, Mecatrónica y Automatización, Bogotá, Colombia, May 2021. (Accepted and Presented)

2. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov, Corinne Alonso. *DTW k-means clustering for fault detection in photovoltaic.* In XI Congreso internacional Ingeniería Mecánica, Mecatrónica y Automatización, Cartagena, Colombia, May 2023. (Submitted)

3. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov, Corinne Alonso. *Detection and classification of faults aimed at preventive maintenance of pv systems.* In XI Congreso internacional Ingeniería Mecánica, Mecatrónica y Automatización, Cartagena, Colombia, May 2023. (Submitted)

1.1.5.2 National conferences

1. Edgar Hernando Sepúlveda Oviedo. *Extraction de signatures et prédiction de l'état de santé des centrales photovoltaïques.* In Journée annuelle de l'école doctorale Geets, Toulouse, France, April 2022. (Accepted and Presented)

1.1.5.3 Workshops

1. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov, Corinne Alonso. *Fault detection and diagnosis for PV systems using machine Learning*, Poster, In 9th NextPV workshop, online edition, November 2020. (Accepted and Presented)

2. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov, Corinne Alonso. *Acquisition de données, et prédiction de l'état de santé de systèmes photovoltaïques.* Oral presentation. In Workshop DO, Mauvezin, October 2021. (Accepted and Presented)

3. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov, Corinne Alonso. *Advanced machine learning methods, for the detection of fine faults in PV systems, aimed to preventive maintenance.* Oral presentation. In 10th NextPV workshop, Bordeaux, France, January 2023.

1.1.5.4 Scientific journal articles

1. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Corinne Alonso and Marko Pavlov. *Feature extraction and health status prediction in PV systems.* Advanced Engineering Informatics, vol. 53, page 101696, 2022. (Published in journal Q_1)

2. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov and Corinne Alonso. *Artificial intelligence based fault diagnosis in photovoltaic systems Part I: A Bibliometric survey.* Renewable and Sustainable Energy Reviews, vol. 00, page 00, 2022. (Submitted in journal Q_1)

3. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov and Corinne Alonso. *An Ensemble Learning-Based Fault Detection and Diagnosis for PV modules.* Sustainability, vol. 00, page 00, 2022. (To appear in journal Q_1)

4. Edgar Hernando Sepúlveda Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov and Corinne Alonso. *Artificial intelligence based fault diagnosis in photovoltaic systems Part II: A topic modeling approach.* Renewable and Sustainable Energy Reviews, vol. 00, page 00, 2022. (Submitted in journal Q_1)

1.1.5.5 Patent

1. Patent Feedgy/LAAS-CNRS (in process)

1.1.5.6 Awards obtained

- Prize for the best oral presentation in the session on electrical engineering and energy management in *Journée annuelle de l'école doctorale Geets.*

Chapter 2

Fault Diagnosis in Photovoltaic Systems

Contents

2.1 Photovoltaic industry 42
2.2 PV system components 45
2.2.1 PV generator 45
2.2.2 Wiring and Junction Box 53
2.2.3 Inverter 53
2.2.4 Protection system 54
2.3 Formal fault dictionary 55
2.3.1 Main causes of faults 56
2.3.2 Multilevel fault classification 57
2.3.3 Frequency of occurrence of faults in the PV system 73
2.3.4 Impact of faults in terms of power loss and human safety . . 77
2.4 Conventional Fault Detection Methods 82
2.4.1 Visual methods 83
2.4.2 Image-based methods 83
2.4.3 Electrical detection methods 86
2.4.4 Protection Device Based Technique 89
2.4.5 ARC Fault Detector Techniques 89
2.5 Discussion and Conclusions 89

This section is divided into three main parts. First, to understand the object of our study, a brief remember on the constituent components of a conventional PV system is done. Second, a formal dictionary of faults is proposed that contains four types of identified fault sources: external causes, material interaction, component aging or caused by other faults, which is named cause-effect circle. This dictionary is built with a new multilevel classification of system faults based on the type of fault, the component where it occurs (cell, module, array, protection system or box junction), whether it is structural, electrical, caused by abnormal increases in temperature (hot spot), poor connections, or shading (by obstacles or dirt). This classification based on an intense bibliographic review, contains a description of each fault based on its meaning together with an illustrative support, the frequency

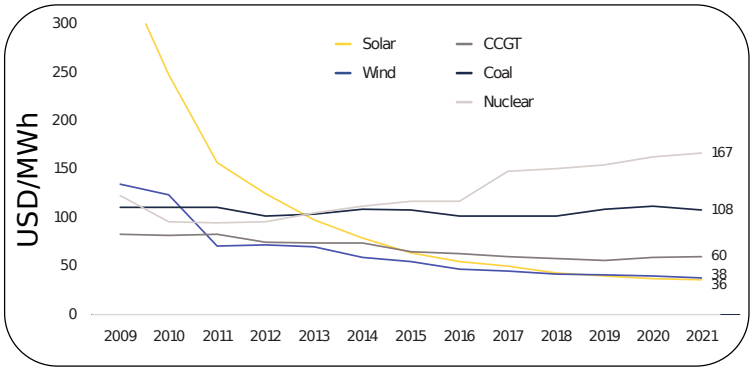

Figure 2.1: Solar electricity generation cost in comparison with other power sources 2009-2021 [SPE 2022].

of occurrence and impact in terms of human safety and loss of power are exposed. Finally, the third aspect discussed in this chapter deals with conventional fault detection methods in PV systems, dividing them into five broad categories: Visual Methods, Image-based Methods, Electrical Detection Methods, Protection Device Based Technique, and ARC Fault Detector (AFD) Techniques.

2.1 Photovoltaic industry

Although the principle of photovoltaic (PV) conversion is well-known since its discovery by the French physicist Edmond Becquerel in 1839 [Ameur 2021, Zhang 2021a] and the design of associated sensors is particularly well mastered in several technology ways to create efficient PV cells, the solar PV industry allowing a large diffusion in the world has really grown since the 1970s under the pressure from the fossil fuel crisis. According to the International Energy Agency (IEA) [IEA 2007a], another notable evolution of the photovoltaic industry induced an important growth in efficiency from 15 % to 20 % between 1991 and 2007. It can be noticed in this period that the increase of the photovoltaic industry and its performances is highly connected to the computer and semiconductor industries improving their technology processes [Vighetti 2010]. Components from microelectronics declared non-compliant are remelted to be reused and supply the raw material sector of the PV industry, considerably reducing the cost of manufacturing the PV cells.

This cost reduction is a key factor, which has presented one of the main advantages of solar energy over its competition such as Wind, Coal, CCGT (combined cycle gas turbine) and Nuclear. The cost of solar power has been lower than that of fossil fuel generation and nuclear power for several years, and is now even lower than wind as presented in Figure 2.1 taken from the study [SPE 2022].

In the same Solar Power Europe study [SPE 2022], it is exposed that the decrease in large-scale solar costs has progressed by 3% more compared to the previous year and is increasingly moving away from other conventional generation technolo-

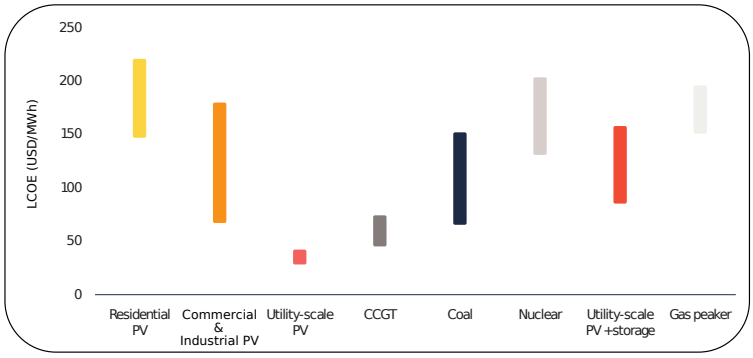

Figure 2.2: Solar electricity generation cost in comparison with conventional power sources 2021 [SPE 2022].

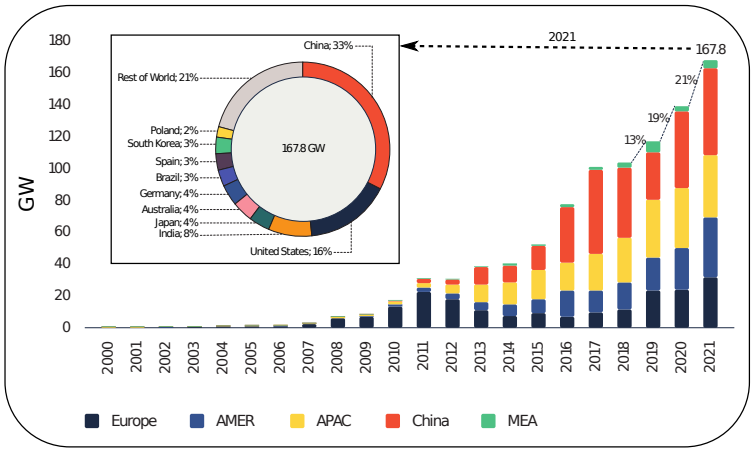

Figure 2.3: Annual solar PV installed capacity 2000-2021 and top 10 countries solar share 2021 [SPE 2022].

gies. Now as large-scale power generation is examined, photovoltaics are remarkably cheaper than any of the other technologies as can be seen in Figure 2.2.

In the same Figure 2.2, it can be seen that residential, commercial and industrial PV installations still have a way to go, but it is expected that in a short time their costs will also be lower than those of other technologies. Along the same way, in [SPE 2022] it is exposed that more and more countries are installing hybrid renewable energies, using various renewable sources plus battery storage to achieve flexible solutions to their energy needs. The installation of these hybrid energies has further promoted the installation of PV energies. If the annual photovoltaic installed capacity is analyzed from 2000 to 2021, it is possible to note that in 2021, 167.8 GW of solar capacity were connected to the grid worldwide [SPE 2022]. This represents a growth of 21 % with respect to the 139.2 GW added in 2020, as shown in Figure 2.3.

In the same Figure 2.3, China can be seen as the largest market in the world, followed by the United States. And in fact, if the cumulative installed photovoltaic solar capacity in the world is analyzed, it grew by 22% to 940.0 GW at the end of

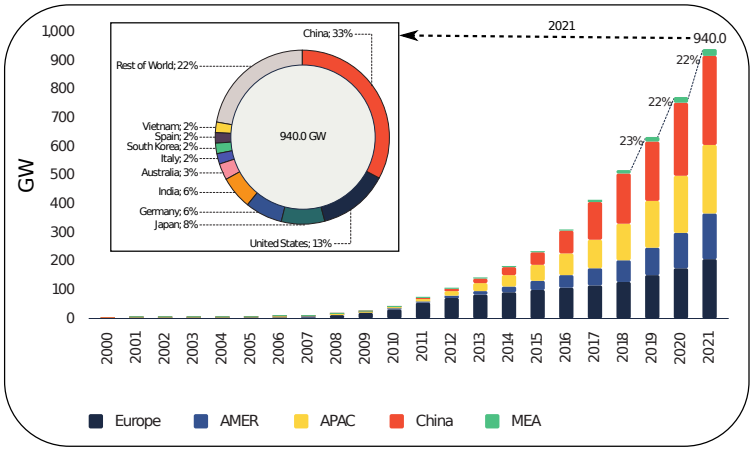

Figure 2.4: Total solar pv installed capacity 2000-2021 and Top 10 solar pv markets total installed shares 2021 [SPE 2022].

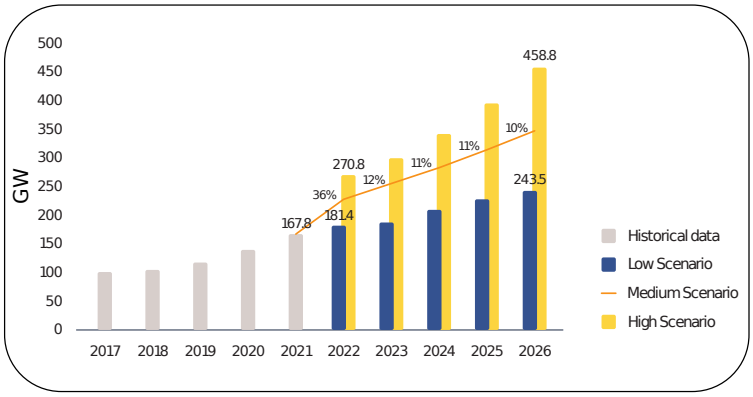

Figure 2.5: World annual solar pv market scenarios 2022 - 2026 [SPE 2022].

2021, compared to 772.2 GW in 2020, as can be seen in Figure 2.4. This means that total solar energy has increased more than 500 times since the turn of the millennium, when the grid-connected solar era began with the launch of Germany's feed-in tariff law [SPE 2022].

When comparing the values of accumulated installed photovoltaic solar capacity for the year 2021 and 2010, an increase from 41.3 GW to 940.0 GW (approximately 1 TW) can be observed, which represents an impressive increase of approximately 2176.0 %. In the same Figure 2.4, a comparison of individual countries can be observed, where it stands out that although there were movements in the first 6 positions (with respect to Figure 2.3). China again followed by the United States, Japan, Germany, India and Australia. Finally, it is interesting to know what are the prospects of the photovoltaic industry for the coming years. In [SPE 2022] a predictive analysis of the photovoltaic industry up to the year 2026 is performed. In this analysis, 3 market scenarios are proposed as presented in Figure 2.5.

As it can be seen in Figure 2.5, in the medium scenario, by 2022 new installed capacities are expected to reach 228.5 GW by the end of 2022, which represents a

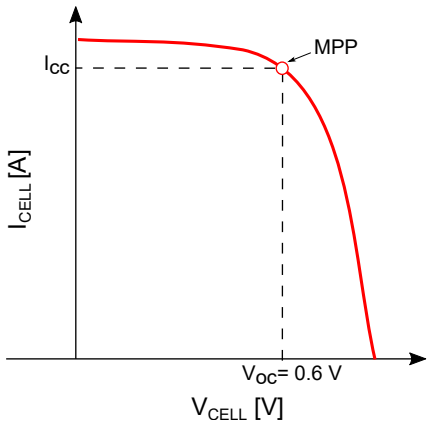

Figure 2.6: Current-voltage characteristic curve I(V).

growth rate of 36% over the 167.8 GW installed in 2021. The low scenario estimates a drop in demand to 181.4 GW by the end of 2022, which, as mentioned [SPE 2022], is really improbable considering the strong demand of solar energy in recent years. Finally, the high scenario forecasts up to 270.8 GW of solar additions in 2022. In addition, as it can be seen in the same Figure 2.5, it is expected that in 2026 the installed capacities will be between 243.5 GW , in the worst case, and 458.8 GW in the best case, being 1.7 the installed capacity in the best case in 2022. After knowing the context of the photovoltaic industry, the following is presented a description of the components that are part of the previously mentioned PV systems.

2.2 PV system components

This section briefly introduces the structure of the photovoltaic system comprising the following elements: the PV generator, the wiring and the junction box, the inverter, and the protection system.

2.2.1 PV generator

The PV generator corresponds to the unit that produces electrical energy in the form of direct current. The generator converts solar energy into electrical energy using the photovoltaic cell as the basic unit. The association of several photovoltaic cells in series/parallel gives rise to a photovoltaic generator that has a non-linear current-voltage I(V) characteristic that presents a maximum power point (MPP) as shown in Figure 2.6.

If the cells are connected in series, the voltages of each cell add, increasing the total voltage of the generator. If the cells are connected in parallel, their current adds to obtain a more important current supplied [Bun 2011a, Cid Pastor 2006a].

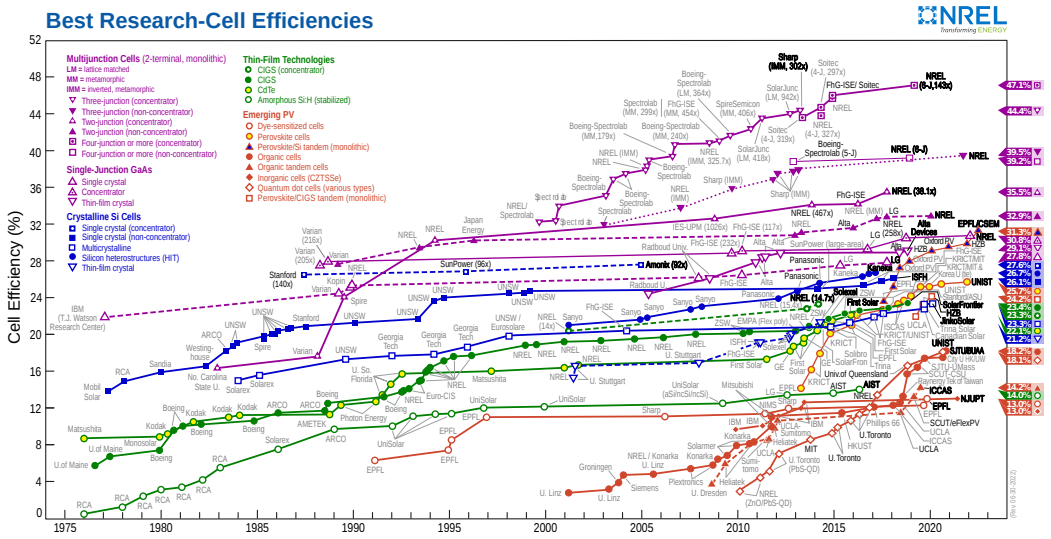

Figure 2.7: Historical evolution of technology market share and future trends [EPIA 2011].

2.2.1.1 PV Cell

The solar cell is the basic unit responsible for converting sunlight directly into electricity. Photovoltaic cells can be built with different materials. Some of these materials are still in the research phase. Among the emerging technologies that are showing very promising results are dye-sensitized [Upadhyaya 2013], organic polymers [Dou 2013] and perovskite [Zhou 2014] cells. These new PV cell technologies reduce the cost of photovoltaic energy by several orders of magnitude [IEA 2007a]. Due to the great variety of PV cells, extensive studies are carried out evaluating the efficiency of all solar cell technologies [NREL 2022]. In [NREL 2022], the National Renewable Energy Laboratory (NREL) exposes a study on the efficiency of PV cell technologies on the market and the results are presented in Figure 2.7.

As it can be seen in Figure 2.7, multijunction cell technologies have an accelerated growth in efficiency. In the same Figure 2.7, it can be seen that cells based on thin film technology present efficiencies approximately equal to those of silicon. However, they still represent a small percentage of the annual production [Correia 2021]. Because the most commercialized cells are the crystalline Si cells, this thesis focuses on the faults present in this type of crystalline Si cells, specifically in polycrystalline. On the other hand, it is important to mention that studies similar to those presented in this investigation can be applied to the multiple technologies presented in Figure 2.7.

Knowing the technology with which PV cells are manufactured is extremely important since the voltage generated by a PV cell is strongly linked to the value of the gap of the material from which it comes. This generated voltage can vary between 0.3 V and 0.7 V depending on the material used and its arrangement, as well as the temperature of the cell and its aging [Cid Pastor 2006b]. For example,

for crystalline and amorphous silicon type cells it is 0.6 V. Similarly, the current of a PV cell is a function of the cell surface and for the same surface, it depends on the cell efficiency. According to [NREL 2022], the efficiency of monocrystalline cells is between 26.1% and 27.6%, while that of polycrystalline cells is between 21.2% to 23.3%.

A deeper analysis in terms of efficiency comparing multiple PV technologies is carried out in [Green 2022]. In [Green 2022] the technologies of Figure 2.7 are compared as a function of Efficiency (%), Area (cm^2), Voc (V), Jsc (mA/cm^2), Fill factor (%), where VOC represents the open- circuit voltage and JSC the short-circuit current density. Table 2.1 shows some of the comparison results between PV cell technologies exposed in [Green 2022].

Table 2.1: Confirmed single-junction terrestrial cell and submodule efficiencies measured under the global AM1.5 spectrum (1000 W/m^2) at 25°C

Classification	**Efficiency** (%)	**Area** (cm^2)	**Voc** (V)	**Jsc** (mA/cm^2)	**Fill factor** (%)
Silicon					
Si (crystalline cell)	26.7 ± 0.5	79.0 (da)	0.738	42.65	84.9
Si (crystalline cell)	26.3 ± 0.4	274.3 (t)	0.7502	40.49	86.6
Si (DS wafer cell)	24.4 ± 0.3	267.5 (t)	0.7132	41.47	82.5
Si (thin transfer submodule)	21.2 ± 0.4	239.7 (ap)	0.687	38.50	80.3
Si (thin film minimodule)	10.5 ± 0.3	94.0 (ap)	0.492	29.7	72.1
III-V cells					
GaAs (thin film cell)	29.1 ± 0.6	0.998 (ap)	1.1272	29.78	86.7
GaAs (multicrystalline)	18.4 ± 0.5	4.011 (t)	0.994	23.2	79.7
I nP (crystalline cell)	24.2 ± 0.5	1.008 (ap)	0.939	31.15	82.6
Thin film chalcogenide					
CIGS (cell) (Cd-free)	23.3 ± 0.5	1.043 (da)	0.734	39.58	80.4
CIGSSe (submodule)	19.8 ± 0.3	665.4 (ap)	0.688	37.96	75.9
CdTe (cell)	21.0 ± 0.4	1.0623 (ap)	0.8759	30.25	79.4
CZTSSe (cell)	11.3 ± 0.3	1.1761 (da)	0.5333	33.57	63.0
CZTS (cell)	10.0 ± 0.2	1.113 (da)	0.7083	21.77	65.1
Amorphous/microcrystalline					
Si (amorphous cell)	10.2 ± 0.3	1.001 (da)	0.896	16.36	69.8
Si (microcrystalline cell)	11.9 ± 0.3	1.044 (da)	0.550	29.72	75.0
Perovskite					
Perovskite (cell)	23.7 ± 0.5	1.062 (da)	1.213	24.99	78.3
Perovskite (minimodule)	21.4 ± 0.4	19.32 (da)	1.149	23.41	79.6
Dye sensitised					
Dye (cell)	11.9 ± 0.4	1.005 (da)	0.744	22.47	71.2
Dye (minimodule)	10.7 ± 0.4	26.55 (da)	0.754	20.19	69.9
Dye (submodule)	8.80 ± 0.3	398.8 (da)	0.697	18.42	68.7
Organic					
Organic (cell)	15.2 ± 0.2	1.015 (da)	0.8467	24.24	74.3
Organic (minimodule)	14.5 ± 0.3	19.31(da)	0.8518	23.51	72.5
Organic (submodule)	11.7 ± 0.2	203.98 (da)	0.8177	20.68	69.3

In Table 2.1 the abbreviations mean the following: (ap), aperture area; (da), designated illumination area; (t), total area; a-Si, amorphous silicon/hydrogen alloy; CIGS, $CuIn_{1,...,y}Ga_ySe_2$; CZTS, Cu_2ZnSnS_4; CZTSSe, $Cu_2ZnSnS_{4,...,y}Se_y$; DS, directionally solidified (including mono cast and multicrystalline); nc-Si, nanocrystalline or microcrystalline silicon [Green 2022].

Other studies such as the one by [Dirnberger 2015] have studied the impact

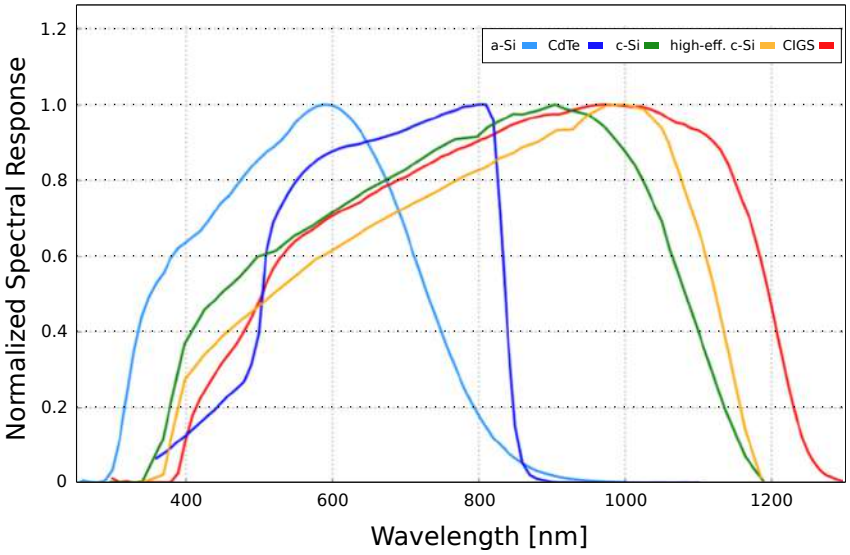

Figure 2.8: Typical, normalized spectral response data for single junction PV technologies, as used for calculation of spectral mismatch factors. [Dirnberger 2015].

of solar spectral radiation on the performance of different photovoltaic technologies. Figure 2.8 shows a comparison between single junction PV technologies such as amorphous silicon (a-Si), Cadmiumtelluride (CdTe) crystalline silicon (c-Si), high-efficiency crystalline silicon (high-eff c-Si), and chalcopyrites $(Cu(In_xGa_{1,\ldots,x})(S_ySe_{1,\ldots,y})_2$, named CIGS [Dirnberger 2015].

The CIGS module used for tracing Figure 2.8 has a small band gap, there are CIGS modules with higher band gaps that have the same normalized spectral response as crystalline silicon [Dirnberger 2015]. In the same study [Dirnberger 2015], it is mentioned that for the CIGS module little spectral impact was observed during the summer months and a positive spectral impact in the winter. In addition, they were able to conclude that the CIGS technology showed a higher energy output with an annual spectrum effect of approximately +1.8% compared to crystalline silicon which showed +1.5%.

Today, most commercial solar cells are photodetectors that use a solid semiconductor material to form a p-n (positive-negative) junction onto which light is incident. This incidence on the semiconductor material excites the flow of electrons that cross the junction by the electric field created when the p-n junction is formed [Jenkins 2017].

Finally, the semiconductor's p-n junction submitted to solar irradiation is connected to an external circuit, where the flow of electrons across the junction creates direct current (DC) electricity. Figure 2.9 represents the physical principle of operation of a PV cell.

As solar radiation increases, the number of electrons increases in the PV cell. In turn, the increase in the flow of electrons increases the current flow generated by the photovoltaic solar cells. Therefore, the short-circuit current Isc is directly

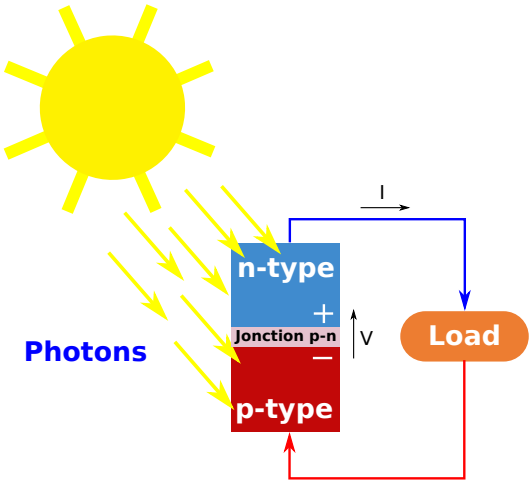

Figure 2.9: Operation of a PV cell [Jenkins 2017].

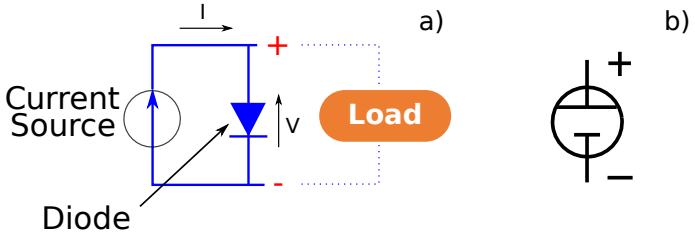

Figure 2.10: Representation of a solar cell. a) Simple equivalent circuit of an ideal solar cell. b) Symbolic representation

proportional to the solar radiation G for a given PN junction temperature.

The open circuit voltage V_{oc} of the photovoltaic cell is determined by the electric field created in the depletion region of the p-n junction, independent of solar radiation but depending on intern PN temperature [McEvoy 2013]. Therefore, the well-known operation of an ideal solar cell is described as follows:

$$I = I_{ph} - I_0 \left[e^{\frac{V_{cell}}{V_T}} - 1 \right] \tag{2.1}$$

where I_{ph} is the photocurrent directly proportional to irradiance, I_0 is the diode saturation current, V_{cell} is the solar cell terminal voltage, and V_T is the thermal voltage described as follows:

$$V_T = \frac{KT}{q} \tag{2.2}$$

where K is Boltzmann's constant (1.38 × 10-23 J/kelvin), T is the absolute temperature, q is the charge of the electron (1.602×10^{-19} C). Equation 2.1 leads to the simple equivalent circuit of a solar cell, which can be represented by a current source in parallel with a diode, shown in Figure 2.10.

The equivalent circuit named one diode PV model in Figure 2.10 can include

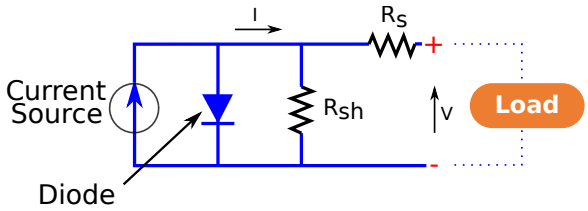

Figure 2.11: Equivalent circuit of a solar cell with series and shunt resistance.

shunt and series resistors to represent the losses within the PV cell under the configuration in Figure 2.11.

The shunt resistor R_{sh} represents current leakage through the p-n junction around the edge of the cell and the effect of sensing and impurities in the junction region. The series resistor Rs represents the resistance of the bulk semiconductor, the metal contacts, and the connection of the contacts to the semiconductor material [McEvoy 2013]. Including the effect of these losses results in the Equation 2.3, where n is the ideality factor, also known as the quality factor or sometimes emission coefficient, that represents the combination of electrons and holes in faults in the junction region [Caprioglio 2020, Yatimi 2014]. According to [Ryu 2019], the value of n varies between 1 and 2. The ideality factor n is considered one of the important parameters describing the performance of photovoltaic cells [Tarabsheh 2011] as it means the measure of how closely the device follows the ideal p-n junction behavior [Muhammadsharif 2017]. For this reason, several papers showed different methods for the calculation of n of PV cells [Yordanov 2013, Bouzidi 2012].

$$I = I_{ph} - I_0 \left[e^{\frac{q(V+R_sI)}{nkt}} - 1 \right] - \frac{V + R_sI}{R_{sh}} \tag{2.3}$$

Individual solar cells are used to power small devices such as electronic calculators or some home appliances. However, there are configurations of PV cell sets that aim to increase the voltage they generate to power applications that require larger amounts of energy. This set of solar cells are electrically connected in series with a single by-pass diode and encapsulated in an environmental protection laminate named a PV module [Bressan 2014, Berasategi Arostegi 2013].

2.2.1.2 PV Module

A photovoltaic module is the smallest set of individual solar cells electrically connected to each other. This set of solar cells are assembled in the module, electrically connected in series, and encapsulated with a protective material to be protected against corrosion with oxygen and humidity.

This encapsulation has several functions of protection. It provides protection against shock, humidity, corrosion, dust and more generally direct contact with air. [Hadj Arab 1989]. In addition, the encapsulation can help to control the intern temperature of the PV cells, which will allow a good dissipation to the outside. This point is important to achieve a good electrical conversion efficiency helping to

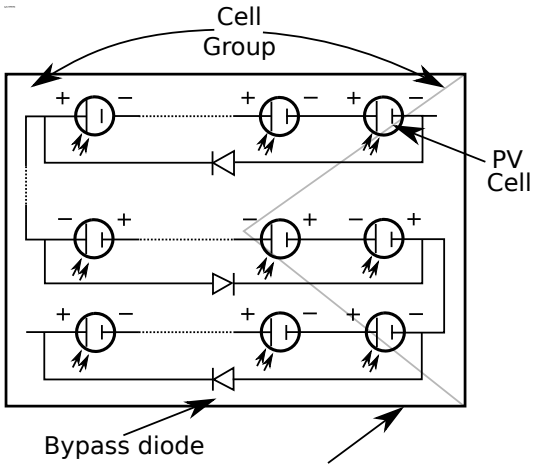

Figure 2.12: Details of a PV module constituted of PV cells. Cells are connected in series to form PV modules and are associated with secure elements named By-pass diodes.

evacuate the part of the incident solar energy that is not transformed into electrical energy and increase the intern temperature [Hadj Arab 1989]. Finally, this encapsulation isolates the user from possible electric shocks [Vighetti 2010, Bun 2011a]. The number of photovoltaic cells in series that are in a PV module depends on the requested power. Generally a PV module contains a variable number of 36, 40, 54, 60, 72 and even 92 cells in series [Bun 2011a]. A possible cell configuration is presented in Figure 2.12.

As it can be seen in Figure 2.12 the cells of a module are associated in several groups. Then, each group is connected in antiparallel (series) with a diode, named a bypass diode (there are usually 18 cells for a bypass diode). Some less common configurations propose connecting each individual cell to a bypass diode [Suryanto Hasyim 1986] or modifications to where the diode is connected [Silvestre 2009, Díaz-Dorado 2010]. The purpose of this diode By-pass is to block when the voltage of the photovoltaic cells that it groups is positive and let the current pass otherwise.

Connecting a set of PV cells under the configuration of a PV module like the one in Figure 2.12 increases the power. In these configurations with PV cells in series, the current remains the same while the voltage is multiplied by the number of cells in series. In an analogous way to the connection of a set of PV cells to increase the power generated, there is a configuration in which a set of PV modules is connected named a PV string.

2.2.1.3 PV string

A PV string is made up of a set of photovoltaic modules in series to achieve the voltage level required by the application. The string is equipped with a protection diode named an anti-reverse diode, the purpose of which is to block the flow of a

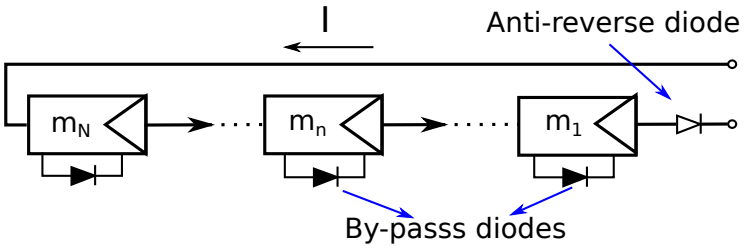

Figure 2.13: string PV

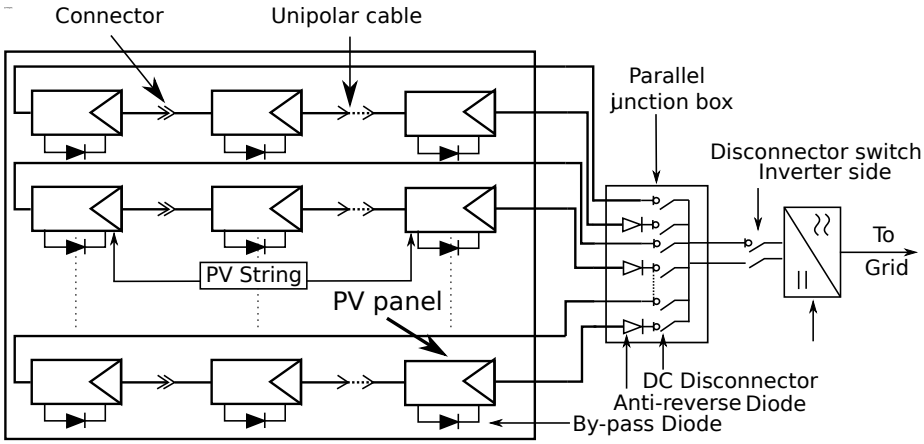

Figure 2.14: Array PV

reverse current in the PV string [Bun 2011a]. An example with N panels connected in series is presented in Figure 2.13.

2.2.1.4 PV Array

A PV array is a complete DC power generation unit. This unit consists of several photovoltaic strings, each made up of the same number of modules. The strings are assembled in parallel in order to increase the current, and thus obtain the desired power for the photovoltaic installation [Bressan 2014, Berasategi Arostegi 2013]. The PV array can be as small as several modules, or large enough to cover acres like a utility photovoltaic plant [Fadhel 2018].

Figure 2.14 corresponds to one of the most used configurations named series-parallel (SP) configuration.

As it can be seen in the same Figure 2.14, a photovoltaic installation also has two important components. First, the parallel junction box [UTE 2008] that designates the connection box of the different photovoltaic strings in parallel and where the anti-reverse diodes and DC disconnectors are located. It is necessary to clarify that the parallel junction box is different from the junction box of the photovoltaic modules that is located on the back of the PV panels and that includes bypass diode type protections. Second, one or more inverters to be able to deliver electricity to the grid and other components. For more technical information on a PV field please

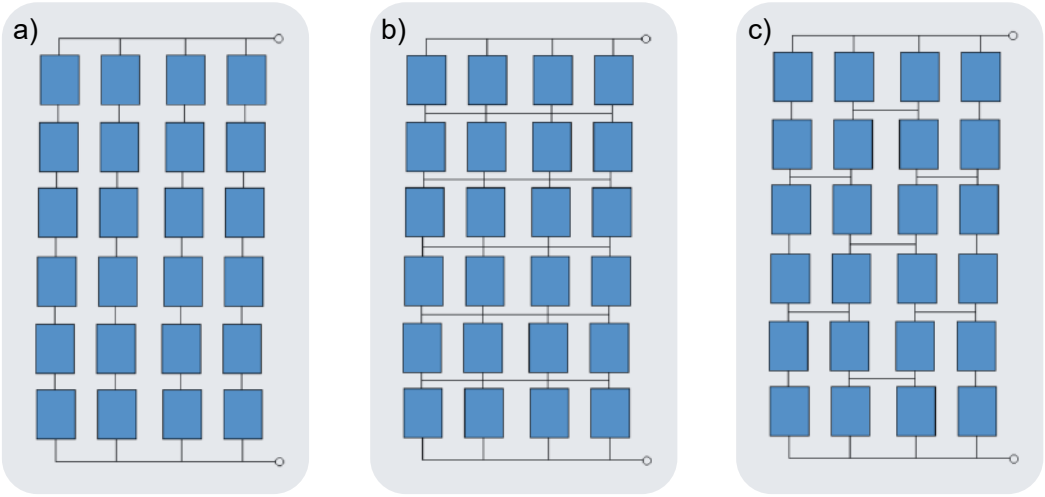

Figure 2.15: Examples of configuration of a PV array. a) Series-Parallel (SP) configuration; b) Total-Cross-Tied (TCT) configuration; c) Bridge-Linked (BL) configuration [Andrianajaina 2017]

see [UTE 2008].

There are also Total Cross Tied (TCT) and Ridge Linked (RL) type configurations [Andrianajaina 2017, Kaushika 2003]. An example of the three configurations is presented in Figure 2.15.

This work is developed on this type of configuration since the total cross connection (TCT) and bridge link (BL) type configurations, although they show to improve field performance, are not widely used due to their low economic viability [Picault 2010].

2.2.2 Wiring and Junction Box

The connections between the PV modules are made using unipolar cables as shown in Figure 2.14. Generally the use of double insulated single conductor cables is recommended to reduce the risk of ground fault or short circuit [Verhoeven 1998]. As a safety measure, the use of plug-in connectors that simplify the installation procedure and reinforce protection against the risk of electric shock is recommended. As these PV arrays generally contain several PV strings, these panels are connected in parallel with the use of a junction box. As it can be seen in Figure 2.14, this junction box can contain protection elements such as fuses, switches and disconnectors.

2.2.3 Inverter

The role of the inverter is to extract the maximum power from the photovoltaic generator if it is dotted with MPP Trackings and intern power structures with for example the dissociation of different strings. Its main function is to convert DC current and DC voltage into alternative ones to be able to inject it into the grid or several appliances. To extract the maximum power from the photovoltaic generator,

the inverter uses an MPPT (Maximum Power Point Tracker) search algorithm. More detailed information about the architecture of the inverters including their topologies can be found in [Akbari 2021, Asif 2021, Sathik 2021, Chamarthi 2020, Syed 2020, Bun 2011a].

2.2.4 Protection system

In PV fields, there are two types of diodes that protect the operation of a photovoltaic array: bypass and anti-reverse diodes. These diodes are made of a semiconductor material, in most cases silicon, with two terminals connected [Bun 2011a, Vighetti 2010]. The function of these diodes is to allow electricity to flow in one direction but not the other, protecting the cells against their operation in the reverse regime [Bun 2011a]. These two diodes are discussed below.

2.2.4.1 Bypass diode

Each PV module is equipped with bypass diodes that prevent the destructive effects of hot spots caused by non-uniform irradiation, protect the weaker PV cells it groups against reverse bias or other types of faults. A bypass diode is connected in parallel but with opposite polarity to groups of between 18 and 24 solar cells inside the photovoltaic module. In other words, this diode is blocked when the voltage of the photovoltaic cells it groups is positive and it lets the current pass otherwise.

2.2.4.2 Anti-reverse diode

The anti-reverse diode is placed in a photovoltaic field at the end of each of its PV strings, as seen in Figures 2.14 and 2.13. This diode allows the passage of the current that leaves the photovoltaic string towards the junction box and blocks the passage of the incoming current to the photovoltaic string. This type of diode is mainly installed in two situations. In the first place, when the strings of the photovoltaic field present different voltages due to the existence of anomalies. In this scenario, the lower voltage strings consume the currents provided by the higher voltage strings. Second, under the absence of the energy produced by the photovoltaic field, due to the absence of sunlight such as during the night. This scenario occurs when the PV array has batteries that can begin to discharge, turning the PV array into dissipation mode. However, the use of these diodes introduces a production loss due to the voltage drop caused by this diode during normal operation of the photovoltaic field. In addition, these diodes can fail, so periodic inspection is required.

As stated, there are multiple components within a PV array that can fail. Therefore, it is essential to have a good understanding of the causes of common PV array faults and conventional detection methods to ensure continuous and optimal PV array production.

2.3 Formal fault dictionary

This dictionary attempts to retrieve most of the information available for understanding faults in PV systems. However, it is first necessary to define the concept of fault adopted by this research. As mentioned in [Jordan 2017], defining the term “fault” in a consistent and meaningful way in PV systems is really challenging. The IEC 60050-191 standard defines the term fault as "the termination of the ability of an item to perform a required function" [IEC 1990]. In other domains maybe this is quite a transparent and clear definition. However, in a PV system it may not be so clear, leading to several different uses during the last decades in the photovoltaic field.

Some institutions, a reference in the field of photovoltaics, such as the Electric Power Research Institute, define the term as a decrease in maximum power of more than 50% in a module that could not be repaired in the field [S 1993][12]. More recently, the International Energy Agency defined the term “fault” as the irreversible degradation of a module resulting in power loss or a safety issue [Köntges 2014a]. This definition is adopted for the construction of this dictionary and the development of the complete research. The documents discussed in this section contain information about faults and generally degradation modes that cause observable changes in the appearance, performance and security of a module.

Using this definition, it is possible to state that PV systems are susceptible to faults in any of their components. Some faults are due to cell deterioration, cracks, overheating, humidity penetration, degradation of interconnections, corrosion of connections between cells. Likewise, in other scenarios the faults are caused by modules of different performance, broken encapsulation, short circuit, or inverted modules. If faults are analyzed on the junction box side, faults caused by electrical circuit breakage, short circuit, destruction or corrosion of connections may occur. Finally, on the part of the diodes, faults occur due to the destruction of the diodes, absence or non-operation of the diodes, reversal of the polarity of the diodes during assembly or poorly connected diodes.

Due to the wide number of faults that can occur in a PV system, this research proposes a comprehensive study of the most known faults in PV systems. This study is presented under a scheme of a "Formal Fault Dictionary". The dictionary proposed in this section makes a significant contribution to the effective and automatic detection of faults in PV systems. This dictionary contains four sections. First, the main causes of faults are presented. Second, a new multilevel classification is proposed and explained along with each element of it. Third, the frequency of occurrence of faults is divided into degradation faults and sudden faults throughout the life of a PV system. Finally, the fourth section of this dictionary exposes the impact of faults in terms of loss of power and risk to human safety.

2.3.1 Main causes of faults

In the literature, it is conventionally mentioned that faults can be caused by external factors, the interaction of materials or the aging of components [Pillai 2018a]. In a complementary way, in [Aghaei 2022, Li 2021c, García-Gutiérrez 2019] , it is proposed to analyze the causal relationship between the same faults. This analysis is carried out using a scheme named the cause-effect circle. These 4 types of origins are adopted for this research and explained below.

2.3.1.1 External causes

External causes include human error and faults caused by environmental conditions. Human error generally occurs during installation or transportation of PV system elements. Transport is the first critical stage of the life cycle of PV modules [Strohkendl 2010], since there can be shocks or vibrations that generate breakages or microcracks in the PV cells [Köntges 2011b]. Equally critical is the installation process. One of the biggest causes of glass breakage is clamping during installation [Dietrich 2008]. Another cause is screws that are too tight, clamps that are too short or too narrow that generate mechanical stress or cracks in the glass. Another common human error is incorrect wiring of connectors. When the connectors are poorly adjusted or crimped, an open circuit, a line fault, a ground fault or a loss of power can be generated [Gallardo-Saavedra 2019]. In other cases, the connectors are installed near flammable materials, where arc faults can cause fires.

Outdoor PV systems are exposed to strong environmental conditions that can generate permanent or non-permanent faults [AbdulMawjood 2018a]. Permanent faults such as cracks or detachments in the frames or glasses of the photovoltaic module [Madeti 2017b] can be generated due to lightning strikes [Falvo 2015], intense snowfall [Köntges 2014b] or hail [Makarskas 2021]. These cracks allow oxygen and humidity to enter the photovoltaic module, causing corrosion of the electrical circuits. Non-permanent faults (of short duration) in these PV systems are generated by the appearance of tree leaves, contamination, sand or dust, excrement or dirt in general, shadows from buildings, clouds, among others. Shading is one of the most studied faults because it generates localized heating or a hot spot [Molenbroek 1991]. If the temperature of the cell exceeds 150°C, it can be irreparably damaged and even start a fire [Fadhel 2018].

2.3.1.2 Material interaction

The combination of different materials used in a photovoltaic module together with environmental factors such as humidity, heat, UV radiation, etc., can generate degradations on the surface of the module and its electrical behavior. Visible faults such as encapsulant discoloration (yellowing or browning), corrosion, cell cracking or delamination may be observed. Most of the PV cells found in the market, i.e. crystalline silicon (c-Si) and thin film type, are constructed of front glass layers, encapsulation layer, solar cell/substrate thin film and backsheet [Li 2021c]. The

interactions between these layers can also generate various types of faults. Likewise, electrical degradations such as cell disconnection, short circuits and PID faults (Potential Induced Degradation) can be generated.

2.3.1.3 Component aging

Another component that strongly affects the performance of PV systems is the natural aging of components. This natural degradation can cause problems such as discoloration of the encapsulant, welding faults, detachment of the module frame, formation of air bubbles on its back face. This degradation of the photovoltaic system is a continuous process over time, which can be caused by factors such as mismatch (cells that are not perfectly identical), the penetration of humidity that generates corrosion [Kuitche 2014] or material degradation caused by UV light. As mentioned in [Manganiello 2015] between aging and mismatch there is a "closed loop" link, since aging generates mismatch which in turn accelerates aging mainly due to thermal effects reducing the production and useful life of the PV system.

2.3.1.4 Cause-effect circle

Finding a single cause for faults in photovoltaic systems is quite complex. In fact, the occurrence of a fault is usually accompanied by a degradation in other properties (mechanical, chemical or electrical) of the photovoltaic module. This degradation in turn aggravates the original fault and/or generates other faults. For this reason, knowing the cause-effect relationships between faults and their impact is vital to improve the design, data acquisition and, consequently, the detection of faults in this type of system. A detailed analysis of these interactions is presented in [Manganiello 2015] using a causal loop, where the mismatch is recognized as the intermediate fault mode caused by operating parameters such as temperature, voltage, or current and external factors such as environmental and human error. For more information on this type of causal relationship please refer to the reference [Manganiello 2015].

In this work, only the most frequent faults associated with solar cells, the module, the array, the protection system, wiring and junction box are analyzed. This analysis allows to build for the first time a formal dictionary of faults in photovoltaic systems that is explained below. As an introductory part of the dictionary and based on the main causes of faults in photovoltaic systems, a new multilevel fault classification is proposed in the following section.

2.3.2 Multilevel fault classification

Carrying out a correct classification of the faults allows to improve the understanding of the similarities and differences between the photovoltaic faults. This is vital for fine-tuning fault detection in these types of systems. Different classifications are proposed in the literature. In [AbdulMawjood 2018a] faults are classified as permanent, intermittent and incipient, taking into account the duration and severity

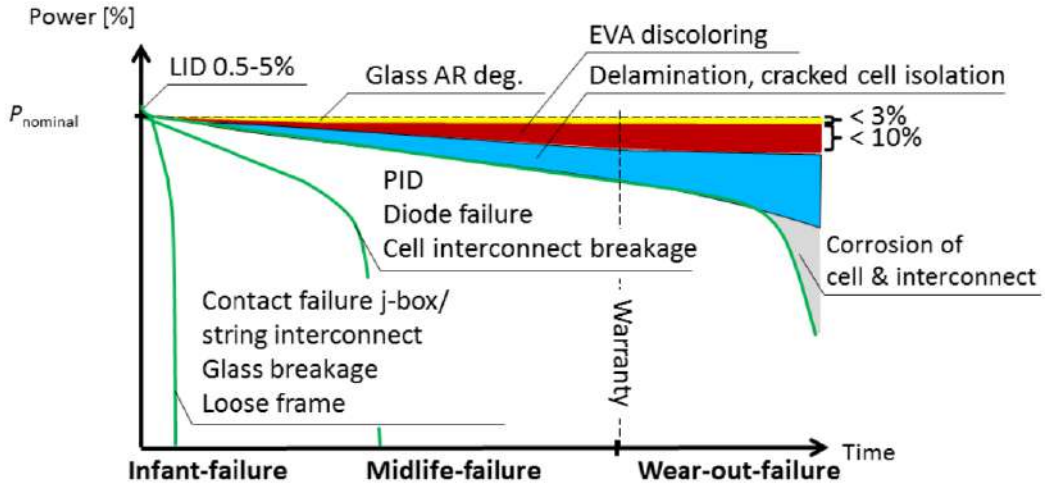

Figure 2.16: Aging mechanisms leading to PV module degradation [Dross 2017].

of the fault. The problem with this classification is that sometimes incipient faults such as delamination or corrosion become permanent. In [Pillai 2018a], faults are classified as physical, environmental, and electrical based on the cause or nature of the fault. However, some faults may meet both criteria simultaneously. For example, a line fault, ground fault, bypass diode faults, or abnormal degradation could be classified as physical, electrical, environmental damage, or a combination of these [Li 2021c]. In [Dross 2017], faults are classified as infant-faults, midlife-faults, and wear-out-faults. The classification exposed in [Dross 2017] is presented in Figure 2.16.

The difficulty of the classification in Figure 2.16 is that the time of occurrence of these faults can be strongly linked to the environmental conditions where the PV system is located and to human errors that allow the appearance in different periods. In [Triki-Lahiani 2018a] the faults are classified as module faults, inverter faults and others, that is, depending on the element where the fault occurs. The problem with this proposal is that it contains a category named "others" that is highly ambiguous. As a possible solution to this ambiguity problem, in [Li 2021c] it is proposed to modify the scheme to classify faults based on the moment where the fault appears in the categories: cell level, module level and at the array level.

Despite the interesting contributions of the aforementioned works, there are still gaps that do not allow an easy and direct understanding of the aspects that cause faults in PV systems. As a contribution to solve this problem, this research proposes a new classification, product of an arduous analysis of the literature, based not only on the elements in which the fault is found (cell, module, array) or the causes (causes external, material interaction, component aging, or cause-effect circle), but rather integrates the two aspects and also differentiates the faults that occur in the protection system, wiring and the junction box. Finally, this classification also considers the nature of the fault such as: structural *(S)*, electrical *(E)*, caused by a

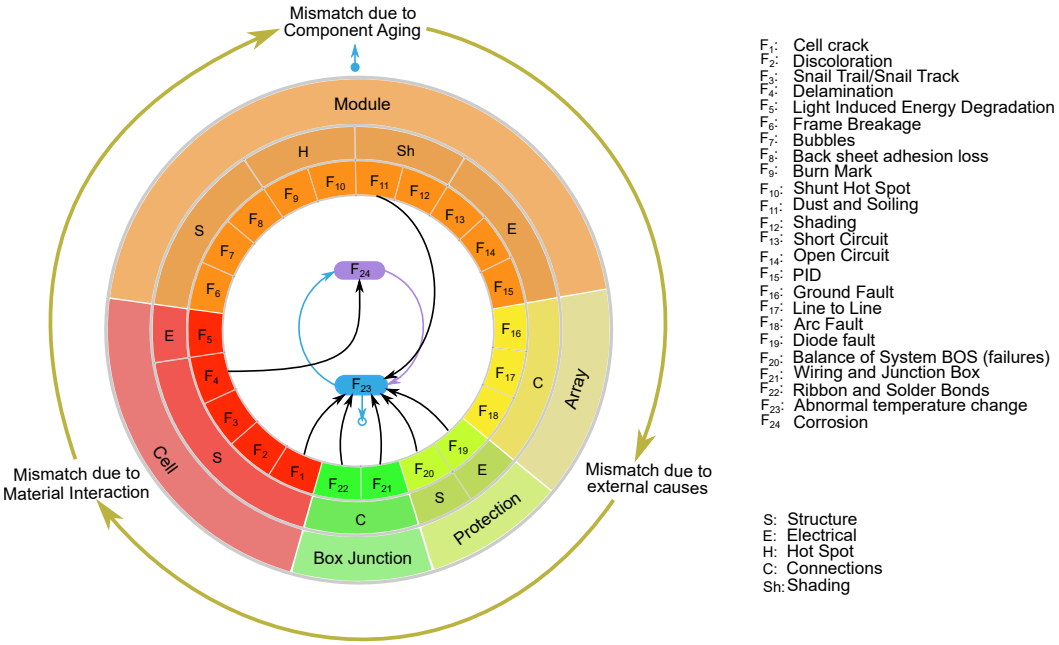

Figure 2.17: Proposed multilevel classification of faults in PV systems

hot spot *(H)*, connection problems *(C)* or shading *(Sh)*. Since this research focuses on PV array faults, faults related to DC conditioning units (such as converter, inverter, battery bank) are excluded. Also, faults that occur on the AC side of a stand-alone or grid-connected PV system are not considered. Those faults on the AC side can be detected and isolated using the protection standards specified in [IEC 2017a, IEC 2016a]. Figure 2.17 shows the general scheme of the proposed classification.

As it can be seen in Figure 2.17 all faults lead to abnormal degradation of the PV system. In other words, the abnormal degradation is an integral reflection of several faults [Manganiello 2015, Masmoudi 2016] that can reduce the output power by up to 50% [Bastidas-Rodriguez 2013, AbdulMawjood 2018a]. Likewise, it can be observed in Figure 2.17 that the operation at high temperature (F_{23} : Abnormal temperature change) promotes the formation of faults due to the interaction of the materials and premature aging of the components. For example, high-temperature protection diodes age rapidly and behave like impedances or contacts that degrade and cause open-circuit type faults. In the same way, in Figure 2.17, it is possible to observe the causal relationships with the fault named Corrosion (F_{24}). This fault occurs when the encapsulant fails to protect the PV elements against environmental corrosion. Natural corrosion in the absence of water is slow, that is, the greater the adhesive force, preventing the access of water to the corrosion sites and the drift of acetic acid towards the corrosion sites, the stronger the protection against corrosion [Kempe 2007, Walker 2001]. These corrosion phenomena degrade the state of the PV system and favor the appearance of delamination, and even overheating of the cell.

Figure 2.18: Example of Cell crack

This once again reinforces the importance of a deep understanding of the faults that occur in PV systems. To this end, the set of faults represented in Figure 2.17 is described in detail below.

2.3.2.1 Cell-level faults

Cell-level faults refer to PV faults that affect a single PV cell. However, on some occasions this type of fault can extend to adjacent areas over time, generally without affecting the entire surface of the PV module [Li 2021c]. Faults are grouped into structural and electrical based on the nature of the fault. The set of faults of this category are presented below.

2.3.2.2 Cell crack (Structure):

Cell cracks are the mechanical stress-induced cracks in the silicon substrate of PV cells or thermodynamic stresses induced by thermal cycling that can occur at any level of the PV cell's lifetime. Normally these cracks are invisible to the naked eye and can be originated in the stages of production, transportation (poor encapsulation or vibrations), installation or operation (internal mechanical problems) [Cristaldi 2015, Köntges 2014b] and aggravated by increases in temperature due to other faults or environmental conditions [Manganiello 2015].

Cracks in PV cells are not always expressed with the same geometry, they can vary in length and orientation. Some of these cracks can be caused during the production process, as some manufacturers try to increase adhesion between the layers to avoid delamination or corrosion problems, but this in turn generates internal stresses that can crack the cell, especially in thin film cells.

The level of impact on power loss is directly related to the "inactive" area of the cell. For this reason, the behavior of the modules with cracked cells can become similar to the case of dust, soiling or partial shading covering the cells, because the cracks reduce the photo-generated current [Meyer 2004]. In addition, these cracks (microcracks) can be the cause of local hot spots [Köntges 2011a] that reduce the production of the PV system due to thermal effects.

Figure 2.19: Example of Discoloration

Discoloration (Structure): The discoloration is generally related to PV modules that use EVA (ethylene vinyl acetate) as the encapsulating material. Discoloration faults refers to the yellowing, browning or darkening of photovoltaic cells. Discoloration faults cause a change in the transmission of solar radiation that reaches the cell surface, reducing energy efficiency. As a possible solution, some manufacturers are using thermoplastic polyolefin as an encapsulant reducing discoloration by up to 9 times [Adothu 2019].

In the literature it is reported that the appearance of this fault is strongly linked to exposure to ultraviolet light and operation at high temperatures ($t > 55°C$) [Köntges 2014b]. One of the symptoms that this fault presents is the more significant intern temperature increase in its central part than in the edges during normal operation[Parretta 2005]. In addition, this fault is cited as the main factor inducing aging of PV cells [Rabii 2003, Parretta 2005]. Discoloration particularly shows up in older systems. However, in more recent systems it appears in warmer climates, but to a lesser degree [Jordan 2017].

Snail track or Snail Trails (Structure): It is a gray/black discoloration of the silver paste of the front metallization of screen-printed solar cells. This discoloration spreads across the surface of the cell giving the illusion of a snail trail, hence its name. Snail trails generally occur 3 months to 1 year after the installation of the photovoltaic modules. The origin of this fault is not clear, but in some documents it is mentioned that it may be due to silver particles containing sulfur, phosphorus or carbon [Li 2021c]. Other works indicate that cell cracks, EVA film additives, chemical additives used on the cell surface, or humidity can accelerate the formation of snail trails [Kim 2016]. This fault can propagate through the PV cell but at a very slow rate [Köntges 2014b], or saturate directly after the first occurrence.

Delamination (Structure): This fault represents the loss of adhesion between the glass, the encapsulant, the active layer and the back layer [Munoz 2011]. When the cell is a thin film, the transparent conductive oxide (TCO) can also flake off from the adjacent glass layer [Li 2021c]. The main causes of this fault are related to environmental limitations such as high humidity and temperature, exposure to UV

Figure 2.20: Example of Snail track or Snail Trails

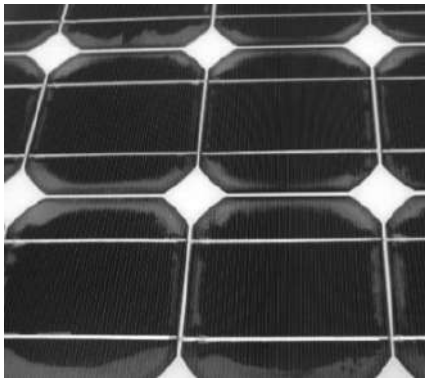

Figure 2.21: Example of Delamination [Omazic 2019]

radiation, abnormal increase in cell temperature, salt accumulation, contamination, cell movement or external factors. [Manganiello 2015, Dumas 1982, Oreski 2010]. This type of fault generally causes high levels of corrosion [Li 2021c].

Among the best known methods to detect this fault are thermography, ultrasonic scanning and X-ray tomography. The irregularity of the surface can be quantified using a reflectometer. Delamination cannot be completely avoided [Zimmermann 2013], leading to changes in the electrical performance of the PV module [Park 2011]. In addition, this type of fault can be aggravated in conditions of high humidity and temperature [Kempe 2006]. Extensive studies of this type of fault are developed over long periods of time (22 years) in c-Si photovoltaic modules exposed to outdoor conditions [Dunlop 2006, Kaplanis 2011].

Light Induced Energy Degradation (LID) (Electrical): LID is a natural degradation caused by a physical reaction as a result of the p-n junction of a photovoltaic cell. This fault is expressed as a reduction in the short circuit current and the open circuit voltage of the solar cell [Dross 2017, Lindroos 2016]. According to the EN 50380 [IOS 2017] standard, this fault must be taken into account by manufacturers for the power rating of the PV cell. For this reason, some classifications do not take this fault into account in their classifications. The occurrence of this fault can be between a few hours, days or even weeks. However it always occurs in

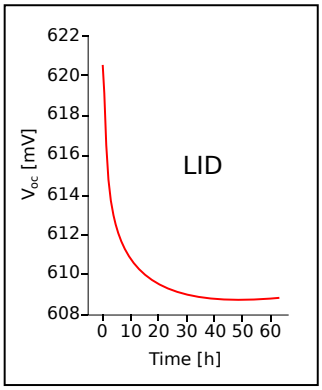

Figure 2.22: Example of Light Induced Energy Degradation (LID)

Figure 2.23: Example of Frame breakage [Köntges 2014b]

the early stages of panel operation [Köntges 2014b].

2.3.2.3 Module-level faults

At the module level, the faults presented in the PV cells are inherited. In addition, a set of faults classified in the categories of shading, structure and electrical are added. The detailed presentation of these faults is given below.

Frame breakage (Structure): Not only the layers of the PV cells can be separated. The frame can also detach or break, generating in many cases delamination and allowing the passage of humidity, giving rise to corrosion and electrical risks. One of the causes of the frame breaking is the heavy load of snow or dust, which will creep downhill and intrude into the gap between the frame and the glass [Li 2021c]. Another cause may be linked to sealing faults, typically silicone, or installation errors [Munoz 2011] that can cause deformation of the module, detachment of the photovoltaic glass frame, and therefore a reduction in the power produced.

Bubbles (Structure): This is also a specific form of delamination that is awarded at the module level. This type of fault produces an optical reflection that reduces the output power. Additionally, this fault causes humidity penetration, which then leads to various chemical and physical degradations and heavy levels of corrosion. Generally this type of fault is related to thermal decomposition [Pern 1997]. Bubbles can appear on both the back and the front of the module. These bubbles form an air chamber in which the gas temperature is lower than

Figure 2.24: Example of Bubbles [Kim 2021]

Figure 2.25: Example of Back sheet adhesion loss (BSAL) [de Oliveira 2018]

in the adjacent cells [Munoz 2011]. This air chamber worsens the heat dissipation capacity of the nearby cell, so the latter overheats, exhibiting a higher temperature than the adjacent cells [Ndiaye 2013].

When the bubbles appear in the front part, there is a reduction in the radiation that reaches the photovoltaic cell, thus creating a decoupling of light and increasing reflection. In other scenarios, the bubbles can rupture and damage the back sealing surface, allowing humidity ingress and thus generating corrosion processes leading to a reduction in series resistance [Kaplani 2012], which is considered the most frequent mode of degradation of the photovoltaic panel [Cristaldi 2015, Ndiaye 2013, Köntges 2014b, Schirripa Spagnolo 2012].

Back sheet adhesion loss (BSAL) (Structure): This fault refers to the loss of adhesion of the back sheet of the module, which is the protection of the electronic components from external factors and the safety of DC voltages. This fault depends directly on the type of sheet material [Köntges 2014b, Schirripa Spagnolo 2012, Sharma 2013, Solórzano 2013] and causes effects similar to those of delamination, and is also aggravated by sudden changes in temperature, humidity, mechanical stress, etc. This loss of adhesion of the sheet exposes the active electrical components and especially when it happens near a junction box or edge of the module [Novoa 2015].

Burn mark (Hot Spot): This type of fault usually originates due to the presence of partial shading + bypass diode fault or other mismatch fault (such as low resistance fault in c-Si). All of these causes result in power consumption

Figure 2.26: Example of Burn mark [Omazic 2019]

in the mismatch area rather than generation, and this in turn increases the local cell temperature and induces burn marks on the cell surface. However, similar burn marks can be generated by arc faults. The most likely cause of this fault permanently is the presence of an open circuit or faulty cell which produces less current and leads to power dissipation [Schirripa Spagnolo 2012, Hu 2014].

Other causes for this type of fault are dirt and dust accumulation [Massi Pavan 2011, Pigueiras 2014, Kalogirou 2013, Adinoyi 2013], cell degradation, incomplete edge insulation [Ndiaye 2013] due to transparent module materials or due to manufacturing tolerance and non-uniform insulation, among others [Köntges 2014b, Schirripa Spagnolo 2012]. That is, in general it occurs when some cells of the PV module have different I(V) curves [Massi Pavan 2014]. Depending on the level of impact of the Burn mark, it can even generate delamination or melting of the material. For the detection of this type of fault, methods based on infrared images are generally used [Simon 2010].

Shunt hot spot (Hot Spot): Partial shading could cause the cell to switch to a reverse biased voltage state and thin film cells are extremely sensitive to this phenomenon. The module current is concentrated in the bypass path and leads to the shunt hot spot. The behavior is quite different from the c-Si hot spot (burn mark). In this case it is the by-pass diode that cannot limit the reserved voltage. It is not likely to cause overheating, but it will cause the glass to break and increase the risk of electric shock. If the cause of the shunt hot spot is temporary, on some occasions the shunt hot spot is temporary, but generally because this phenomenon persists, the affected solar cells are permanently affected [Yang 2010].

Dust and Soiling (Shading): This fault is caused by the deposit of snow, dirt, dust, bird droppings and other particles that cover the surface of the photovoltaic module [Nguyen 2015, Patel 2008]. It is reported that dirt or solid shade (permanent shading) can cause a 10% to 70% reduction in power generation [Maghami 2016]. This solid or homogeneous shading has a balanced reduction in irradiation in the photovoltaic panels [Pillai 2018a, Solórzano 2013, Ji 2017]. This partial shading comes from scenarios such as passing clouds, smoke, dust, or other

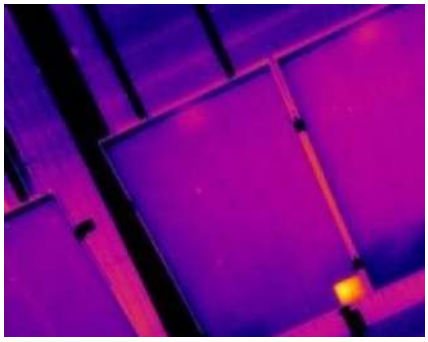

Figure 2.27: Example of Shunt hot spot [Aghaei 2015]

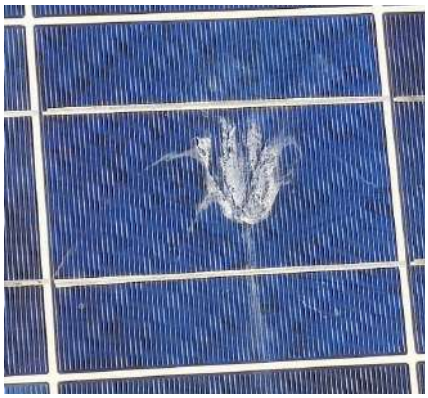

Figure 2.28: Example of Dust and Soiling

temporary effects. Usually this type of fault causes corrosion and hot spots.

Shading (Shading): Shading and partial shading (PS) are generally caused by poor planning of the PV system, leaving the system under a shadow from buildings, chimneys, or other elements that are not taken into account [Nguyen 2015, Patel 2008]. Depending on the object that causes the shadow, this type of fault can be classified as hard or soft, or permanent/temporary. For permanent shading there is a reduction in the output voltage that affects the power generation output of the photovoltaic arrays [Triki-Lahiani 2018a]. In the case where the irradiation is non-homogeneous, that is, resulting in unbalanced reduction of irradiance in the panels [Pillai 2018a, Solórzano 2013], the fault covers a part of the PV module and is named partial shading [Stettler 2005].

As demonstrated in an experimental research [Dreidy 2013], in certain conditions partial shading activates the bypass diodes affecting the voltage and not the current. That is to say, the "bypassed" cells do not have the opportunity to contribute voltage, which is why there is generally a reduction of the panel voltage by 1/3, a noisy high frequency I(V) curve and, consequently, a reduction of the output power [Triki-Lahiani 2018a]. However, it is an important mechanism since it avoids the reduction of the current that would be had without the diodes. When a PV cell is in shading fault it behaves like a resistor that begins to raise

Figure 2.29: Example of Shading

its temperature resulting in hot spots and permanent degradation of the PV array [Koutroulis 2012, Pillai 2019b, Ji 2017]. The reduction in power is directly proportional to the amount of unequal shadow created.

Short circuit (SC) / Open circuit (OC) (Electrical): Corrosion and damage to the structure are the main causes of open circuit or short circuit of the module. These types of fault lead to different levels of power loss or system shutdown and increase the risk of electrical shock or even electric arc. Another factor that particularly generates the short-circuit fault is the aging of the system [Dhanraj 2021]. In particular, the open circuit can be generated, in addition to the causes mentioned above, by multiple connections made by manufacturers between photovoltaic modules or PV cells. Due to the aging of low quality electrical cables, some disconnection may occur in the circuit preventing the panel from producing electrical energy [Dhanraj 2021].

Another reason for disconnection is due to poor soldering at cell string interconnections. The short circuit current and peak power decrease due to the open circuit fault, while the open voltage remains close to its normal value [Gokmen 2012]. Another reason for the appearance of the open circuit is the disconnection of the connection between two current-carrying conductors. This can happen due to: cyclical thermal stress, environmental effects, or damage during installation or maintenance, etc., [Zhao 2015a]. These two faults cause the output power of the photovoltaic module to drop significantly; however, the operating voltage of the PV module remains almost identical.

PID (Electrical): The PID fault is a fault that mainly affects modules made of silicon and thin layers where eddy currents are generated due to the lack or degradation of the ground connection. There are 3 types of PID faults (PIDc, PID-d and PID-s). The PID-s type is the most frequently observed. It can even lead to total fault of the PV module. The PID-s is mainly due to the migration of Na ions. Na ions are derived from the anti reflective coating under negative bias conditions. These ions penetrate crystal faults and result in large cell deflection and degrade efficiency. It should be noted that PID is more common for PV modules with EVA (ethylene vinyl acetate) as the encapsulating material.

With a PID-resistant material, such as polyolefin, this fault has almost dis-

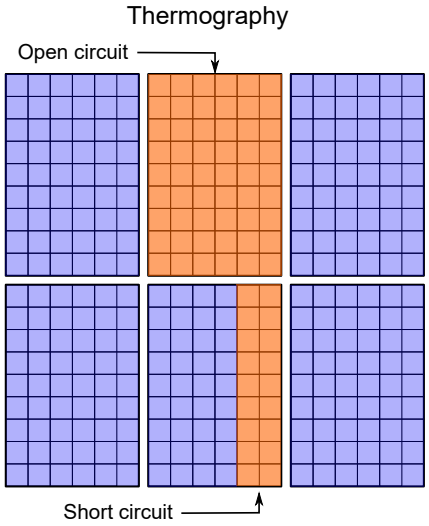

Figure 2.30: Example of Short circuit (SC) / Open circuit (OC)

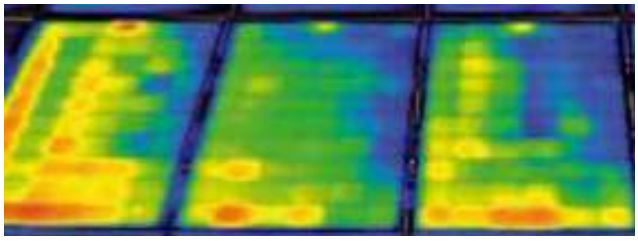

Figure 2.31: Example of PID [Köntges 2014b]

appeared [López-Escalante 2016]. This fault deteriorates the fill factor (ff) of the characteristic I(V). This deformation can be modeled by a degradation of the internal parameters of the cell, decrease in resistance Rshunt (Rsh) and increase in the ideality factor of the diode(s). It occurs primarily at negative voltage with respect to ground potential and is accelerated by environmental conditions, system configuration, module design parameters, high system voltages, high temperatures, high humidity [Pingel 2010] and even after meltdown of the anti-reflective (AR) layer and the corrosion of the conductive layer of the cell [Köntges 2014b].

2.3.2.4 Array-level faults

For optimal energy performance, PV modules are interconnected to form a PV array [Deshkar 2015, Rani 2013]. However, this is where connection-related problems such as ground fault, line fault and arc fault occur. The detailed presentation of these faults is given below.

Ground fault (GF) (Connections): This fault is caused by an unintentional low impedance path between one or more current carrying conductors (CCC) and the established ground connection [Chen 2018a, Zhao 2011a, Appiah 2019a, Zhao 2013a, Zhao 2011b]. For a grounded PV system, this fault causes a high cur-

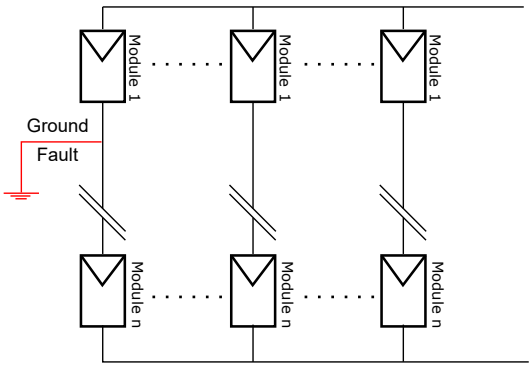

Figure 2.32: Example of Ground fault (GF)

rent to flow through an intentional path [Alam 2015a]. For an ungrounded system, this fault causes a residual magnetic field to be generated between the forward and reverse current flow [Alam 2015a]. In both scenarios this fault causes a change in insulation resistance and a lasting loss of power.

As a consequence of this fault, subsequent fault currents, output voltage disturbance or drop, and sudden changes in the I(V) characteristics of the PV array are observed [AbdulMawjood 2018b]. This type of fault can result in other types of hazards such as electric shock and fire hazards [Falvo 2015, Zhao 2015a]. As a mitigation measure, it is common for the metal parts of the PV array to be grounded using grounding conductors (EGC) [Pillai 2018a, Bower 1994].

Among the most known causes of this fault are the incidental short circuit between the normal conductor and ground, cable insulation fault, GF inside the PV module, a fault in the cable insulation due to manufacturing faults, overheating or aged cables [Zhao 2015a, Flicker 2015]. Other causes of this fault are discussed in [Forman 1982, Zhao 2012]. Detecting this type of fault in ungrounded systems is really challenging because such ground faults do not provide enough leakage currents for detection and localisation during system operation [Karmacharya 2018]. Due to the severity of these faults, most PV systems are equipped with ground fault detection and fault current interruption [Alam 2013b].

Line to line fault (LLF) (Connections): The LLF type fault arises from an unintentional low resistance path between two current carrying conductors (CCC) with different electrical potentials [Zhao 2011a, Zhao 2015a, Yi 2017d, Zhao 2013b, Zhao 2015b]. It usually occurs due to poor insulation of the array connectors or cables, an accidental short between the CCCs, mounting fault, or external damage [Mellit 2018b]. This fault leads to a high reverse current (depending on the potential difference of the location where the LLF occurs) flowing down the faulty path and generating a loss of power. There are two types of LLF [Pillai 2018a, Cotterell 2012]: intra-strand and cross-strand. This type of fault can occur between modules belonging to the same array or between two adjacent arrays (bridge fault).

Also, a line-to-line fault can occur between array cables of different potential, without involving any grounded point. The LLF fault is expressed as a reduc-

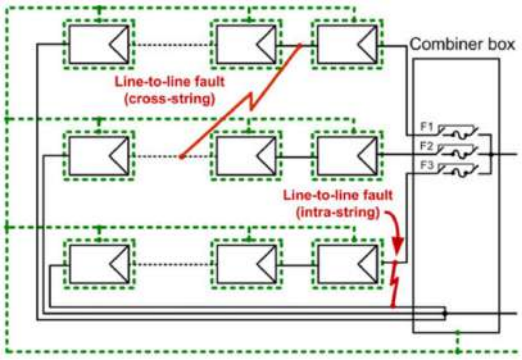

Figure 2.33: Example of Line to line fault (LLF) [Alam 2015a]

tion in open circuit voltage, but the short circuit current may remain the same. This change in voltage modifies the behavior on the I(V) characteristic curve of the photovoltaic array. Detection of this fault is also very difficult and they are often misunderstood as short-circuit faults in grounded PV systems, since the fault current is determined by the voltage differential between two fault points [Pillai 2018a]. This difficulty makes this fault go unnoticed and represents significant losses [Zhao 2012, Alam 2013b].

Arc fault (AF) (Connections): Multiple external factors could lead to discontinuity or insulation fault of current carrying conductors (CCC) and establish an air path for an arc fault [Spooner 2008, Xia 2015, Chen 2018a, Johnson 2012c]. There are two types of arc faults (AF). Series AFs are typically caused by weld separation, connection corrosion, cell damage, rodent damage, or abrasion from numerous sources. Parallel AFs (intra-array, cross-array, and parallel to ground) result from insulation faults in current-carrying conductors [Spooner 2008, Xia 2015].

In rare cases, parallel AF can also occur between two points in the same array, as well as between ground and a point on any of the PV array's current-carrying conductors. In general AF faults can occur at almost any connection point or structure in the PV array. When this fault occurs, an extremely high transient temperature is generated that can burn the metallic coating of the modules. In addition, it generates high-frequency components that cause serious non-linear distortions in current and voltage of the array or multiple arrays, and sudden drops in output current and voltage [Alam 2015a]. These types of faults have a high probability of producing serious fire threats and security risks [Johnson 2012c, Johnson 2012a, AbdulMawjood 2018b, Alam 2013b].

Two methods are conventionally used to detect this type of fault. The first is based on the average value of the DC current in a conductor. This method adds a small impedance in series with the circuit and measures the resulting voltage. The second method is based on the measured value of AC current in a conductor, this approach is relatively easy, due to the oscillatory nature of an AC current, a transformer can be used as the sensing element. More details of both approaches are presented in [McCalmont 2013].

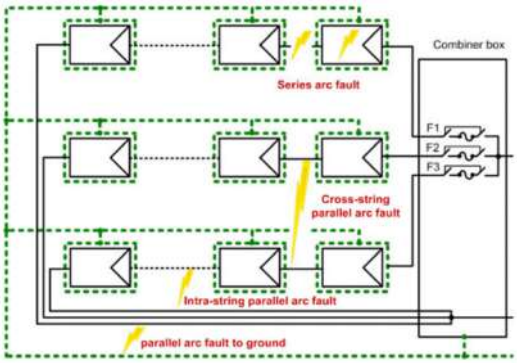

Figure 2.34: Example of Arc fault (AF) [Alam 2015a]

2.3.2.5 Protection-level faults

At the level of the PV array protection system, there are faults linked to the diodes and to the balance elements of the system such as fuses, ground detectors, etc. Generally, the faults present in these elements are of a structural type. Detailed presentation of these faults is given below.

Diode fault (Electrical): Diode faults are divided into two types: By-pass diode and blocking diode faults. By-pass diode (BPD) fault is the most common and is caused by excessive current level and inadequate or insufficient heat sink. The lack of airflow in the junction box is also crucial for diode fault, particularly in the case of fast shade-sun-shade transitions [Kato 2015]. When a BPD diode is burnt, it can cause a short circuit or an open circuit of the diode reducing the power produced by the PV module.

Also, it is not possible to prevent reverse bias heating of the solar cells during shading conditions, resulting in hot spots, discoloration, burn marks and in worst case fires. This type of fault can be easily detected in the I(V) and P(V) characteristic curves of a photovoltaic module since the open circuit voltage and the maximum power of the set drop significantly [AbdulMawjood 2018b]. Faults related to blocking diodes (BKD) do not allow to protect the system against reverse current [Zhao 2011a]. As with the By-pass diode, the electrical faults associated with these blocking diodes are: diode short circuit and diode open circuit and strongly accelerated when the PV module/array is partially shaded for a long period [Rezgui 2014, Kato 2015]. Known causes of these faults include diode disconnection or reverse mounting of the diode [Köntges 2014b].

It is interesting to mention that when the temperature rises in the diodes, it means that diodes are working correctly [Schirripa Spagnolo 2012]. Due to its function as a protection system, it is vital to diagnose faults in the diodes of the PV system, considering that these faults increase in environments with hot and humid climates [Duman 2021]

Balance of system (BOS) (Structure): BOS component faults are considered the main reason behind the existence of non-producing modules in the PV field. A BOS component fault can lead to reduced production. BOS elements in-

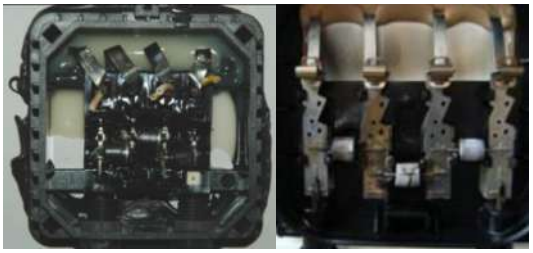

Figure 2.35: Example of Diode fault [Chang 2015, Köntges 2018]

Figure 2.36: Example of Balance of system (BOS) [Flicker 2016]

clude cables/wires, switches, enclosures, fuses, ground fault detectors, fuses, etc., which can put an entire PV array out of order [Cristaldi 2015].

2.3.2.6 Wiring and junction box-level faults

Faults of this type directly affect the connections between the photovoltaic modules (single-pole cables) and the junction box. As possible causes, structural insulation problems are reported that end up generating a ground fault or short circuit or electric discharge [Verhoeven 1998]. A detailed presentation of these faults is given below.

Junction box fault (Connections): The faults observed in the junction box (JB) are usually caused by poor fixing, faulty electrical wiring, broken connection, electrical power overload, repair of the cable during installation, repair of the connector, prolonged exposure to heat or poor installation practices of the connector. In a photovoltaic system, humidity or internal arcing between contacts can cause wear and/or melting of the solder on the junction box (JB) or PV array connections [Chang 2015, Solórzano 2013, Köntges 2014b, Schirripa Spagnolo 2012]. Likewise, the fretting corrosion that occurs in the JB can lead to a rapid increase in the contact resistance [Mellit 2018b]. This would ultimately damage the modules, the array, and even stop the production of the PV system. Some actions and suggestions to avoid JB reliability risk are given in [Chang 2015]. In [Sánchez-Friera 2011], a study is presented where all crystalline silicon photovoltaic modules tested after 12 years of operation in southern Europe had junction box faults.

Ribbon and Solder Bonds Degradation and Broken Interconnect (R and SB) (Connections): Degradation of broken solder bonds and tapes and

Figure 2.37: Example of Junction box fault [Köntges 2014b]

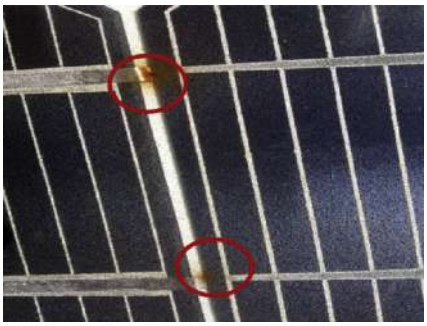

Figure 2.38: Example of Ribbon and Solder Bonds Degradation and Broken Interconnect (R and SB) [Rajput 2019]

interconnections is caused by continual thermal cycling that creates continual expansion and contraction of solder bonds. As a consequence, the solder dissociates further over time, increasing the chance that the tape and solder bonds will crack [Munoz 2011, King 1999]. On the other hand, excessive heating of a part of the cell can cause degradation of the solder joint, and even melting of the solder [Kaushika 2007]. This fault is included in this section considering it as a wiring fault, however, without loss of generality it could also be classified as a PV module fault.

Knowing the classification and description of the most frequent faults in PV systems is vital to be able to design and implement efficient fault detection systems. However, when two or more faults have the same electrical or thermal signature, knowing the frequency of appearance of the faults can add an extra factor for the correct identification of the detected faults. For this reason, this research performs an analysis of the possible frequency of occurrence of faults in PV systems.

2.3.3 Frequency of occurrence of faults in the PV system

To guarantee the performance of PV plants, it is essential to design effective prevention and data acquisition strategies. Consequently, to design advanced systems, the first needed information is to know the statistical frequency of occurrence of faults in this type of system or at least an estimation. This factor makes it possible to identify, at a given moment, the type of fault that occurs in the system. The occurrence of faults depends mainly on three factors: *i)* Environmental conditions

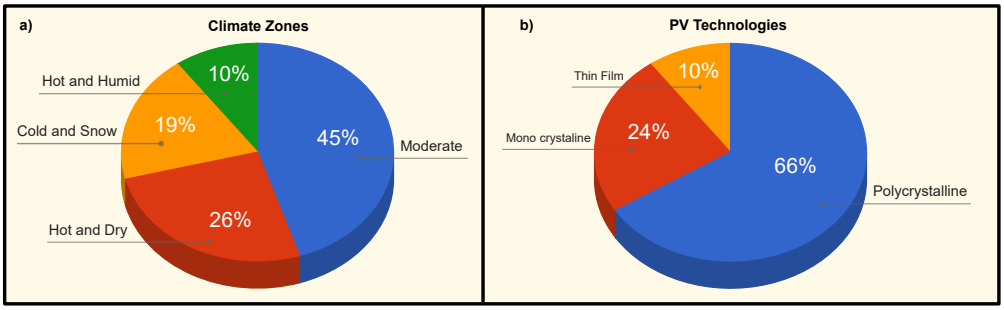

Figure 2.39: Percentages of objects studied in [IEC 2016b]. a) Objects studied according to climate zones. b) Objects studied according to PV technologies.

of the PV plant; *ii)* PV cell technology; and *iii)* Age of the PV plant. A complete study carried out with different weather conditions, and on modules with PV cells of different technologies to identify the possible occurrence of faults over time is carried out by the International Electrotechnical Commission (IEC) over a period of more than 20 years [IEC 2016b]. The proportions of the objects studied in that study are presented in Figure 2.39.

As it can be seen in Figure 2.39, photovoltaic modules from moderate climate zones or Polycrystalline or Multicrystalline (mcSi) silicon technology constitute the majority of the objects surveyed. In the same way as demonstrating the relationship between the appearance of faults and weather conditions, Jordan et al. [Jordan 2017] analyze a set of faults in different weather conditions. The faults analyzed are discoloration, hot spots, By-pass diode faults, cell cracks and PID. Each of these faults is analyzed in moderate, hot and humid, and desert climates on 457, 2.718, and 1.451 panels, respectively. The results of the study are condensed in Table 2.2.

Table 2.2: Report on the rate of faults observed in PV modules according to the climate [Jordan 2017].

Fault	**Percentage of affected PV panels (%)**		
	Moderate	**Hot and humid**	**Desert**
Decoloracion	0	9.9	3.3
Hot spot	11.7	26.1	1.1
Diode Fault	0	21.7	0.1
Cell Cracks	0.5	5	1.7
PID	9.7	1.2	0

As it can be seen in Table 2.2, all faults are severely aggravated in hot and humid climates, for example in the case of the Hot spot doubling its appearance rate and even worse in the case of fault of the diodes where the possibility of appearance goes from 0% in a moderate climate and 0.1% in a desert climate to 20% in a humid and hot climate. Likewise, it is interesting to analyze the frequency of appearance of faults in detail.

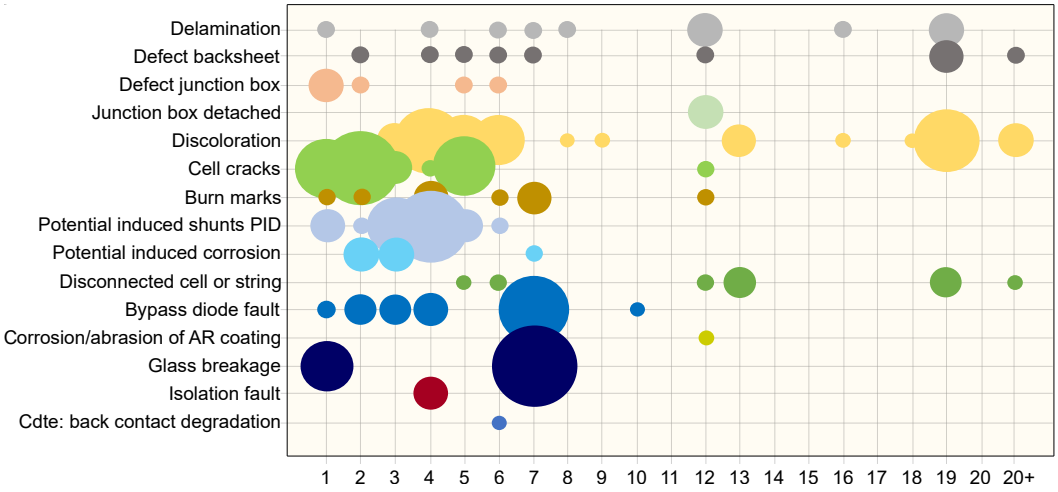

Figure 2.40: Occurrence distribution of degradation faults

For this reason, the frequency of occurrence of faults is analyzed below. The occurrence of the faults presented in the following figures is adapted from the material presented in [Köntges 2017]. Figure 2.40 shows the possible occurrence of all faults due to internal factors such as delamination, bypass diode fault, discoloration, etc. These types of faults are named "degradation faults". The size of the bubbles in the following figures gives an idea of the frequency of faults.

As it can be seen in Figure 2.40 the cell crack usually occurs from the beginning, strongly during the first 5 years and is occasionally detected after 12 years of operation of the photovoltaic array. The PID shunt (PID-s) also occurs from the first years, however, the probability of occurrence increases during the third and fourth years of operation of the PV array. The burn marks due to abnormal and localized increases in temperature (hot spot), occur from the beginning of the operation of the PV array and up to 7 years mainly.

In the same Figure 2.40, it can be seen that the disconnection (for cells or strings) starts from year 5 and covers the entire period of operation (up to 20 years or more). Discoloration begins after year 3 and extends over the years to the end of the PV plant's life, with very heavy accumulation after 18 years. Diode fault can occur during the first 10 years of operation. Likewise, faults linked to the wiring and the box junction must be closely supervised during the first 6 years of operation. It is also important to note that at 12 years there is a wide set of faults, for which it would be an interesting period to pay special attention to the behavior of the PV array.

In a complementary way, Figure 2.41 shows only the occurrence of detected degradation faults that cause measurable energy loss.

As it can be seen in Figure 2.41 the discoloration related to energy loss reaches a high accumulation after 18 years of operation. It is interesting to note from Figures 2.40 and 2.41 that junction box faults, delamination, while frequent, have

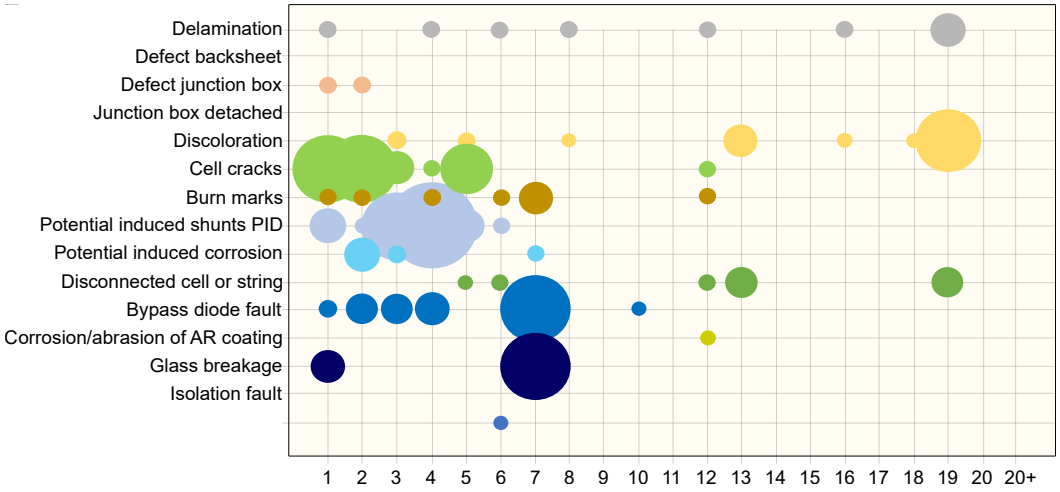

Figure 2.41: Occurrence distribution of degradation faults that cause measurable energy loss

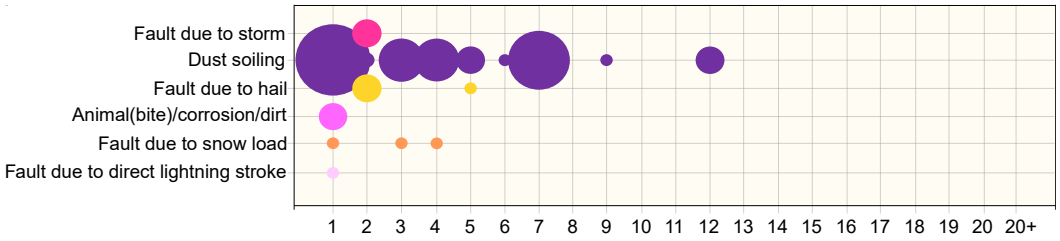

Figure 2.42: Occurrence distribution of sudden faults

negligible impact on power output. Another important fact to note is that the loss of energy due to faults such as delamination, although it occurs from the first year, evolves and takes years to be truly measurable. Additionally, in Figures 2.40 and 2.41 it can be clearly seen how the first 7 years of operation are the most critical, since a wide and severe array of faults occurs. The information condensed in Figures 2.40 - 2.41 allows not only to take preventive measures and identify the possibilities of occurrence, but also to indicate which photovoltaic faults should be prioritized for the different stages of operation of the photovoltaic array.

On the other hand, Figure 2.42 shows the possible occurrence of all faults that occur suddenly due to an external factor such as hail, snow, lightning, etc. These faults are named sudden faults.

As shown in Figure 2.42, sudden photovoltaic faults are more related to climatological causes or factors such as animal attacks. As it can be seen in the same figure, the formation of dust soiling occurs more frequently and extends over several years compared to the other types of faults. The behavior of the other faults in this figure has a very high random component to be able to draw coherent conclusions. In a complementary way, Figure 2.43 shows only the occurrence of sudden faults

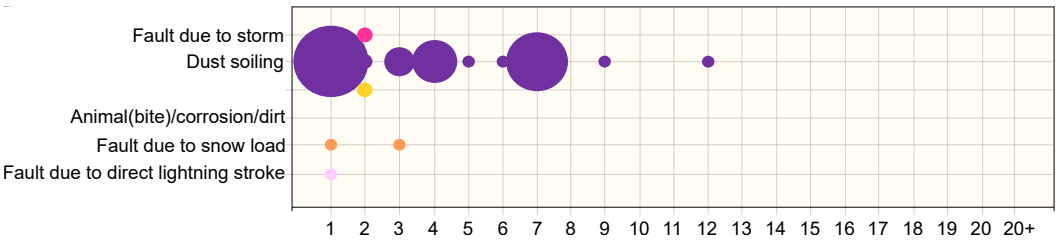

Figure 2.43: Occurrence distribution of sudden faults that cause measurable energy loss

that cause measurable energy loss.

In Figure 2.43, it can be seen that the loss of power due to faults caused by storms, hail or animal attacks is really difficult to measure. All this is because each case would need to be examined individually to determine the level of impact of the fault on the integrity of the PV array.

It is interesting to note that in any of Figures 2.40-2.43, the fault type "snails tracks" or "snails trails" is studied. This is because this fault is generally associated with microcracks in the cells that do not generate a significant loss of power. In the literature it is described that this type of fault occurs after approximately 3 months to 1 year of exposure to the open air of the PV array [Li 2021c]. This fault primarily affects crystalline silicon cells and often occurs at the edges of cells [Fadhel 2018]. In addition, it is very likely that this type of fault is the main cause of more complex and severe faults in PV arrays. Only the aforementioned faults are analyzed since this research considers them to be the most widespread and studied in the literature.

As it can be seen in the study by the International Electrotechnical Commission (IEC) [IEC 2016b], analyzing the impact of faults in terms of power loss is of utmost importance to guarantee the correct operation of the PV plant. For this reason, and after observing the results of Figures 2.40-2.43, in the following section a detailed analysis of the impact of faults in terms of power loss is carried out.

2.3.4 Impact of faults in terms of power loss and human safety

As the previous sections have shown, faults in PV systems can result in both a human safety hazard and a loss of power [Johnson 2012c, Johnson 2012a, AbdulMawjood 2018b, Alam 2013b].

Hazard to human safety refers to risks to personnel working in the facility or to passersby. Kontges et al. [Köntges 2014b] proposes to classify faults into 3 categories according to the risk to human safety: (A) Faults that have no effect on safety, (B) faults that may cause a fire, electrical shock, physical hazard, or a second fault that poses a hazard to personnel working on the PV installation, and finally, (C) faults that cause a direct security problem. The main difference between category (B) and (C) is that in category (B) the PV module should be replaced immediately to avoid a subsequent fault that directly affects the personnel working

on the PV installation, while in category (C), the PV module can sometimes remain in place until replaced.

Likewise, it is important to know the level of energy loss due to the occurrence of a fault. This is vital to develop adequate strategies that establish a priority for preventive or corrective maintenance, and therefore guarantee the optimal performance of the photovoltaic system. However, defining the level of power loss due to a fault in a standard way is one of the biggest challenges in the PV domain. This challenge is due to the fact that the measured degradation, caused by the same type of fault, varies from one photovoltaic installation to the other depending on aspects such as: the severity and propagation of the fault, the weather conditions of the installation site, the operating time of the PV system, photovoltaic module technology, among others.

Different studies have proposed classifications based on the level of power loss of the PV system due to faults. Kontges et al. [Köntges 2014b] proposes a classification of six categories of power loss based on their evolution over time, defined as follows: (A) Power loss below detection limit < 3 %, (B) Power loss degradation exponentially over time, (C) Degradation of power loss linearly over time, (D) Degradation by power loss saturates over time, (E) Degradation in steps over time, and (F) Various types of degradation over time.

Alternatively, Kuitche et al. [Kuitche 2014] proposes a classification of 5 categories with values from 1 to 10. Being, 10 the fault that will cause the non-operation of the system or non-compliance with government regulations, 8-9 the fault will cause the non-functionality of the system, 6-7 the fault will result in the deterioration of part of the system performance, 3-5 the fault results in a slight deterioration of part of the system performance, and finally 1-2 corresponds to faults that do not cause a perceptible effect. This classification is adopted and modified to a discrete scale in the work of Jordan et al. [Jordan 2017] for two reasons. The first reason is that the correlation of specific degradation modes with certain energy losses remains an active field of research. The second reason is that in this way the classification can be used to minimize potential bias and allow better discrimination of the various modes of degradation.

In the same way, as the previously exposed classifications, this work proposes a new classification. It is important to clarify that the information contained in this new classification corresponds to the common tendency of the energy loss of the faults of Section 2.3.2. Each fault is analyzed based on three criteria: *i)* Main consequence; *ii)* Danger to human security; and *iii)* Severity of power loss. These three aspects are vital to evaluate the necessary action to take when the fault occurs in coherence with the recommendations of the standard IEC 61730-1 [IEC 2016b]. The description of the criteria used for the classification is presented in Table 2.3.

Table 2.3: Definition of potential risks. The risks are divided into human safety risks and system power loss risks.

Type	Label	Description
Safety Risk	**L**	Faults that represent a very low or zero risk to humans.
	M	Moderate risk of fire, electric shock, physical hazard or new moderate faults.
	H	High risk of generating fire, electric shock, physical danger or causing new severe faults
Power Loss	**1**	Imperceptible power loss or below detection threshold $< 3\,\%$,
	2	Linear power degradation over time or a slight deterioration in system performance.
	3	Saturated degradation over time with constant power loss after a threshold time. It deteriorates part of the performance of the PV system.
	4	Hybrid degradation over time. Irregular evolution (difficult to establish only a pattern of degradation). The fault causes non-functionality of the PV system.
	5	Exponential power degradation over time. The fault causes the non-operation of the system or non-compliance with government regulations.

The classification presented in Table 2.3 may not be optimal, but it is a contribution to the understanding of the behavior (energy losses and human risk) of the faults reported in Section 2.3.2. The classification is presented in the Table 2.4.

Table 2.4: Summary of the impact of common PV faults. Classification based on the element, the main consequence, the human safety risk and the loss of power.

PV fault			Safety Risk			Power Loss				
Level	**Fault**	**Consequences**	**L**	**M**	**H**	**1**	**2**	**3**	**4**	**5**
Cell	Cell crack	Hotspot generation, corrosion and loss of tightness, deterioration of cells, decrease in shunt, decrease in performance.	⊛	⊛		⊛	⊛	⊛	⊛	⊛
	Discoloration	Premature aging, hot spot.		⊛				⊛	⊛	
	Snail Trails	Localized heating (hot spot).		⊛		⊛	⊛			
	Delamination	Localized heating (hot spot).		⊛				⊛	⊛	⊛
	LID	Formation of recombination active faults.	⊛	⊛	⊛	⊛				
Module	Frame breakage	Delamination and passage of humidity into the cell.		⊛	⊛			⊛		
	Bubbles	Optical reflection that reduces the output power, thermal decomposition, corrosion, reduction in series resistance.		⊛				⊛	⊛	⊛
	BSAL	Exposes active electronic components and causes similar consequences as delamination.		⊛		⊛				

Continued on next page

Table 2.4 – continued from previous page

PV fault			Safety Risk			Power Loss Evolution				
Level	**Fault**	**Consequences**	**L**	**M**	**H**	**1**	**2**	**3**	**4**	**5**
Module	Burn Mark	Power consumption in mismatch area instead of generation, local cell temperature increase and induces burn marks on cell surface.		⊛				⊛		
	Shunt hot spot	It does not generally cause overheating, but it will cause the glass to break and increase the risk of electric shock.		⊛				⊛		
	Dust and Soiling	This fault causes corrosion and hot spots.	⊛	⊛			⊛	⊛		
	Shading	The consequences depend on the level of shading. If it is totally shaded, it generates the same effects as dust and soiling, if it is partially shaded, it directly activates the by-pass diodes and can convert an area of the panel into a passive element that consumes energy.	⊛	⊛			⊛	⊛		
	SC/OC	These types of faults lead to different levels of power loss or system shutdown and increase the risk of electrical shock or even arc flash. Accelerates the aging of the system.		⊛	⊛			⊛		
	PID	It can lead to total fault of the PV system and accelerates corrosion.	⊛				⊛			
Array	GF	Depending on the type of system, this fault can cause a residual magnetic field to be generated between the forwardand reverse current flow or a high current to flow through an intentional path.			⊛			⊛		
	LLF	Fault is expressed as a reduction in open circuit voltage, but the short circuit current may remain the same.			⊛			⊛		

Continued on next page

Table 2.4 – continued from previous page

PV fault			Safety Risk			Power Loss Evolution				
Level	Fault	Consequences	L	M	H	1	2	3	4	5
Array	AF	When this fault occurs, an extremely high transient temperature is generated that can burn the metallic coating of the modules. In addition, it generates high-frequency components that cause serious non-linear distortions in current and voltage of the array or multiple arrays, and sudden drops in output current and voltage.			⊛	⊛				
Protection	Diode	It can cause a short circuit or an open circuit of the diode reducing the power produced by the PV module.	⊛	⊛		⊛		⊛		
	BOS	This fault can put an entire PV system out of operation.		⊛	⊛		⊛	⊛	⊛	⊛
JB and wiring	Junction Box	This fault can generate the disconnection of the PV array, temperature increase (hot spot) or even short circuits or open circuits or arc problems.		⊛		⊛				
	R and SB	As a consequence, the solder dissociates further over time, increasing the chance that the tape and solder bonds will crack.		⊛				⊛		
Temperature	Abnormal Temperature Change (ATC)	It promotes the formation of faults due to the interaction of the materials and premature aging of the components		⊛	⊛	⊛		⊛	⊛	⊛
Material	Corrosion	The corrosion phenomena degrade the state of the PV system and favor the appearance of delamination, and even overheating of the cell.		⊛	⊛	⊛		⊛	⊛	⊛

As it can be seen in Table 2.4, most faults have an impact level M for human safety, that is, a medium security risk of fire, electric shock or physical danger or generate new faults that end in the aforementioned phenomena. However, faults such as open circuit, short circuit, ground fault, line-to-line fault, arc fault, among others, can cause an impact of level H. This impact indicates that there is a high risk for human security.

Regarding power loss, it can be seen that some faults are not categorized into a single category. This is because the level of power loss depends directly on the level of impact and the propagation of the fault. The information condensed in Table 2.4 shows that it is necessary to analyze not only its current impact but also its evolution. Likewise, it is interesting to analyze faults such as Snail Trail/ Snail Tracks that, although they do not generate an instantaneous power drop when they appear, evidence is found that the isolation of the parts of the cracked cells in a PV cell with snail trail, can speed up more than it would in a cell without the snail tracks. Likewise, it is interesting to analyze faults such as Snail Trail/ Snail Tracks that, although they do not generate an instantaneous power drop when they appear, evidence is found that the isolation of cracked cell parts in a PV module that contains snail trail, may be accelerated more than it would be without snail tracks [Kim 2016]. That is, this type of fault can accelerate multiple faults that could significantly reduce the production of a PV module.

During the summer and in hot climates snail tracks/trails seem to occur faster [Kim 2016].

Intuitively, it could be concluded that faults with a significant impact on safety and energy loss are those that must be detected in time as part of monitoring the status of the photovoltaic system. This would make sense if the problem of fault detection is approached from a corrective maintenance point of view. However, behaviors such as the Snail Trail fault is extremely interesting, since detecting this type of fault would allow the loss of power to be anticipated. These types of detections are widely studied in the literature because the electrical signal of a panel with a snail trail is very similar to that of a panel without a fault. Despite this, detecting these faults would really be a challenge oriented to preventive maintenance.

In this same way, to understand the limitations of detection using conventional methods, the most relevant methods of conventional fault detection are presented below.

2.4 Conventional Fault Detection Methods

Fault detection and diagnosis (FDD) is fundamental to guarantee the normal operation of a photovoltaic system at least during its operational life of 25 years without significant production losses. In addition, FDD is vital to avoid human and power loss risks such as those discussed above in section 2.3.4 that can even cause fires [Mellit 2018b]. The main function of these FDD methods is to certify with the greatest certainty that faults have occurred and that the system is no longer working in its optimal operating range. For this, FDD techniques are based on a priori knowledge, estimates and field measurements [Livera 2019a].

In this research we have grouped the conventional methods into five main categories: Visual Methods, Image-Based Methods, Electrical Detection Methods, Protection device Based Technique and ARC Fault Detector (AFD) Techniques. This classification does not take into account machine learning techniques as they are ad-

dressed in detail later. The description of the 5 categories of conventional methods is discussed below.

2.4.1 Visual methods

Visual inspection is a method that can be carried out before or after the commissioning of the PV system [Fadhel 2019b, AbdulMawjood 2018a, Madeti 2017b]. When testing before commissioning each PV module is inspected before and after it is subjected to environmental, electrical or mechanical stress or laboratory stress tests to predict its response and performance [Fadhel 2019b, AbdulMawjood 2018a, Madeti 2017b]. The most common tests performed are thermal cycling, humidity freezing cycling, moist heat exposure, UV radiation, mechanical loading, hail impact, weathering and thermal stress [Köntges 2014b]. The National Renewable Energy Laboratory (NREL/IEA), with the support of the US Department of Energy, has developed a verification protocol for the visual inspection of photovoltaic modules in the field [Packard 2012]. The visual inspection must be carried out at 1000 lux and from different angles to avoid reflections as mentioned in the IEC-61215 standards [IEC 2005b, Madeti 2017b].

Faults reported in the literature as detectable using visual inspection include: discoloration, bubbles, delamination, burn marks, shading, broken glass or cells, dirty, missing or damaged wiring, interconnections, rust or corrosion, snail trails and damaged or broken parts [AbdulMawjood 2018a, Kirchartz 2009]. Despite the large number of faults detected with visual inspection, its main drawbacks are the large-scale cost, the method's direct dependence on human capabilities, long detection times, and the risk exposure of electric shocks to inspectors. [Madeti 2017b]. Due to these conditions, this method is impractical in the case of large installations where it would be necessary to cover kilometers of panels. For this reason, Mellit et al; [Mellit 2018b] affirms that visual inspection is more appropriate for small-scale PV installations where it can be frequent and cost-effective.

Also, causes of reduced performance and certain faults are not always visible. For example, Kato et al. [Kato 2011] conducted a study on 1,272 monocrystalline photovoltaic modules during 4 years of operation to assess the impact of faulty bypass diodes and showed that, in most cases, this type of fault, despite its severity, does not produce traces of visible burns. For this reason more sophisticated techniques are developed.

2.4.2 Image-based methods

The most widely used image-based methods are: infrared/thermal methods, ultrasonic inspection, electroluminescence imaging and the lock in thermography (LIT).

2.4.2.1 Infrared/Thermal methods

Infrared (IR) or thermal imaging is an efficient and systematic diagnosis of solar cell faults [Cubukcu 2020]. To do this, an IR camera is used for scanning

the photovoltaic array while it is in operation. The objective is that the camera measures the temperature differences in the cells and modules. Generally these thermal signature faults are related to: (a) module wiring or interconnect faults, (b) hot spots due to internal short circuits, faulty bypass diodes, change in series resistance value, cell mismatch, snail tracks, and (c) cell cracks [Madeti 2017b, AbdulMawjood 2018a, Cubukcu 2020, Hong 2022b].

According to Madeti et al. [Madeti 2017b], there are two types of IR images: (i) reverse polarization images and (ii) forward polarization images. For the detection of faults with thermal signatures, it is the most suitable method in large photovoltaic plants. A cell-level fault detection and classification study using thermal imaging is described in [Vergura 2015] and [Guerriero 2016]. These studies use an IR camera that records the images for further identification and fault location treatment. In [Vergura 2015], the temperature variation and the average value of each cell are calculated, then the cells are classified according to their average temperature value (detection criteria and classification of the level of affectation of the cells).

In the same way, fault detection works based on thermal gradient analysis, edge recognition for faulty cell detection, photovoltaic module recognition using thermal images to differentiate the temperature between the metal structure and nearby solar cells are proposed [Guerriero 2016]. Cubukcu et al. [Cubukcu 2020] classify and locate faults from heated junctions (fuses/wires/breakers), hot spots, faulty strings, heated junction boxes, and broken modules. In [Hong 2022b], using Infrared/Thermal Methods, the following faults are detected: module degradation, cell cracks/microcracks/snail tracks, cell breakage, loss of cell material, potential induced degradation, shadowing, cells in short circuit/open circuit, defective soldering, system connection fault, and bypass diode fault.

Furthermore, this method is reported in the literature to be able to detect faults in wiring, diodes, junction boxes, connectors, and other [Hong 2022b]. Another advantage is that it can be done without affecting the complete operation of the system and it does not need extra sensor installations since it only uses the IR camera. Finally, this method can be used from small PV arrays to large PV power plants [Cubukcu 2020, Hong 2022b] at a lower cost than other methods [Hong 2022b]. Although this method is very powerful for detecting faults with thermal signatures, it is conditioned to work under controlled conditions. In other words, the correct positioning of the IR camera, the distance from the photovoltaic array, and an overlap between consecutive images must be guaranteed [Guerriero 2016].

2.4.2.2 Ultrasonic inspection

It is one of the main techniques for the detection of cracks, holes and detached lamination structures in PV modules [AbdulMawjood 2018a]. The technique has two inspection methods: the transmission method and the pulse-echo method. A more detailed explanation of these methods is presented in [AbdulMawjood 2018a, Hund 1995]. In general, the first method is capable of locating the faults and their level of impact on the PV system. The second method measures the ultrasonic

pulses reflected by the faults, which provides the fault depth value in addition to the fault size and location information [Hund 1995].

2.4.2.3 Electroluminescence imaging

Electroluminescence imaging is used to detect faults in photovoltaic modules such as fine cracks, and to identify solar cells with different conversion efficiency due to an increase in series resistance of cells and/or reduction in the parallel resistance of cells [Alves dos Reis Benatto 2020]. This method works on the principle of injecting ramp voltage to the module and the electroluminescence of the product reveals non-uniformity and faults [AbdulMawjood 2018a]. That is, the solar cells are powered by a defined external drive current while the camera takes an image of the emitted photons. Damaged areas of a solar module appear dark or radiate less than healthy areas because after a certain voltage level, the glow becomes visible and reveals faults and cracks that reduce cell efficiency. The main problems with this method are its high cost, and the need to take the system offline to do it [Koch 2016, Madeti 2017b]. Among the advantages of the method is that it provides a diagnosis in a short period of time due to its sensitivity [Kirchartz 2009] and that the high resolution of the images allows certain faults to be detected with greater precision than with thermal images, especially in the case of cracks, microcracks and contact faults [Ebner 2010]. However, its implementation is difficult as it requires a high resolution camera and a high pass filter, a set of calibration tests in a dark environment and requires production interruption [Fadhel 2019b]. For these reasons, this method is more practical at the scale of photovoltaic modules than at a large scale.

2.4.2.4 Lock in thermography

Lock in Thermography is a non-contact, non-destructive fault detection and location technique [Bachmann 2012]. This technique uses an excitation device with lock-in capability bundled with a power supply to inject pulse current with different modulation into the PV module. The injected current results in an increase in the temperature of cells with faulty shunts. This increase of temperature is captured using an infrared camera that produces amplitude and phase images, locating the position and indicating the nature of the fault [AbdulMawjood 2018a, Hong 2022b]. This method is usually done with the PV system offline and in dark conditions [AbdulMawjood 2018a]. However, this test can be performed at any time of the day since lighting is not an issue [Hong 2022b]. This method is used mainly to detect solder bond faults [Asadpour 2020]. Because LIT is a non-destructive characterization method, LIT is very effective during pre-characterization to further investigate the origin of physical faults in PV cells [Cao 2020].

2.4.3 Electrical detection methods

An increasing number of diagnosis methods for fault detection in PV systems are based on the analysis of electrical parameters. These methods require direct electrical, irradiation and meteorological measurements and also make use of $I(V)$ characteristic curve data, signal generators, circuit and simulation model, among others. In this section, only electrical methods are reviewed and discussed. Electrical methods can be classified into four groups, discussed below.

2.4.3.1 Climatic data Independent

Climatic Data Independent (CDI) is a group that includes detection techniques that do not involve climatic data such as solar radiation, temperature, humidity, and wind speed. Among the most recognized techniques are the time domain reflectometry (TDR), and the earth capacitance measurement (ECM). Both techniques use LCR meters (to measure the parameters of the photovoltaic circuit: inductance, capacitance and resistance) and signal generation through injection [AbdulMawjood 2018a, Hong 2022b]. TDR is used for discontinuity, impedance change or fault measurements [Schirone 1994, Lu 2021b, Takashima 2006]. In the TDR method, the waveform changes, then the delay between the injected signal and the reflected signal is used to detect the existing degradation. It is recommended to perform TDR periodically to inspect for degradation of the PV array [AbdulMawjood 2018a].

Other studies have used the ECM method. In [Takashima 2006, Takashima 2009] it is used to estimate the disconnection position in the photovoltaic string without the effect of irradiance change. According to [AbdulMawjood 2018a], in the ECM method, the position of the disconnected module (n) in string of M modules can be estimated by:

$$n = \frac{C_x}{C_D} * M, \tag{2.4}$$

where, C_D represents normal string capacitance and C_x represents capacitance of the faulty string.

2.4.3.2 I(V) characteristics analysis

The current-voltage I(V) characteristics of the PV module or PV string can be used to evaluate the performance/health of a PV system. When the PV system is healthy, the characteristics have a particular pattern, which changes during a fault. Depending on the level and pattern of change, it is possible to estimate the type and severity of a fault in a PV system [AbdulMawjood 2018a]. To increase the accuracy of this method, it is recommended before starting the diagnosis to check the other electrical elements such as the junction box, protection systems, etc., in order to identify in advance possible faults such as disconnection or degradation that can reduce output power or alter the I(V) curve of the string [Lin 2017]. This

is the best known fault detection approach. This approach measures the healthy current-voltage I(V) characteristics of the PV system and compares them to the faulty current-voltage I(V) characteristics [Hong 2022b, Hu 2015]. This comparison approach is used to detect faults such as ground faults, short circuits, faulty connections, partial shading, series resistance losses and potential-induced degradation [Stellbogen 1993, Fadhel 2019a, Spataru 2015].

In order to improve fault diagnosis, multiple features extracted from I(V) characteristic measurements are proposed. In [Miwa 2006a, Miwa 2006b], the authors calculate the feature $(-dI/dV)-V$ for partial shadow detection based on the peaks found. Other authors propose fault detection based on the analysis of the shunt and series resistances (Rsh and Rs) and the fill factor (FF) that are determined from the I(V) characteristics [Kaplanis 2011]. The fill factor (FF) is defined as follows:

$$FF = \frac{P_{max}}{I_{sc}V_{oc}}, \tag{2.5}$$

Where P_{max} is the maximum power, I_{sc} is the short circuit current and V_{oc} is the open circuit voltage. In [Bressan 2016] it is shown that the calculation of the first and second derivatives of the I(V) characteristic allows us to identify the number of active bypass diodes to detect the shading fault. This same fault is detected by comparing the I(V) characteristics in normal mode and in fault mode in the work of [El Basri 2015].

In [Chao 2008] the I(V) features are combined with another PV parameter under observation. [Daliento 2016] uses a correlation function and a matter element model. That method, instead of using only the I(V) characteristics of the photovoltaic module, takes its first and second derivatives to detect By-pass diodes and series resistance faults. The I(V) characteristics analysis (IVCA) method can detect a wide number of faults. For example, in [Fezzani 2015], 12 types of faults are classified among which are shorted bypass faults, reversed bypass, shorted one cell, shorted module, reversed module, modules connected with resistance, module shade, shade with By-pass open, shade with Inverter By-pass, shade with By-pass in shunt and module shade with resistor in series. Five types of anomalies are identified using IVCA in [Chine 2015]: short-circuit current reduction, open-circuit voltage reduction, change in output current, a change in output voltage, and a change in the number of I(V) spikes. After expert analysis, the authors in [Chine 2015] found that the anomalies are due to 12 faults. In [Hachana 2016] they combined a meta-heuristic technique named artificial bee colony to analyze P(V) and I(V) curves and detect full and partial shading, By-pass diode faults, connection faults and a shorted substring fault. Other works have detected cracks in the photovoltaic surface using IVCA [Wang 2016a]. Likewise, some works have explored the use of the Euclidean norm between the I(V) characteristics of a normal PV array and the faulty PV array to detect interconnection resistance and various PV shading conditions [Ali 2017].

Due to its simplicity, IVCA is extensively studied in different publications, however, it is a difficult method to implement without cutting the production of the

PV system.

2.4.3.3 Power Loss Analysis Technique

This technique is based on the analysis of the energy losses of the photovoltaic system. Generally, in this method, PV system parameters are calculated from data acquisition together with climate information data [Hong 2022b]. All this information is then used to simulate the behavior of the photovoltaic system in real time. Once the simulation is complete, data monitored under real working conditions on the DC side of the system is compared to simulated results to detect system power losses and classify faults. In [AbdulMawjood 2018a], two indicators are proposed to identify the variation of the DC variables with respect to their simulated values. These indicators are the current and voltage ratios as follows:

$$RC_{pv} = \frac{I_{real}}{I_{simulated}} \tag{2.6}$$

$$RV_{pv} = \frac{V_{real}}{V_{simulated}} \tag{2.7}$$

Some of the approaches using this method have been able to detect faults such as faulty module or string, partial shadowing, aging and MPPT fault [AbdulMawjood 2018a]. Another approach proposes fault detection comparing simulated and measured variables and classifying detected faults by comparing current and voltage values with a set of error thresholds [Silvestre 2013]. Using this same Power Loss Analysis (PLA) Technique, [Solórzano 2013] can detect faults such as shadowing, hot spot, module degradation and power loss due to wiring problems. In general, this method is tested multiple times showing interesting results [Stauffer 2015, Shimakage 2011, Dhimish 2016]. However, as mentioned in [AbdulMawjood 2018a], although this detection method is easy, it has many difficulties to correctly detect and classify faults, generating false alarms, when unpredictable changes in irradiance or other weather conditions occur.

2.4.3.4 Comparison between Measured and Modeled PV System Outputs Technique

In the same way as the PLA method, the current and voltage measurement method is proposed: the main difference with the PLA method is that simulation models are used to predict the power outputs of photovoltaic systems for later comparison [Drews 2007], not to compare in real time as in the case of the PLA. Generally PV cell parameters are defined according to an electrical analogy like those presented in [McEvoy 2013, AbdulMawjood 2018a]. These parameters are adjusted using real world data and the output of the PV array is predicted based on the fitted model. In order to further fit the model, weather conditions or panel surface temperature can be added [Chouder 2010]. Alternatively, some authors have proposed models developed in PSIM to create the PV model instead of using a conventional mathe-

matical model and proposed an extended correlation function for fault identification [Chao 2008].

2.4.4 Protection Device Based Technique

As a protection measure against risks to human safety, photovoltaic systems are equipped at the string level with devices that instantly break the electrical circuit to stop faults in progress. For example, ground fault detection and interrupting (GFDI) devices are installed in PV systems to isolate faults that manifest by creating a path between a current-carrying conductor and earth named ground faults [Hernández 2009, Flicker 2013]. To isolate line-to-line (L-L) and line-to-ground (L-G) faults, residual current devices (RCD) are installed [AbdulMawjood 2018a]. This device can detect the difference in current passing through string terminals or array output terminals, and thus react to isolate the element.

2.4.5 ARC Fault Detector Techniques

The arc-fault detector or arc-fault circuit interrupter (AFCI) is required due to NEC regulation 690.11 [AbdulMawjood 2018a]. These devices have two components: *i)* arc fault detectors (AFDs); and *ii)* interrupt devices (IDs). For correct operation, care must be taken in the location and experimental verification of the devices [Johnson 2012b]. Most manufacturers use AC noise to determine arc flash on the DC side of PV. Alternatively, in [Johnson 2011a], the authors investigate the frequencies that appear in the voltage and current on the DC side of PV systems that can be used for arc fault detection. As mentioned in [AbdulMawjood 2018a], when performing an analysis based on frequency, to avoid false detection results, it is necessary to take into account considerations such as: the signals generated during partial shade and the level of irradiance that have frequencies below 1000 Hz, frequencies above 100 kHz are low energy and include various effects from nearby antennas and RF phenomena or switching frequencies of most inverters, DC/DC converters and power conditioners are in the range from 10kHz to 50kHz. In the same manner, the test study at Sandia National Laboratory (SNL) [Johnson 2011b] shows that the best frequency band to detect noise from arc faults is 1-100 kHz, and it is not recommended to select a single frequency band for fault detection [Johnson 2012b].

2.5 Discussion and Conclusions

The wide set of conditions and scenarios in which a fault can occur means that the choice of the fault detection method depends on the available knowledge about the system (system components), sometimes its size (impact level of the fault), its characteristics (electrical, thermal characteristics, etc.), its origin and the type of fault to diagnose. Because of this, it is evident that challenges around fault detection in the photovoltaic system still persist.

Furthermore, fault detection conditions in PV systems become more difficult when taking into account the influence of meteorological conditions [Yi 2017c] and its continuous changes in conditions, nonlinear system outputs [Fadhel 2018], the presence or not of maximum power point tracking (MPPT) devices [Zhao 2013a], the occurrence of multiple simultaneous faults or primary faults that generate more severe secondary faults. In fact, there are also huge scenarios in this field where the electrical behavior of panels with faults is very similar to that of panels without faults [Hariharan 2016a, Sepúlveda Oviedo 2022]. Taking all this into consideration, it is evident that the approaches presented in this chapter reach their own limits when the aforementioned conditions are present or when large amounts of high-dimensional data (Big data) are introduced in the new data acquisition systems.

Therefore, this research considers that fault detection strategies in PV systems can be improved to obtain more efficient systems with predictive fault detection functions and be applied to a wide range of photovoltaic plants with various data acquisition systems. In addition, this research pays special importance to the detection of faults with signatures similar to those of healthy panels, taking into account the multilevel classification presented in this chapter and the relationships between faults in the cause-effect circle. This is primarily important, since detecting a fault that is the cause of multiple severe faults would greatly increase the level of accuracy of fault detection systems.

This thesis considers that improvements can be strongly supported by artificial intelligence approaches due to the ability of artificial intelligence (AI) to handle high-dimensional multivariate data and extract hidden relationships within data in complex and dynamic environments [Wuest 2016]. Additionally, new approaches must be able to detect faults not only when the impact is severe but even from the beginning or when their signatures are similar to those of a panel without fault, as in the case of the Snail Trail. These approaches are vital to avoid destructive consequences and hazards to personnel who come into contact with the PV system. Taking into account the aforementioned aspects, fault diagnosis in photovoltaic systems is discussed in detail below in Chapter 2.

Chapter 3

Fault Diagnosis in Photovoltaic Systems using Artificial Intelligence

Contents

3.1 Description of Methodology (Smart B2TE) 92
3.1.1 Document recovery 93
3.1.2 Bibliometric analysis 95
3.1.3 Topic Modeling 98
3.1.4 T-distributed stochastic neighbor embedding 103
3.1.5 Expert qualitative content analysis 106
3.1.6 Summary of the relevant information retrieved. 109
3.2 Methods based on artificial intelligence techniques 110
3.2.1 Supervised Learning 112
3.2.2 Semi-Supervised Learning 120
3.2.3 Reinforcement Learning 122
3.2.4 Unsupervised Learning 123
3.3 Promising Research Topics and Challenges 125
3.4 Discussion and conclusions 128

As discussed in Chapter 2, there are multiple faults susceptible to occur in a PV plant. For many decades faults have been detected using conventional methods presented in Section 2.4. However, these methods are focused on some faults, a major part of faults are not previously detected before high power impact. With the development of monitoring systems, the amount of data obtained from the instrumentation of a PV plant highlights the limitations of conventional diagnosis methods with the discovery of new faults and new detection methodologies. As an alternative to overcome these issues, the use of artificial intelligence was recently proposed. Artificial intelligence (AI) can handle high-dimensional multivariate data and extract hidden relationships within data in complex and dynamic environments [Wuest 2016]. That is why this thesis dedicates an entire chapter to make a comprehensive state of the art that exposes the identification of methods used, promising emerging themes and current limitations. This chapter serves as the basis for positioning the contributions of this thesis in terms of artificial intelligence.

For the construction of the state of the art, this chapter proposes an innovative methodology made up of different existing methods. Precisely, we choose to use firstly a statistical method named **B**ibliometric used in several other domains. This method is improved working collaboratively with a second hybrid method based on two machine learning algorithms: i) **T**opic modeling; and ii) **T**-distributed Stochastic Neighbor **E**mbedding t-SNE, hence the name given to this **smart** methodology (Smart B2TE). As a result of these two methods, a set of clusters are obtained, which are then combined and subjected to the third method named Expert Qualitative Content Analysis guided by experts. This analysis allows extracting relevant information such as: the identification of machine learning methods used, promising emerging issues, current limitations, among others. This methodology can be extrapolated to other domains, for reducing the subjectivity existing in conventional reviews and positions the reader at the forefront of understanding aspects of effective fault diagnosis in photovoltaic systems using artificial intelligence. It is important to mention that a conventional review of the state of the art would have been extremely difficult to complete on this volume of documents.

3.1 Description of Methodology (Smart B2TE)

In recent years, multiple review works have tried to concentrate as much information as possible on the different fault diagnosis methods in PV systems [Hong 2022b, Berghout 2021b, Pillai 2019b, Massa 2021, Ramírez 2021, . 2019, Abubakar 2021, Li 2021a, Mellit 2018b, Pillai 2018a, Triki-Lahiani 2018a, Livera 2019a, Madeti 2017c, Navid 2021b, Afrasiabi 2022b]. However, these reviews use conventional state-of-the-art review methodologies such as structured review, model/framework review, meta-analysis, theoretical examination and systematic reviews, among others. All these methodologies are limited by the human capacity analysis of numerous documents. Furthermore, it is difficult to provide an accurate and unbiased overview of the conceptual and intellectual framework of a scientific field.

This smart B2TE methodology is a useful tool to reduce problems of subjectivity presented in reviews done on a conventional state of the art. The contribution of this chapter is precisely to solve the problem discussed above. For this, the new hybrid methodology, combining Bibliometric analysis, Topic modeling, t-SNE and an Expert qualitative content analysis, is used to make our state of the art the most complete and objective possible. This proposed study corresponding on different stages is illustrated in Figure 3.1.

As it can be seen in Figure 3.1, the methodology consists of five steps or stages. First with the **Document recovery** stage, documents from the *Scopus* and *WoS* databases corresponding to studied areas are retrieved. Then, in a second time, **Bibliometric analysis** stage processes these documents to extract relevant information, such as keywords, among others. In parallel, **Topic modeling** stage processes the documents as a single corpus using the abstract, introduction, title,

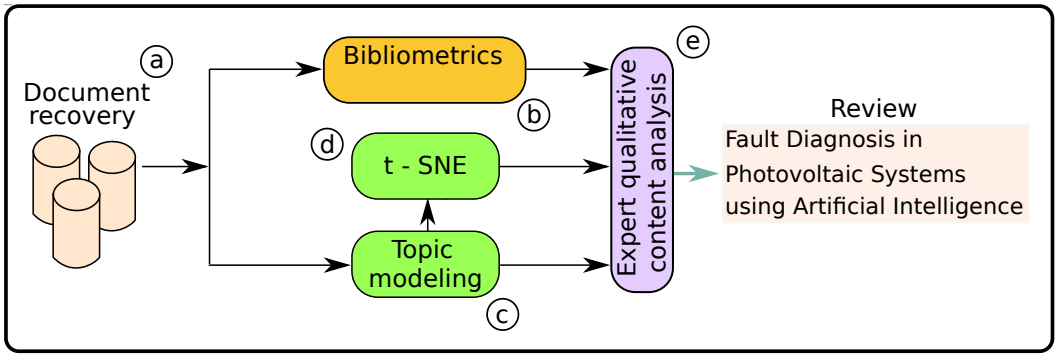

Figure 3.1: 5-stage methodology. a) Document recovery, ; b) Bibliometrics analysis; c) Topic modeling; d) T-SNE and e) Expert qualitative content analysis used for construction of the global review on Fault Diagnosis in Photovoltaic Systems using Artificial Intelligence.

keywords, and conclusion sections. After this stage, **Topic modeling** is carried out using a machine learning method for text mining called the Latent Dirichlet Allocation (LDA) method . The results are used as input to the new stage named **t-SNE** stage. The aim of the **t-SNE** stage is correct visualization of the distribution of the documents within the topics. The last stage needs the results of Bibliometric analysis, **Topic modeling** and **t-SNE** stages. It is named **Expert qualitative content analysis** stage.

This Expert qualitative content analysis is the stage constituting the general review on the use of artificial intelligence for diagnosis of faults in photovoltaic systems, and provides information on the main artificial intelligence methods used in fault diagnosis in PV plants. Also, this stage provides information about **Promising Research Topics**, which exposes new research trends in fault diagnosis using artificial intelligence. In turn, these results of promising research topics can be used to once again apply the same proposed methodology. On the other hand, this stage also gives an idea about **Challenges**. It expresses the hot spots that block the development of research in this area and that are strongly linked to promising topics. Each stage of the proposed methodology and its type of results is detailed below.

3.1.1 Document recovery

A data corpus with sufficient accuracy and robustness is necessary to guarantee reliable conclusions when it is used in construction of a large range of state of the art [Lim 2021, Wang 2021a]. First a search equation is built based on keywords with the Boolean structure presented in Figure 3.2. Building a search equation using keywords is cataloged as the best way to start a coherent and consistent systematic search [Almeida 2018].

The set of documents used in this article are retrieved with the search equation on July 16, 2022 from the Scopus and WoS databases widely used for state-of-

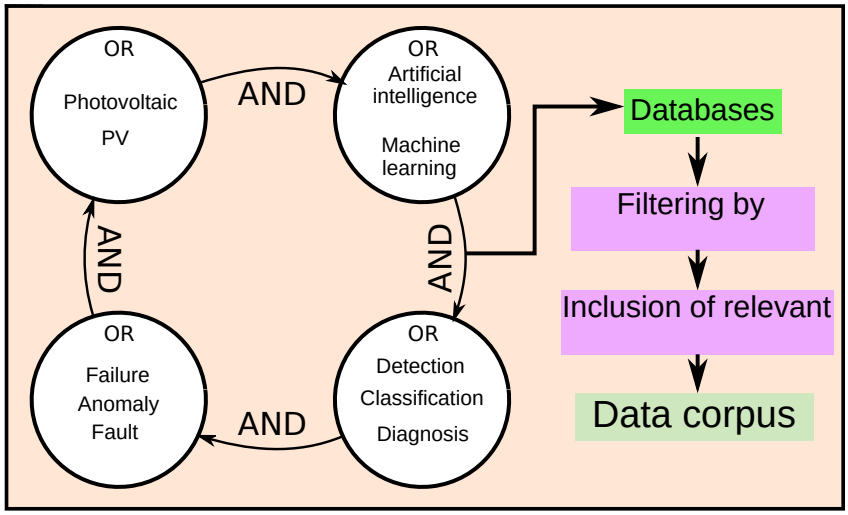

Figure 3.2: Steps of the document recovery stage for the construction of the data corpus: i) To build the search equation using keywords and logical operators; ii) To use the search equation to retrieve the documents in the Scopus and WoS databases; iii) To filter documents by selection criteria; iv) To extend the number of retrieved and filtered documents, adding filtered documents from other sources; and v) Construction of the data corpus.

the-art review studies [Shen 2021, Zhao 2018]. 266 records of Scopus and 206 of WoS are obtained. These results are merged and duplicates are removed using Zotero software [Rakshikar 2015]. In the Scopus and WoS databases some of the documents retrieved are written in languages other than English, or have thesis-type documents. With the aim of building a data corpus homogeneous, a series of filter criteria are applied to the retrieved documents as seen in Figure 3.2. The list of inclusion criteria is: *i)* Article Published in peer Reviewed Journal; *ii)* English Language; *iii)* Period 2010 - 2022; *iv)* Conference Papers; and *v)* Book chapter. After applying the filter criteria, a filtered set of documents from other databases and provided by experts are added to the filtered documents to complete a total of 625 peer-reviewed documents that make up the final data corpus used for this study. Figure 3.3 shows the main statistics of the 625 peer-reviewed documents.

As can be seen in Figure 3.3, research in fault diagnosis in photovoltaic systems using artificial intelligence has grown greatly since 2015, increasing the number of publications by 3.75 times the number of publications in 2010. This increase may be directly related to the clean energy transition, which is of vital importance to mitigate the problems of climate change. Following the trend of recent years, it is possible to conclude that the research of new artificial intelligence algorithms applied to renewable energies and especially to PV energy will continue with an exponential growth during the next decades. This reinforces the idea of continuing to investigate in depth artificial intelligence issues to detect faults in PV systems. On the other hand, another interesting aspect to analyze in Figure 3.3 is the publication category. It is evident that most of the publications belong to doc-

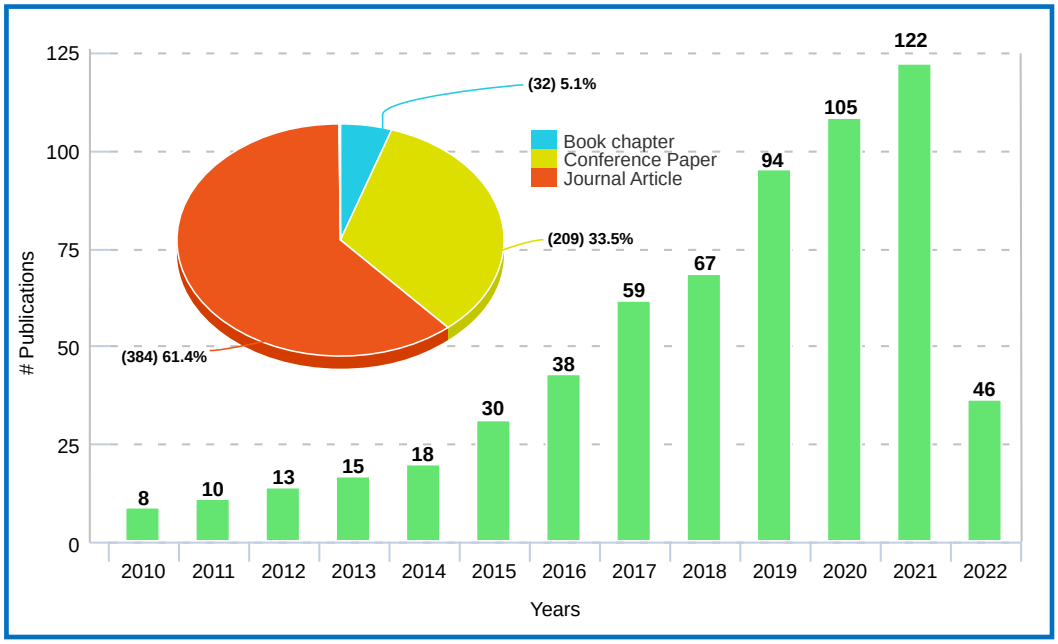

Figure 3.3: Statistics of the retrieved documents. Distribution of retrieved documents by type. Number of leaked documents published per year between 2010 and 2022.

uments in the categories of conference papers and journal articles. Once the data corpus is complete, bibliometric analysis and topic modeling methods are applied. Carrying out the Bibliometric analysis and the topic modeling on this data corpus guarantees a structured, replicable, transparent and iterative study, preserving only the relevant documents and reducing the subjectivity present in traditional reviews [Tranfield 2003]. Bibliometric analysis, topic modeling and **t-SNE** stages are explained below in a generic way and their final results are analyzed in **Expert qualitative content analysis** stage.

3.1.2 Bibliometric analysis

Bibliometric analysis of the information allows an objective and replicable review of the data corpus. This method is applied to the analysis of indicators (authors, countries, citations, keywords, etc.) and allows to measure the quantitative contribution of different aspects within a given area of knowledge [Keiser 2005, Zhang 2010, Bjurström 2011]. This means that Bibliometric facilitates the process of identifying popular topics in present, past or future [Zhou 2007]. Therefore, it is a crucial tool to review the state of the art of a research field from micro level (scientist and institute) to macro level (national and global) [Mao 2015, Calderón 2020].

The results of treatment of big data using Bibliometric follow an established protocol, that is, translate the analysis through nodes, creating links between nodes and networks with less or more great potential. This type of graphic representation is

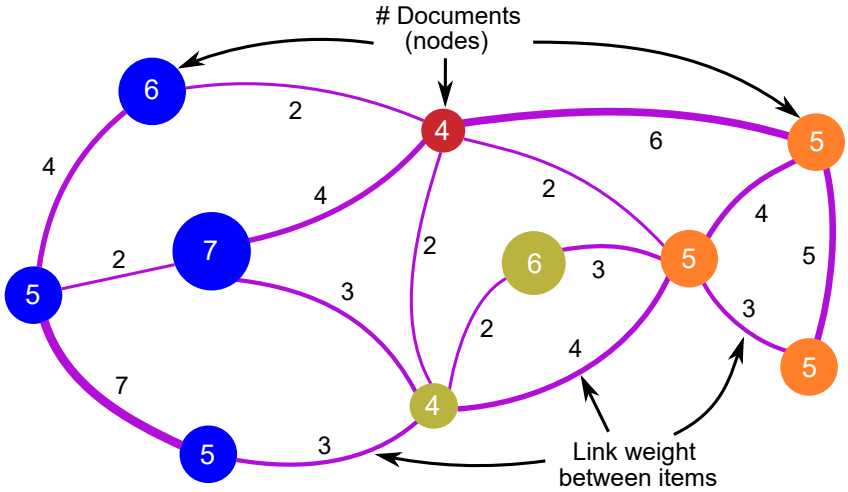

Figure 3.4: Example of a graphical representation Bibliometric network.

especially adapted for engineering and research applications [Reyes-Belmonte 2020]. The principle of nodes and networks used to visualize results of Bibliometric analyses is named map or Bibliometric network. To illustrate how a result of an analysis can be done by a Bibliometric map, an example of scheme is shown in Figure 3.4.

To illustrate what can be a graphical representation of a Bibliometric network, Figure 3.4 is an example of a network of 10 nodes. Each node has its own size represented in the form of circles, proportional to the number of entities with similar attributes noted by the value inside the node. Each color represents a cluster. The thickness of the lines is proportional to the value on the line and spacing (distance) of the circles represent intensity of collaboration. As mentioned in [Du 2014], this type of Bibliometric analysis and visualization has become an indispensable instrument to measure scientific progress in a scientific field in different application domains and intensities of collaborations [Zupic 2015, Al Mamun 2022, Akinlolu 2020]. Specifically, in our case dedicated to the field of solar energy, Bibliometric is widely used in areas such as organic solar cells [Dong 2012, Qadir 2019], energy transition [Zhang 2021d], rooftop photovoltaic fields [Shen 2021], analysis of different Maximum Power Point Tracking methods [Hadke 2021] or more globally [Garg 1993, Du 2014, Dong 2012, Qadir 2019, Azad 2022, Dominković 2022, Zhang 2015, Calderón 2020, Yu 2020, Mao 2018]. In all these studies, the great potential of Bibliometric to identify emerging research fields through statistical analysis of keywords of articles is demonstrated. These bibliometric approaches can determine the main topics in an area of knowledge by analyzing the keywords of the articles and significantly reducing the subjectivity of conventional reviews.

For the realization of Bibliometric maps, distance-based and graphic-based methods can be distinguished. Different distance-based map construction techniques are found in literature. In [van Eck 2010a] which is one of the most popular techniques in the field of Bibliometric, it is proposed to use entities called multidimensional scaling to represent within a geometric space of few dimensions, prox-

imities between a set of objects [Borg 2005]. As an alternative, techniques such as VOS mapping [van Eck 2007b, Van Eck 2007a, van Eck 2006] are also proposed. These last techniques show better results than those presented by multidimensional scaling [van Eck 2010b]. Another technique based on distance called $VxOrd$ is proposed in [Davidson 2001, Klavans 2006], this technique shows a high performance for construction of maps with large numbers of elements (7000 or more). Finally, in [Kopcsa 1998] a new method implemented in a software called BibTechMon is proposed.

Similarly, several works on elaboration of graph-based maps are published. In [Kamada 1989], it is proposed to construct graphs using the Euclidean distance between two nodes or vertices in the drawing as the “graph-theoretical" distance. Then two vertices are connected by a “spring" of the calculated distance. The arrangement of vertices is calculated based on the total elastic energy of the system. As an alternative technique, in [Fruchterman 1991] it is proposed a heuristic aimed at achieving uniform edge lengths. They propose this technique by modifying the spring embedding model using an analogy with forces in natural systems. Other authors [Moya-Anegón 2007, Vargas-Quesada 2007] propose combinations of the work presented by [Kamada 1989] with the pathfinder network technique [Schvaneveldt 1988]. There are also some versions of software like CiteSpace [Chen 2003] specialized in this type of maps, or some like Pajek from [De Nooy 2018] that combine distance-based and graph-based approaches.

For the Bibliometric analysis carried out in the PV research area a distance-based Bibliometric mapping is performed using Bibliometric visualization and mapping software developed by Nees Jan van Eck and Ludo Waltman called VOSviewer[1] due to its interesting distance-based Bibliometric results [van Eck 2010b] and its widely known and used in the field of Bibliometric [Wang 2018b, Wang 2021b, Pan 2019, Bai 2020]. This induces not only an analysis of relationships but also of relationship level between different entities. Construction of Bibliometric maps using VOSviewer is based on the co-occurrence matrix. Co-occurrence matrix analyzes counts of concurrent entities within a collection of units. In the co-occurrence matrix the items (authors, institutions, countries, etc.) form row and column headings and the intersection of the row and column represent the co-occurrence [Zhou 2022].

To carry out the Bibliometric analysis three times steps are needed. In the first step, the similarity matrix based on the co-occurrence matrix is computed. In the second step, a map applying the VOS mapping technique to the similarity matrix is performed. Finally, the map is translated, rotated and reflected to find consistent results.

As mentioned above the similarity matrix is calculated from the normalization of the co-occurrence matrix. This normalization depends on the total number of occurrences or co-occurrences of elements. This means to obtain a correct normalization a similarity metric must be selected. Among the most common normalization metrics are the cosine [Huang 2008] and the Jaccard index [Shtovba 2020].

[1] See Vosviewer

However, VOSviewer uses a measure of similarity known as the strength of association [Van Eck 2007a, van Eck 2006], proximity index [Peters 1993, Rip 1984], or probabilistic affinity index [Zitt 2000] due to the advantages over other measures of similarity [Eck 2009]. In these cases, this index defines the similarity s_{ij} between two elements i and j as:

$$s_{ij} = \frac{c_{ij}}{w_i w_j} \tag{3.1}$$

where c_{ij} corresponds to the number of co-occurrences of items i and j. w_i and w_j denote either the total number of occurrences or the total number of simultaneous occurrences, under the assumption that the occurrences of items i and j are statistically independent.

Once this similarity matrix is built, VOS mapping technique is applied. It is important to mention that VOS technique can have one or more solutions of global optimum. In cases of multiple solutions, it is important to apply translation, rotation, or reflection operations to ensure that VOSviewer produces consistent results. More detailed information about this method is presented in [van Eck 2010a]. For this chapter, the Bibliometric analysis has been carried out only taking into account the keywords. However, the analysis carried out by bibliometric approaches continue to be merely quantitative and are limited to the list of keywords provided by the authors and they do not consider the context in which the keywords are found. For this reason, in this thesis the complementary use of topical modeling is proposed.

3.1.3 Topic Modeling

In an alternative way to analyze keywords in context, some authors propose the use of machine learning tools for text mining such as topic modeling and more specifically such as Latent Dirichlet Allocation (LDA) [Delgosha 2021, Mustak 2021]. These tools are vital to achieve a more accurate and unbiased general understanding of the current state of a scientific field. Moreover, these tools are of great help to discover hidden topics in large amounts of text. [Nielsen 2019, Jiang 2016]. In addition, they allow analyzing not only keywords defined by the author, but also keywords extracted from full texts or fragments. Topic modeling is considered more flexible and efficient than alternative approaches such as document clustering [Kuhn 2018] and is widely used in different areas of knowledge to verify the scientific trajectory [Jelodar 2019, Bastani 2019, Chen 2020b, Jiang 2016]. However, the use of topic modeling for the exploration of research topics in energy has recently started to be explored [Saheb 2022a].

In this study, topic modeling is used to generate so-called "topics" that can be compared in an analogous way with the clusters obtained through bibliometric [Lu 2012, Yau 2014]. This research considers the term "topic" as a fundamental object to describe the intellectual structure of an area of knowledge. Topic modeling is a quantitative statistical-based method that extracts semantic information and

evaluates substantial data from large collections of texts [Jiang 2016]. In addition, topic modeling has proven to be a productive approach to find hidden (semantic) structures in BIG DATA [Jelodar 2019].

The first topic modeling is proposed by Hofman [Hofmann 1999] in 1999 with the model named probabilistic latent semantic indexing (pLSI). In this model, each word in a document is a sample of a mixture model, where the components of the mixture are random variables that can be seen as representation topics. Years later, in [Blei 2003] a method based on the three-layer Bayesian model named Latent Dirichlet Assignment (LDA) is proposed. This method is based on a one-parameter reduced model of the Dirichlet distribution. The LDA method is widely used and extensions are proposed, such as the Correlated Topic Models (CTM) [Blei 2007] or the Hierarchical Dirichlet Process (HDP) [Teh 2006], to reduce computation time and required memory.

There are multiple approaches to perform topic modeling from employing such as Latent Dirichlet Allocation (LDA) [Mustak 2021], Latent Semantic Analysis (LSA) [Foltz 1996] or clustering with the k means algorithm [Tijare 2022]. For the topic modeling carried out in this research, the unsupervised generative probabilistic method LDA is selected. LDA method is widely used in natural language processing, text mining and social network analysis, information retrieval in various fields including medical sciences [Zhang 2017], software engineering [Gethers 2010], political science [Chen 2010, Greene 2015], among others [Blei 2003, Jelodar 2019].

3.1.3.1 Latent Dirichlet Allocation

The LDA approach works on the premise that documents are made up of random mixes of a number of latent topics K. In turn, each topic k_p, $p = 1, \ldots, K$ is characterized by a specific probability distribution of words [Xie 2020, Daud 2010]. The formal description of the LDA approach mentions that given a data corpus D constituted by M documents, where a document m_i, $i = 1, \ldots, M$ contains a number of words or size of the vocabulary N_{m_i}, LDA models D according to the following generative process [Blei 2001]:

- *Choose a multinomial distribution φ_k for a topic k_j, $j = 1, \ldots, K$ from a Dirichlet distribution with parameter β.*
- *Choose a multinomial distribution θ_{m_i}, for document m_i, $i = 1, \ldots, M$ from a Dirichlet distribution with parameter α.*
- *For each of word w_n, $n = 1, \ldots, N_{m_i}$ in a document m_i, $i = 1, \ldots, M$:*
  - Sample a topic $k_{j,n}$ from θ_{m_i}.
  - Sample a word $w_{j,n}$ from φ_k conditioned on the z_n topic selected.

Therefore, the probability of observed data $p(D \mid \alpha, \beta)$ is calculated and obtained from a corpus as follows:

$$a(\theta_{m_i}) = \prod_{m_i=1}^{M} \int_{\theta} p(\theta_{m_i} \mid \alpha),$$

$$b(\theta_{m_i}) = \prod_{n=1}^{N_{m_i}} \sum_{k_{j,n}} p(k_{j,n} \mid \theta_{m_i}) p(w_{j,n} \mid k_{j,n}, \beta),$$

$$p(D \mid \alpha, \beta) = a(\theta_{m_i}) * b(\theta_{m_i}) \; d\theta_{m_i}, \tag{3.2}$$

Where, β and α are the Dirichlet-multinomial pair for topic-word distributions (hyperparameters) previously defined for topic and document respectively. φ and θ are the Dirichlet-multinomial pair for the corpus-level topic distributions. The variables θ_{m_i} are document-level variables, sampled when per document. $k_{j,n}$, $w_{j,n}$ variables are word-level variables and are sampled when for each word in each text-document. This topic modeling method is applied to the 625 retrieved documents. A pre-processing described below is necessary before using the LDA model.

3.1.3.2 Document pre-processing

The processing of the information must be carried out on the corpus of the documents. In this paper, the corpus contains information of the abstract, keywords, title, conclusions and introduction sections of the 625 retrieved documents. The sections are considered because they contain the most succinct summary of the key ideas [Delgosha 2021]. Once this corpus is selected, a 6-stage preprocessing is performed. First, a **Tokenization** that breaks the text into sentences and the sentences into words. Second, all words are lowercase and punctuation is removed. Third, the **stopwords** and emails are removed. Stopwords are those words that you want to filter so that they are not taken into account in natural language processing. The list of stopwords is carefully made by experts in the domain of interest. Fourth, the remaining words are **lemmatized**, turning third person words into first person and past and future verbs into present. Fifth, words are **stemmed**, reducing the words to their root. Finally, **bigrams** and **trigrams** are built. Bigrams are two words that appear together frequently in the document. Trigrams are three words that occur frequently. For example, "battery_energy_storage", "renewable_energy", "grid_connected", "standalone_system" etc.

For topic modeling there are programming languages like *R*, *Julia*, *java* and *Python*. For this investigation Python 3.7.9 is used, due to its large number of specialized word processing libraries, accelerated computation time and simple [Saheb 2022b] implementation. For pre-processing, the Natural Language Toolkit (NLTK), Scikit-learn, Pandas, Mallet, Seaborn, Matplotlib and Numpy packages are used.

3.1.3.3 Implementation of the LDA method

Topical modeling in Python can be done using the Gensim or Mallet libraries. For this study, the LDA model is obtained using the Mallet package, due to its high performance compared to Gensim [Ebeid 2016]. This Mallet library contains efficient sample-based implementations of latent Dirichlet mapping, Pachinko mapping, hierarchical LDA, and memory usage optimizations [McCallum 2002]. Topic modeling using LDA is an unsupervised technique, which means that it is not known before running the model how many topics exist in the corpus. In addition, it is not possible to rely on previous studies to determine the appropriate number of topics for our investigation since it is the first time that this method is used for the analysis of the literature on fault diagnosis using artificial intelligence in photovoltaic systems. For this reason, it is necessary to build multiple LDA models with different values of the number of topics (K) and choose the model that offers significant and interpretable topics. It is necessary to clarify that selecting a very high number of topics can sometimes provide more granular subtopics, and even topics in which the same keywords are repeated, which means that (K) is probably too large. That is why in this article the use of the topic coherence technique to estimate the appropriate number of topics is proposed.

3.1.3.4 Topic Coherence Measurement

Coherence can be defined as the quality or consistency of the relationships between the words in a topic [Morstatter 2018]. Topic coherence measures take the set of top words N_{m_i}, $i = 1, \ldots, M$ in a topic k_p, $p = 1, \ldots, K$ and add a confirmation measure over all pairs of words. Multiple techniques to measure the coherence of topics are proposed. In coherence based on point mutual information (PMI) the probability of a word belonging to a topic is estimated based on word co-occurrence counts using a sliding window that moves over the corpus [Church 1990]. Mimno et al [Mimno 2011] uses an asymmetric confirmation measure between headword pairs or smoothed conditional probability. In [Aletras 2013], context vector-based topic coherence named Pointwise Mutual Information (PMI) is introduced. This method creates a context vector of a word w using word co-occurrence counts determined by context windows containing all words placed ± 5 tokens around occurrences of the word w. An extension of this method named $NPMI$ defines the elements of these normalized context vectors improving coherence of topics with human topics [Bouma 2009]. In the same work, the authors proposed to restrict the co-occurrences of words to those words that are part of the same topic (top word space).

Other metrics for evaluating coherence are evaluated in [Röder 2015]. In that same study, it is shown that the metric (C_V), which is a combination of the indirect cosine measurement with the $NPMI$ and the Boolean sliding window, presents the best results compared to various metrics. In [Syed 2017], the metric (C_V) is shown to achieve the highest correlation with all available human subject classification data in that study. In the same study they explain that the metric (C_V) is based

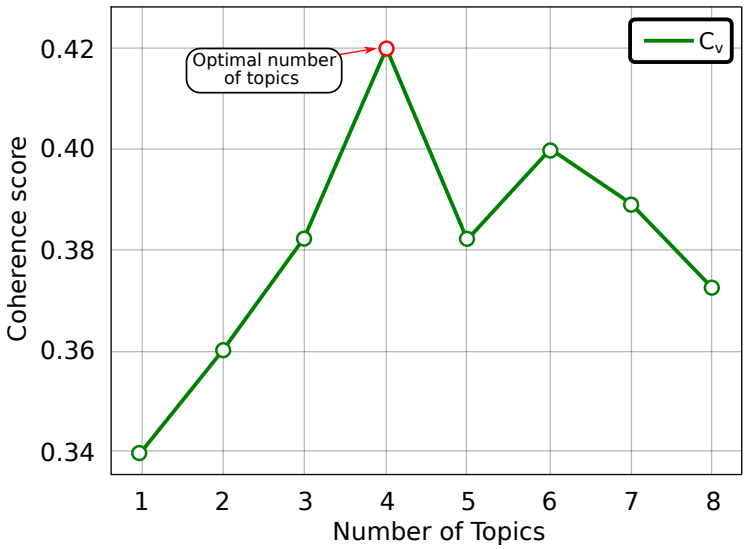

Figure 3.5: Topic coherence analysis based on the coherence score using the metric (C_V). The optimal number of topics is 4.

on four parts: *i)* segmentation of the data into pairs of words; *ii)* calculation of probabilities of words or pairs of words; *iii)* calculation of a confirmation measure that quantifies how strongly one set of words supports another set of words; and finally *iv)* aggregation of individual confirmation measures into an overall coherence score. A formal presentation of the metric (C_V) is described in the same document [Syed 2017]. Figure 3.5 shows the topic coherence values using the metric (C_V) for the LDA models as a function of their number of topics.

As seen in Figure 3.5, the coherence score increases with the number of topics, with a decrease between 4 and 8. This coherence analysis provides a good picture for the selection of the appropriate number of topics. As mentioned above, choosing the right number of topics also depends on expert analysis to avoid selecting topics that may have repeated keywords in the topic. The result in Figure 3.5 suggests that the data is better explained with a model that incorporates $k = 4$ topics with a value of $C_V = 0.42$. Manual checks are used to ensure the validity and robustness of the model. From now on, this model is named *Optimal LDA Model*. A more detailed analysis of the content of each topic is carried out below.

3.1.3.5 Analysis of words per topic

Table 3.1 shows the 20 examples of representative words for the 4 topics in the *Optimal LDA Model*, along with the number of documents on each topic.

As can be seen in Table 3.1, the number of documents in topic 2 is substantially higher than in the other topics. Intuitively, it would be thought that each document contained in each topic k_p, $p = 1, \ldots, K$ belongs only to that topic. However, the fact that the document is classified in a topic does not mean that within the document only words related to that topic are identified. That is, the same document

Topic (# Documents)	Examples of representative words per topic
1 (238)	supervised, classification, diagnosis, feature extraction, feature selection, datum, neural, network, training, tree, kernel, segmentation, label, pattern, transfer, extreme, bayesian, fusion, statistical
2 (150)	semi_supervised, graph_based, optimization, identification, online, hybrid, behavior, network, dynamic, wavelet, threshold, smart, modeling, scenario, transmission, noise, optimize, management, minimum, signature
3 (102)	reinforcement, monitoring,time, operation, production, prediction, cost, maintenance, lifetime, camera, thermal, topology, dynamic_programming, estimation, expect, signal_processing, supervision, platform, timely, forecast
4 (190)	unsupervised, shade, location, clustering, forecasting, irradiance, voltage, current, wind, temperature, characteristic, impedance, arc, circuit, point_tracking_mppt, mismatch, partial, dimensionality, abnormal, anomaly

Table 3.1: 20 examples of representative words for the 4 topics in the *Optimal LDA Model*, as well as the number of documents in each topic

can contain terms from different topics, but it is assigned based on the topic that groups the largest number of terms of the document.

Once an initial understanding of the content of the topics is established, it is necessary to visualize what the intra and inter-topic relationships are. For this, the t-distributed stochastic neighbor embedding (T-SNE) method is proposed as a high-dimensional data visualization tool [van der Maaten 2008a].

3.1.4 T-distributed stochastic neighbor embedding

This method T-distributed stochastic neighbor embedding (T-SNE) is a variation of the Stochastic Neighbor Embedding method (SNE) introduced in [Hinton 2002]. The SNE converts the distance between two points in high-dimensional space to a conditional probability that represents the similarity of the two points in high-dimensional space. Then the SNE matches the conditional probability between two data points in high-dimensional space to the conditional probability between two map points in low-dimensional space.

The conditional probability between two points x_i and x_j is denoted $p_{j|i}$ and represents the probability that x_i would pick x_j as its neighbor. Since this research is only interested in modeling pairwise similarities, the value $p_{i|i} = 0$ is set [Liu 2021a]. The conditional probability $p_{j|i}$ can be defined using a Gaussian kernel as follows:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)}, \tag{3.3}$$

where $\|x_i - x_j\|$ is the Euclidean distance between data points x_i and x_j and σ_i^2 is the variance of the Gaussian that is centered on x_i. Taking into account that the data density is likely to vary, it is likely that there is no single value of σ_i^2 that is optimal for all data points in the data set [Zhang 2021b]. For example, in dense regions, a smaller value of σ_i^2 is often more appropriate than in more sparse regions. Any particular value of σ_i^2 induces a probability distribution, P_i, over all of the other

data points that has an entropy which increases as σ_i^2 increases. Taking this into account, the SNE performs a binary search for the value of σ_i^2 that produces a P_i with a fixed *Perplexity* that is specified by the user. As mentioned in [Zhang 2021b] perplexity can be understood as a measure of the effective number of neighbors. The *Perplexity* is typically set to a value between 5 and 50 [Zhang 2021b] and is calculated as follows:

$$Perp(P_i) = 2^{H(P_i)}, \tag{3.4}$$

where $H(P_i)$ is the Shannon entropy of P_i measured in bits. $H(P_i)$ is calculated using Equation (3.5) [Liu 2021a].

$$H(P_i) = -\sum_j p_{j|i} log_2(p_{j|i}), \tag{3.5}$$

For low-dimensional counterparts the SNE computes a conditional probability for the map points y_i and y_j. These two points y_i and y_j are the equivalent or corresponding representation of the data points x_i and x_j respectively. This conditional probability is denoted as $q_{j|i}$, with variance $1/\sqrt{2}$ and calculated as follows:

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{i \neq k} \exp(-\|y_i - y_k\|^2)}, \tag{3.6}$$

As for the probability $p_{i|i}$, $q_{i|i} = 0$ is set. If the points y_i and y_j correctly model the similarity between the high-dimensional data points x_i and x_j the conditional probabilities, $p_{j|i}$ and $q_{j|i}$, should be equal. In this way, SNE tries to find a low-dimensional data representation that minimizes the discrepancy between $p_{i|i}$ and $q_{i|i}$. This discrepancy can be measured by the Kullback-Leibler divergence considering all data points using a gradient descent method. The cost function $C_{P_i \| Q_i}$ to be minimized is given as follows:

$$C_{P_i \| Q_i} = \sum_i \mathrm{KL}\left(P_i \parallel Q_i\right) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \tag{3.7}$$

where P_i y Q_i represent the conditional probability distribution over all given data points x_i and map y_i respectively. It is important to mention that the Kullback-Leibler divergence is not symmetric, so the different error types in the pairwise distances in the low-dimensional map are not weighted equally [Zhang 2021b]. This translates to a large cost of using widely spaced map points to represent closely spaced data points and a small cost of using closely spaced map points to represent widely spaced data points. Taking this cost function problem into account, Van der Maaten and Hinton [van der Maaten 2008a] present a modification of SNE named t-SNE. Also, this variation is born to solve the problems of the cost function that is difficult to optimize in the SNE [van der Maaten 2008a]. t-SNE uses a symmetric version of SNE to estimate pairwise similarities in low- and high-dimensional spaces. t-SNE introduces the Equation (3.8).

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, \tag{3.8}$$

where p_{ij} is the probability that x_i chooses x_j as its neighbor such that $p_{ij} = p_{ji}$ for $\forall i, j$ (symmetry) and n is the number of data points. Equation (3.8) ensures that the condition of Equation (3.9) is met for all data points x_i.

$$\sum_j p_{ij} > \frac{1}{2n}, \tag{3.9}$$

Equation (3.9) ensures that each data point x_i makes a significant contribution to the cost function. Also, T-SNE uses a Student-t distribution rather than a Gaussian to compute the similarity between two points in the low-dimensional space. That is, for map points y_i and y_j T-SNE uses a Student-t distribution with one degree of freedom to compute the Equation (3.10) [Zhang 2021b].

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}, \tag{3.10}$$

which is the probability that y_i chooses y_j as its neighbor. As in the case of p_{ij} and p_{ji}, $q_{ij} = q_{ji}$ for $\forall i, j$ conserving the property of symmetry. These new conditions transform cost function $C_{P_i \| Q_i}$ from Equation (3.7) into function $C_{P \| Q}$ from Equation Equation (3.11).

$$C_{P \| Q} = \mathrm{KL}\left(P \parallel Q\right) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{3.11}$$

where P and Q are the joint probability distributions in high- and low-dimensional spaces, respectively. Equation (3.11) can be seen as an alternative to minimize a single Kullback-Leibler divergence between a joint probability distribution, P, in high-dimensional space and a joint probability distribution, Q, in low-dimensional space.

Using Equations (3.8), (3.10) and (3.11) the probability that a document belongs to a topic is calculated. The grouping results of the clusters are measured in terms of how compact and clearly separable the classes are using a metric such as the silhouette coefficient[Rousseeuw 1987]. The silhouette coefficient φ is a measure of how similar a sample is to the samples of its own cluster compared to the samples of other clusters [Xiang 2021]. The value of the silhouette coefficient ranges from $-1 \leq \varphi \leq 1$. A value of 1 indicates that the sample is well assigned to the cluster i.e far from its neighboring clusters, a value of 0 indicates that the sample can also be assigned to another cluster and finally, a value of -1 indicates that the sample is not correctly assigned i.e it is far from the other samples of its own cluster or somewhere in between the clusters. The average value of silhouette coefficient of all the samples n_e of a cluster g is represented by $\bar{\varphi}_g$ and is defined according to [Eler 2015] as:

$$\bar{\varphi}_g = \frac{1}{n_e} \sum_{i=1}^{n_e} \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \tag{3.12}$$

where $a(i)$ is the average distance between the sample e_i, $i = 1, \ldots, n_e$, and the rest of the sample in cluster g and $b(i)$ is the minimum distance between object e_i and the rest of the objects in all other clusters except g. A low value of $a(i)$ indicates a good compactness of the cluster (low intra-cluster distance). A high value of $b(i)$ indicates a good cluster separability (high inter-cluster distance). Once the methods are fully explained, their final results are analyzed in the next stage.

3.1.5 Expert qualitative content analysis

In this section the results of the bibliometric analysis, topical modeling and t-SNE are examined in order to extract as much information as possible to build the state of the art on Fault Diagnosis in Photovoltaic Systems using Artificial Intelligence. First, the result of the bibliometric analysis of the keywords is presented in Figure 3.6. Figure 3.6 presents a cluster Visualization Map of keywords with the 6 keyword clusters labeled according to the approach to which they belong.

This bibliometric analysis showed that all the clusters identified coincide with the classification presented by Rodrigues et al. [Rodrigues 2017]. This previous classification is supported by the main contribution of the algorithms contained in each group extracted with the "master algorithm" tool. The concept of "master algorithm" is introduced by Domingos [Domingos 2015] in his book entitled "*The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*". The author considers the "master algorithm" as the unifier of machine learning, and then as, hypothetically, a machine learning algorithm capable of perfectly understanding the behavior of any system. Taking into consideration these two previous authors approaches, the names of each cluster are assigned as respectively *Symbolic approach*, *Regression approach*, *Bayesian approach*, *Analogy-based approach*, *Connectionist approach* and *Evolutionary approach*. Our approach confirms that each of these clusters has its own master algorithm with its properties and drawbacks.

3.1.5.1 Cluster 1 - Symbolic approach

This cluster groups algorithms whose master algorithm is identified as inverse deduction, also called induction. The approaches of this cluster try to reach a specific conclusion based on pre-existing knowledge previously learned [Domingos 2015]. For this reason, these machine learning approaches cannot start "from scratch". Because the inverse deduction is very expensive from the computational point of view, the application of these algorithms in massive data sets until now rests very difficult to generalize. Some of the most recognized algorithms in this cluster are well-known as Decision tree, Random Forests and Fuzzy Logic [Antonanzas 2016].

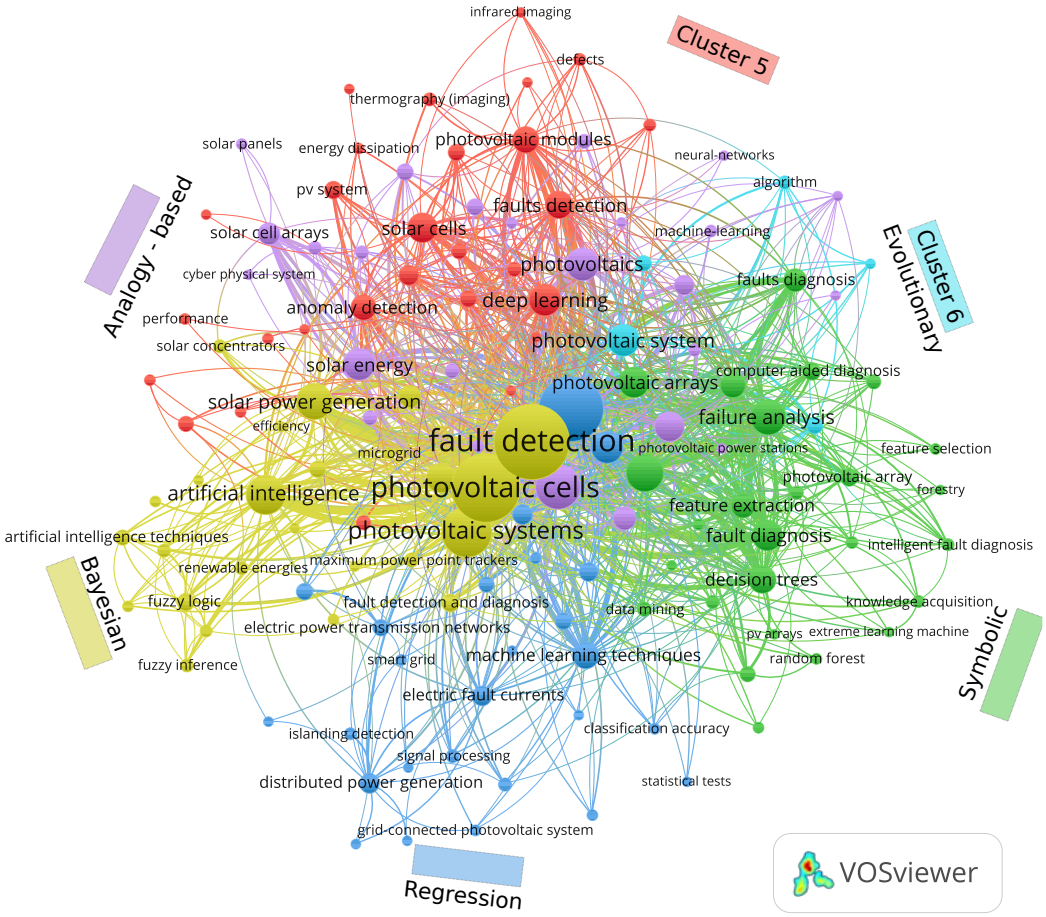

Figure 3.6: Visualization Map of co-occurring keywords identifying 6 main clusters. Each cluster corresponds to a type of approach.

3.1.5.2 Cluster 2 - Regression approach

This cluster groups algorithms based on data that are not learning algorithms but regressive algorithms and depend on historical data according to [Domingos 2015]. Some of the most recognized algorithms in this group are ARIMA (Auto-Regressive Integrated Moving Average), Linear Regression models, Principal Component Analysis (PCA) and statistical machine learning approaches [Antonanzas 2016].

3.1.5.3 Cluster 3 - Bayesian approach

This cluster groups algorithms identified by the master algorithm tool as the probabilistic Inference. These algorithms have common properties with their capacity to reduce the uncertainty of the new knowledge using the probabilistic event inference algorithm [Rodrigues 2017]. These types of algorithms recognize the inherent uncertainty and incompleteness of all types of knowledge. In these algorithms, at each known event a probability is assigned. Then if the data supports a hypothesis, the hypothesis becomes more weight. If the data contradicts it, less weight is assigned to the hypothesis as in [Domingos 2015]. Some of the best known methods used in this cluster are Naïve Bayes and Monte Carlo [Antonanzas 2016].

3.1.5.4 Cluster 4 - Analogy-based approach

This cluster groups algorithms which are called by master algorithms, kernel machines. These algorithms analyze similarities between old and new data by using the nearest neighbor kernel machine algorithms capable of doing analysis of their environment and try to generalize by help of similarity. This group of algorithms presents similar results to neural networks [Domingos 2015]. One of the main drawbacks that exists with these algorithms is the high dependence on the size of the data set, the calculation time and the complexity of the programming that can quickly become important. Among the best known algorithms in this group, the support vector machines and the K-Nearest Neighbor algorithms are the most used as mentioned on [Antonanzas 2016].

3.1.5.5 Cluster 5 - Connectionist approach

This cluster groups algorithms whose master algorithm is called the backpropagation. These algorithms are able to emulate the functions of the brain by creating artificial neurons and connecting them in a neural network using an input layer, one or more hidden layers, and an output layer. Input neurons are taken by the hidden neurons which generate an output able to be read by other neurons which are performed with the same function [Domingos 2015]. Artificial neural networks and extreme learning machines are some of the most recognized algorithms in this cluster [Antonanzas 2016].

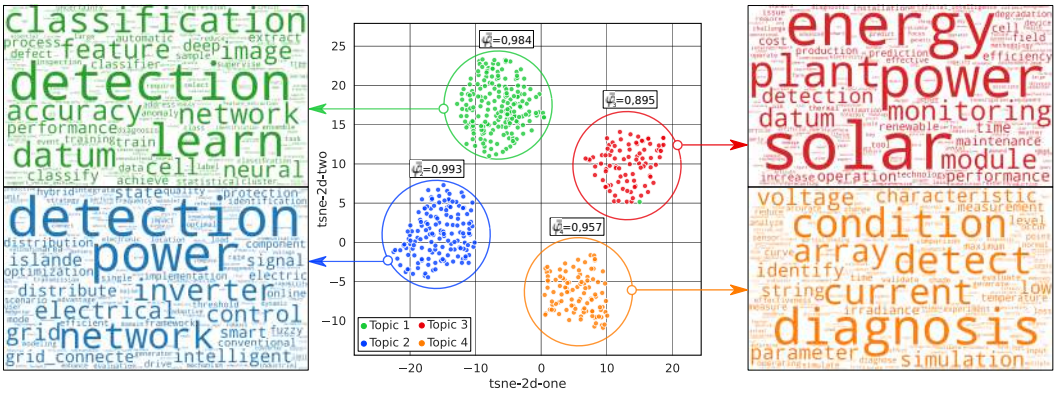

Figure 3.7: Distribution of Document Word Counts by Dominant Topic. In the center the 2D visualization T-SNE of the 4 topics is presented. On the sides, word clouds for each topic are presented. Each of the average values of silhouette coefficients $\bar{\varphi}_g$, $g = 1, \ldots, 4$ is presented above each topic.

3.1.5.6 Cluster 6 - Evolutionary approach

These algorithms work on imitation of the evolutionary process of real genomes and DNA based on Darwin's principles, where performance is measured by the fitness of offspring [Rodrigues 2017]. These algorithms have a set of individuals which each one competes with each other, mix, mutate and then only the fittest are not discarded as best genes with evolutionary biology [Domingos 2015]. Among the most popular algorithms belonging to this cluster are the genetic algorithm and genetic programming [Antonanzas 2016].

Then the results obtained with topic modeling and T-SNE are analyzed. The cluster of documents in a 2D space using the T-SNE algorithm, the silhouette coefficient of the clusters and the word clouds obtained by topic modeling are presented in Figure 3.7. This figure helps to explore the inter-topic distance and therefore as well as how the topics are related to each other, including the most relevant words to infer the topic from Keywords.

As can be seen in Figure 3.7, all the clusters or topics obtained average values of silhouette coefficients close to 1. This indicates that the method effectively grouped the documents correctly. After analyzing the words contained in each theme, it is identified that the work carried out in fault diagnosis in PV plants can be classified into 4 types of machine learning (see 3.7) and the 6 families of algorithms (see Figure 3.6). A recapitulative diagram that contains the two classifications, the physical variables and algorithms most used for fault diagnosis is presented in the following section.

3.1.6 Summary of the relevant information retrieved.

After an in-depth study, 11 different variables are identified that are used to train machine learning models. The variables are: Irradiance, ambient temperature, wind speed, module temperature, current, voltage, power, energy, humidity, cell

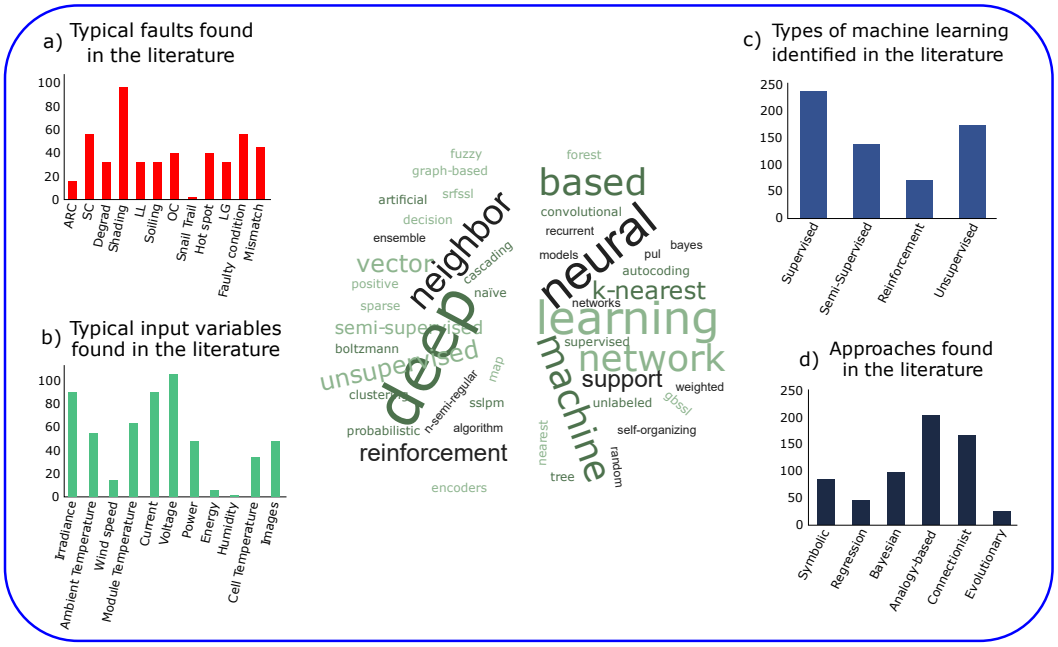

Figure 3.8: Summary scheme of the information found with the Smart B2TE methodology. a) Typical faults studied in the literature; b) typical variables used for fault diagnosis; c) types of machine learning found with bibliometric analysis; and e) types of machine learning found with topical modeling and T-SNE.

temperature and images. Similarly, the 12 most studied faults are identified: arc faults (ARC), short circuit (SC), degradation (Degrad), Line to Line (LL), Soiling, Open Circuit (OC), Snail Trail, Hot spot , Mismatch and other fault conditions. All this information together with the two proposed classifications are shown in Figure 3.8.

As seen in Figure 3.8, multiple faults are diagnosed using multiple physical variables and machine learning approaches. To give a more specific view of the work developed in the field, Section 3.2 exposes multiple articles on each of the 4 types of Machine learning. This classification has been adopted in this thesis because it is widely known in the field of artificial intelligence.

3.2 Methods based on artificial intelligence techniques

Artificial intelligence (AI) methods or techniques are gaining strength as a new research direction in the energy sector, mainly due to the ability to manage a large amount of data without the need for a physical model of the system [Massa 2021]. Without loss of generality, AI can be defined as an area of computer science whose goal is to create and train intelligent machines that simulate human behavior based on available data [IRENA 2019, Licciardo 2018, Shobha 2018]. More specifically, Machine Learning (ML) is considered part of AI [Gupta 2020, Ibrahim 2020], which is classified in the literature by subgroups as Supervised Learning (SL)

[Massa 2021], Semi-Supervised Learning (SSL)) [Pise 2008], Reinforcement Learning (RL) [Massa 2021, Barja-Martinez 2021] and Unsupervised Learning (UL) [Massa 2021]. Among them, SL and UL are widely researched to be used in fault diagnosis problems, while RL is mainly applied to plant control problems and control faults [Massa 2021]. In general, in fault diagnosis using machine learning, algorithms are designed and trained to learn and improve the relationship between the input and output parameters of a photovoltaic system and thus use them to discriminate between healthy and faulty behavior [Pillai 2018b]. Training data can be constituted from experimental collections or with the help of accurate PV models even in few cases the both of them. However, outliers to be detected when faults occur require high accuracy from training and prediction. As mentioned in the literature [Pillai 2018b], although MLT-based fault diagnosis helps to overcome the difficulty of defining thresholds to achieve a more accurate fault detection and improve fault classification, it persists the following disadvantages. First, accuracy depends on the quality of the training data used. Second, specific training data including faults are needed, especially for fault occurrences. But, it is extremely difficult to collect. Third, machine learning algorithms that use PV models are in major part linked to the fact that the precision depends entirely on the photovoltaic model used. Finally, on PV plants, training data are not easy to obtain linked to the difficulty of access and then can be considered non available globally. In addition, how to obtain training data varies from one PV plant to another depending on the type, size, and geographic location.

Despite all these limitations, as mentioned in [Livera 2019b], ML techniques have shown multiple advantages such as symbolic reasoning, flexibility and the ability to explain the results and are able to analyze and identify patterns in nonlinear, large, complex, and even incomplete data. In addition, compared to the standard linear models for optimization methods, ML methods have compact solutions for multivariable problems without the need to know the internal parameters of the system [. 2019]. For these methods, only one training process is required to directly determine the output parameters. These solutions are obtained without solving any nonlinear mathematical equations or making statistical assumptions as in conventional optimization methods.

In order to give an overview of the current methods used for fault diagnosis in PV plants, articles detected with the process described in the previous section as the more interesting in the PV diagnosis and grouped into different clusters shown in Figures 3.6 and 3.7 are studied below, with various methods of machine learning with distinctive principles and structures, belonging to each of the 4 types of machine learning (SL, SSl, RL and UL), depending on the selected algorithm, the type of fault detected, the input variables between other relevant aspects for fault diagnosis. In each type of machine learning, priority is given to the documents with the highest number of citations.

3.2.1 Supervised Learning

Supervised learning models are the most widely used in machine learning [Baharin 2014]. In the photovoltaic domain, one of the main drawbacks of these models and consequently one of the main challenges is to make the model work correctly in different seasons of the year, taking into account that the amount of data needed for training is large and the model is built on a case-by-case basis [Youssef 2017].

In the supervised learning scheme it is necessary to have a database with inputs (predictors) and outputs (labels or targets) [Berry 2019]. Then, the documents found in this topic try to discover the relationship between inputs and outputs in the formation process [alias Balamurugan 2011]. These algorithms first produce a function that assigns data to labels. And this function is used to predict the label of unlabeled data. Among the best known algorithms are Naïve Bayes (NB), Decision tree (DT), Random forest (RF), Artificial neural network (ANN), Probabilistic Neural Network (PNN), Extreme Learning Machine (ELM), Ensemble learning (EL), Deep learning (DL), Support vector machine (SVM) and K-nearest neighbor (KNN). Articles with the best known algorithms are presented below.

3.2.1.1 Naïve Bayes

Naïve Bayes (NB) corresponds to a probabilistic machine learning classifier, built based on Bayesian theory. The main disadvantages of a Naïve Bayes model are: 1) all features are considered independently and then if correlation between features is not considered; 2) performance can be negatively affected when the difference between the training and test data is slightly different. Despite these limitations, NB is used in [Eskandari 2021] to classify line-to-line faults, line-to-ground faults, and normal conditions. In [Maaløe 2020], simulated training data from 10 shadowing fault modes is used to train a NB model to accurately differentiate between different operating behaviors using the characteristics of conventional I-V curves. In [Niazi 2019], a NB model is used for the identification of faults and degradation in photovoltaic modules in a stand-alone photovoltaic plant. In [He 2021], a distributed PV array fault diagnosis method is proposed based on Naive Bayes model fine tuning for PV array fault conditions such as open circuit, short circuit, shading, abnormal degradation and diode abnormal shunt. The data of the maximum power point of the PV inverter and the weather data are used as training data. The approach proposed in [He 2021] is validated by simulation.

3.2.1.2 Decision tree

In Decision Tree (DT), the learning process is carried out by transforming the input data into a tree form. DT is considered a non-parametric model. DT does not need scaling or normalization in the PV fault identification process. Furthermore, DT is considered an interpretable machine learning structure. What does not happen with the ANNs or the DL that are a black box. Among the drawbacks of DT are:

1) a small change in the data set can cause a significant change in the structure of DT; and 2) DT training is complicated and requires more computing time than other methods. Despite this, DT is widely used. In [Benkercha 2018], DT is used to detect line-to-line faults and short-circuit faults in PV arrays. In [Zhao 2012], a DT model is trained to detect faults at the matrix level, showing 99% accuracy in detecting line-to-line faults, shadow faults, and ground faults.

3.2.1.3 Random Forest

Random Forest (RF) is one of the strongest shallow-based classifiers. RF is an extended version of DT and consists of several DTs. The RF output is determined by the class label with the largest number of trees. Unlike other shallow builds, the RF cannot be easily overturned with a large number of features. However, each DT in an RF requires additional memory space, which increases the required memory space. It is important to mention that each DT in RF works independently. In [Chen 2018b, Dhibi 2020], RF is used to detect mismatch faults and line-to-line faults. In [Hajji 2021], a PCA-based multivariate time series feature selection and extraction technique is developed. The data represents 5 fault modes belonging to the line-to-line and shaded categories in various components of a photovoltaic system. Several classifiers are used in that study, but the results showed that RF has the ability to provide the best results. In [Han 2019], the use of RF to detect AC and DC faults at an early stage in grid-connected PV systems is proposed.

3.2.1.4 Artificial Neural Network

The Artificial Neural Network (ANN) is one of the most widely used machine learning methods. In principle, the ANNs are made up of input, hidden, and output layers, and generate a function based on the relationship between the inputs and the learning weights. The relationship of the outputs as a function of the inputs of an ANN network is defined by the activation function. In the field of fault diagnosis in PV plants, ANNs such as the multilayer perceptron (MLP) are used to detect and locate DC arc faults in shading faults in [Mekki 2016]. In [Haykin 1999] the superior performance of MLP against approximation problems is validated. In that study the MLP is trained with data from faulty conditions in real time in a PV system and they are trained using a robust reverse propagation (RP) algorithm. The radial basis function (RBF) to detect faults in the photovoltaic array [Benkercha 2018]. In [Chine 2016a], an experimental study of 775 patterns from various data sets is performed. For the training of the ANNs model a comparison between the MLP and RBF architectures is performed.

The artificial neural network (ANN) to classify different faults in the photovoltaic array [Li 2012]. In [Jiang 2015], an ANN is used whose training data is collected using simulation analysis of a Single diode PV model (SD PV) model. In [Leva 2019], the prediction error of an ANN model is compared to a threshold for fault detection. In [Syafaruddin 2011] an ANN is used to identify fault locations in

a single PV string. In [Syafaruddin 2009], it is used as a three-layer artificial neural network (ANN) to detect short-circuit faults in a photovoltaic system. Compared with the RBF radial-based neural network, the proposed structure has a simple structure and better precision [Syafaruddin 2009]. In addition, the proposed algorithm also has the ability to locate faults. Other studies such as [Li 2012], it trains the ANN model with temperature data, the MPP voltage and the current of the PV modules to identify 4 different health states labeled as: normal, degraded, short-circuit fault and shading. In [Chine 2016b], under established conditions of solar irradiation and temperature of the photovoltaic (PV) module, a series of attributes such as current, voltage and number of peaks in the current-voltage (I–V) characteristics of the strings are calculated using a simulation model. Next, an ANN model is used to isolate and identify eight different types of faults. This work also shows the implementation of the developed method on a field-programmable gate array (FPGA) using a Xilinx System Generator (XSG) and an Integrated Software Environment (ISE). In [Syafaruddin 2011], it is stated that a single artificial neural network (ANN) is not adequate to provide an accurate solution for fault diagnosis in PV plants. Therefore, in that work several ANNs are proposed, then the voltage terminal of the module based on automatic control is established. The method proposed in this work is able to detect the exact location of the short-circuit condition of the photovoltaic modules in the array.

Other ANN adaptations are proposed. A physics artificial neural red (PHANN) to predict the power output of normal photovoltaic systems [Leva 2019]. In [Jones 2015], the Laterally Primed Adaptive Reference Theory (LAPART) based on ANN is introduced. In that work [Jones 2015], LAPART is used to identify faulty general conditions in a photovoltaic system. Real and simulated data are used for both training and validation of the model.

As can be seen, ANNs are widely used for fault diagnosis in PV plants. However, as demonstrated in [Afrasiabi 2022b], ANNs have serious difficulties in working with correlated signals, such as output voltage and current in a photovoltaic system. Other disadvantages of ANNs, described in [Liu 2022], are: 1) The neural network requires a large number of data samples to train, and the generalization of the networks is difficult to guarantee; 2) the computation cost increases due to the iterative learning process of the neural network; and 3) The network structure and hyperparameters are difficult to determine, which is mainly based on the experience of experts and extensive experiments.

Probabilistic Neural Network: The Probabilistic Neural Network (PNN) is a type of ANN that uses a sigmoid activation function instead of a linear or exponential activation function. The main advantage of the PNN is to use a probabilistic procedure and improve accuracy compared to conventional ANNs. The main disadvantage is that a PNN model works slower than an ANN model, and requires additional memory space compared to other machine learning methods. In [Garoudja 2017b], PNN is used to detect line-to-line faults on the DC side of PV systems. As input to the PNN model, common I(V) signals on the DC converter side are used in En [Garoudja 2017b]. The training data is obtained by simulation.

On the other hand, in [Akram 2015], a probabilistic neural network (PNN) is used to detect open-circuit faults and line-to-line faults in a PV array.

In [Akram 2015], a health monitoring method for photovoltaic (PV) systems based on probabilistic neural networks (PNN) is proposed that detects and classifies open and short circuit faults in real time. That approach proposes a model of PV systems that only requires data sheet information from manufacturers reported under normal operating cell temperature (NOCT) and standard operating test (STC) conditions. The proposed model considers variables such as the ideality factor, the series resistance and the thermal voltage.

Extreme Learning Machine: An ELM is constructed based on a single-layer feed-forward neural network. In ELM, two main steps need to be performed, including random initialization and linear parameter solving. In the first stage, the weights and biases of the single-layer feedforward neural networks are initialized randomly. In this type of neural networks, a non-linear activation function is needed. In [Chen 2017], it is proposed to use ELM for PV parameter identification and PV array fault detection. The four different fault types considered include degradation, short circuit, open circuit, and partial shadowing. In [Wu 2017] an optimal ELM is presented to detect short-circuit faults and aging of photovoltaic energy. In [Huang 2020], a semi-supervised ELM is developed to detect faults in photovoltaic panels taking into account the impact of dust. As mentioned in [Huang 2020], detecting dust can increase the aging process and the temperature of photovoltaics during operation. The disadvantages of the ELM are its high sensitivity to noise. On the other hand, ELM might fail in a multi-class classification problem where the difference between the classes is low.

3.2.1.5 Ensemble learning

In EL, multiple machine learning methods are combined instead of using just one. EL can be performed with a high level of accuracy in case the combination of machine learning structures is organized in a proper way. However, training and blending can be difficult and EL can take a long time during the training process. In [Dhibi 2021] , an EL consisting of the kNN, DT and SVM algorithms is proposed. In [Yang 2022], a voting-based ensemble learning algorithm with linear regression, a DT and an SVM called ($EL - V_{LR-DT-SVM}$) are proposed for fault diagnosis on PV plants. Normalized training data (voltage-current characteristics) are captured under different weather conditions. In [Badr 2021], an EL composed of a DT, kNN, and SVM is proposed to detect permanent faults (Arc Fault, Line-to-Line, Maximum Power Point Tracking unit fault, and Open-Circuit faults), and temporary (Shading) under a wide range of climate datasets, fault impedances, and shading scenarios. In that same research, Bayesian Optimization is proposed to assign optimal hyperparameters to fault classifiers. This approach is evaluated with experimental and simulation cases; showing interesting results. In [Eskandari 2020a], a methodology similar to that exposed in [Huang 2020] is proposed for PV-based fault detection. In this study, an EL model consisting of the KNN, SVM and Naive

Bayes (NB) algorithms is proposed. The final decision is established by a voting process as in [Yang 2022], in unique training environments, which are concluded by a voting process to pass a final decision. This approach is used to detect line-to-line faults. In the same way, in [Eskandari 2021], using the SVM, NB and Logistic Regression (LR) classifiers for line-to-line fault detection.

In [Kapucu 2021] it is proposed a fault diagnosis method based on ensemble learning (EL). This method uses the grid-search with cross-validation method to extract features from signals such as: the voltage, current and power of the photovoltaic string, solar radiation, the temperatures of the rear surface of the photovoltaic module read from four DS18B20 temperature sensors, the average value of these four temperatures, the ambient temperature, the humidity and the cell temperature. Optimization methods are used to define the parameters of each individual learning algorithm used. This approach presents interesting results in terms of classification performance and generalization capacity. In [Guo 2020], it is proposed a diagnosis approach based on EL. This model uses SVM, LR and RF algorithms. The parameters of the algorithms are iteratively modified by analyzing the history of the PV plant at the end of each diagnosis cycle. The results obtained in [Guo 2020] show that using a set of algorithms increases the accuracy of the classification.

3.2.1.6 Deep learning

In the 1980s, John Hopfield and David Rumelhart popularized deep learning (DL) in training brain-inspired algorithms [Rao 2021]. The DL focuses mainly on feature representations and mappings [Berghout 2021a]. As a result, the larger the feature space, the more meaningful the representations are. Among the proposed DL technologies are Deep Belief Networks (DBN) [Hinton 2006], Convolutional Neural Networks (CNN) [Lecun 1998], Deep encoders (DA) [Afrasiabi 2021], Deep Boltzmann machine (DBM) [Afrasiabi 2021] and recurrent neural network (RNN) [Afrasiabi 2021]. As mentioned in [Berghout 2021a], in general, DL models are not well suited to sensor-based condition monitoring of PV plants. However, despite this serious limitation, multiple works are carried out on fault diagnosis in PV plants.

Convolutional Neural Networks: The main disadvantage of CNN is that it cannot fully understand the temporal characteristics [Afrasiabi 2022a]. However, it is widely used for fault detection in PV plants. In 9180283, a CNN is proposed on a set of recorded features (ie, I(V), solar radiation, and temperature). This approach is capable of classifying 10 different types of faults. The CNN algorithm is consolidated with a Residual Activated Recurring Unit (Res-GRU) to provide dynamic online training capability. In [Aziz 2020], a CNN is also proposed to classify multiple faults in photovoltaic arrays. In [Pierdicca 2020], thermal images recorded through infrared sensors installed on unmanned aerial vehicles (UAVs) are used to train a hybrid mask region-based CNN model for fault classification of a photovoltaic system. Three fault modes are studied (i.e., one anomaly, non-contiguous

cells with anomalies, and contiguous cells with anomalies). In [**?**], a retrained CNN algorithm for image classification (ie VGG16) is used to extract features from thermal images obtained from UAVs. After extracting the appropriate features based on a generative model, the assignments are passed through a discriminative CNN algorithm to achieve the approximation process. Five different degradation fault modes (i.e., burn marks, delamination, discoloration, glass breakage, and snail trails) are studied.

In [Moradi Sizkouhi 2021], an encoder-decoder architecture is implemented to train a fully connected CNN to detect shadows caused by bird droppings. The CNN is trained with aerial images necessary to train, test and validate the proposed network. The labeling of the collected images depended on the analysis of the output current of the PV system. In [Manno 2021], thermographic images obtained from ground facilities and UAVs are used. These images are used to train a CNN to locate and identify Hot Spot faults. In [Hong 2022a], CNN is used to detect line-to-line, open-circuit, and short-circuit faults in PV arrays. In [Aziz 2020], a CNN network trained with 2D scalograms of photovoltaic system data is proposed. This CNN is proposed in two configurations: one derived from a pretrained AlexNet CNN in which the last three layers are tuned to provide a six-way classifier, and another where the results of a pretrained AlexNet layer (fc7) are used with a classifier classic (RF and SVM). The faults considered detectable with the proposed approach are Partial Shading, Line to Line, Open Circuit, Arc fault and faults (Line to Line and Open Circuit) in Partial Shading. That study mentions the importance of aggregating the MPPT data (Imax and Vmax) to obtain good precision. The two approaches presented demonstrated high levels of performance.

The FCNN mask is used by [Mehta 2017] to predict dirt category and location in solar modules. It also predicts power losses in the module. In [Akram 2019], a CNN architecture is used, along with various data augmentation strategies as suggested in [Chen 2019] to deal with data scarcity. This approach kept up the prediction speed in real time. This approach is used to detect faults in Electroluminescence (EL) images of PV cells. Isolated and transfer deep learning both can be used for successful detection of faults in infrared images of PV modules [Akram 2020]. The study [Akram 2020] collected IR images dataset after performing experiments on normal and faulty modules. For isolated learning, they used a light CNN network that is trained from scratch. For transfer learning, they developed model techniques. Therein, a base model is pre-developed on another dataset of EL images, whose knowledge is transferred to target model trained on IR image dataset. The transfer learning scheme with model development approach performs relatively better. They also discussed different types of faults appearing in IR images of solar panels.

YOLO CNN network [Redmon 2015] is used by [Greco 2020] for detection of hotspots in infrared images of PV modules. Skip connections are employed for concatenation of features extracted from initial layers with refined features from following (last) layers. The developed approach can segment the PV modules from images and detect hotspots. Therein, first the single (separate) modules are detected from images and then an ID number is given to each panel according to

the original frame. Following this, hotspots are identified in each single module. The experimental results indicate the robustness of the proposed method. Moreover, this approach does not require heavy fine tuning and also achieves real time speed. Faster R-CNN [Ren 2015] is used by [Wei 2019] for detection of hotspots in thermal infrared images of PV modules. The pre-trained model weights are used and fine-tuned on infrared images dataset for subject task. In addition, they also used an image processing based scheme containing Hough line transformation and canny edge detection processes for hotspots detection. The Faster R-CNN approach achieved excellent results; however, it has relatively very high computational cost.

Faster R-CNN [Ren 2015] is used by [Wei 2019] for hot spot detection in thermal infrared images of photovoltaic modules. The pre-trained model weights are used and fitted on the infrared imaging data set for the subject task. In addition, they also used an image-processing-based scheme containing Hough line transform and clever edge detection processes for hotspot detection. The Faster R-CNN approach achieved excellent results; however, it has a relatively very high computational cost.

Deep encoders: In [Liu 2021b], DA is used to detect short-circuit faults, partial shading, and degradation faults. In [Lu 2019a], domain adaptation combined with deep convolutional generative antagonistic network (DA-DCGAN)-based methodology is proposed, where DA-DCGAN first learns an intelligent normal-to-arc transformation from the domain data. source domain. Then, by generating fictitious arc fault data with the learned transformation using the normal data of the target domain and employing domain adaptation, a robust and reliable fault diagnosis scheme for the target domain can be achieved. The proposed method is implemented in an embedded system and tested in real time according to the UL-1699B standard [UL 2018]. The experimental results demonstrate the high performance of the DA-DCGAN approach.

Deep Boltzmann machine: In [Tao 2020], DBM is used to diagnose open circuit faults, short circuit faults and partial shading

Recurrent neural network: RNN-based networks include long-term memory (LSTM) [Alrifaey 2022, Schmidhuber 2015] and closed recurrent neural network (GNN) [Van Gompel 2022], they can realize temporal features from time-varying parameters and can be used in photovoltaic fault detection problems. However, RNNs have serious problems in fully understanding the spatial features [Afrasiabi 2019]. Despite these different RNN networks are used for fault detection in PV plants. In [Appiah 2019b], automatic LSTM capable of extracting significant features with greater learning capacity over time is developed. In this study, the authors used I(V) signal analysis to address the condition monitoring problem of photovoltaic systems. In [Veerasamy 2021], an LSTM combined with Discrete Wavelet Transform (DWT) is used as a feature extraction step, to detect high impedance faults (HIF). The results of the LSTM are compared with other based methods such as: SVM, Naïve Bayes, Decision Tree, showing a better performance.

3.2.1.7 Support Vector Machine

SVM-based machine learning is considered very efficient in terms of classification accuracy and memory consumption [Afrasiabi 2022b]. For this reason, like neural networks, ANNs are widely used. In [Harrou 2021], an improved kernel function for the SVM model is proposed. That SVM model is used to detect string fault, partial shading, short-circuit fault, line-to-line fault, and module degradation for PV systems. In [Baghaee 2020] an SVM is developed to detect grid anomalies, including islanding conditions and grid faults for photovoltaic systems. In that work, the detection of islands and faults in the network are considered abnormal conditions and binary SVM is used to detect them. In that same work [Baghaee 2020], it is stated that with an increasing number of data and a reduced margin of separation between classes, the precision of SVM is significantly reduced. In [Ali 2020] an SVM model trained with infrared thermography (IRT) images is proposed. This model is capable of classifying panels into three categories: healthy, non-faulty hotspot, and faulty. This work centered his efforts on proposing a novel preprocessing phase for infrared thermography (IRT) images. This image feature extraction results in 41 features: 3 RGB, 12 contrast, 12 correlation, 3 energy, 1 oriented gradient histogram, and 10 local binary patterns. [Ali 2020] is able to demonstrate that the proposed feature extraction significantly improved the precision results compared to classification algorithms such as: such as quadratic discriminant analysis (QDA), naïve-Bayes (n-Bayes), k nearest neighbor (KNN), bagging ensemble (BE).

In [Eskandari 2020b] an SVM model is used to detect line-to-line (LL) faults. The hyperparameters of the SVM model are selected using a Genetic algorithm (GA). This model uses characteristics extracted from the IV curve data resulting from a simulation model of a photovoltaic plant. The results show that the Gaussian kernel is optimal for the SVM model. In [Demant 2014], photoluminescence images having cracks and normal images are classified using an SVM model. There, the shapes of the cracks are defined by the location of the gradient and the orientation histogram (GLOH) [Mikolajczyk 2005].

In [Phua 2019], an SVM classifier on image features is used to perform an automatic classification of fault modes in screen printed and plated PV cell' RGB images resulting from stylus impact with metallic conductors and different forces. In that approach the feature histogram of oriented gradients (HOG) [Dalal 2005] and robust accelerated features (SURF) [Bay 2008] are used to extract features from the images which are then used to train the SVM classifier. In the same way, in [Demant 2016] the diagnosis of cracks in photoluminescent and infrared images of photovoltaic modules is carried out using local descriptors to train the SVM model with the radial-based kernel.

In [Chouay 2021] the diagnosis of 8 types of common faults that occur in the photovoltaic generator is proposed using a multi-stage approach. They first use the energy loss analysis approach to identify the presence of potential faults by comparing measured and expected generated energy. The second part compares the extracted PV characteristic and the reference one to identify the type of fault.

Finally, the last part uses an SVM model that classifies the faults. The simulation results show that the method is able to identify and distinguish faults with 100% accuracy.

3.2.1.8 k-Nearest Neighbor

The kNN is a non-parametric supervised machine learning model. In the kNN model, multiple classes are classified according to k neighbors and distance metrics. The kNN model is trained based on the distance metric as the loss function and k as the threshold to determine neighbors. In [Harrou 2019a], kNN is improved based on a set of moving average thresholds and is applied to detect short-circuit faults, open-circuit faults, and partial shading. The main advantage of the kNN model is its easy implementation, since it only requires determining the k neighbors and the distance function. On the other hand, the kNN depends to a great extent on the preprocessing of the data (scaling, normalization, etc). Furthermore, kNN is very sensitive to noisy values and with an increasing number and variety of inputs, computing the distance function can be computationally difficult.

3.2.2 Semi-Supervised Learning

This learning scheme typically uses a small amount of labeled data and a large set of unlabeled data for the learning process. The articles dealing with this approach first use a supervised learning algorithm trained on the labeled training set. Then, to deal with the unlabeled data these articles consider two options: i) use the supervised algorithm to predict the unlabeled data and its most reliable predictions are added to the training set or ii) use an unsupervised learning algorithm to produce data samples with new labels [Zhang 2021c] and add these labels to the labeled training set for the supervised learning algorithm. Some articles of the best known algorithms (Graph-Based Algorithms, Semi-supervised method based on probabilistic models, Positive unlabeled learning and N-semi-regular Fuzzy Semi-Supervised Learning) are presented below in detail.

3.2.2.1 Graph-Based Learning

In [Momeni 2020], a fault diagnosis approach in PV plants based on a semi-supervised learning process called graph-based learning (GBSSL) is proposed to extract hypotheses about the labels of unseen samples following a type of analysis, based on in a previously labeled data set. Two types of PV faults related to different cases of line-to-line faults are investigated using the same methodology to analyze the measured I(V) signals.

In [Zhao 2015b] it is proposed a graph-based semi-supervised learning model. That method uses a few training data that is labeled and normalized for better visualization. The model proposed in this approach not only detects the fault, but also identifies the possible type of fault to speed up system recovery. This model

presents one of the most interesting results in terms of fault detection in PV systems with high detection accuracy and fault classification.

In [Momeni 2020], it is presented as a method for identifying, classifying, locating, and correcting faults. The proposed method is an expansion of the diagnosis space of the graph-based semi-supervised learning algorithm with a larger number of class labels. This approach temporarily isolates the fault as soon as it is identified and located to keep the system running without interruption. A large set of data over a wide range of environmental conditions is collected to train the system. The proposed method demonstrates a high performance in fault detection under different weather conditions.

3.2.2.2 Based on probabilistic models

In [Maaløe 2020], it is proposed a semi-supervised learning based on probabilistic models (SSLPM), which performs condition monitoring in a photovoltaic system with high precision and only a small fraction of labeled data. The modeling approach utilizes all the unsupervised data by jointly learning a low-dimensional feature representation and a classification model in an end-to-end fashion. The feature representation detects new internal condition monitoring states, demonstrating a practical way to update the model for better monitoring and fault detection. The results of the proposed approach are compared with purely supervised approaches, and significant improvements in detection are achieved.

3.2.2.3 Positive unlabeled learning

In [Jaskie 2021], it is exposed as an approach based on a little-known area of semi-supervised machine learning named positive unlabeled learning (PUL) that can effectively learn solar fault detection models using only a fraction of labeled data. Based on this area, they propose a new feedback enhanced positive unlabeled learning algorithm that increases the accuracy and performance of the model in situations such as solar fault detection when few sensor functions are available. Likewise, the results are compared with supervised approaches and important improvements in classification accuracy are obtained.

3.2.2.4 N-semi-regular Fuzzy Semi-Supervised Learning

In [Murugesan 2020], it is proposed an N-semi-regular Fuzzy Semi-Supervised Learning (SRFSSL) System based on an N-semi graph with a few labeled, normalized training data for improved visualization. This model not only detects faults but also provides insights into the probable fault structure to facilitate corrective maintenance of the network. With this model, PV systems can learn to independently monitor and identify PV faults under environmental changes over time. Efficient detection and classification results on experimental and real data are obtained.

3.2.3 Reinforcement Learning

In this scheme of learning there are 4 essential elements: agent, environment, action and reward. The agent is the algorithm that is trained to do a task. The environment is the ecosystem in which the agent performs the tasks. The action is the movement made by the agent. Finally, the reward is the evaluation of an action that can be negative or positive. In the documents found in this topic, agents learn their ideal behavior in a particular situation based on previous experience [Coronato 2020]. Learning in this model is continuous through interactions with the environment and collecting of information to carry out the agent activity [Xu 2014]. Among the best known algorithms are dynamic programming (DP), Monte Carlo (MC) methods, Q-Learning, State-action-reward-state-action (SARSA) and Deep reinforcement learning (DRL). The reinforcement learning is used to train agents to complete tasks such as robots. The reinforcement learning problems include complex training, optimal weight initialization learned.

3.2.3.1 Deep reinforcement learning

In [Dai 2021], a fault diagnosis method in deep reinforcement learning (DRL) is proposed. In this approach first the compressed detection algorithm is used to fill in the missing PV data. Then state, action, strategy, and return functions from the environment are obtained. Using the interaction rules and other factors, the fault diagnosis model is established and the deep neural network is used to approximate the decision network to find the optimal strategy. The efficiency and precision of the method are verified by simulation. The results found in simulation show that this approach is an interesting alternative for fault detection in PV systems.

3.2.3.2 Other methods

In [Chen 2018a], it is presented as a fault detection algorithm that uses multiple meters to measure different PV system output signals. The time correlation of the faulty signal and the signal correlation between different meters in a vector autoregressive model in the modeling of the signal after the change. This article proposes an interesting approach to detect faults whose behavior is not previously known. For this, it uses a change detection algorithm based on the generalized local likelihood ratio test. The approach is validated in simulation and the results obtained demonstrate high adaptability and rapid detection when dealing with various types of faults in photovoltaic systems.

In [A.h.mohamed 2015], a new approach is proposed that uses a genetic algorithm that optimizes the topology of artificial neural networks (ANN). The method is used for the diagnosis and repair of photovoltaic (PV) energy systems dynamically online as a case study. The results obtained are compared with photovoltaic diagnosis systems based on traditional and fuzzy neural networks. These results demonstrated interesting study improvements for applications in PV systems.

In [Lin 2015], a monitoring system is presented that collects the output current and voltage of each photovoltaic module, and the temperature and irradiation of the place of installation of the PV system. With this data, a photovoltaic model with healthy and fault cases is built. This data is used for fault detection by training a back propagation neural network (BP) fault diagnosis model optimized by a genetic algorithm. As in [A.h.mohamed 2015], using a genetic algorithm together with another algorithm improves the detection results of common PV array faults with high accuracy.

In [Zheng 2017] it is proposed a hidden Markov model (HMM). In this approach the initial value of the HMM has a great influence on the model. The final results depend on the value of this parameter to achieve a local or global minimum in the training process. For this reason, Zheng et al. [Zheng 2017] propose to combine the hidden Markov model (HMM) with a genetic algorithm to optimize the initial value of the HMM. This approach is used in photovoltaic inverter fault diagnosis. First, the genetic algorithm implemented in Matlab determines the optimal initial value of HMM. Second, iterative training is done using a Baum-Welch algorithm. Finally, a Viterbi algorithm is used for fault identification. The results are compared with conventional methods and an increase of 13% is obtained.

As can be seen in the articles exposed in Section 3.2.3.2, due to the little scientific production that exists around reinforcement learning for fault diagnosis, many of the articles are grouped automatically by the proposed methodology (Smart B2TE), although they do not deal directly with the implementation of reinforcement learning, they do deal with methodologies that have already been explored as direct improvements of reinforcement learning [Abbas 2022, Yoon 2019, Bakker 2007]. This further reinforces our hypothesis that this methodology is not only capable of finding obvious relationships, but also hidden ones.

3.2.4 Unsupervised Learning

This learning scheme contains only one unlabeled dataset, i.e. whose relevant result is not clear [Seaton 2021]. Articles using this approach attempt to discover data patterns and relationships in the data. In this approach the data are compared based on their similarity scale to classify them into groups. Among the best known algorithms are K-means clustering, Hierarchical clustering and Fuzzy-c-means, among others.

The unsupervised machine learning is more used in the complex system that is required to handle complex patterns or processes reference data training, the algorithms learn inherent structure from the input data. There are many drawbacks in the unsupervised learning models such as clustering problem, and association problem. Some based on clustering methods such as Fuzzy C-means [Tsai 2015] at the solar cell level or hierarchical clustering at the solar panel level are proposed [Sepúlveda Oviedo 2021, Tsai 2015]. Some articles of the best known algorithms are presented below.

3.2.4.1 Based on clustering

In [Liu 2021b], a fault diagnosis method for PV arrays based on a stacked auto encoder and clustering algorithm is proposed, which can automatically extract features and use a small number of labeled data samples to extract features for fault diagnosis. In this approach the stacked auto encoder first extracts the features from the current-voltage curves. Then t-distributed stochastic neighbor embedding is used to perform dimensionality reduction, improving the performance of the clustering algorithm. Finally, the cluster centers and clusters are obtained by the clustering algorithm and the membership function is used for fault diagnosis. Experimental and simulated data are used to validate the model and classification accuracies of 97.3% and 98.3%, respectively, are obtained.

3.2.4.2 Based on weighted K nearest neighbor

In [Reddy 2021], it is presented as a two-layer machine learning scheme based on weighted K nearest neighbor (WKNN) and decision tree (DT). The WKNN method detects the fault in the line and then the DT classifies it as pole to pole (PP) or pole to ground (PG) improving preventive and corrective maintenance. The features used are the network voltage and current. This approach presents precision results of up to 100%.

3.2.4.3 Unsupervised cascading algorithm

In [Zhao 2019b], it is presented as a two-methods pronostic, composite approach based on data extracted from a SCADA supervision platform. The first method is a hierarchical context-aware anomaly detection algorithm using unsupervised learning. The second method is a multimodal anomaly classification algorithm. The proposed solution is validated in two solar parks of sizes 39.36 and 21.62 MWp. The results of fault detection in the field demonstrate the effectiveness, robustness, and cost and computing efficiency of the proposed approach.

3.2.4.4 Sparse autocoding

In [Manohar 2019], it is proposed a protection scheme based on sparse autocoding (SAE) and deep neural network that discriminates between matrix faults and symmetrical line faults. The features used are the voltage-current signals recovered from the retransmission buses which are then converted to grayscale images. These images are used as input to the SAE to perform unsupervised function learning. The proposed method is compared with techniques based on artificial neural networks, support vector machines and decision trees in islanding mode and connected to the microgrid network. Furthermore, the approach is validated using simulation data in real time with the OPAL-RT digital simulator. In all scenarios, this approach presents great accuracy in fault detection.

3.2.4.5 Self-organizing map neural network

In [Vyas 2016], it works on the formation of unintentional islanding in the integration of distributed photovoltaic solar generation with a distribution network. To this end, the application of a self-organizing map neural network for the preemptive detection of unintentional islanding by classifying the discovered islanding precursor from other power system events is proposed. The approach is validated for classification of a three-phase short-circuit fault at the point of common coupling. Under other scenarios tested, a high detection accuracy is also obtained.

3.3 Promising Research Topics and Challenges

In this section, a synthesis of the main promising research topics is provided. First, as an evidence mentioned in several parts of the articles, in particular the ones described in Section 3.1.5, the use of *hybrid machine learning techniques* becomes more essential each day to increase the level of accuracy in fault diagnosis methods in PV systems. In some cases, authors propose to mix two or more algorithms from different clusters to obtain joint benefits. These types of approaches called hybrid machine learning techniques [Rodrigues 2017] show that combining several approaches improves the precision on results both in diagnosis and prognostic systems [Sepúlveda Oviedo 2022, Antonanzas 2016]. The most cited articles using hybrid techniques is [Dhimish 2018b] with 131 citations, it proposes a new fault diagnosis algorithm for photovoltaic (PV) systems with an innovative approach combining radial basis function (RBF) ANN network and both Mamdani, Sugeno fuzzy logic systems through a new interface. [Lu 2019a] with 42 citations makes a choice to orient its work to detect DC arc faults. It proposes a domain adaptation combined with deep convolutional generative adversarial network (DA-DCGAN)-based methodology. This method is implemented in an embedded system and tested in real time according to the UL-1699B standard. A real 1.5 kW rooftop photovoltaic grid-connected system is used to validate this approach.

In general, these hybrid techniques show a significant increase in the accuracy and robustness of fault diagnosis algorithms. For this reason, it seems one of the most interesting ways to be supported by researchers to continue the development of this type of new hybrid techniques taking combined advantages and obtaining better diagnosis results. This is especially relevant with the recent important progresses of both software and hardware technologies capacities. Today, it is realistic that researchers imagine to develop methodologies even complex that can be embedded in real systems pushing back actual limits of these systems on sampled measures and real execution conditions. Likewise, the authors are encouraged to develop hardware works for the acquisition of quality data oriented to fault diagnosis [Spanias 2017].

Analyzing the algorithms of Subsection 3.1.5.1, it can be seen that random forest and decision tree algorithms are the most used. In general, the Symbolic approach presents good results. There exist few experiments using other less explored algorithms from the same cluster, for this reason we recommend increasing research on

other algorithms such as the C4.5 Decision Tree or the binary Tree Bagging.

In Subsection 3.1.5.2, most of the works are oriented to use the PCA algorithm as a dimensionality reducer for a subsequent data treatment with a classifier. However, there are other options such as signal processing using ARIMA models or regression models that allow fault diagnosis based on prediction. Prediction-based fault diagnosis compares the predicted values with the real values and, based on a set of limit parameters establishing a priori, the existence of a fault is determined, including the level of fault done in these systems or the type of fault.

In Subsection 3.1.5.3, the results of these algorithms are really very interesting, especially on the advantages obtained with methods based on Semi-Supervised Learning based on Graphs. Moreover, the authors are encouraged to further investigate and apply this type of algorithm in combination with the data normalizations proposed in those articles. Data normalization can open a window to increase the size of databases and to the generalization of fault diagnosis models applied on PV systems.

Analyzing the results of Subsection 3.1.5.4, a conclusion can be that both kNN and SVM are interestingly applied. It is recommended to continue working with these algorithms due to their results, but it is also recommended to investigate other methods of this cluster such as the local outlier factor or some kernel modifications for the SVM.

In general the algorithms of the section 3.1.5.5 require large amounts of data and therefore a high computational potential. However, with the accelerated technological evolution which is no longer a limitation. The authors are strongly encouraged to further investigate these algorithms and explore algorithms such as “Extreme Learning Machine" that show very accurate results, or methods based on pre-trained neural networks to reduce the calculation time of training.

In Subsection 3.1.5.6 it is confirmed that there are very few articles using these algorithms. The authors are encouraged to explore more generally metaheuristic algorithms that allow estimating parameters of the models and thus adjust the fault diagnosis models. The most used algorithms in this area of fault diagnosis in PV systems are related to genetic algorithms, however there are multiple algorithms that show more accurate results and with less calculation time for the same purposes [Banerjee 2022].

If an analysis is carried out regarding the 4 types of machine learning, it is possible to notice that the results of supervised learning (topic 1), it can be seen that classic algorithms such as random forest, decision trees, support vector machine (SVM) or neural networks are widely used. For this reason, researchers are invited to explore the use of other algorithms from this broad family of algorithms belonging to this topic such as; K-nearest neighbors, Adaptive Boosting, Discriminant Analysis, Naive Bayes, Kernel SVM, Linear Support Vector Classification, stochastic gradient descent, binary decision tree, deep neural networks, recurrent neural networks, generative adversarial networks, Generalized linear model, Gaussian Process Regression, elasticNet lasso, ridge regression SVR, Ordinary Least Squares regression, among others, which are scarcely used in fault diagnosis in PV

systems.

With the results of Semi-Supervised Learning (topic 2), it can be seen that there are very few documents that deal with this approach. The few works that use this approach demonstrate important contributions to fault diagnosis in PV systems and also demonstrate that they are easily implementable in real life. In addition, they do not need large amounts of labeled data, so their real applications are greater. Researchers are strongly encouraged to continue exploring the algorithms mentioned in the section 3.2.2, as well as to propose new approaches based on co-training, re-Weighting, deep belief network, restricted Bolzmann Machine or to propose new semi algorithms. In topic 3 of reinforcement learning, all the articles published are very recent. This means that it is a hot topic in research on fault diagnosis in PV systems. For this reason, researchers are invited to continue exploring the algorithms mentioned in the section 3.2.3 and explore the use of other algorithms such as agent based systems, heuristic methods, dynamic programming, thompson sampling, upper Confidence bound, temporal difference methods, Markov decision process, Q-Learning, policy optimization, collaborative adaptive and fuzzy-Q.

In topic 4 of unsupervised Learning, the great difficulty to detect that unsupervised algorithms have to detect and classify faults in PV systems can be observed. This difficulty has given rise to the creation of well elaborated algorithms such as DyD^2 used in satellite fault diagnosis but that could be extrapolated to the energy domain [Dorise 2022]. Additionally, this work proposes researchers to explore algorithms such as DBScan, Hierarchical Cluster Analysis [Sepúlveda Oviedo 2021], K-medoids, Fuzzy C-Means, Gaussian Mixture, meanShift VBGMM, miniBatch Kmeans, spectral clustering GMM, one-Class -SVM, isolation forest, principal component analysis, locally linear embedding, t-SNE, autoencoders, isomap, spectral embedding, auto-associative NN, Manifold, boltzmann machine, kernel density, gaussian mixtures, among others.

Another interesting aspect to work on is data normalization. This normalization allows comparison of data from different PV plants. Consequently, standardization allows the fault databases to be increased more quickly. Taking into account that, among the challenges is building a labeled database of all types of faults in order to effectively train PV systems. Likewise, the construction of a fault dictionary is proposed where the characteristics that allow differentiating between faults are recorded. On the other hand, the fault diagnosis results are closely linked to the quality of the data acquired in the PV plant. For this reason, one of the greatest challenges is linked to the construction of data capture systems aimed at fault diagnosis, that is, that guarantee the quality of the data and the richness of the signals. This allows comparing fault diagnosis using different sampling times and determining the influence of this on the accuracy of fault diagnosis and classification, especially in cases of faults with similar signatures.

Similarly, the accuracy of all error diagnosis algorithms is directly linked to the quality of the data which is trained. This is why authors are encouraged to pay more attention to signature extraction and selection processes as in the papers [Sepúlveda Oviedo 2022, Dhibi 2020]. Due to the complexity of generating labels

to train systems with all types of faults, it is possible to identify that an interesting possibility is to work on the diagnosis of anomalies where no fault labels are needed. In this approach, also called outlier diagnosis or novelty diagnosis, the aim is to identify rare items, events, or observations that deviate significantly from the majority or appear inconsistent with "healthy" or baseline data. In anomaly diagnosis on photovoltaic systems there are already some important and well-cited records about it [Pereira 2018, Hu 2017b, Zhao 2019b, Liu 2017, Benninger 2020]. For this reason, we also believe that it would be interesting to strongly initiate research into the diagnosis of anomalies in the PV domain.

Research on fault diagnosis in photovoltaic systems must evolve towards the acquisition of relevant data for monitoring and fault diagnosis in real conditions. This means that model-based training of mathematical analogies of systems should be increasingly avoided, in favor of training systems based on data, models built from data, or mathematical models that are fitted to real data. This ensures a more adequate representation of the system and therefore a better and more adapted response from machine learning systems.

Finally, some new algorithms present interesting results for fault diagnosis in PV systems, such as the one entitled “Online Fault Diagnosis for Photovoltaic Arrays Based on Fisher Discrimination Dictionary Learning for Sparse Representation" [Xi 2021]. A set of challenges linked to these promising research topics are identified and are presented below.

3.4 Discussion and conclusions

The art review methodology presented in this chapter, and supported by the visual representation of Figures 3.6 and 3.7, takes advantage of the latest advances in natural language processing and statistics to provide details about research behavior in an area of knowledge. With this methodology, 625 articles on fault diagnosis in photovoltaic systems using artificial intelligence are analyzed for the first time. These articles are classified into 6 families of algorithms (Symbolic approach, Regression approach, Bayesian approach, Analogy-based approach, Connectionist approach and Evolutionary approach), grouped in turn into 4 types of machine learning (Supervised Learning, Semi-Supervised Learning, Reinforcement Learning and Unsupervised Learning).

Multiple articles from these classifications are studied to demonstrate their respective advantages, limitations, and possible future research. This methodology is a tool that can be easily extrapolated to any scientific area from the perspective of Big Data. It is important to mention that this research did not take into account articles that only present the general framework of the subject without experimental research. The main objective of this research is to promote the development of approaches and tools for fault diagnosis in photovoltaic systems. For this reason, this research focuses its efforts on providing an objective, coherent and meta-analytic vision of current research on sustainable artificial intelligence applied to the energy

domain.

On the other hand, the results of this chapter encourage researchers to work on hybrid machine learning and comparison studies of these methods in fault diagnosis in photovoltaic systems. Similarly, research into photovoltaic system forecasts must be encouraged to provide condition-based preventive maintenance and to reduce investment payback times. Even though industrial demands tend to be directly oriented to complete PV arrays, monitoring at the photovoltaic panel level still has a long way to go and must be built in parallel with machine learning algorithms. In addition, not only should research be carried out to improve monitoring systems, but also work hard on feature engineering that can extract the necessary characteristics to differentiate multiple types of faults in PV plants.

The results of the innovative approach presented in this chapter confirm, as expected, that researchers should continue on the path of developing new algorithms and tools for fault diagnosis in PV systems using artificial intelligence and provide a highly objective, consistent approach. and above. and meta-analytic view of current research on sustainable AI in energy.

Taking into account all the aforementioned aspects, in the next chapters 4-6 vital issues of data acquisition and feature engineering are addressed, aimed at improving the accuracy of fault diagnosis in PV plants.

Chapter 4

Conventional Data Acquisition in PV plants

Contents

4.1 Characteristics of a data acquisition system 131
4.1.1 Measured parameters 133
4.1.2 Data acquisition system 137
4.1.3 Wired communication 137
4.1.4 Wireless communication 138
4.1.5 Controller 139
4.1.6 Sample rate 139
4.1.7 Data preprocessing 140
4.1.8 Signal treatment methods 142
4.1.9 Existing data acquisition systems 144
4.2 Tigo industrial and commercial data acquisition system . . 148
4.2.1 Components and Connection Scheme 148
4.2.2 Instrumented PV plant 149
4.3 Discussion and Conclusions 152

As it was exposed in Chapter 3, the performance of machine learning algorithms depends directly on the quality of the data with which they are trained. For this reason, this chapter focuses on the analysis of the characteristics that a data acquisition system for photovoltaic systems must have, thinking of a subsequent fault diagnosis. In this way, Chapter 4 presents a review of current data acquisition systems in photovoltaic systems, their limitations, advantages, disadvantages and development challenges. In addition, an industrial and commercial Tigo platform used in this thesis is presented to test the limitations of fault detection in PV plants using widely commercialized data acquisition systems. The components of the Tigo data acquisition system and the instrumented PV plant are also described in this chapter.

4.1 Characteristics of a data acquisition system

Due to the intermittent nature of solar energy, the power output of a photovoltaic system can increase or decrease dramatically causing increased stress on the grid

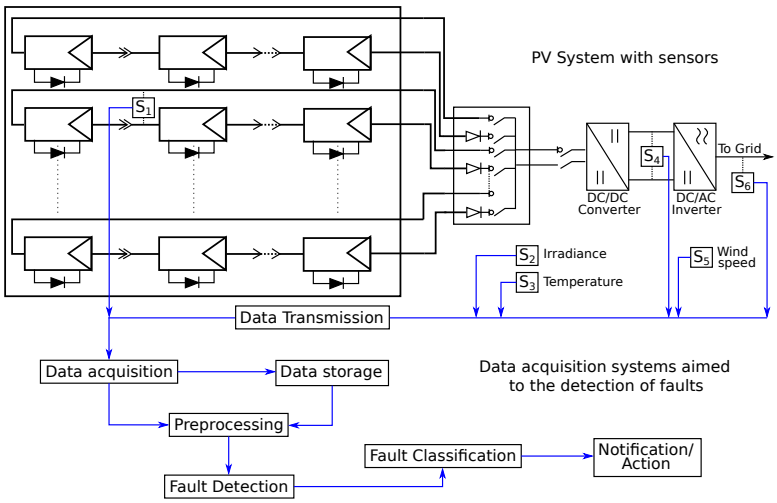

Figure 4.1: A typical data acquisition system aimed at the detection and classification of faults in PV plants. PV plant composed of PV strings, junction box, and inverter. Detection and classification system fed by directly preprocessed data or preprocessed data stored in a database.

or even outage [IEEE 2003]. In addition, the existence of faults in these systems makes continuous data acquisition of PV plants vital to monitor performance and be able to act in case of production losses. An important aspect of a reliable, robust and usable data acquisition system for fault detection is the inclusion of simultaneous, high-quality and relevant measurements of environmental conditions (solar irradiation, ambient temperature, wind speed, humidity, etc.) and the electrical operating data of the photovoltaic system [IFC 2015, Triki-Lahiani 2018a]. Other documents have reported that another important aspect is the relationship between the accuracy of fault detection and the data sampling rate [Cross 2018].

Generally, a typical data acquisition systems aimed at the detection and classification of faults in PV plants is composed of two parts: *i)* instrumentation installed in the PV plant (sensors); and *ii)* a set of elements responsible for the data acquisition and transmission, storage and preprocessing of data and storing the algorithms for detection and classification of faults as shown in Figure 4.1.

To carry out the instrumentation of a PV plant, it must be taken into account that, as shown in Figure 4.1, the locations of the sensors can vary, depending on the parameters and the type of faults that are being considered. As mentioned in [Hong 2022b], sensors can be located in PV modules, PV arrays, between PV array and DC/DC converter, between DC/DC converter and DC/AC inverter, or after the DC/AC inverter. Regarding the meteorological sensors, these can be located based on factors such as the inclination of the PV string, the temperature of interest (ambient or surface), etc.

In the same Figure 4.1, it can be seen that the transmission of the data measured by the sensors is coupled to the data acquisition device that can take two actions:

1) To send the data directly to pre-processing; or *2)* To send it to storage for further processing. In either case, once the preprocessing is done, the data is sent to the detection and fault classification stages consecutively. Finally, once the fault is detected and classified, a notification is sent allowing for a corrective action to be taken.

However, in order to monitor all these electrical and meteorological variables of the PV plant, dedicated components are used, such as sensors, data acquisition systems, data communication systems, and dedicated algorithms for data analysis. The appropriate size of the data acquisition system depends directly on the size of the plant, the criticality of the system and the operation and maintenance costs [Cristaldi 2015]. Likewise, it is necessary to pay attention to the signal communication mode of the sensors [Triki-Lahiani 2018a]. Adopting wired sensors in small plants is inexpensive and less complex. However, wireless networks are more suitable for medium and large plants.

The architecture of the data acquisition systems aimed at the detection and classification of faults in PV plants can be divided into three levels. At the first level are the sensors. The quality of the data sent by these sensors must be guaranteed for the construction of an accurate and reliable database. At the second level is data treatment, which includes measurements and pre-processing of data (for example, filling in missing data or removing outliers) using specific hardware and communication networks. The final level is the most flexible and adapted depending on the measured variables of the PV plant. This final level consists of the implementation of analytical techniques that lead to evaluating and estimating the performance of photovoltaic systems, that is, to the detection and classification of faults [Cristaldi 2016].

Below is an in-depth analysis of the main aspects of data acquisition systems, including sensors, data acquisition and controllers, data transmission and data storage, as discussed in [Madeti 2017d]. In general, it presents a brief description of the key aspects of the architecture of typical data acquisition systems aimed at the detection and classification of faults in PV plants.

4.1.1 Measured parameters

Knowing the layout of the measurement elements (sensors) and the information processing (pre-processing and fault detection and classification) does not guarantee the correct operation of the PV plant. It is necessary to carry out a rigorous study of the parameters to be monitored. There are international standards that establish those parameters for monitoring PV plants. For example, the British standard BS IEC 61724 [IEC 1998] provides guidelines for analyzing the performance of photovoltaic systems. The list of variables proposed in the British standard BS IEC 61724, the guidelines of the European Joint Research Center [Blaesser 1995] and the guidelines of the National Renewable Energy Laboratory (NREL) [Kurtz 2013] are shown in Table 4.1. In the same table the precision and units of each variable are presented.

Table 4.1: Parameters to be measured according to BS IEC 61724 [IEC 1998].

General parameter	Specific parameter	Notation	Unit	Accuracy
Meteorology	In-plane irradiance	G_I	Wm^2	±3%
	Ambient temperature	T_{am}	$°C$	$\pm 1°C$
	Wind speed	S_w	ms^{-1}	$\pm 0.5\, ms^{-1}$
	Wind direction	D_w	degrees	±5°
Photovoltaic array	Output voltage	V_A	V	±2%
	Output current	I_A	I	±2%
	Output power	P_A	P	±2%
	Module temperature	T_m	$°C$	$\pm 1°C$
Energy storage	Operating voltage	V_S	V	±2%
	Current to storage	I_{rs}	I	±2%
	Current from storage	I_{FS}	I	±2%
	Power to storage	P_{rs}	kW	±2%
	Power from storage	P_{FS}	kW	±2%
Load	Load voltage	V_L	V	±2%
	Load current	I_L	I	±2%
	Load power	P_L	kW	±2%
Utility grid	Utility voltage	V_U	V	±2%
	Current to utility grid	I_{ru}	I	±2%
	Current from utility grid	I_{FU}	I	±2%
	Power to utility grid	P_{ru}	kW	±2%
	Power from utility grid	P_{FU}	kW	±2%
Back up sources	Output voltage	V_{BU}	V	±2%
	Output current	I_{BU}	I	±2%
	Output power	P_{BU}	kW	±2%

It is important to note that of all the parameters listed in Table 4.1, the most prominent and generic parameters are solar radiation, wind speed, temperature, voltage, current and consequently photovoltaic power, while the other parameters depend on the configuration. An illustration of the correspondence of the electrical and environmental data to be measured according to the British standard BS IEC 61724 [IEC 1998] is presented in Figure 4.2.

Other authors have proposed the acquisition of additional variables such as rainfall and humidity for evaluating the performance of a PV plant [Livera 2019b]. Rainfall measurements can also be used to estimate module cleanliness and thus for estimation of soiling losses. Humidity measurements may be more useful in research related to the degradation of photovoltaic materials. The following sections present the most outstanding and generic parameters mentioned in Table 4.1.

4.1.1.1 In-plane irradiance or Total irradiance

Irradiation is recognized as the main component that affects the power output of PV plants [Li 2021c]. In-plane irradiance (G_I), recorded in the same plane as the PV array. The most used devices to measure irradiation are thermopile pyranometers, photovoltaic reference devices (reference cells and modules) and photodiode sensors [Livera 2019b, Friesen 2018]. There are two types of pyranometer: thermopile and photodiode pyranometer. A thermopile pyranometer measures the irradiance in the range of 300 to 2800 nm with a flat spectral sensitivity, while the photodiode measures a portion of the solar spectrum between 400 nm and 1100 nm [Stoffel 2012].

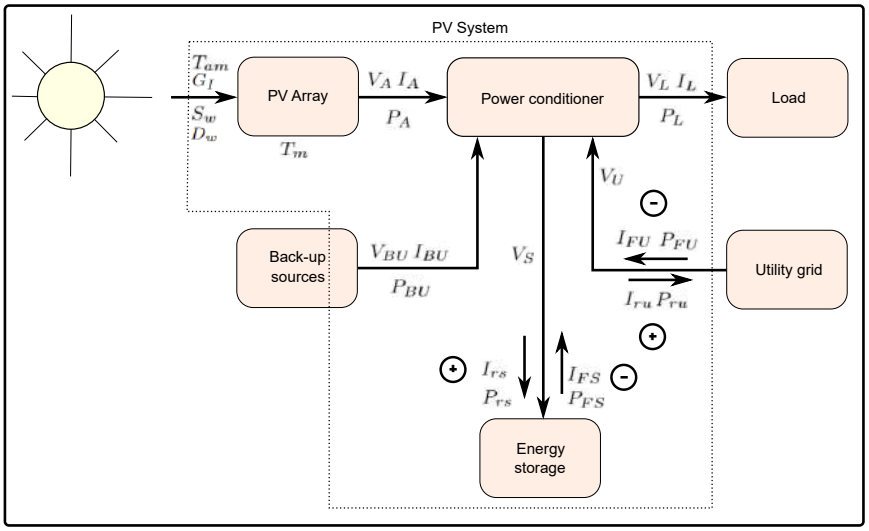

Figure 4.2: Diagram of the most relevant parameters to be measured in real time, according to the IEC 61724 standard [IEC 1998].

If reference cells or modules are used, they must be calibrated and maintained in accordance with the IEC 60904-2 [IEC 2015] or IEC 60904-6 [IEC 1994b] standards. In addition, it must be ensured that the reference cell is made of the same material as the PV module [Dunn 2012]. Likewise, it must be taken into account that the location of the sensors must be representative of the irradiance conditions of the PV plant [IEC 1998]. It is recommended that the accuracy of irradiation sensors, including signal conditioning, should be between 3% and 5%. For performance analysis or health monitoring of PV modules, the In-plane irradiance (G_I) of a PV module is commonly adopted [Friesen 2018]. In [Polo 2017] a performance analysis of a small-scale plant is presented. In that study, a pyranometer and modules in short-circuit conditions are used as irradiance sensors. The results show a higher precision when using short-circuited cadmium telluride "CdtTe" modules as irradiation sensor, which can be an interesting option for small-scale photovoltaic arrays.

4.1.1.2 Ambient and module temperature

Temperature is defined as the second most influential parameter in the power output of PV plants especially when the temperature is above the value of 25 $°C$ defined in the Standard Test Condition (STC) [Li 2021c]. The problem in measuring these two variables is the location of the sensors. To capture this data, it must be ensured that the sensors are able to represent the conditions of the panel. The ambient temperature (T_{am}) data is measured by outdoor protected temperature sensors, located away from direct solar radiation. According to the IEC 61724 standard, the accuracy of these temperature sensors, including signal conditioning, must be better than 1 K or $\pm$ 0.5 $°C$.

It is also possible to measure temperature using thermistors, thermocouples, and resistance temperature devices (RTDs). If the aim is to measure the module temperature (T_m), there are generally two methods. The first method is contact based; in this method a sensor is attached to the back sheet of the module [Livera 2019a]. The second method is non-contact or indirect measurement. In this method the temperature (T_m) is estimated from ambient temperature (T_{am}) and irradiance (G_I) [Kratochvil 2004], or extracted from the relationship between open circuit voltage (V_{oc}) and (T_m) [IEC 2011]. The last method relies on the use of calibrated infrared cameras [IEC 2017b].

4.1.1.3 Wind speed and direction

Wind speed (S_w) should be measured at the same array conditions, that is, ideally at the same height of the PV plant and avoiding obstacles that introduce noise into the measurements [IEC 1998]. Wind speed sensors accuracy should be such that the error is no higher than $0.5ms^{-1}$ for values measured for less than $5ms^{-1}$, and no higher than 10% of reading for wind speed values greater than $5ms^{-1}$. The wind direction of the wind should be measured under the same array conditions [IEC 1998]. The accuracy of the wind direction (D_w) sensors should be $\pm 5^{\circ}$ degrees.

4.1.1.4 Voltage and current

The values of voltage (V_A, V_S, V_U and V_{BU}) and current ($I_A, I_{rs}, I_{FS}, I_L, I_{ru}, I_{FU}$ and I_{BU}) are some of the main electrical measurements related to a PV plant both on the DC and AC side [Madeti 2017d]. According to the IEC 61724 standard [IEC 1998] the accuracy of current and voltage sensors, including signal conditioning, must be better than 1% of reading. Conventionally the capture of this data is done using shunts or current transducers or voltage transducers [Livera 2019a]. The advantage of shunts is that they do not require an additional power supply, unlike current transducers. However, on the other hand, shunts require significant sampling frequency control to measure alternating current [Triki-Lahiani 2018a]. Current transducers have excellent accuracy, low temperature drift, good linearity, optimized response time, good immunity to external interference, no insertion loss, and no current overload loss [Shariff 2015]. The voltage transducers also have excellent accuracy, low thermal drift, low common mode disturbance, good linearity, and good immunity to external interference. Transducers are the most suitable solution to capture high precision measurements. But, it is necessary to find a balance between the quality of the data and the cost of the transducers, which, depending on the case, can be very high.

4.1.1.5 Power

Like current and voltage, power is a vital variable to describe the behavior of a PV plant [Madeti 2017d]. The power data ($P_A, P_{rs}, P_{FS}, P_{ru}, P_{FU}, P_{BU}$) can be measured on the DC or AC side or on both sides. It can be measured directly

by power sensors or calculated in real time as sampled voltage and current values [IEC 1998, Fuentes 2014]. Power sensors on the AC side take into account power factor and harmonic distortion. DC input power on stand-alone systems can have considerable AC ripple, so a DC wattmeter is more suitable for measuring DC power. Accuracy of power sensors, including signal conditioning, must be better than 2% of reading. Another conventional way of obtaining the power measurement is calculating it in real time with the sampled voltage and current values [Livera 2019b]. Power is typically measured on both the DC and AC sides of the PV plant.

4.1.2 Data acquisition system

The elements of the data acquisition system are responsible for collecting and storing data for fault analysis. In PV plants, the elements of the data acquisition system are used to monitor performance and control operations. According to Livera et al. [Livera 2019b], the elements of the data acquisition system are generally composed of five parameters: data transfer mechanism, monitored data, controllers, sampling intervals, and program development software. Examples of the elements of the data acquisition system are discussed in [Madeti 2017d] and [Shariff 2015] including data transfer mechanisms, wired, wireless and power line communication systems.

Data may be transmitted using wireless or wired technologies [Ansari 2021]. The choice of the best transmission medium is subject to the configuration of the equipment, the location, the data traffic, the cost of maintenance and the type of control system (centralized or decentralized) [Heilscher 2020]. As criteria for identifying the best data transmission medium, the following must be taken into account: latency, data speed, reliability, security, interoperability and scalability [Michael Mills-Price 2014]. The explanation of the two communication technologies (wired and wireless) is given below.

4.1.3 Wired communication

Several types of wire connection are available for transmitting digital data. First is a coaxial cable as one of the most well-known transmission media. This kind of cable has low resistance, low error rate and good data transmission rate of 10 Gbps. Its limitations are the distance covered and the display [Madeti 2017d]. Second, the MODBUS rtu RS 232 or RS 485 protocol can be used [Cristaldi 2010, Ben Belghith 2014, Tejwani 2014, Shariff 2013]. This protocol has some limitations, such as the distance covered, and is less favorable than wireless systems for monitoring large photovoltaic plants. In third place, the fiber optic cables are positioned. These cables can be used for long distances and offer a high data transmission rate of 100 Mbps to 200 Mbps [Madeti 2017d]. However, it should be clarified that it is an expensive solution. Finally, the Power Line Communication (PLC) is positioned. The PLC is considered a better option for data transmission compared to Zigbee, WiFi, Bluetooth and RS 485 [Han 2014]. The

PLC has a data transmission speed of around 200 Mbps.

4.1.4 Wireless communication

Unlike wired communication, there are multiple wireless data transmission options presented in the literature.

The first is satellite transfer. This means of data transmission is described as a slow and expensive mechanism [Drews 2007, Drews 2004].

Second is the widely known Global System for Mobile (GSM). This mechanism is described as reliable and accurate [Ben Belghith 2014, Tejwani 2014, Shariff 2013]. In the GSM mechanism, data is transmitted through Short Message Service (SMS) or General Packet Radio Service (GPRS). This type of transmission allows the sending of data at high speed and large volumes of data. Furthermore, this protocol features low retransmission and low data loss rates. The main drawback of this protocol is that it has a high cost of operation, since users must pay for this service [Triki-Lahiani 2018a].

The third is the transmission of data by radio frequency that allows sending and receiving information at very large distances. This mechanism is an interesting alternative in areas without telephone lines. However, it must be taken into account that the implementation can be complicated due to the authorisation of use of a given transmission frequency and its cost [Triki-Lahiani 2018a].

The fourth is the wireless local area network (WLAN) [Madeti 2017d]. The advantage of the WLAN network is that it covers a large area; it is flexible in data transmission and can communicate without future restrictions. The difficulty with this transmission mechanism is that it has lower bandwidth and lower quality of service due to interference.

The fifth corresponds to the File Transfer Protocol (FTP) server. This mechanism is an option for data transmission via GSM-GPRS. Conventionally, this mechanism is embedded in a kind of PC board that can be connected via USB or serial cable [Triki-Lahiani 2018a].

A sixth mechanism found in the literature is Bluetooth [Hua 2009] is a simple network but does not cover long distances, at most 100m.

A seventh mechanism explored is the Zigbee device. This device is considered as the best solution and the most economical alternative for data acquisition systems. This mechanism allocates a special time slot to avoid data collision, and its topology allows the integration of other wireless nodes, making it upgradable to support large network capacity.

An eighth mechanism is Wi-Fi and WiMax, which have also been used for data transmission [Triki-Lahiani 2018a, Michael Mills-Price 2014] and which have a high data transmission speed, but are more expensive than other technologies such as Bluetooth or Zigbee [Triki-Lahiani 2018a].

A ninth is the Internet/Ethernet protocol commonly known as TCP/IP is considered the most convenient, especially for real-time data acquisition systems [Bressan 2013].

Finally, IEEE 2030.5-2018 (Intelligent Power Profile Application Protocol) has also been proposed as an effective mechanism for data transmission [Michael Mills-Price 2014].

Apart from the aforementioned aspects, for a correct data acquisition, special attention must be paid to the controller.

4.1.5 Controller

This device is the interface between the sensors and the end users. As a controller, some systems propose the use of microcontrollers or data acquisition boards or modules [Cristaldi 2010, Bayrak 2013]. The advantage of boards and modules is that they are easier to program, however, their price is higher than that of a [Triki-Lahiani 2018a] microcontroller. On the other hand, a microcontroller is characterized by the resolution of the analog digital converter (ADC), which represents the most important factor for the accuracy of the monitored data. However, it is important to note that the quality of the data captured in the microcontroller is directly related to the resolution of the ADC [Triki-Lahiani 2018a]. In addition to the controller, attention should be paid to the development programming language of the data capture and processing software. For the programming of microcontrollers and acquisition systems in general, languages such as C and MATLAB [Triki-Lahiani 2018a] are used. One of the most popular options is LABVIEW, mostly in academic research, but not only, due to its high licensing costs. LABVIEW is generally used as system design software providing comprehensive measurement and control tools [Cristaldi 2010, Bayrak 2013].

4.1.6 Sample rate

Attention should be paid to the selection of sensor technology, resolution and data aggregation. According to [Hong 2022b], the most important requirements to consider for sensors are their accuracy, reliability, stability, calibration, maintenance, design simplicity and sensor cost, as well as installation conditions. Another important aspect for data acquisition aimed at fault diagnosis and data acquisition is the data sampling frequency.

In the literature, various sampling intervals from seconds to one hour are proposed. But, according to IEC 61724 [IEC 1998], the sampling interval should be selected based on the types of parameters, depending on the measured variables (see Table 4.1) and on whether the photovoltaic system is connected to the network or not. According to the IEC 61724 standard [IEC 1998] for the irradiation, temperature, wind and electrical output parameters, the data sampling period must ideally be less than 3 seconds (s) and in cases of hardware limitations priority should be given to more than 1 minute (m). It is also mentioned that for the parameters of dirt, rain and humidity the maximum sampling interval is 1 minute (min). Lastly, it mentions that, for parameters with a long sampling time, they must have a sampling frequency between 1 and 5 minutes [IEC 1998]. Despite the recommendations

proposed in the IEC 61724 [IEC 1998] standard, it is also mentioned that the data acquisition period should be sufficient to provide representative PV operating data and environmental conditions for fault diagnosis routines.

Other technical reports such as the International Energy Agency (IEA) technical report [Woyte 2014] suggest that the data should be sampled every second or faster and the averaged values should be stored every 5-15 min. Also, tracking data availability must be 99% or higher. Data availability of less than 95% indicates a poor quality data acquisition system.

Another recommendation is provided by SolarPower Europe [Europe 2016] in their report entitled best O&M practices, where they mention that irradiance should be stored at maximum average intervals of 15 minutes, and a fine resolution is achieved with averages of 1 minute. For this reason, it is mentioned that satellite-based irradiation measurements should have a sampling time of at least 15 minutes min. It is also mentioned that the electrical measurements of the input DC voltage and current must be sampled with a resolution of less than $1s$ and also averaged over a range of less than 1 min. Along the same lines, in the Australian PV data acquisition guidelines [Copper 2013b], the maximum recommended sampling interval for all averaged parameters should be set to 1 s. The data can then be averaged and logged over the logging interval. Also, it is recommended in the same guidelines that in case power is calculated from sampled current and voltage measurements, the sampling interval should be significantly less than 1 s. The World Meteorological Office (WMO) recommends that irradiation observations be made at a sampling interval less than $1/e = 0.368$ of the time constant of the measuring instrument. The time constant of a sensor can be understood as the time it takes, after a step change in the measured variable, for the instrument to register 63.2% of the step change in the measured parameter. Lastly, it is recommended that the sampled data for each measured parameter should be processed into time-weighted averages.

All these reports agree that the recording interval depends on the final use of the measured data of the photovoltaic system. The data acquisition period must be sufficient to provide operational data that is representative of PV performance and environmental conditions. The minimum continuous data acquisition period should be chosen according to the collected data (see Table 4.1).

The availability of high-quality data along with the use of proper algorithms for fault diagnosis will ensure reliable and stable operation of the photovoltaic system, which is crucial for the reduction of costs associated with operation and maintenance and time of system inactivity. In addition to ensuring all the aforementioned aspects, it is necessary to take into account a stage of processing the captured data.

4.1.7 Data preprocessing

The data captured by the sensors in a PV plant may or may not be pre-processed. That depends on the requirements of the applied PV fault detection and classification method. Data pre-processing is crucial in detection and classification problems.

The main objective of this pre-processing is to guarantee the constitution of a robust and complete database [Hong 2022b]. For this, the pre-processing must be designed to remove noise, extract features, remove or minimize outliers, deal with missing attributes/values, repair broken data, among others, [Famili 1997]. The main data pre-processing methods are classified as data transformation, data unification, data cleaning, information gathering, and data augmentation [Famili 1997, Li 2021c]. An explanation in detail and applied to the cases of interest of this thesis is presented in Chapter 6.

4.1.7.1 Data Transformation

Data transformation can involve filtering, sorting, editing, and denoising. These transformations are performed mainly for two reasons: *1)* to find a more suitable representation for fault detection and classification analysis; *2)* to combine different PV data formats into an identical one.

4.1.7.2 Format unification

This process is carried out because in the photovoltaic system they are obtained through different sensors and acquisition systems. Therefore, in some cases it is possible to have very different formats in terms of variation intervals, duration, sampling periods, etc., [Li 2021c]. This is why it is recommended to standardize the data. Multiple proposals are found in the literature for this. For electrical signals in the time or frequency domain, common operations are resampling or window slicing [Lu 2019b]. For the analysis of the characteristic curves $I(V)$ resampling with a different number of points or current-voltage distribution [Chen 2017] is required. Another recommendation for the correct detection and classification of faults, when several variables of the photovoltaic system are used, is to perform the normalization [PATRO 2015] to standardize the range of variation of the characteristics in [0, 1] or [-1, 1]. If the detection and classification of errors is performed on images, the operations of resizing [Karimi 2020], RGB separation [Aghaei 2016] and dimming [Meng 2018] are recommended.

Some acquisition platforms may contain interference or invalid information that is removed by filters [IEC 1998]. Along the same lines, advanced signal processing methods are proposed, such as wavelet denoising [Chikh 2015] or smoothing [Juxing 2020], to remove unwanted noise, interpolation, Kalman filtering, autoregression or moving averages [Turrado 2014, Demirhan 2018]. For images, denoising is applied which removes external interference and restores the real image [Fan 2019].

4.1.7.3 Data augmentation

Taking into account that statistical and machine learning techniques are more effective when they have a large amount of data that is sufficiently representative of all operating modes, data augmentation is proposed. This process is really a great

challenge because the conditions must be guaranteed to increase the data that do not hide faults and that are consistent with real scenarios of the PV plant. There are two major drawbacks to data augmentation. The first is an insufficient amount of electrical signals, images or in general data captured from the plant PV, due to the limited number of PV modules. The second is the appearance of unbalanced data, that is, the amount of electrical signals, images, etc., of healthy modules and faulty modules is different [Perez 2017]. These two obstacles can significantly hinder the learning performance of machine learning models. Therefore, data augmentation is widely used and studied [Shorten 2019, Karimi 2019].

Fault detection analysis can be performed directly on the raw signals captured by the system, however sometimes in order to extract small details that allow to differentiate the classes, a signal transformation that transforms the signal can be useful.

4.1.8 Signal treatment methods

There are two main types of Signal treatment or processing methods: *1)* Signal transformation methods and *2)* feature extraction methods.

4.1.8.1 Signal transformation methods

The signal transformations can be used to extract information from time series data in the frequency domain for further analysis. Some of the most well-known methods of signal transformation are presented below.

Fourier transform: The Fourier transform (FT) is used to extract the frequency components of a signal from its original domain to the frequency domain [Bracewell 1986]. Among the best known variants of the Fourier transform are the continuous Fourier transform, the Fourier series, the discrete Fourier transform and the fast Fourier transform (FFT) [Nussbaumer 1981]. The FFT algorithm is published in 1965 [Duhamel 1990] and is conceived to reduce the order of complexity of computational tasks such as the Fourier transform from N^2 to $Nlog_2$, where N is the size of the problem. In the photovoltaic domain, the FFT is widely used [Riza Alvy Syafi'i 2018]. In some of these works, it is used to extract the frequency content of the current to detect arc faults (AF) in a photovoltaic array [Riza Alvy Syafi'i 2018, Sudiharto 2017, Fitrianto 2019b]. It is important to mention that this type of fault is very difficult to detect in the time domain [Mukherjee 2017]. This type of fault is characterized by a sudden dip and thus a rapid change in system current, therefore frequency domain analysis is proposed as a solution [Mukherjee 2017]. This fault is widely studied because it has caused various fire hazards in photovoltaic systems worldwide [Haeberlin 2007]. For this reason, arc fault oriented detection systems focus mainly on this type of transformation [Murtadho 2020, Fitrianto 2019a]. This type of method has also been used in the identification of harmonic loads in power quality problems [Sudiharto 2017].

Additive and multiplicative decomposition: The additive and multi-

plicative decomposition attempts to extract the trend and seasonal factors from the time series. This approach is widely used for prediction of future values [Mbuli 2020, Saxena 2017, Patidar 2019, Prema 2015]. The seasonal effects obtained can be used to create and present fitted values of the original data. At present, we are not aware of fault detection methods that use this type of decomposition. However, taking advantage of its interesting features for prediction processes, it could be inferred that there is ample potential to develop prediction-based detection algorithms.

Wavelet transform: The principle of the wavelet transform is to decompose an input signal into subsets. Each subset is made up of a time series of coefficients that characterize the evolution of the signal in the corresponding frequency band [Heil 1989]. That is, the wavelet transformation uses a function named the mother wavelet that decomposes the signal into different frequency components that make up a family of functions that are translations and dilations of a mother function [Gómez-Luna 2013]. Therefore, the wavelet transform decomposes the signal into a series of wavelet components where each one is a time-domain signal covering a specific frequency band [Zhao 2000, Karimi 2000, Pang 2010].

In [Wang 2013] it is mentioned that wavelets are especially effective in functions with discontinuous or abrupt changes such as power system fault signals. But they also mention that the challenge lies in the correct choice of the mother wavelet since the performance of the transformation depends on the choice of the mother wavelet function and the translation and expansion coefficients to adjust the time and frequency resolutions [Li 2021c]. Wavelet transform can be classified into two types: *i)* Continuous Wavelet Transform (CWT); and *ii)* Discrete Wavelet Transform (DWT). CWT uses an infinite number of scales and locations, and DWT uses a finite set of wavelets [Graps 1995].

Like the Fourier transform, the wavelet transform is a linear transform by [Faziludeen 2013]. The difference with FFT is that it allows the temporal location of different frequency components of a given signal [Karimi 2000].

The DWT transform is widely used in the fault detection domain [Fong 2015, Qibin Zhao 2005, Faziludeen 2013, Wang 2018a]. Specifically in the photovoltaic domain, the DWT transform is adopted to extract features from AC current (IAC) to identify AF (Arc Fault) [Wang 2016b], over AC voltage (VAC) and AC current (IAC) to classify Line to Line Fault (LLF) and Ground Fault (GF) [Manohar 2017]. In [Wang 2018a] it is used together with a Long short-term memory (LSTM) algorithm to forecast solar irradiance models. In [Wang 2013] the detection of arc faults in photovoltaic systems is carried out and it is concluded that the conventional methods that are based on the recognition of patterns in the time domain, or the detection of amplitude in the domain of frequency by using a Fourier transform do not work effectively for arcs because the signal to noise ratio is low and the arc signal is not periodic.

4.1.8.2 Feature extraction methods

Feature extraction methods are used to extract local features from raw time-domain measurements. Some of the measures that can be extracted as features for detection and classification are peak curvature, crest factor, signal-to-noise ratio (SNR), root mean square (RMS) level [Scharf 1991]. Other measures are proposed for cases where the signal is the I(V) curve. In these cases, parameters such as the open circuit voltage (V_{oc}, the short-circuit current (I_{sc}, voltage and current at the maximum power point (V_{MPP}, (I_{MPP}), fill factor (FF), equivalent series resistance R_s, shunt resistance (R_{sh}) [Garrido-Alzar 1997]. Other widely extracted features for fault detection are presented below.

Power spectral density: Power Spectral Density (PSD) describes the power of the signal as a function of its frequency in units of W/Hz [Maral 2004]. This measurement is widely used in the detection of different types of faults [Ahmadi 2011, Li 2015, Al Ahmar 2010, El Bouchikhi 2013, Ayaz 2014, Heidarbeigi 2008, Gritli 2012, Martinez 2017, Gritli 2013, Oviedo 2011, Bellini 2006, d. J. Rangel-Magdaleno 2009, Anil Kumar 2016]. Some works use a mix of PSD with the FFT [Wescoat 2020] or with the Wavelet decomposition [Cusido 2008]. However, in the domain of fault detection in photovoltaic systems there are very few articles, among which the one by [Bharath KurukuruF 2020] stands out. In that article they use the Wavelet composition, and from it they extract a series of characteristics, among which is the PSD. That information is used as input to the neural network-based classifier.

Autocorrelation: Autocorrelation is a statistical tool that expresses the correlation of a signal with a delayed copy of it as a function of the delay [Yang 1998]. This type of statistical tool has been used before in different fields. In [Fucheng 2015] it is used in vibration signals for gearbox fault detection. Similarly, vibration signal analysis is used to detect faults in reciprocating compressors [Pichler 2016]. In [Zhang 2013], they present a bearing fault detection method based on kurtosis-based adaptive band-stop filtering (KABS) and iterative autocorrelation. Some articles, such as those written by [Dey 2019] and [Xu 2021], use this tool to perform bearing fault detection. Other works, such as the one presented in [Rafiee 2009], perform fault detection by autocorrelation on the wavelet decomposition coefficients. Specifically, in the photovoltaic field, it is used to detect ground faults using time domain spread spectrum reflectometry [Alam 2013a]. In this article, they set a threshold for the difference between the autocorrelation peaks and examine the peaks before and after the fault.

Finally, a comparative chart of some existing data acquisition systems in the PV domain is presented below.

4.1.9 Existing data acquisition systems

Considering that the costs of renewable energy technologies have not yet come down enough for grid parity to be universally achieved without subsidies, there is

still room for technological improvement and cost reduction. In this way of building new technologies mainly focused on data acquisition PV plants, multiple approaches have been proposed in recent years. One of the first PV plant data acquisitions was introduced by Blaesser [Blaesser 1997]. This on-site data acquisition system performed the measurement of the characteristics of the PV plants and allowed the collection, analysis and presentation of operational data. The main drawback of that system is that it is extremely expensive, more than 10% of the cost of the PV plant.

Then, with the decline in prices in data acquisition hardware, PV plant data acquisition is heavily applied to small PV installations. One of the first papers in this area is presented in [Mukaro 1998, Mukaro 1999]. Those low-cost systems monitored solar radiation and general environmental conditions of the PV plant. That system is built using an 8-bit microcontroller that drives an analog-to-digital converter (ADC) and stores data in a serial EEPROM until it is loaded into a laptop. In this data acquisition system, the data is sampled and stored at intervals of 10 min. They considered low power consumption as they minimized power consumption by keeping the microcontroller in a low-power mode between measurement intervals. This stage is necessary since the data acquisition system is powered by a rechargeable battery. The great advantage of this data acquisition system is that it is very suitable for data acquisition meteorological or environmental parameters in remote stations, particularly in developing countries. Due to its versatility, that work is adopted in [Mukaro 2008], where it is shown that an operator with a laptop is all that is required to collect data acquired from systems scattered around an area of interest.

Another effort to build data acquisition systems is presented by Koutroulis and Kalaitzakis [Koutroulis 2003]. The main disadvantage of the work presented by them is the dependency on a PC, the use of commercial software (Labview™) and the requirement of mains power supply. These parameters strongly limit the development of a data acquisition system for PV plants since it makes the system more expensive and, therefore, limits its diffusion and use. In the literature it is possible to find other data acquisition efforts that are based on microcontrollers. Some of the drawbacks with these systems is that they use a low resolution ADC connected to an amplifier stage, which defines each input to a specific sensor [Kamunda 2007, Benghanem 2009a], or others depend on a PC [Benghanem 2009a, Demirtas 2008, Benghanem 1998b, Mahjoubi 2012], commercial software [Benghanem 2009b, Benghanem 2010], or do not follow IEC standards for managing accuracy or obtaining data. Not following the standards counteracts the achievement of low cost, portability and low power consumption, among other advantages [Purwadi 2011, Ikhsan 2013]. Table 4.2 presents some examples of existing data acquisition systems with their main features.

Table 4.2: Conventional PV data acquisition system developed

Year	Ref	Applications	Measurements	Platform	Software	ADC resolution	Sampling interval	Data transmision	Data storage
1998	[Mukaro 1998]	Solar radiation	Irradiance	ST62E20	Turbo C++	8−−bit	10 min	Wired RS232	EEPROM
1998	[Benghanem 1998a]	PV plant	Irradiance, ambient temperature, PV module voltage and currents, battery voltage, battery charge current, load current	Processor MC 68B09	−−	−−	−−	Wired RS232C	−−
2001	[Wichert 2001]	RES System	Ambient temperature, irradiance, voltage, current and power	DataTaker DT50	Labview	−−	5 min	Wired RS232	PC
2003	[Koutroulis 2003]	RES System	Ambient temperature, irradiance, voltage, current and power	PCI−−6024E	Labview	12−−bit	1 min	Wired RS232	PC
2005	[Papadakis 2005]	RES System	Ambient temperature, irradiance, wind, humidity, voltage, current and power	Commercial DAQ unit	VB, SQL Server 2000	−−	1 min	RF	PC
2005	[Forero 2006]	Grid connected PV plant	DC current, DC voltage, AC current, AC voltage, energy, power, ambient temperature, solar radiation, IeV curve	FP DAQ board	Labview	16−−bit	30s	Wired RS232	PC
2008	[Demirtas 2008]	RES System	Wind speed, the panel positions of the solar module, the currents and the voltages of the solar and the wind systems	PIC 18F452 and PIC 18F2550	C#	10−−bit	−−	Wired RS485	PC
2011	[Purwadi 2011]	PV LED Street Lighting	Daily energy, total energy used, charging or discharging status and fault condition	Atmel Atmega8	−−	−−	−−	−−	EEPROM
2011	[Anwari 2011]	Small PV installation	PV array voltage, PV array current, ambient temperature, solar irradiance	PIC16F877a Microcontroller	Labview	8−−bit	−−	Wired RS232	−−
2012	[Mahjoubi 2012]	PV Water Pump	water flow rate, ambient temperature, global irradiation (inclined and horizontal), voltages, Currents	HOBO	HOBO	−−	−−	−−	−−
2012	[Eke 2012]	PV module temperature	PV module temperature	PC	Qbasic routine on PC	12−−bit	1 min	Wired RS232	PC
2012	[Andreoni-López 2012]	Grid connected PV plant	Solar radiation, wind speed, voltage and current of PV modules	DSP	Labview, C	16−−bit	−−	ZigBee	PC
2014	[Tina 2014]	SAPV system (fridge application)	Currents, voltages, temperatures of the fridge (internal and external), Temperatures of PV plant, ambient temperature, humidity, irradiance	ET−−7017 produced by ICPDAS	Matlab	−−	−−	Wired (Modbus Protocol)	Database

Continued on next page

Table 4.2 – continued from previous page

Year	Ref	Applications	Measurements	Platform	Software	ADC resolution	Sampling interval	Data transmision	Data storage
2015	[Devaraju 2015]	Weather station	Humidity, atmospheric temperature, wind speed, wind direction, rainfall, solar radiation, surface temperature, ambient temperature, atmospheric Pressure	PIC16F887 Microcontroller	Weather Monitoring Station (Embedded C Langage)	14 bit	30 s	ZigBee	MySQL server
2015	[Shariff 2015]	Grid connected PV plant	Solar radiation, ambient temperature, module temperature, PV voltage, PV current, grid voltage and grid current	PIC18F8720	C/Netbeans	12−−bit	1min	Zigbee	EEPROM
2017	[Villagrán 2017]	Enviroment parameters	Temperature, humidity, wind	LP3500	Dynamic C		1min	GSM	SD
2017	[Chao 2017]	Fault diagnosis	Temperatures, irradiance, voltages, currents and curves I−−V/P−−V	PIC18F8720	Solar Pro Software	12−−bit	−−	ZigBee	PC
2017	[Rezk 2017]	SAPV system (fridge application)	Temperatures, irradiance, voltages, currents and curves $I(V)/P(V)$	DAQ board NI USB−−6009	Labview	12 bits	−−	−−	−−

In order to demonstrate the direct relationship between the quality of the data, the quantity of variables measured and the sampling frequency, an industrial and commercial data acquisition system is selected. The characteristics of Tigo's industrial and commercial system are explained below together with the PV plant instrumented with said platform.

4.2 Tigo industrial and commercial data acquisition system

The Tigo data acquisition system used in this thesis is part of a more complex data acquisition and optimization platform for PV plants. Tigo platform is designed and built by the Tigo company born in 2007 in Silicon Valley, California. The Tigo company is one of the world leaders in power electronics at the module level aimed at increasing energy production, improving safety and reducing the operating costs of solar installations. The Tigo platform proposes the solution named Plate-forme TS4 that offers three main functionalities. First, they offer an optimization service to increase energy performance by reducing impact and shading compensation. Second, it provides a data acquisition service aimed at reducing operational expenses with module-level performance visibility. This service is the one that is used as a data acquisition system of the PV plants for fault detection. Among all the data acquisition systems on the market, this solution is selected for this research for its independence of the inverter brand and the availability of access to data via API. Finally, the Tigo platform has a safety system that reduces the voltage in the module if necessary, increasing safety and complying with electrical codes.[1].

The components of the Tigo data acquisition system are briefly described below.

4.2.1 Components and Connection Scheme

In this thesis, Tigo data acquisition system is used mainly to capture the current signal of each PV panel separately. This is possible because each module is connected in parallel with an optimizer and then all the optimizers are connected in series. The signals obtained with this method can be analyzed for fault detection and for obtaining conclusions that can be extrapolated to the string level. This is possible if it is taken into consideration that the current of each of the panels can be represented, without loss of generality, as the current of a string that contains a single panel. Figure 4.3 shows a representative diagram of the connection between a panel and a Tigo optimizer.

In Figure 4.3, it can be seen the current I_P that corresponds to the PV panel electric current and the current I_{string} that corresponds to the current shared by all the sets (panel + optimizer). The current of each panel PV_i is captured with the Tigo data acquisition system in the form of a time series denoted by $I_{i\{1:n_I\}} = \{i_{i,t_1}, \ldots, i_{i,t_{n_I}}\}$, where n_I is the number of samples (12 panels) of the $i-th$ time

[1]For more details about Tigo, click here

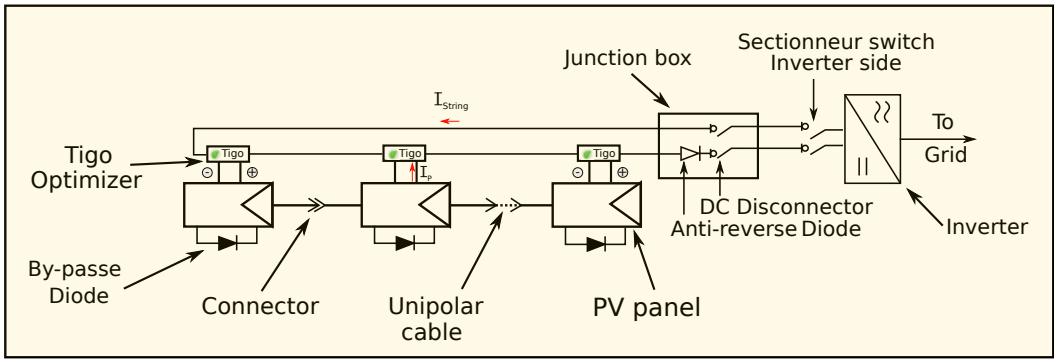

Figure 4.3: Tigo data acquisition system connection diagram

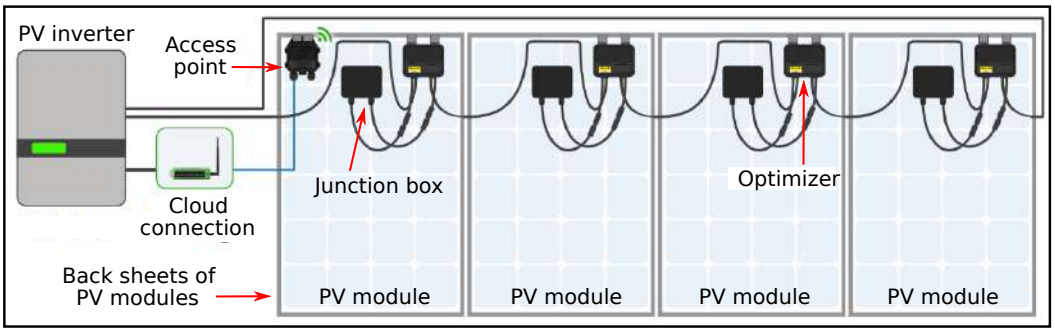

Figure 4.4: Tigo platform connection diagram

series that has a sampling period of one minute and $t_i, i = 1, \ldots, n_I$, is the date of the sample.

As can be seen in the same Figure 4.3, each panel is equipped with a reference Tigo optimizer TS4-R-O (R is for Retrofit and O is for Optimizer) on its back sheet. Tigo PV panel Optimizer is an MPPT device that individually controls each PV panel, to achieve maximum performance. To do this, the optimizer constantly monitors the maximum power point. The values are then sent to the Tigo server where they are stored. A connection diagram of the Tigo data acquisition system with all the constituent elements is presented in Figure 4.4.

The database stored in the cloud can be accessed through the Tigo Web platform presented in Figure 4.5.

This Tigo data acquisition system is used to instrument 12 PV panels from a single PV string. The characteristics of the PV plant implemented with this data acquisition system are presented below.

4.2.2 Instrumented PV plant

The photovoltaic system (experimental platform) is located in Toulouse in the department of Haute Garonne (31), in the region of Occitanie, at the address 5 Avenue du Colonel Roche, 31400, Toulouse in the LAAS-CNRS laboratory in the building named ADREAM. The geographical location at the level of France, Toulouse and

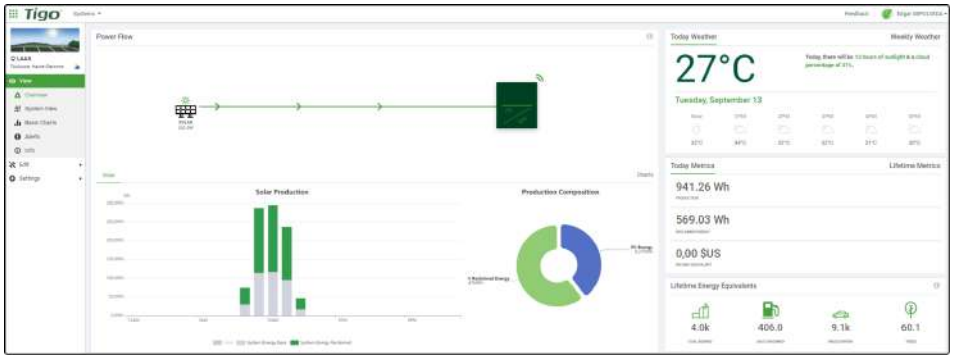

Figure 4.5: Tigo supervision website

Figure 4.6: Geographical location in France

LAAS is shown in Figure 4.6.

Figure 4.7, shows the location of the experimental platform on the terrace of the ADREAM building.

This PV plant is constituted of $n = 12$ PV panels with reference *SLK60P6L* from Siliken California with a nominal power of $250\,Wp$ connected in series forming a PV string. The main parameters of these PV panels are given in Table 4.3 under standard test conditions (STC) ($1000\,W/m^2$, $25°C$). Each PV panel is composed of 60 poly-crystalline silicon cells grouped into 3 sub-strings of 20 cells.

The 12 PV panels are used to represent three different health status. In addition, the PV panels are located spatially close, and therefore subjected to similar meteorological conditions to guarantee the possibility of comparing their electrical conditions with each other. Figure 4.8 represents the spatial location of the three

Figure 4.7: Experimental platform in the LAAS - CNRS, Toulouse - France

Symbol	Quantity	Value
P_{MPP}	Maximum Power (W)	250
I_{MPP}	Current at $P_{MPP}(A)$	8.21
V_{MPP}	Voltage at $P_{MPP}(V)$	30.52
I_{SC}	Short-circuit Current (A)	8.64
V_{OC}	Open-circuit Current (A)	37.67
S	Area of the module (m^2)	1.64

Table 4.3: PV module specifications at STC.

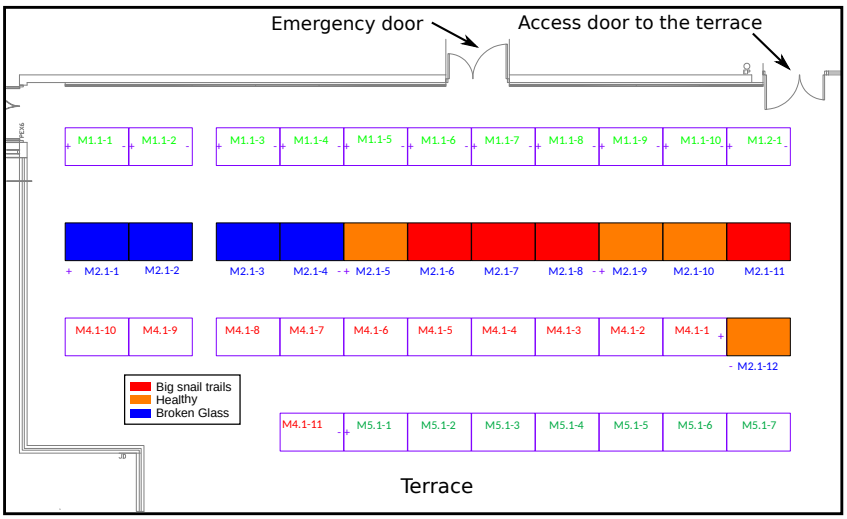

Figure 4.8: Distribution of the panels of the string PV used (experimental platform) with different health statuses: healthy (yellow), broken glass (blue), and big snail trails (red).

types of health status of the 12 PV panels.

As can be seen in Figure 4.8, the first health status are the panels with broken glass indicated in blue, followed by the red panels that correspond to the panels with Snail Trails. Finally, the orange panels correspond to the healthy panels. The location of the healthy panels is conceived with the aim of having at least one healthy individual in the same or similar conditions as the panels with faults. All panels have the same inclination and they are instrumented with a Tigo data acquisition system.

In the Tigo supervision website it is also possible to see the PV panels in the same position in which they are installed on the terrace of the ADREAM building as can be seen in Figure 4.9. Also, this representation allows to verify in real time the production of the panels.

As can be seen in Figure 4.9, the magnitude in the reduction of the current between the panels with Snail Trail (red color) and the healthy panels (orange color) is minimal, even the panel with the highest current is a panel with a Snail Trail. Due to this type of behavior, this fault is little studied, and even less, studied as an objective of early detection. Although the Tigo data acquisition system captures data every minute, its design is not aimed at fault detection, so it does not pay

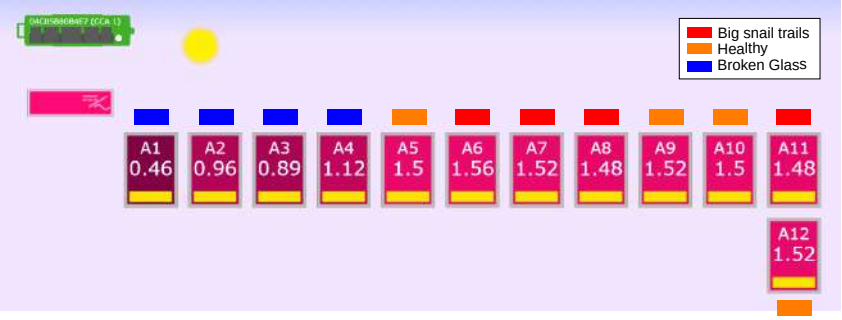

Figure 4.9: Physical location of the panels on the Tigo supervision website.

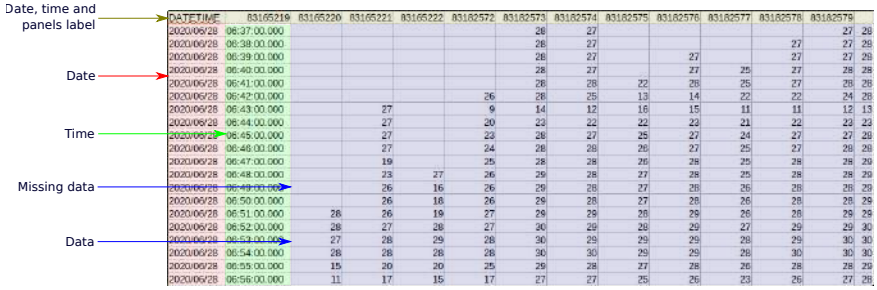

Figure 4.10: Example of problems found in the database generated by the Tigo data acquisition system.

special attention to the quality of the stored data. For this reason, it is necessary to perform a data processing step to return the relevant data for fault detection. To correctly process and clean the data, it must be taken into account that the Tigo platform begins to capture electrical current data only as soon as the panel begins to produce, which brings some associated issues. First, the Tigo platform does not record current and voltage 24 hours a day. Second, due to different weather conditions, the time at which the Tigo platform starts recording electrical current changes from day to day. Third, the Tigo platform begins to record voltage data before the plant begins to generate actual production. This is a serious problem since in case of performing a fault detection on the voltage variable, the probability of obtaining false results is very high. In Figure 4.10 it is possible to observe the behavior of the voltage database at the beginning of the day.

The same behavior illustrated in Figure 4.10 occurs at the end of the day.

Taking into account these limitations of commercial and industrial platforms, and following the parameters recommended in Section 4.1 of this thesis, a new photovoltaic data acquisition system, named Solar Vitality, is proposed in Chapter 5, and a weather station is also attached to that Solar Vitality system.

4.3 Discussion and Conclusions

In this chapter, a comprehensive review of existing photovoltaic data acquisition systems reported in the literature is presented in terms of the sensors and acquisi-

tion systems used. In addition, the most used parameters in the data acquisition of PV plants are analyzed, among which are: voltage, current, solar radiation, temperature and wind speed. In the field of data acquisition systems, this chapter covers the controllers that are used for the data acquisition system, types of data transmission methods, data storage, and data analysis. In this same way, it is identified that an effective data transmission system is essential to ensure data quality, especially considering the particular operating conditions to which these data acquisition systems are subjected. Different means of data transmission are discussed (wired, wireless and powerline, etc).

The choice of the best transmission medium is completely linked to the operating conditions of the data acquisition system. Multiple parameters can be evaluated to determine the best data transmission medium. First, the coverage area and length of the distance between the sensors and the data acquisition system must be taken into account. For example, coaxial cables cannot operate over long distances compared to fiber optic cable. However, for short distances it is shown that many studies prefer coaxial cables that show great performance. On the other hand, the WLAN protocol can cover only a small area of about 20 km^2, compared to GPRS-GSM, but can transmit data over distances of hundreds or thousands of kilometers over the Internet. The power line communication (PLC) protocol transmits information over hundreds or thousands of meters using the existing wired infrastructure without any additional installation. However, there are problems of aging of the wiring and external causes that can affect the sending of data over scenarios of hundreds of meters. In addition, when the wiring is extensive and the sensor output signal is in voltage, significant signal losses can occur, so it is recommended that the system be designed for current sensor output when the signal must travel long distances by cable.

Another aspect to take into account is the speed of sending data. For example, the maximum speed at which data can be transmitted using a coaxial cable is around 10 Gbps. Fiber optic cable is only capable of transmitting data from 100 Mbps to 2000 Mbps. On the other hand, the WLAN protocol is much faster than coaxial cable or fiber optics, but it has strong limitations when there are physical obstructions between devices, radio interference, simultaneous communication of multiple devices on a network, and distance between devices. The PLC has a maximum sending capacity of 200 Mbps and the slowest protocol is GPRS-GSM with approximately 40 to 50 Kbps.

Regarding the signal decomposition and signature extraction methods, these methods must be tested individually depending on the faults to be detected. Some faults may be widely visible with Fourier transforms while others will go unnoticed. It is evidenced that the wavelet-based transform is the most used and versatile due to the large number of faults that are detected in the literature.

Chapter 5

Diagnosis-oriented Data Acquisition (Solar Vitality)

Contents

5.1 Motivation 156
5.2 Characteristics of Solar Vitality 157
5.2.1 Portable data acquisition requirements 157
5.2.2 Measured parameters 158
5.2.3 Hardware 158
5.2.4 Software 164
5.2.5 Electrical power supply 165
5.2.6 Assembled prototype and electrical connection diagram to the PV system. 166
5.2.7 Product evolution cycle and test scenarios 167
5.3 Discussion and Conclusions 177

As explained in Chapter 4, current PV plant monitoring systems are not designed for fault diagnosis, much less, they are aimed at detecting faults whose occurrence is very fast, which requires high speeds of data sampling. As a proposal to solve that problem, this Chapter 5 presents a new diagnosis-oriented data acquisition system named Solar Vitality. Solar Vitality is designed with special emphasis on its precision or uncertainties under the IEC 61724 standard [IEC 1998]. Solar Vitality uses the Arduino open source electronic development board to solve the current problem of data acquisition photovoltaic (PV) systems together with a Raspberry PI4. Solar Vitality can be used to monitor faults in PV systems from residential to utility power PV plants, in developed countries and especially in remote areas or regions in developing countries. Solar Vitality meets all the relevant requirements in terms of accuracy included in the International Electrotechnical Commission (IEC) standards for photovoltaic systems, with measurements every 15 milliseconds, including 11 analog inputs to measure up to 4 independent photovoltaic strings and the meteorological parameters of irradiation, ambient temperature and wind speed. Solar Vitality is completely autonomous in terms of power supply, portable and easily coupled to different topologies of photovoltaic systems. Solar Vitality is tested in different scenarios and with different topologies of photovoltaic systems in real production conditions. Solar Vitality is tested in continuous

operation for more than 6 months, presenting a robust operation even in the harsh environmental conditions of summer and winter in France.

Solar Vitality is capable of capturing the electrical behavior of the photovoltaic system with a high sampling frequency (every 15 milliseconds) and meteorological variables such as ambient temperature, irradiation and wind speed. The design and construction of Solar Vitality is aimed at meeting two objectives. The first is to demonstrate the influence of the sampling frequency in the detection and classification of faults in photovoltaic systems. The second is to demonstrate and quantify the effect of meteorological variables on the photovoltaic system. Next, the explanation of Solar Vitality is presented. The results of the data collected from the meteorological and electrical sensors indicate that the new system is reliable and exhibits performance comparable to that of commercial systems. Solar Vitality is of special interest for both research and industry. Finally, Solar Vitality is easily customizable for the specific needs of each project and photovoltaic system and can even be extended to other domains.

5.1 Motivation

Solar Vitality captures the electrical and meteorological behavior of the PV plant. Capturing these two behaviors is vital to improving the accuracy and the number of different faults detected in PV systems [Blaesser 1997]. Most data acquisition systems have focused on capturing the electrical behavior of the system, since for meteorological behavior they use data compiled by national or European institutions [NAS 2022, Atl 2022, PVG 2022]. This information is useful to know in a general way the meteorological conditions of operation of a region or zone. However, when thinking about fault detection in PV systems, the use of these satellite data has several drawbacks. First, this data cannot replace the specific data taken on the site. Second, there are many places where these databases are not available or are in the process of being compiled. Third, although a wide range of weather databases are available, they are generally expensive, highly sophisticated, and not easily manageable [Fuentes 2014].

As a result, further development of data acquisition systems is required to collect and process electrical data and meteorological data in operation, under the premise of obtaining measured values using accurate and easy-to-handle. Taking these aspects into account, this research presents a new diagnosis-oriented data acquisition system (Solar Vitality) that can be used to instrument PV plants of any size, on a PV panel, PV string or PV array level. It can also be used to capture data from PV plants with different configurations. These characteristics are vital to ensure and facilitate rapid and continuous development. In addition, Solar Vitality has a wide flexibility and can be adapted to each specific case (both research and industrial applications in developed and developing regions). Solar Vitality will allow the PV community to move faster in some of the research areas that have required comprehensive PV data acquisition but are limited by cost and technology issues.

5.2 Characteristics of Solar Vitality

Solar Vitality is based on the hypothesis that the capture of a set of variables such as voltage, current, irradiation, ambient temperature and wind speed, increases the number of detected faults and also the detection accuracy under different environmental conditions and with panels of different technologies. This hypothesis is based on the causal relationships, presented in Section 2.3.2, for different faults. In addition, Solar Vitality is built following the recommendations of the British standard BS IEC 61724 [IEC 1998], the guidelines of the European Joint Research Center [Blaesser 1995] and the guidelines of the National Renewable Energy Laboratory (NREL) [Kurtz 2013]. Also, Solar Vitality is designed to be portable and energy autonomous.

The design of Solar Vitality is detailed below, including the following parts: *i)* Portable data acquisition requirements; *ii)* Measured parameters; *iii)* Hardware; *iv)* Software; *v)* Electrical power supply; *vi)* Assembled prototype and electrical connection diagram to the PV system; and finally *vii)* Product evolution cycle and Test scenarios.

5.2.1 Portable data acquisition requirements

Portable data acquisition systems must meet a number of stringent requirements that are not necessarily present in traditional laboratory systems [Fuentes 2014]. For example, the portable systems are generally intended to be connected in harsh environments that must be taken into account when designing the data acquisition system. Some of the conditions that must be taken into account are: extreme temperatures, humidity, dust, shock and vibration. In the photovoltaic domain, according to the IEC 61215 standard [IEC 2005a], the equipment must withstand temperature ranges that vary between -40°C and 85°C in the worst of the cases [IEC 2005a].

Another difficulty that this type of system must face is the need to admit a combination of particular sensors, and adequate memory/storage capacities to adequately record the behavior for considerable times for the diagnosis of a fault or to obtain relevant conclusions of the behavior of the system. In addition, these portable data acquisition systems face major challenges in integrated signal conditioning issues, such as gain and filtering, and data acquisition sampling rate that determine system accuracy.

The main purpose of portable data acquisition systems is that, once set up, they can measure, record and display data without the intervention of an operator or a computer. Therefore, these systems must be compact, lightweight units that can be powered using two different configurations. The first configuration is based on one or several batteries and the second is to connect an external cable to a DC or AC power source (taking into account the voltage passed to DC).

To size the battery or battery bank it is necessary to identify the necessary processor. A low power consumption processor should be preferred and combined

with a suitable storage system to avoid selecting a processor with capacities beyond what is necessary.

As these portable data acquisition systems are left to run unattended for days or possibly years at a time (depending on system requirements), they must be able to have the option of storing the data on-site, with large memory sticks, or send them to the cloud to be downloaded by remote computers. In addition, they must have an intuitive user interface for remote configuration and control of the device.

All of these complexities make the design of these systems a real challenge, but just as powerful yet compact data acquisition devices that play an important role in verification testing and data acquisition of critical systems. In the design of Solar Vitality, a design supported by a low-cost processor is favored, whose limitations are overcome by different devices that meet the aforementioned objectives, as well as the IEC requirements [IEC 1998], without significantly increasing the cost.

5.2.2 Measured parameters

As observed in Figure 4.2, a photovoltaic system is constituted of different elements such as solar modules, batteries and regulators in the case of autonomous systems, inverters in the case of grid connection, AC and DC wiring, electrical safety devices and protection devices. The relevant variables to be measured associated to these elements, proposed in the British standard BS IEC 61724 [IEC 1998], the guidelines of the European Joint Research Center [Blaesser 1995] and the guidelines of the National Renewable Energy Laboratory (NREL) [Kurtz 2013], are shown in Table 4.1. However, the number of variables can be reduced if it is decided to calculate the power as a function of voltage and current and current sensors that can distinguish the direction of electron flow are also used. Under these conditions the number of variables can be reduced from 24 to 14: 2 temperatures (T_{am}, T_m), 1 irradiance (G_I), 4 voltages (V_A, V_L, V_{BU}, V_S), 3 directional currents (I_A, I_L, I_{BU}) and 2 bidirectional currents (I_{rs}, I_{FS}). It is also important to note that, as mentioned in the British standard BS IEC 61724 [IEC 1998], the number of variables can be increased or reduced depending on the primary objective of the data acquisition system. For our case, focused on fault detection of PV strings, the number of variables is reduced to only 5 (T_{am}, G_I, S_w, V_A, and I_A). The variables measured with Solar Vitality are divided into electrical and meteorological.

5.2.3 Hardware

The description of the different PV measurements, according to the type, range, and precision requirements for each one of them according to, in coherence with the IEC standards [IEC 1998], is given below.

5.2.3.1 Sensors

To comply with the recommendations of the British standard BS IEC 61724 [IEC 1998], the accuracy of current and voltage sensors, including signal condi-

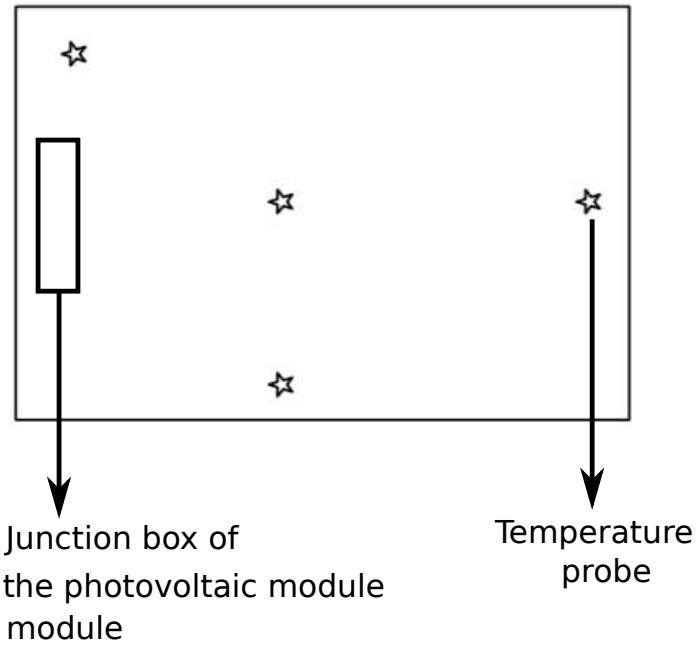

Figure 5.1: Recommended locations to properly measure the temperature of a PV module according to the ISO 16077 standard [ISO 2013]

tioning, must be better than 1% of reading. Measuring the current (I_A) can be achieved using shunts or current transducers. The transducers are selected due to their high accuracy even if this drives up the cost of Solar Vitality. This decision is made due to the fact that Solar Vitality is oriented to the detection of faults and therefore the high quality of the data must be guaranteed, and more so considering that data with behaviors very similar to those of a healthy panel are explored in this thesis.

For temperature measurement, only the ambient temperature T_{am} is selected. The module temperature T_m has two major difficulties that can introduce a high level of noise to the diagnosis system. First, it is necessary to determine the best location for the temperature sensor on the back face of the module, that is, it is necessary to determine which section of the module is the most representative in terms of temperature. To our knowledge, there is no consensus on the ideal placement of one single sensor. However, the ISO 16077 standard [ISO 2013] recommends 4 sensors located in the locations presented in Figure 5.1.

Despite the recommendation of the ISO 16077 standard [ISO 2013] presented in Figure 5.1, it is not easy to guarantee that this positioning captures the overall temperature of the PV module. Second, it is possible that during the installation of the data acquisition system, the module temperature sensor is not accurately placed within the back face of the module by the installer, generating a bias. In other words, the repeatability of the instrumentation setup may be difficult to achieve under real operating conditions. Due to the fact that the level of uncertainty is

very high, the variable module temperature T_m has not been integrated in Solar Vitality. Instead, the ambient temperature T_{am} is chosen for measurement, with the aim to use these measurements along with irradiance and wind measurements to explain the combined temperature effect on module electrical parameters. The premise behind the choice is that ambient temperature is more uniform across the PV plant, allowing for the use of fewer sensors, and that it allows for a more repeatable instrumentation, reducing uncertainty. For the selection of the ambient temperature sensor, it is taken into account that, according to the British standard BS IEC 61724 [IEC 1998], the accuracy of these temperature sensors, including signal conditioning, must be greater than 1 K or 0.5 $°C$. Based on this a PT 1000 temperature probe is selected.

For the irradiation measurement, it must be taken into account that the sensor is located with the same inclination of the PV string [IEC 2005a]. With this, it is possible to guarantee that a representative sample of the irradiation that is affecting the string panels is really being collected. In addition, special attention must be paid to locating the sensor in a representative place in the PV system, avoiding shading problems that lead to false fault alerts. As a sensor, a reference cell made of the same material as the PV system being analyzed is selected. This guarantees that the reaction of the material to the incidence of sunlight is very similar (taking into account that there may be internal problems of the material, or problems related to external causes). This reference cell also follows the guidelines of the IEC60904 [IEC 1994a] standard. Its accuracy, including signal conditioning, has also been verified to be better than 5% of reading.

Finally, for the measurement of wind speed, two main aspects must be guaranteed [IEC 1998]. First, that the sensor will always be in a representative place in the PV system. Second, the sensor must always be vertical to guarantee the correct measurement of the wind speed. In addition, the selected sensor has superior accuracy between the ranges of $0.5ms^{-1}$ for values measured for less than $5ms^{-1}$ and 10% of reading for wind speed values greater than $5ms^{-1}$. Taking these aspects into account an anemometer is selected.

Once the meteorological sensors are selected, two extra conditions must be guaranteed. The first is that they must be easily coupled to the PV system from a small scale such as a residential installation to a large scale installation such as a utility PV plant. The second condition is that the irradiation sensor must be adaptable to adjust to the inclination of the PV system to be measured. To fulfill these two conditions, the 3 meteorological sensors are coupled in a meteorological station whose structure is delicately thought out and modular to be adaptable to different PV system conditions. The weather station is presented in Figure 5.2.

As can be seen in Figure 5.2, the station allows the movement of only the irradiation sensor to coincide with the inclination of the PV plant. It also guarantees the verticality of the wind speed sensor, and the shading conditions for the temperature sensor, complying with the norms of the IEC 61724 standard [IEC 1998]. Finally, the weather station is equipped with an electrical box. In this way, it is much more efficient to replace the sensors without having to affect the data acquisition

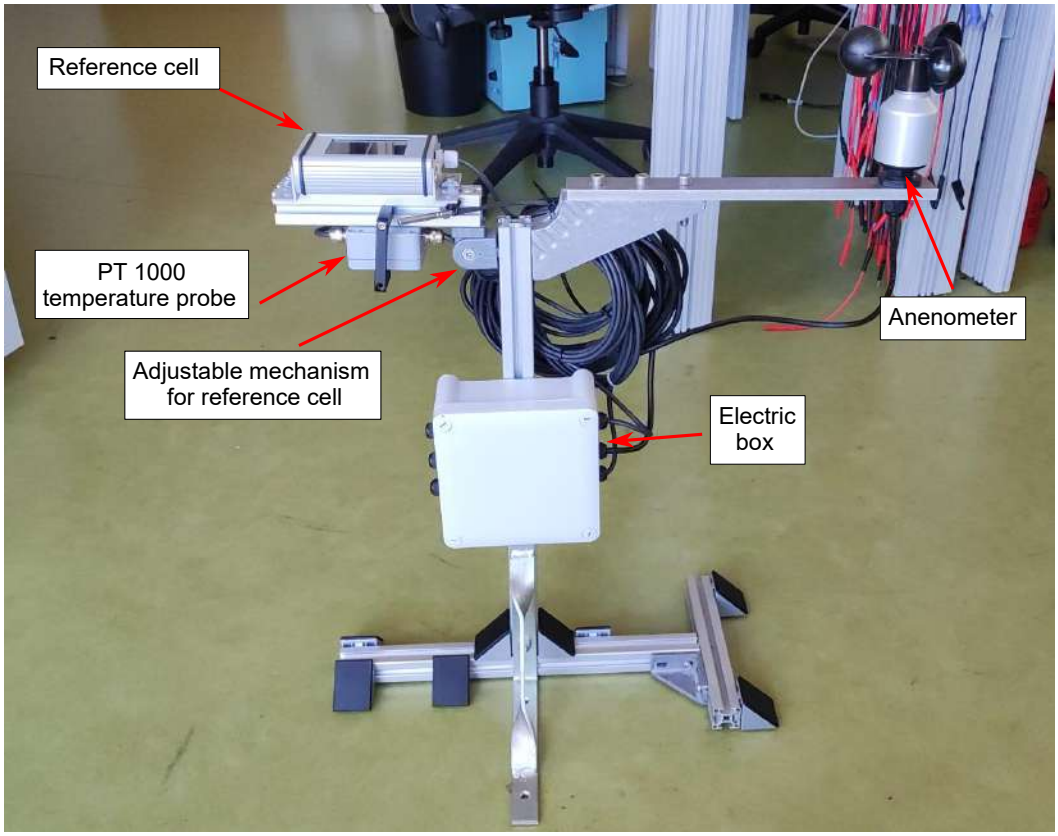

Figure 5.2: Portable and adjustable weather station

and fault detection system. The detailed description of the components and the evaluation of the sensors are presented in Annex A.

5.2.3.2 Electronic boards

This section presents the hardware of Solar Vitality proposed in this research. The characteristics of the hardware used for signal capture and processing are presented below.

The first premise of Solar Vitality is the use of open source hardware. This premise allows the development of a system suitable for developing countries or research applications that must use these high-power industrial devices or systems. Among the many microcontroller-based platforms that are available on the market, MSP430 launchPad, STM8L Discovery, Libelium's Waspmote, and Arduino Mega are well known among developers [STM 2022, Was 2022, Meg 2022, Tex 2022].

In general, the prices of microcontrollers are below 10€ for Texas Instruments and STMicroelectronics. On the other hand, the most expensive is the Waspmote with prices around 135€, which makes it not viable for this project. Among all the options, Arduino stands out for the simplicity of its hardware, that is, easy to duplicate without expensive means, and its software based on C/C^{++} and with numerous already developed and highly advanced libraries for data acquisition in comparison with the other three. It is also ideal for the growth and continuous improvement of the prototype because Arduino is conceived in a modular way. That is, you can find a wide range of modules to interconnect to adapt Solar Vitality to any design, such as Bluetooth, Wi-Fi, LAN, GPS, GPRS, etc., or have the possibility of developing new modules with a specific purpose [Hac 2022]. It is important to note that this characteristic is not easy to find on all data acquisition platforms or systems. Although it is true that there are other cheaper data acquisition platforms or systems, it is very difficult to find support, libraries or modules for such systems or platforms. For this reason, it is necessary to develop all programs and libraries from scratch. These developments from scratch take long development and testing times to verify their robustness.

An Arduino is a small programmable electronic board in which there is a microcontroller. This board is used to capture data from sensors and/or control different electrical components. The name "Arduino" is, in fact, the trade name for the open source prototyping platform. This type of Arduino platform is widely used in tasks such as home automation, robot control, on-board computing, etc. One of the biggest advantages of this board is the high level of the microcontroller programming language. The Arduino language is a C/C^{++} based language. A comparison of the 3 best known Arduinos is exposed in Table 5.1.

Table 5.1: Comparative table of characteristics of different Arduino models

Characteristics	Arduino Uno	Arduino Mega	Arduino Due
microcontroller	8-bit ATmega328P 16 MHz	8-bit ATmega2560 16 MHz	32-bit AT91SAM3X8E 84 MHz
Flash Memory	32 KB	256 KB	512 KB
RAM Memory	2 KB	8 KB	96 KB
EEPROM	1 KB	4 KB	–
Analog Inputs	6	16	12
Digital Inputs	14	54	54
Power Supply	7-12 V	6-20 V	7-12 V
Pin input voltage	5 V	5 V	0-3.3 V
Pin output voltage	0-5 V	0-5 V	0-3.3V

As can be seen in Table 5.1, the Arduino Due is superior for the majority of the characteristics. However, Solar Vitality is designed to perform the reading of multiple PV strings (6 string or more) and also the climatological sensors. Therefore, the selection criteria is the number of analog pins available on the board. Another difference between the Arduino Mega and the Arduino Due is the EEPROM memory capacity. The Arduino Mega has 4KB while the Arduino Due has none. Although the Arduino Mega is limited to 4 kilobytes (KB), it is more than enough to store data for a short data acquisition period (a few hundred seconds). These short times are enough to perform sliding window calculations on the data, or filters. However, if the objective is to store the data, it is an insufficient capacity since 12 parameters stored every minute would exceed 1 kB in approximately 6 min or a little less. To solve this storage problem there are mainly two options. The first is to use memory board expansion modules, each with a capacity of up to 2 gigabytes (GB). The second option, which is adopted in Solar Vitality, is to send data by serial means to another board that will store, pre-process and detect PV system faults. With this in mind, the Arduino Mega [Arduino 2020] data acquisition board is selected together with another external data processing board that works as a central computer.

The aforementioned conditions added to the low cost of the Arduino Mega make this development board interesting, both for research and for industrial application in developing and developed countries. In addition, the software to program this board is free to download and the hardware reference designs are available under an open source license and users are free to adapt them to their needs. Although as can be seen in Table 5.1, the board can work with external power (USB cable, AC/DC adapter or battery) from 6 V to 20 V. However, it is recommended to operate the board in a range of 7 V to 14 V to prevent the board from becoming unstable or the voltage regulator from overheating and damaging the board [Meg 2022]. Finally, another interesting feature of these development boards is their sleep mode function that significantly reduces power consumption [Ard 2022].

Arduino MEGA has 16 analog inputs with an 8 bit resolution Analog Digital Converter (ADC). The ADC has an input range of 5 V. For this reason, in Solar Vitality it is necessary to add a signal adaptation stage to decrease the output voltage of the sensors $(0-10V)$ to $(0-5V)$. This adaptation allows Solar Vitality

to continue within the recommendations of the IEC61724 standard [IEC 1998].

In the same way of developing data acquisition platforms or data acquisition systems, some works have proposed the use of the Raspberry Pi board [Ferencz 2018]. In this work, the Raspberry Pi board demonstrated great performance results and storage of large amounts of data on low-cost local servers. That same principle is adopted in this research: a Raspberry Pi 4 board, with 8Gb of RAM memory, is selected to receive the data sent by the Arduino Mega and then store them in a MySQL database with the phpMyAdmin database manager. The Raspberry Pi 4 can be defined as a computer stripped down to its simplest form with an ARM processor board almost the size of a credit card. The Raspberry Pi supports running several variants of the free GNU/Linux operating system and compatible software.

This type of interaction between the Arduino Mega and the Raspberry Pi has already been studied before showing a high performance [Wali 2018]. However, as mentioned in [Moreno 2020] the serial communication between the Arduino Mega and the Raspberry Pi can have transmission errors that should be monitored. To control these errors, a python script is integrated into the Raspberry Pi that tracks these errors online and corrects them based on the history of the stored data. No displays are taken into account in the design for three reasons: reduce costs, reduce power consumption, and simplify the software and hardware design.

5.2.4 Software

The software developed in Arduino is based on time slices on which a Kalman filter is carried out to avoid data acquisition noise. This software is based on the Arduino base language which is an open source language based on C/C^{++}. This developed software prohibits interaction with the user to guarantee its robustness and avoid unauthorized modifications. Then this filtered data is sent via serial to the Raspberry Pi which processes it. The code developed inside the Raspberry Pi is made up of two scripts that are running in parallel. The first is in charge of receiving the data, and storing it.

Data storage can be done under three options. The first option is to build the database and tables automatically if they do not exist and store the data continuously inside a MySQL database. The second option is to generate a file with all the SQL statements that is stored in a removable memory. Once this memory is extracted, the data can be loaded into any database for further analysis. The last option is to generate the data in a comma-separated CSV file. To define the option to use and the system configuration parameters, it is only necessary to modify a json.config file including the particular information of the analyzed PV system. In this file one can find the site information, the topology of the PV system and the data storage conditions.

In that same json.config file is the configuration for the fault detection algorithm. That script containing the fault detection algorithm is automatically executed in parallel with the data storage code. This fault detection script is in charge of executing the code by time slices and of sending the alerts in case of detecting a

fault.

In accordance with the IEC61724 standard, filtered data from the Arduino Mega is captured at 15 millisecond sampling intervals. This high sampling rate is selected well above the speeds recommended by the standard with the aim of increasing the richness of the signals and under the hypothesis that increasing the sampling frequency can increase the ability of the system to differentiate the states of the strings, even if the behaviors are similar to those of a healthy string.

In addition, thanks to the json configuration file, this developed software is easily adaptable to new variables when the hardware, the channels or the topology of the PV system change. Following the same IEC61724 standard, both the files with the SQL statements and the CSV files and the storage in the MySQL database generate a database with the same structure. They are databases in single-byte ASCII code.

The goal of making the files in this format is to make the files largely immune to computer architecture incompatibilities and to facilitate data exchange between organizations. Therefore, Solar Vitality can be linked online with many others for centralized remote control. In this prototype presented, the structure of the files first includes the date and time in the first column in *Datetime* format, that is, year, month, day, hour, minutes, seconds, milliseconds. The following eight columns represent the eight differential inputs of the ADC expressed in their respective physical variables (irradiation, temperature, wind speed, string, current, etc). The order of the variables will depend on the structure described in the configuration json file. The temperature is expressed in degrees Celsius $^{\circ}C$, the irradiation in Wm^2, the wind speed in m^{-1}, the voltage in Volts (V) and the current in Amps (A).

5.2.5 Electrical power supply

Solar Vitality is portable and autonomous in terms of energy since it can be installed even in large PV plants where there is no connection to the electrical network. First it is necessary to know the total consumption of the system. Table 5.2 shows the list of elements that make up Solar Vitality together with their uncertainty, consumption current and supply voltage.

Table 5.2: Energy consumption and uncertainty of the elements of Solar Vitality.

Task	**Model**	**Power Supply**	**Current consumption**	**Uncertainly**
Data Acquisitior	Arduino Mega	5V (Serial Port)	500 mA max	–
Data Processing	Raspberry PI 4	5.1V	$\sim$ 3.0 A	–
Wind Speed	Thies Clima	24 V	10-46 mA	$\pm$ 3 %
Irradiance	Si-V-420TC	24 V	$<$ 1 mA	$\pm$ 2.0 %
Temperature	Tm-I-4090	24 V	$\sim$2 mA	1 K
Current	IgT-MU	24 V	$\sim$10mA	$\pm$ 0.5 %

Continued on next page

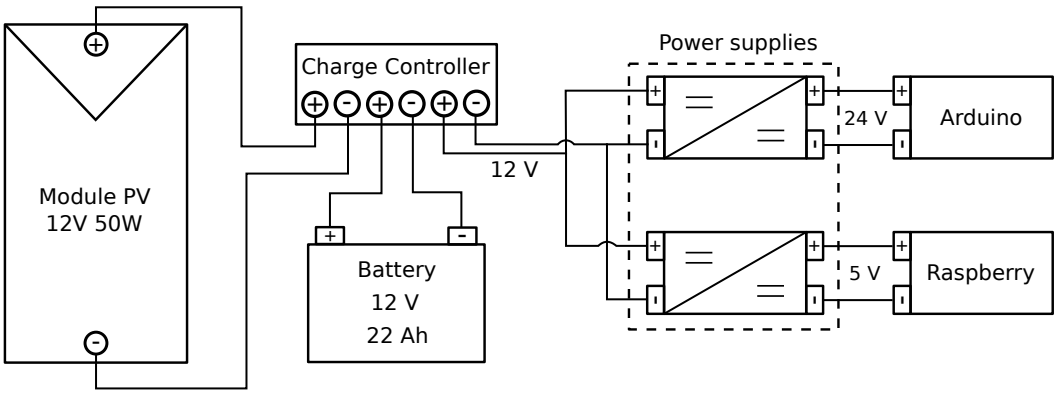

Figure 5.3: Electrical diagram of Solar Vitality

Table 5.2 – continued from previous page

Task	Model	Power Supply	Current consumption	Uncertainly
Voltage	SCK-M-U	24 V	< 8 mA	$\pm$ 1 %

As can be seen in Table 5.2, all elements are within the recommendations of the IEC 61724 standard [IEC 1998]. Based on the information in Table 5.2, an electrical power supply kit is designed and coupled with Solar Vitality. This power kit makes Solar Vitality completely autonomous in terms of energy.

The kit is composed of a 12V-50W PV panel, a 12V-22Ah battery whose charge is controlled by a 20A PWM type solar charge controller. In turn, this solar kit feeds two electrical power supplies: 1) of 24V, responsible for powering the sensors, and 2) of 5V, responsible for powering the Raspberry Pi.

The power supply of the Arduino is done through the serial cable connected to the Raspberry Pi. A diagram of the electrical power supply of Solar Vitality is presented in Figure 5.3.

5.2.6 Assembled prototype and electrical connection diagram to the PV system.

Solar Vitality is conceived to be able to be adapted to PV systems from the residential level to large utility PV plants. To be able to operate Solar Vitality it is necessary to install three main elements. The first is a PV module (12V-50W) that is responsible for charging the battery of Solar Vitality so that it works continuously. The panel can be installed on the structure of the PV system or on an external structure. The second element is the weather station and finally Solar Vitality.

Solar Vitality is built to be indoors in the same place where the PV inverters are located or, failing that, in the place where the junction boxes are. The third element is the weather station. The station is built with a conventional photovoltaic

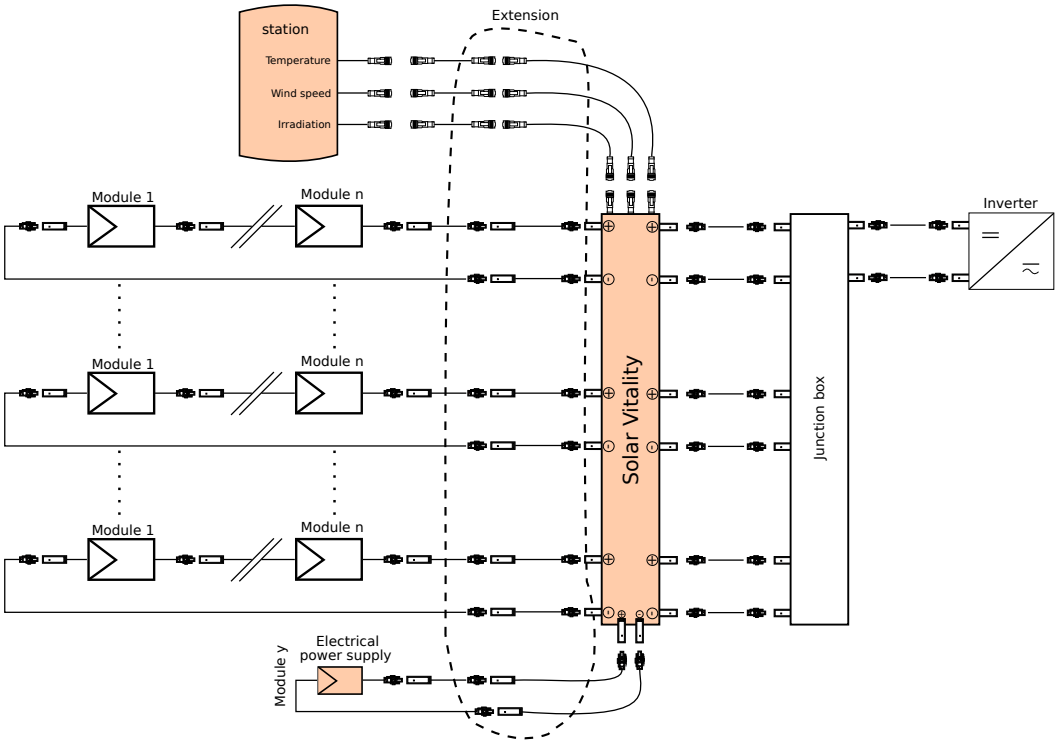

Figure 5.4: Electrical connection diagram. This schematic represents the PV system, the junction box, the PV inverter, the weather station and the external power supply system.

fixing structure, however it is adaptable depending on the surface on which the PV system is installed, that is, on the ground, on the roof, etc. Figure 5.4 shows a complete connection diagram of Solar Vitality and the weather station with the photovoltaic system.

As can be seen in Figure 5.4, the connectors are standard M-type connectors (MC_1, MC_2, MC_3 and MC_4) used in the PV domain. In addition, it is made clear that depending on the topology of the PV system, an extension of the PV cables that connect Solar Vitality with the weather station, the PV power supply module and the PV strings is necessary. This is the latest version of Solar Vitality built, however, Solar Vitality is the result of various debugging and tests carried out in the field. The evolution process of the data acquisition system based on the different instrumented PV systems is presented below.

5.2.7 Product evolution cycle and test scenarios

As in any process of design and innovation of a new product, it is necessary to carry out a process of constant improvement. For this, the product is subjected to different connection, performance and usability tests in connection to different PV systems (real operating environments). Each of these scenarios markedly improved the prototype until the current prototype, which is still undergoing constant testing,

Figure 5.5: Evolution diagram of the proposed product for Solar Vitality.

validation and improvement. Figure 5.5 shows the proposed cycle of evolution of the product.

As can be seen in Figure 5.5 the evolution cycle consists of 6 stages. A description of the process carried out in each stage is presented below.

1. Analysis: In this phase, all the characteristics of the site where Solar Vitality would be installed and the variables of interest to be measured are recovered. Among them, a small study is carried out with the client to find out the requirements and interests to present a solution tailored to their needs. Likewise, in this phase the commercial requirements are met and the potential risks are identified (external factors that prevent the correct functioning of Solar Vitality).

2. Planning: In this phase the scope of the problem is delimited, solutions are identified taking into account the price/functionality ratio. That is, the resources, costs, time and other aspects that must be considered here are evaluated. Then, a development plan is built by identifying and prioritizing features that meet the requirements of the client to build the structure of a project.

3. Design: In this phase, the future physical architecture of Solar Vitality is designed, as well as programming how each and every one of the characteristics of the product should work. Among the aspects that must be analyzed is the language of the database, type of database manager, data storage, communication protocols, processing, etc.

4. Development: In this phase, the new prototype is developed, converting the previous requirements and prototypes (if they exist) into a tangible solution. This phase is the integration of all the changes such as new sensors, new ADCs (analog to digital converters), new code, new types of storage files, etc. At the end of this process a new prototype version of the product is finished.

5. Testing: At this stage, all the tests are carried out, connecting Solar Vitality with the PV system. These tests are necessary to verify and validate the product. In this phase, the product is constantly monitored in search of errors and faults that allow it to verify the correspondence between the real and expected behavior of Solar Vitality.

6. Feedback: Finally, once the system is implemented, a retrospective evaluation is carried out that collects the needs for updates, improvements and necessary changes, to continue improving Solar Vitality. Bearing in mind that in any commercial and industrial product, such as Solar Vitality, it is crucial to maintain and modernize the system regularly so that it can be adapted to future needs.

In this way, the first 5 versions of Solar Vitality are presented below as an example. In each scenario the problems and changes are exposed.

5.2.7.1 First test scenario

In this first scenario, a single module on the terrace of the Adream building of the LAAS - CNRS laboratory is considered. This PV module is instrumented with the first version of Solar Vitality. This first version captured irradiance using a pyranometer and a reference cell. The objective of these redundant measurements of irradiation is to be able to compare the performance of the two sensors to determine the best method of capturing irradiation on the PV system.

Regarding temperature, a surface temperature probe is used. The current and voltage is measured using two transducers with current output that is then transformed into voltage to be captured in the Arduino Mega. Figure 5.6 shows the first version of Solar Vitality.

Although Solar Vitality, in Figure 5.6, correctly collected data with high quality, Solar Vitality presented multiple drawbacks. The response of the pyranometer is not the same as the response of the PV module, in addition, determining the correct position to install the temperature sensor is really a strong inconvenience considering that it must be guaranteed that the cell in which that sensor is located represents the average module temperature. Finally, this version of Solar Vitality had a strong dependence on a PC for the reception, storage and processing of data sent by the Arduino Mega. Finally, this version of Solar Vitality must also be connected to the electrical network all the time for its operation.

Due to the aforementioned limitations, a new version of photovoltaic data acquisition system is built.

5.2.7.2 Second test scenario

In this test scenario two PV strings from a small private PV installation located in Upie, in the Drôme department, in the Auvergne-Rhône-Alpes region, are considered. This installation has two PV strings each with 12 panels of 230W. In this version the surface temperature sensors are changed by an ambient temperature sensor with which the approximate temperature of the modules can be estimated. In addition, a small central computer (Raspberry Pi 4) is added that controls the acquisition, storage and treatment of the data captured by the Arduino. Likewise, a protection system is integrated to avoid all kinds of electrical risks and even possible fires in the system. To reduce the dependency of this version of Solar Vitality on the power grid, a battery and charge controller are added. This battery allowed

Figure 5.6: Installation of the first version of Solar Vitality in the LAAS-CNRS, Toulouse, France

Figure 5.7: Installation of the second version of Solar Vitality in Upie, department of Drôme, France

data capture to be carried out autonomously for two days without problems and the state and behavior of the battery could be followed using the charge controller. In this version of Solar Vitality, irradiation is only captured with a reference cell built with the same technology as the PV system panels. Figure 5.7 shows the second version of Solar Vitality.

This version of Solar Vitality again presented a high quality in the data capture of each PV string. It is also the first step towards the realization of a portable and autonomous photovoltaic data acquisition system in terms of energy. However, when the data from the two strings are analyzed in parallel, a small gap is found when the signals are sampled at rates less than one minute.

Another limitation observed is that the installation of the temperature and irradiation sensor is very complex because it depends directly on the structure of the PV system. Lastly, there is a maximum limit of two days to maintain energy autonomy.

Due to the aforementioned limitations, a new version of Solar Vitality is built.

5.2.7.3 Third test scenario

Although this third version of Solar Vitality is equipped with current sensors to measure 4 strings, in this scenario Solar Vitality is only connected to one string for security restrictions and to test connectivity.

The first modification in this version is that using the information collected from the beta charge controller, the size of a small panel is dimensioned, which is also installed on the terrace of the building and whose function is to feed the battery to prolong the capture period of data.

The tests are developed for a month and no interruption in terms of system power outage is recorded. In addition, the sensors are fixed to the same structure of the PV power supply module, ensuring its easy and adequate fixing. On the

Figure 5.8: Installation of the third version of Solar Vitality in the LAAS-CNRS

other hand, one more variable is added to the system - wind speed. This variable is added because strong changes in temperature measurement and electrical behavior are observed in version two of Solar Vitality that are related to changes in wind speed. On the other hand, considering the acquisition of signals, a second Arduino Mega is added to reduce the gap in electrical signals and a trigger signal ensures synchronization between the two devices for the transmission of signals via serial to the raspberry.

Figure 5.8 shows the third version of Solar Vitality.

Although the meteorological sensors could be kept in a fixed position and the measurements could be captured, this type of installation is difficult to adapt to other topologies and configurations of PV systems. Furthermore, the structure in which the data is stored did not correspond to the British standard BS IEC 61724 [IEC 1998]. In the same way of the data analysis, the sampling frequency is increased to seconds. This increase in the sampling frequency, which is within the recommendations of the IEC 61724 [IEC 1998] standard, revealed the limitations of the ADCs that are used for data capture. A considerable amount of noise and

Figure 5.9: Installation of the fourth version of Solar Vitality in the Delegation 14 of the CNRS, Toulouse, France.

incomplete data is transmitted at that sample rate.

Due to the aforementioned limitations, a new version of Solar Vitality is built.

5.2.7.4 Fourth test scenario

This scenario is tested in a photovoltaic installation belonging to the CNRS delegation 14 in Toulouse, France. This version of Solar Vitality captures data on wind speed, ambient temperature, solar irradiation, 4-string current and the voltage at the input of the 4-string inverter. In this version all the meteorological sensors are coupled in a meteorological station adaptable to different topologies and structures of the PV system. The weather station is adjustable to different orientations to ensure irradiation capture in the same plane as the PV system. In addition, the data is stored in a structured file organized by dates of the "Datetime" type. Likewise, to avoid the problem of the external ADCs that had been used in previous versions, the electronics are modified to increase the sampling frequency, maintaining the quality of the data until measurements are made every 0.5 seconds. This version ensures data quality and continuous data acquisition of the PV system. Figure 5.9 shows the fourth version of Solar Vitality.

Although Solar Vitality complies with the recommendations of the IEC 61724 [IEC 1998] standard, there are some factors that need to be improved. First, the data storage system had to be converted into a versatile system that would allow data to be stored in different formats or on a local server due to the large amount of data that is collected at high sampling rates. In addition, there are strong storage capacity limitations in the CSV format, since one day of data at a frequency of 0.5 seconds generates files of 2 GB per day. Finally, this photovoltaic data acquisition system is conceived for systems where multiple strings are connected in parallel in a single PV inverter. However, there are some special configurations especially in

scientific research environments where each PV string is connected to a single PV inverter. In those scenarios, if one is interested in instrumenting multiple strings, it is necessary to increase the number of voltage sensors and make modifications to the internal electronics.

Due to the aforementioned limitations, a new version of Solar Vitality is built.

5.2.7.5 Fifth test scenario

This stage is installed again on the terrace of the LAAS - CNRS. This last version of Solar Vitality is currently and for more than 6 months monitoring 4 PV strings that are in production and connected to the Adream building.

On this version of Solar Vitality, all input and output connectors are changed to standard MC4-type connectors for PV systems. In addition, the data storage system is modified to generate three types of options, all in accordance with the IEC 61724 [IEC 1998] standard. The first option builds a comma-separated CSV file that is stored on an SD memory with 1 TB storage capacity. The second option is a file that contains a list of SQL statements that is stored in the same SD memory and that can be extracted to build the database on another computer or external server. The last option automatically stores the data on a local server using the MySQL database together with the phpMyAdmin database manager. This storage medium reduces the size of the files from 2 GB, as in the previous version of Solar Vitality, to just 500 megabytes. All of these storage options are fully functional locally and do not require an internet connection or any other cloud communication protocol. The selection of the data storage system configuration is done through a JSON type file that contains all the acquisition parameters, the name of the PV system, the data export conditions, among others. The creation of the MySQL database, related tables and internal configuration is done automatically based on the conditions entered in the JSON file. This avoids direct user interaction with the internal configuration of the data acquisition system.

Another interesting aspect of this latest version is that it also has the fault detection system embedded. Likewise, this last version of Solar Vitality is capable of monitoring the meteorological conditions, the voltage and current of multiple PV strings, maintaining data quality at frequencies of 15 milliseconds. Finally, this system has the option of remote control, which requires Wi-Fi for its operation, but which allows this version of Solar Vitality to be integrated into a centralized supervision and fault detection system. All these features are easily adaptable to different types of PV systems.

Figure 5.10 shows the last version of Solar Vitality.

Solar Vitality connection scheme and the distribution of the strings measured to test the fault detection algorithms is shown in Figure 5.11.

As can be seen in Figure 5.11, the 4 strings are each connected to an independent PV inverter. An example of signals acquired with the last version of Solar Vitality is presented in Figures 5.12 and 5.13.

Other examples of signals captured with earlier versions of Solar Vitality are

Figure 5.10: Installation of the last version of Solar Vitality in the LAAS-CNRS

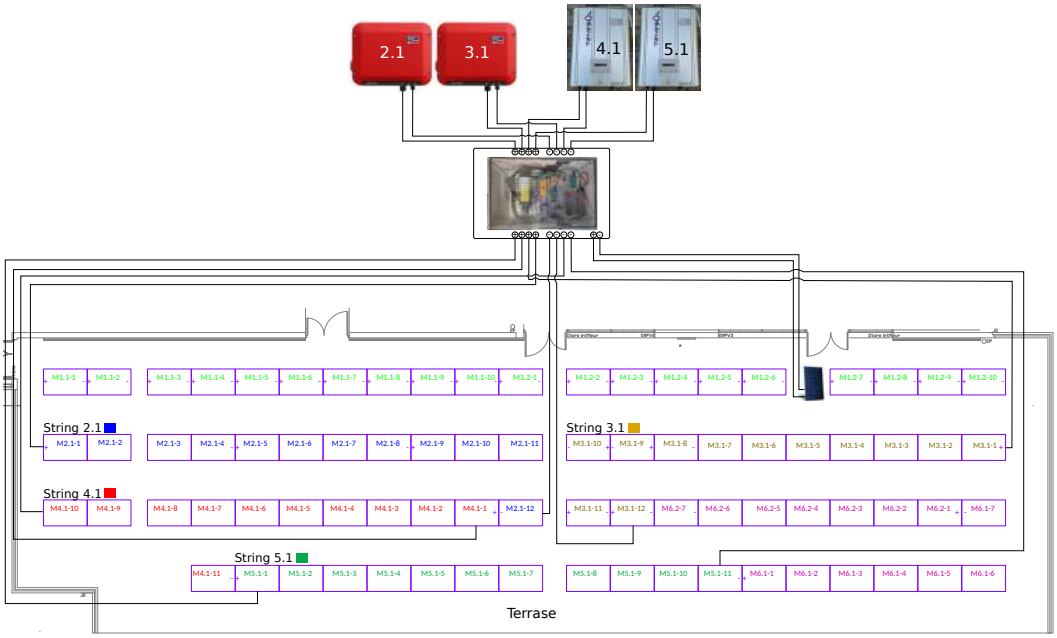

Figure 5.11: The last version of Solar Vitality on the terrace of the LAAS-CNRS. Selected PV strings, power supply panel and weather station.

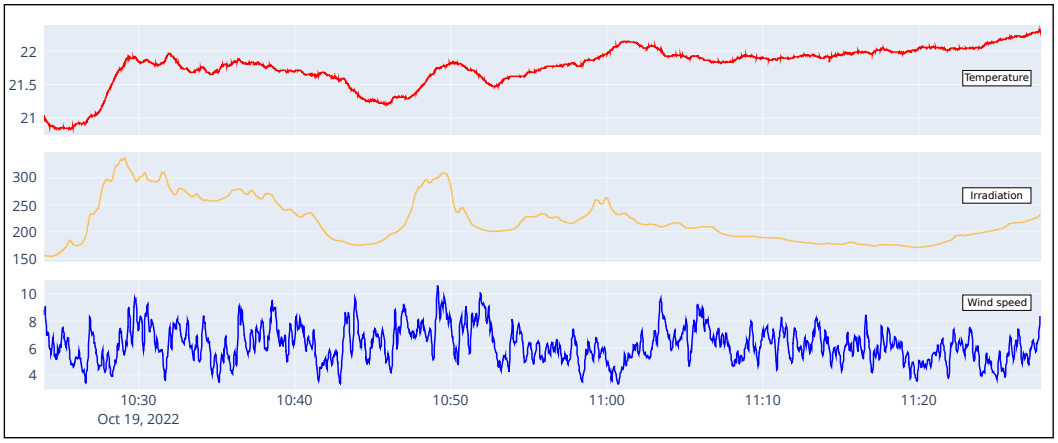

Figure 5.12: Example of measurements obtained with the weather station.

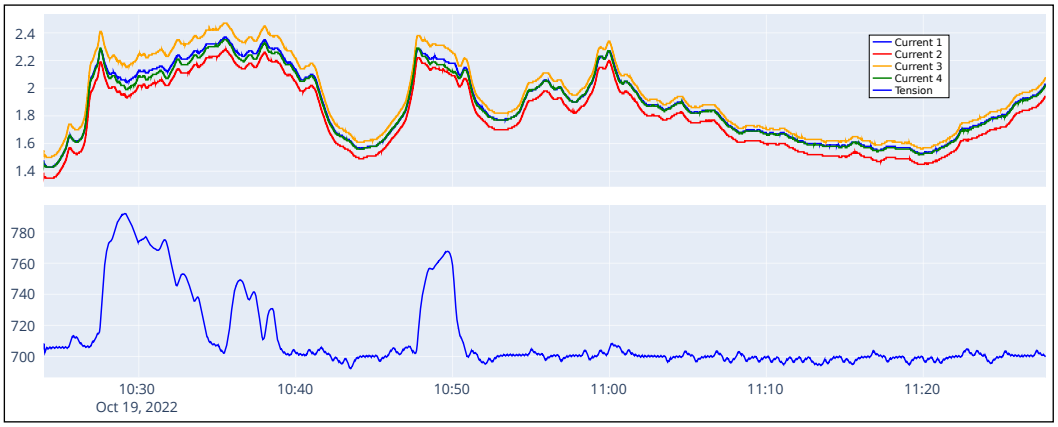

Figure 5.13: Example of current and voltage measurements taken with the last version of Solar Vitality.

presented in Annex A.1

5.3 Discussion and Conclusions

All the knowledge collected in Chapter 4 in this review is vital for the construction of Solar Vitality proposed in this chapter, complying with the recommendations of the IEC 61724 standard [IEC 1998]. Furthermore, this extensive revision allowed for a series of adaptations in Solar Vitality that make it viable and effective for small, medium and large-scale photovoltaic plants, without compromising the desired performance. Among the critical parameters of Solar Vitality, it must be taken into account that it must be guaranteed that the acquisition of all the data sent by Arduino does not exceed the sensor data acquisition time. In addition, it must be ensured that the portable power supply can supply the necessary current to avoid data loss or corrupt data on the Raspberry. When data acquisition is performed at high frequencies such as milliseconds or less, problems such as drift start to become apparent and must be addressed to avoid false fault diagnosis results. In general, Solar Vitality and the meteorological station proposed in this chapter proved to be able to efficiently monitor PV plants. In addition, due to the use of the Raspberry Pi board, it is possible to ship different machine learning algorithms that work in parallel coded in high-level languages such as Python. The platform was put into operation in different PV plants, demonstrating high performance and continuity in operation. Solar Vitality also demonstrated that it is efficient in terms of storing large amounts of data thanks to the format transformations it performs on the captured signals. It also demonstrated great versatility and easy parameterization to be adapted to different topologies of PV plants. However, based on the data captured with the Tigo data acquisition system described in the chapter 4 and the Solar Vitality prototype together with the meteorological station proposed in this chapter, it is necessary to mention that not only because high quality data acquisition and sampling rate can ensure fine fault detection. It is also necessary to carry out in-depth research on signal processing to extract the appropriate features that allow to identify and separate the different health states of the PV plant. That analysis of features processing named feature engineering is covered below in the chapter 6.

Chapter 6

Feature Engineering for Fault Diagnosis

Contents

6.1 Motivation 180
6.2 Pre-processing 181
6.3 Feature extraction 182
6.3.1 Multi-resolution signal decomposition 183
6.3.2 Features based on signal characterization 186
6.4 Feature Selection 188
6.4.1 Correlation based feature selection 188
6.4.2 Variance based feature selection 189
6.5 Feature Transformation 191
6.5.1 Principal Component Analysis 192
6.5.2 Isometric mapping 194
6.6 Discussion and Conclusions 195

Once the data is collected with any of the data acquisition systems presented in Chapters 4 and 5, the information is generally used to train Machine Learning algorithms for the detection of faults. However, as it is demonstrated in Chapter 3, many algorithms are limited when they directly use the captured signals. For this reason, it is sometimes necessary to use advanced techniques to extract features from raw data (electrical signals in the case of this thesis) that allow differentiating the different states of the PV plant. Recently the analysis of extraction, selection or transformation of features is named Feature engineering. In general, the objective of feature engineering is to create new variables that are not in the initial training set (current signals) to simplify and speed up ML model training while increasing model accuracy [PAU 2022]. With the appearance of increasingly complex faults to detect, such as those of the snail trail type, feature engineering has become increasingly important and essential.

First, two feature extraction methods based on signal decomposition in time and frequency and statistical features are proposed. Then, two feature selection methods based on correlation and variance are presented. Finally, two feature transformation methods based on Principal Component Analysis (PCA) and isometric mapping (Isomap) algorithms are exposed. The detailed description of the approaches and the results is presented below.

Figure 6.1: Example of snail trail/track fault

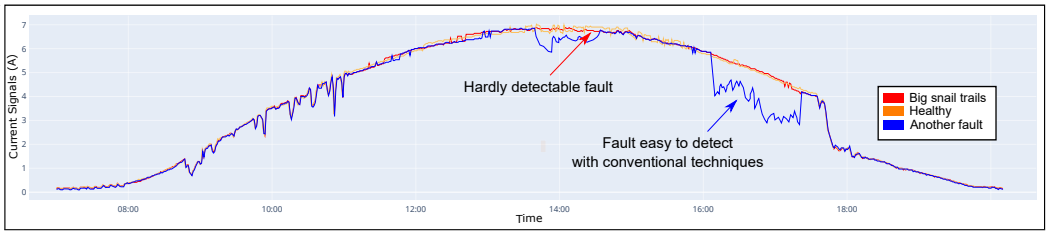

Figure 6.2: Example of current of three PV modules in status of health, Healthy (yellow), other fault (blue) and snail trail (red)

6.1 Motivation

As presented in Chapter 3, in fault detection on photovoltaic systems, multiple methods based on machine learning are proposed [Pillai 2019a, Navid 2021a, Livera 2019b, Carvalho 2019, Madeti 2017b, Alam 2015b, Okere 2020, Alam 2013c, Jadidi 2020, Pillai 2018b, Mellit 2018a, Lu 2018, Hare 2016, Tsanakas 2016, Mellit 2021, Li 2021b, Triki-Lahiani 2018b]. However, most of these works have focused on faults whose energy reduction in the PV system is critical. Damages caused by these faults induce the reduction of the power generated, from very low levels to cause the complete stoppage of the system. This makes sense if solutions oriented to corrective maintenance are thought, however, few works are concerned with developing algorithms oriented to the detection of faults whose electrical signature is equal or similar to that of a non faulty PV panel, i.e. whose energy reduction to the PV system is low. In this category there are faults such as the snail trail (see Figure 6.1), which, although it does not generate a considerable loss of energy, is the cause of multiple damages that can cause fire in the PV plant [Duerr 2016, Li 2021c, Kim 2016, Koentges 2014]. An example of the difference between the current signal of a panel with a fault that generates large power loss (blue line), a panel with a snail trail fault (red line - hardly detectable fault) and a healthy panel (blue line - Fault easy to detect with conventional techniques) is presented in Figure 6.2.

As can be seen in Figure 6.2, achieving fault detection with large power loss could be a trivial and straightforward task using conventional machine learning algorithms [Tsai 2015, Sepúlveda Oviedo 2021, Chouay 2021, He 2021, Sepúlveda Oviedo 2022, Akram 2015]. However, managing to detect almost un-

noticeable faults like snail trails is really a challenge. It is important to emphasize that the sooner a fault such as a snail trail is detected and classified, in order to carry out preventive maintenance on the defective part, the greater the production guarantee. This means that the useful life of the photovoltaic plants is lengthened and the cost of maintenance is largely reduced. The literature assesses that this type of fault (snail trail/Snail track) occurs after approximately 3 months to 1 year of exposure to the open air of the PV array [Li 2021c]. This fault primarily affects crystalline silicon cells and often occurs at the edges of cells [Fadhel 2018]. The origin of this fault is not clear, but some documents mention that it may be due to silver particles containing sulfur, phosphorus or carbon [Li 2021c] that may be included in the manufacturing process accidentally. Furthermore, as presented in Chapter 2, the snail trail fault can cause localized temperature increases (hot spots), non-uniform degradation, corrosion, among others and then more important impacts in degraded performances of power productions. This thesis defines a fine fault as a fault whose electrical signature is similar to the one of a healthy panel (for example the snail trail). All these facts highlight the importance of developing research such as efficient and early detection and classification of faults in these systems to guarantee high performances and in a long term, low-cost continuity of service.

To carry out fault detection in general, it is first necessary to perform data acquisition and pre-processing of the current signals as a function of the time of all the panels. Then, a stage of feature extraction followed by a feature selection stage and/or feature transformation can be performed that allow to differentiate the states of the PV panels.

6.2 Pre-processing

First the electric current signals I_i for a PV panel PV_i, $i = 1, \ldots, n$, are obtained. For a PV panel PV_i, the data takes the form of a time series denoted by $I_{i\{1:n_I\}} = \{i_{i,t_1}, \ldots, i_{i,t_{n_I}}\}$, where n_I is the number of samples of the $i-th$ time series that has a sampling period of one minute and $t_i, i = 1..n_I$, is the date of the sample. However, when the raw signals are obtained from the acquisition system, they are not directly ready for feature extraction. This is because these raw signals can sometimes contain missing or null values that can influence the performance of any algorithm that takes this data as input. Data cleaning is an elementary phase that must precede all other phases. Because no null data was found, but missing current values in some scenarios, this thesis uses the Equation (6.1) to replace the missing current values $i_{i,t}$ as:

$$i_{i,t} = \frac{i_{i,t-1} + i_{i,t+1}}{2}, \tag{6.1}$$

Although the data cleansing process is not complex, it is very efficient and it is an indispensable tool to eliminate most faults that affect the performance of the decomposition algorithms to be applied further. Figure 6.3 presents the PV panel

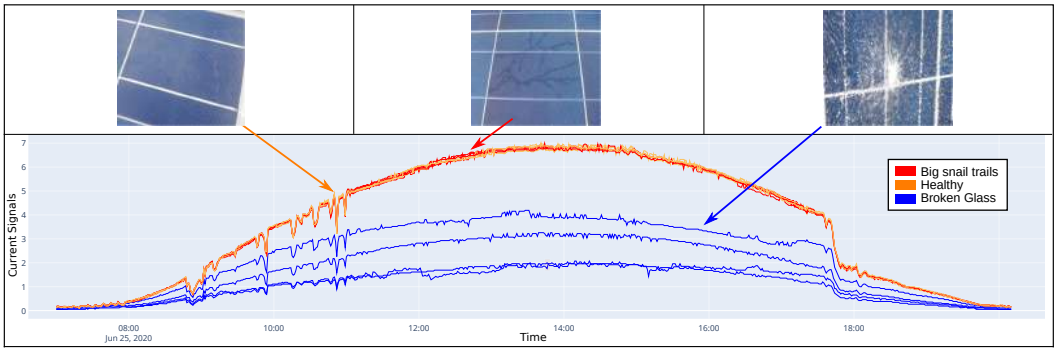

Figure 6.3: Behavior of the current over one day for different health statuses: healthy (yellow), broken glass (blue), and snail trails (red) for a period of 13 hours every minute.

current behaviors, captured with the Tigo acquisition system described in Chapter 4, over one day for different health statuses after data cleaning, along with an illustration of the physical appearance of the faults on the PV module.

The blue color corresponds to the PV panels with a broken glass fault, the yellow color corresponds to the healthy PV panels and the red color to the snail trail fault. The snail trail represents corrosion of the sheet of the encapsulation surface and although it does not significantly decrease the performance of the PV panels, it can be the cause of cracks or micro cracks in the modules that reduce the production of a PV panel. As shown in Figure 6.3, the behavior of the PV panels with a snail trail is very similar to that of healthy PV panels. It is in such complex cases of detection where feature engineering plays a vital and essential role.

Once the current signal of each panel is captured (in the form of a time series) it is necessary to carry out a feature extraction stage that allows extracting details of the signals that are used to discriminate the different health states of the panels.

6.3 Feature extraction

The key to achieving high-precision detection is the availability of a robust, high-quality database. To that point, close attention should be paid not only to the quality of the sensors and, more generally, to all parts of the measurement chain mentioned in Chapter 4, but also to the different types of pre-processing that can be performed on the raw data. Few approaches in the literature have focused on improving feature extraction and trying to select appropriate methods, aimed at ensuring the quality of the data that is used to identify faults that occur in the PV system. The small number of researches is largely due to the high complexity of the data acquisition system required. In fact, in order to diagnose faults that occur in photovoltaic systems, the main difficulties are linked to fault signatures that vary with weather conditions, to the performance of inverters or optimizers, and to other causes that must be taken into account.

This thesis proposes the exploration of two different feature extraction tech-

niques. First, a signal decomposition method is explored that allows an extraction of features that contain information in time and frequency. Then, a feature extraction based on statistical characteristics is exposed. These two techniques have already been explored previously in the literature for fault detection [Ahmad 2018, Kurukuru 2020, Haque 2019, Dadhich 2019, Fatama 2019]. The two proposed methods are explained below.

6.3.1 Multi-resolution signal decomposition

On the way of signal decomposition techniques, Continuous Fourier Transform (FT), Discrete Fourier Transform (DFT) among others are been proposed for fault detection [Harrou 2019b, Pedersen 2019, Belaout 2018a, Lebreton 2022]. However, these transformations only provide information about the frequency. In [Ji 2016] the authors propose the Fourier Transform with a window that provides both time and frequency information. However, the fixed window selection may not always be efficient for detecting critical non-stationary disturbances, such as three-phase faults and short circuits associated with [Ray 2018] transients. Alternatively, in recent years, the wavelet transform (WT) started to gain popularity [Haje Obeid 2017, Bayram 2017, Costa 2015, Sangeetha 2018] due to its multi-resolution time-frequency analysis. This type of decomposition shows better identification characteristics of all types of faults in photovoltaic systems, as long as the presence of noise in the signal is avoided [Ray 2018]. Based on the WT, different modifications are proposed, such as: the Multiresolution Signal Decomposition (MSD) that applies the wavelet decomposition iteratively [Yi 2017a], the Slantlet transform [Ahmadipour 2018b] that is based on a modified discrete wavelet transform with two zero moments and modified temporal localization and the wavelet packet transform (WPT) that performs an iterative decomposition on the high and low frequency coefficients [Ahmadipour 2018a, Kumar 2018].

Every faulty condition in a PV system is associated with a change in the output current. These changes are reflected as variations in the waveform of the output signal compared to a healthy PV panel. Some of these changes are visible in the frequency domain and others in the time domain. In order to analyze these changes simultaneously (time - frequency), Multi-resolution Signal Decomposition is used. The Multiresolution Signal Decomposition is based on the discrete wavelet transform (DWT) that can decompose a signal into levels with different time and frequency resolutions using the wavelet transform iteratively [Wang 2014b].

In the following, DWT is presented in a generic form. In our case study, it is applied to the current times series $I_{i\{1:n_I\}}$ of each PV panel PV_i, $i = 1, \ldots, n$.

DWT is a signal processing technique (linear transformation) like the Fourier transform [Wang 2013]. Some of the differences between these two techniques can be read in [Daubechies 1990]. DWT decomposes the input signal into a variable frequency range that depends on the *mother wavelet* selected as the decomposition pattern [Ray 2018]. The input signal is decomposed into approximate and detailed coefficients that correspond to the high and low frequency components respectively.

DWT is known for its properties to simultaneously analyze frequency and time [Kashyap 2003, Etemadi 2008, Mallat 2008]. As mentioned in [Wang 2013], the wavelet transformation with the proper *mother wavelet* is a useful tool for fault detection and feature extraction. For this reason DWT is widely used in this field [Belaout 2018b, Shaik 2015, Ray 2018, Ray 2016].

Wavelet decomposition uses a *mother wavelet* that decomposes the signal into a set of oscillatory functions named wavelets. Each of these *mother wavelets* is a signal in time that captures a specific frequency band [Zhao 2000, Pang 2010]. There are different well-known families of *discrete mother wavelets* such as: Harr, Meyer, Bior, Daubechies, Rbio, Coiflet and Symmlet. Which are composed of 1,1,15,38,15,17 and 19 *mother wavelets* respectively. Each of these *mother wavelets* has a different computational calculation speed and decomposition quality depending on the particular application.

Some *mother wavelets* are particularly used in the PV domain, for example: Sym8 [Yi 2017a] from the Symmlet family, Harr [Kumar 2018], and db1, db3 - db5, db8, db9, [Haque 2019, Ray 2018, Dadhich 2019, Kurukuru 2020, Ahmad 2018, Yi 2017a, Wang 2013, Wang 2014b] from the Daubechies family. Each of these wavelet families is defined according to Equation (6.2) [Iyer 2013, Dadhich 2019].

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \tag{6.2}$$

where a is the scale or dilation factor, b is the shifting factor, t refers to the timestamp of the input signal, and ψ is defined as the *mother wavelet* [Iyer 2013]. To restrict the values of a and b to discrete values, these factors are defined according to Equations (6.3) and (6.4) [Yi 2017c, Iyer 2013].

$$a = a_0^{-(m_x/2)}, \tag{6.3}$$

$$b = n_x b_0 a_0^{m_x}, \tag{6.4}$$

where m_x and n_x range over $\mathbb{Z}$ and $a_0 > 1$ and $b_0 > 0$ are fixed [Iyer 2013]. The DWT of the discrete signal $X_{\{1:n_X\}}$ that uses the *mother wavelet* $\psi_{a,b}(t)$ of Equation (6.2) is described in Equation (6.5)[Yi 2017c, Dadhich 2019, Kurukuru 2020, Iyer 2013]:

$$\mathrm{DWT}(a,b) = \frac{1}{\sqrt{a}}\sum_{1:n_X} X(t)\psi\left(\frac{t-b}{a}\right), \tag{6.5}$$

For the decomposition of the signal, it is necessary to select the appropriate *mother wavelet* $\psi_{a,b}(t)$. Some works proposed complex algorithms for the optimal selection of the *mother wavelet* [Singh 2006]. In this article, the mother wavelet selection follows the work of Wang et al. [Wang 2014b] that aims at detecting faults in PV systems with wavelet transform. According to Wang et, al., [Wang 2014b] the selected wavelet must comply:

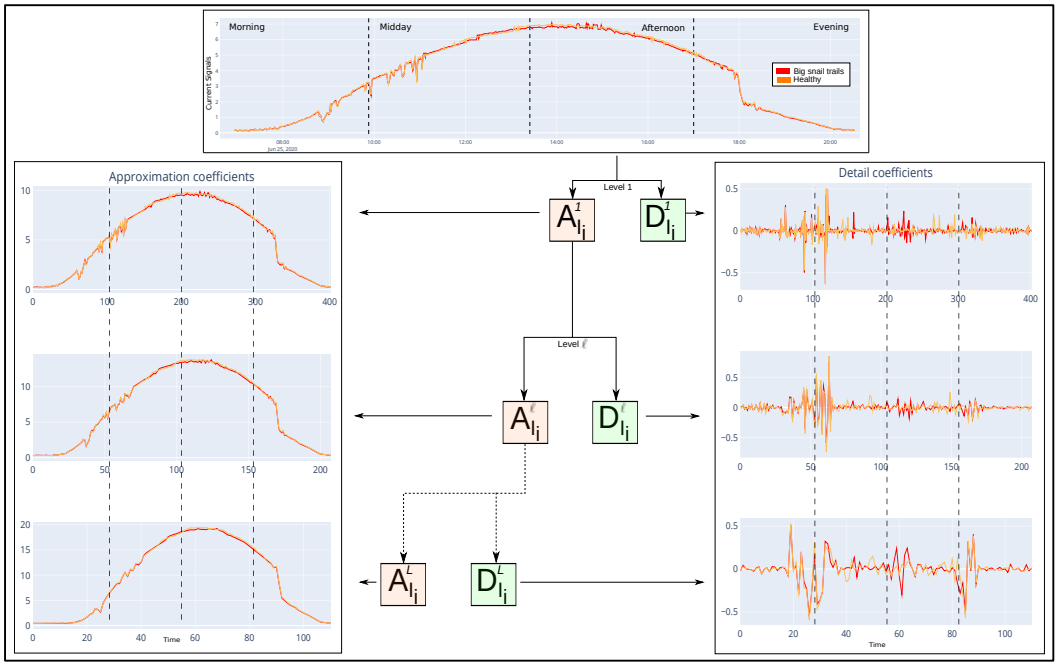

Figure 6.4: Decomposition into 3 levels of the current signal for a panel with big snail trails (red) and a healthy panel (yellow). The approximation and detail coefficients resulting from the decomposition are presented on the left and right of the figure, respectively.

1. To have a sufficient number of vanishing moments to represent the salient features of the anomalies.

2. To provide sharp cutoff frequencies to reduce the amount of leakage energy into the adjacent resolution levels.

3. The wavelet basis should be orthonormal.

Taking into account these considerations and the fact that most of the work in fault detection in PV systems use wavelets of the Daubechies family (db), the entire family is tested and the Daubechies38 (db38) *mother wavelet* is selected due to its computational speed and good decomposition result.

Multi-resolution signal decomposition can be performed at different levels of decomposition. For example, the result of the decomposition into 3 levels for a current signal I_i is shown in Figure 6.4. At each level ℓ of decomposition of a current signal I_i, two signals can be created as the result of the wavelet transform. The first signal corresponds to the approximation coefficients ($A^{\ell}_{I_i}$). This signal receives this name due to the fact that it is an approximation of the “low frequency" components of I_i. The second signal corresponds to the detail coefficients ($D^{\ell}_{I_i}$). This signal represents the “high frequency" corrections of the signal I_i.

The original signal I_i can be reconstructed from the detail and approximation coefficients. The reconstructed signal $I_{i,r}$ is the sum of all the detail coefficients prior

to the last selected level L, with the detail and approximation coefficients of level L. This description is formally presented in Equation (6.6) [Wang 2014b, Jensen 2001].

$$I_{i,r} = A_{I_i}^{L} + \sum_{\ell=1}^{L} D_{I_i}^{\ell}, \tag{6.6}$$

Although the decomposition of the signal can make some details of the time series more evident, a strong time dependency is still preserved. As a solution to eliminate this time dependency, this thesis proposes feature extraction based on signal characterization to pass from a temporal space to a non-temporal. Feature extraction based on signal characterization is explained below

6.3.2 Features based on signal characterization

Statistical features have already been used in previous works to extract relevant statistical properties from PV system data [Li 2021c]. These features increase the variance between the different classes [Kurukuru 2020]. Features such as mean, variance, skewness, kurtosis, entropy, among others are suggested for troubleshooting PV systems [Ahmad 2018, Kurukuru 2020, Haque 2019, Dadhich 2019]. Each of these features has a better or worse performance depending on the type of fault to be analyzed. Generally these features are used as input for different fault classification methods [Sharma 2016, Arunkumar 2019, Hui 2017, Nanopoulos 2001]. Even in other works this extraction of statistical features is combined with a signal decomposition method based on Multiresolution Signal Decomposition [Ahmad 2018, Kurukuru 2020, Haque 2019, Dadhich 2019]. In this subsection, features are first presented for a generic signal. Then the signals that are used for their extraction in our case study are made explicit.

For a given generic signal X represented by a time series $X_{\{1:n_X\}}$, a number of features can be extracted. Note that the selected features retain only some characteristics of the signal, which has an impact on the possible discrimination of different signals. The n_F selected features are chosen to capture several characteristics of a signal. These selected features are considered as they are also used in previous works aimed at fault diagnosis in *PV* systems [Ray 2018, Ahmad 2018, Kurukuru 2020, Haque 2019, Dadhich 2019, Vergura 2009, Ismail 2016, Wang 2018a] and works aimed at fault diagnosis in vibration signals [Arunkumar 2019, Sharma 2016, Xia 2012, Goyal 2020]. Given a time series $X_{\{1:n_X\}}$ of mean μ, these features are :

- *Skewness (F1)*: skewness represents the asymmetry of the data with respect to the mean and is calculated by Equation (6.7) [Esmael 2012].

$$F1 = \frac{\frac{1}{n_X}\sum_{t=0}^{n_X}(X_t - \mu)^3}{\left(\sqrt{\frac{1}{n_X}\sum_{t=0}^{n_X}(X_t - \mu)^2}\right)^3}, \tag{6.7}$$

- *Kurtosis (F2)*: kurtosis measures the peak of the probability distribution of the data. It also allows knowing how prone to outliers is a distribution.

Kurtosis is defined according to Equation (6.8) [Esmael 2012].

$$F2 = \frac{\frac{1}{n_X}\sum_{t=0}^{n_X}(X_t-\mu)^4}{\left(\sqrt{\frac{1}{n_X}\sum_{t=0}^{n_X}(X_t-\mu)^2}\right)^4}, \tag{6.8}$$

- *Variance (F3)*: the variance represents the variability of a series of data with respect to its mean.

$$F3 = \frac{1}{n_X}\sum_{t=0}^{n_X}(X_t-\mu)^2, \tag{6.9}$$

- $P - P_{value}$ *(F4)*: the peak-to-peak distance (p-p) is the distance between the peak with the highest amplitude and the valley with the lowest amplitude.

$$F4 = max(X_t) - min(X_t), \tag{6.10}$$

- *Energy (F5)*: explain the energy contained in the signal, it is conserved regardless of whether it is in frequency or in time [Ray 2018].

$$F5 = \sum_{t=0}^{n_X} X_t^2, \tag{6.11}$$

- *Power spectral density (F6)*: Power spectral density represents the power content of the signal as a function of frequency. The amplitude is normalized per unit frequency as seen below:

$$F6 = \lim_{n_X\to\infty}\frac{1}{n_X}|X_t|^2 \tag{6.12}$$

- *Entropy (F7)*: Entropy is widely used in information theory to evaluate the uncertainty of a signal and even as a tool to identify the quality of the information or inherent surprise of the signal. Entropy can be defined as:

$$F7 = -\sum_{t=1}^{n_X} p(X_t)log(X_t) \tag{6.13}$$

These statistical characteristics can be directly extracted from the raw signal. However, to obtain the maximum richness in the training information for machine learning algorithms, this thesis proposes to perform the multiresolution signal decomposition followed by the features based on signal characterization in the following way. The characterization of the operational condition of a PV panel PV_i is performed with the set of $L+1$ time series $\{A^L_{I_i}, D^\ell_{I_i},\ \ell = 1,\ldots,L\}$, obtained from the L levels multi-resolution decomposition of the corresponding current signal I_i. These time series are segmented in four time slices corresponding to morning, midday, afternoon, and evening. Sliced signals are indexed accordingly by $* \in \{$ *morning*, *midday*, $afternoon$, *evening* $\}$ and we obtain the set $S \in \{A^L_{I_{i,*}}, D^\ell_{I_{i,*}}\}$, $i = 1,\ldots,n_P$, $\ell = 1,\ldots,L$, where n_P is the number of PV panels.

The selected features are then determined for each time series in S, forming a feature vector composed by the feature subvector $FA^L_{I_{i,*}}$ for the approximation coefficients and the features subvectors $FD^\ell_{I_{i,*}}$ for detail coefficients, each sub vector being of dimension n_F. The characterization of every time slice can be summarized in a matrix of dimensions $n_P \times ((L+1) \times n_F)$:

$$\mathbb{F}_* = \begin{pmatrix} FA^L_{I_{1,*}} & FD^1_{I_{1,*}} & \dots & FD^\ell_{I_{1,*}} & \dots & FD^L_{I_{1,*}} \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ . & . & \dots & . & \dots & . \\ FA^L_{I_{n_P,*}} & FD^1_{I_{n_P,*}} & \dots & FD^\ell_{I_{n_P,*}} & \dots & FD^L_{I_{n_P,*}} \end{pmatrix}, \tag{6.14}$$

$$* \in \{\text{morning, midday, afternoon, evening}\}$$

Each row $\mathbb{F}_*(i,.)$ of the matrix $\mathbb{F}_*$ provides the signature of the health state of the PV panel PV_i.

It is important to mention that some of the features in each high dimensionality signature may provide redundant information, which may reduce the performance of data-based diagnosis algorithms [Lambrou 1998, Chen 2014]. Therefore, it is necessary to select the outstanding features by means of an algorithm that identifies a subset of features that preserve the fine details related to a faulty state as represented in the high dimensional space. As mentioned in [Esmael 2012], a high dimensionality data set could be reduced by brute-force with an exhaustive search enumerating and testing all the feature subsets. However, it is more efficient to use feature selection and feature transformation algorithms. For this reason this chapter proposes two types of feature selection and two types of features transformation. First the two feature selection techniques are explained. Then the techniques based on feature transformation are exposed.

6.4 Feature Selection

For a given generic matrix of features $\mathbb{F}$ of dimensions $(n_P \times \eta_b)$, whose η_b columns represent the features that characterize the health status of n_P individuals, a set of $\eta_c^\oplus$ features, where $\eta_c^\oplus \subseteq \eta_b$ features that preserve relevant details for class discrimination can be selected. The selection of the $\eta_c^\oplus$ features is first based on correlation, then on variance analysis. In a first selection step, highly correlated features are discarded. Then the remaining weakly correlated features are given as input to the variance based feature selection algorithm. Without lack of generality, feature selection is presented for the matrix of features $\mathbb{F}_{morning}$ obtained with 3 levels of decomposition. It can be easily extrapolated to L levels of decomposition in any of the 4 time slices of interest.

6.4.1 Correlation based feature selection

Correlation based feature selection allows to choose a subset of η_c uncorrelated relevant features with high predictive value to create solid learning models for the

n_P individuals in matrix $\mathbb{F}$. In the literature, it has been previously mentioned that a feature is redundant if one or more other features are highly correlated with it. The use of the *Pearson's correlation matrix* for these analyzes has been proposed in social science works [Zajonc 1962, Buško 2011, Hogarth 1977]. Correlation based feature selection uses the *Pearson's correlation matrix* to determine the degree of correlation between the initial features η_b. The level of correlation between two features ranges between -1 and 1, with 1 being the highest positive correlation and -1 the highest negative correlation. 0 indicates no correlation at all. The more a feature is correlated to another, the less information it brings while it can introduce noise. Thus, it is recommended to eliminate it [Esmael 2012]. A correlation threshold $\tau_{\mathbb{F}}$ is defined to remove the correlated features that are out of the range $[-\tau_{\mathbb{F}}, \tau_{\mathbb{F}}]$ and form a set of uncorrelated features of cardinal η_c that will be used for class discrimination and that reduce the matrix $\mathbb{F}$ into $\mathbf{F}$.

The selection of the uncorrelated features corresponding to the columns of the matrix $\mathbb{F}_*$ that contain the relevant details of the health states of each PV panel PV_i is performed with a correlation threshold $\tau_{\mathbb{F}_*} = 0.9$. As an example, the correlation based feature selection on the matrix $\mathbb{F}_{morning}$ is presented in Figure 6.5. Figure 6.5a provides the correlation matrix crossing the η_b initial features before feature selection. Figure 6.5b provides the correlation matrix crossing the η_c weakly correlated features after eliminating the strongly correlated features. With this feature selection, the number of features is decreased from 20 to 14 uncorrelated features for the matrix $\mathbb{F}_{morning}$. In other words, the feature dimension is reduced by 40%. Correlation based feature selection is carried out for each matrix $\mathbb{F}_*$, obtaining the matrices $\mathbf{F}_*$, where $* \in \{morning, midday, afternoon, evening\}$.

6.4.2 Variance based feature selection

Now, it is not because features are not strongly correlated that they have a strong discriminating power for a classification problem. For this reason, a feature selection based on variance is also applied. For this purpose, *parallel coordinates* is used. This technique, based on the variability of the features [Steed 2012], is widely used in multivariate data analysis [Johansson 2016]. In the parallel coordinates, there are as many normalized axes as features.

For a given matrix of features $\mathbf{F}$ of dimensions $(n_P \times \eta_c)$, whose η_c columns represent the uncorrelated features that characterize the health status of n_P individuals, there are $\eta_c^{\oplus}$ features, $\eta_c^{\oplus} \subseteq \eta_c$, that preserve relevant details and present significant variance between the n_P individuals. To select the $\eta_c^{\oplus}$ features, the variance of the η_c features is compared between the rows $\mathbf{F}(\mathbf{i}, .)$, $i = 1, \ldots, n_P$ of the matrix $\mathbf{F}$. Those features that do not show a significant variation are not selected to form the final set of $\eta_c^{\oplus}$ features, reducing the matrix $\mathbf{F}$ into the matrix F of dimension $(n_P \times \eta_c^{\oplus})$.

To illustrate the variance based feature selection, Figure 6.6 shows the plot of parallel coordinates on the matrix $\mathbf{F_{morning}}$ resulting from the correlation based feature selection. The horizontal axis of Figure 6.6 represents the η_c weakly corre-

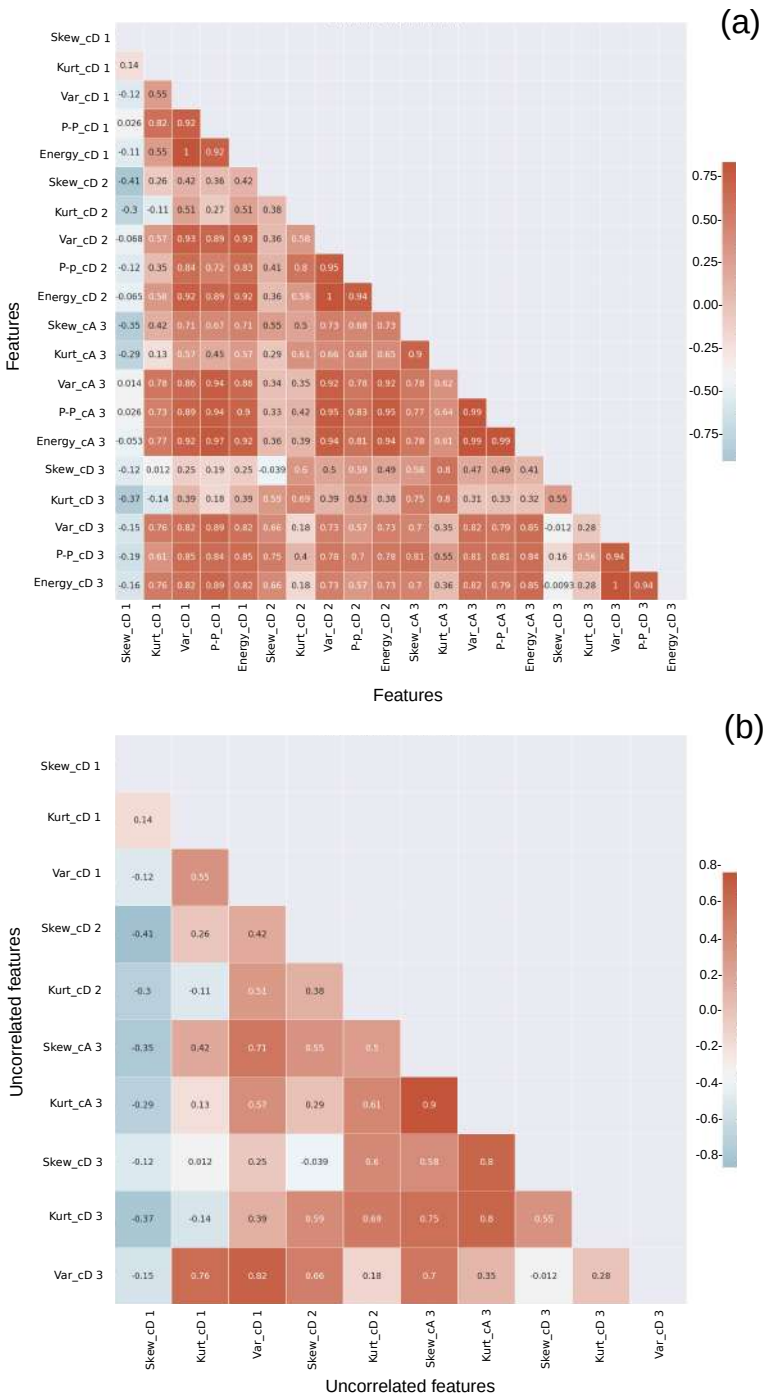

Figure 6.5: Correlation matrices: (**a**) Pearson correlation matrix of the η_b initial features of $\mathbb{F}_{morning}$, i.e., before correlation based feature selection; (**b**) Pearson correlation matrix of the η_c uncorrelated features of $\mathbf{F}_{\mathbf{morning}}$, i.e., after correlation based feature selection.

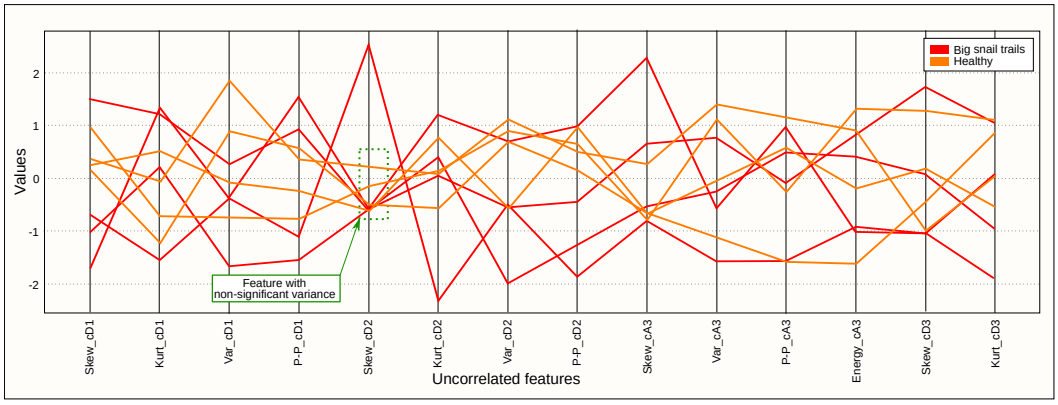

Figure 6.6: Parallel coordinates plot on the matrix $\mathbf{F_{morning}}$. The normalized values of the uncorrelated features η_c are plotted on the vertical axis. The horizontal axis represents the uncorrelated features η_c.

lated features and the vertical axis represents their normalized values. As shown in Figure 6.6, the feature $Skew\_CD2$ (corresponding to skewness extracted on the detail coefficients at level 2) does not provide significant variance to distinguish between the different operating states (healthy and big snail trails). Therefore, this feature is not selected to form the final set of features $\eta_c^{\oplus}$. By discarding this feature, the matrix $\mathbf{F_{morning}}$ that had 14 features is reduced to matrix $F_{morning}$ with 13 features, maintaining the relevant information to identify the different operating conditions of the PV panels.

Variance based feature selection is applied to all matrices $\mathbf{F}_*$, where $* \in \{ morning, midday, afternoon, evening \}$, leading to four reduced feature matrices F_* of 13, 12, 11 and 16 dimensions respectively.

However, alternatively, feature transformation methods can be used.

6.5 Feature Transformation

Due to the high dimensionality (high number of features), of the matrix $\mathbb{F}_*$ the computational cost of this classification is very high. Therefore, feature transformation methods are proposed to reduce the high dimensionality of features with minimal loss of information. This feature transformation creates a compressed version of the original feature matrix $\mathbb{F}_*$ [Jolliffe 2002]. One of the main advantages of this process consists of obtain a drastic decrease of the computational time of the learning algorithms increasing their efficiencies and their capabilities to treat complex big data.

Multiple algorithms for feature selections and dimensionality reduction are proposed in the literature. Some of the best known approaches are Principal Component Analysis (PCA) [Wang 2015, Tipping 1999, Haque 2019, Kurukuru 2020, Esmael 2012, Xia 2012, Basnet 2020, Zhao 2020, Hajji 2021, Onal 2021], Isometric Mapping (Isomap) [Tenenbaum 2000a]; Local Linear Embedding (LLE)

[Roweis 2000, Wang 2015], Singular Value Decomposition (SVD) [Alter 2000]; Multiple Resolution Analysis (MRA) [Yi 2017b, Khoshnami 2018]; among many others [van der Maaten 2008b, Schölkopf 1999, Ross 2008, Usman 2017, Xia 2012, Donoho 2003, Ng 2001, Zhang 2004, Hyvärinen 2000, Huang 2019, Jenatton 2010].

As for the case of feature selection, in feature transformation, for a given generic matrix of features $\mathbb{F}$ of dimensions $(n_P \times \eta_b)$, whose η_b columns represent the features that characterize the health status of n_P individuals, a set of $\eta_c^{\oplus}$ features, where $\eta_c^{\oplus} \subseteq \eta_b$ features that preserve relevant details for class discrimination can be selected. The selection of features $\eta_c^{\oplus}$ using feature transformation can be performed using PCA or Isomap. These two algorithms are selected because they are widely used in the field of feature transformation. As a result of the feature transformation, a transformation of the matrix $\mathbb{F}_*$ into the new matrix $M_{F,*}$ of dimensions $n_P \times U$ described as follows is obtained:

$$M_{F,*} = \begin{pmatrix} C^1_{I_{1,*}} & \dots & C^{u-1}_{I_{1,*}} & \dots & C^U_{I_{1,*}} \\ . & \dots & . & \dots & . \\ . & \dots & . & \dots & . \\ . & \dots & . & \dots & . \\ C^1_{I_{n_P,*}} & \dots & C^{u-1}_{I_{n_P,*}} & \dots & C^U_{I_{n_P,*}} \end{pmatrix}, \tag{6.15}$$

where U is the number of latent components or the number of selected features obtained as results of the PCA or Isomap, depending on the case. Each row $M_{F,*}(i,.)$, $i = 1, \dots, n_P$ of the matrix $M_{F,*}$ provides the signature of the health state of the PV panel PV_i. Each column $M_{F,*}(.,j)$, $j = 1, \dots, U$ of the matrix $M_{F,*}$ provides each of the latent components.

The description of the two algorithms selected for feature transformation, which result in the matrices $M_{F,*}^{PCA}$ and $M_{F,*}^{ISOMAP}$, is presented below.

6.5.1 Principal Component Analysis

Principal Component Analysis (PCA) is a powerful multivariate statistical technique that identifies and extracts uncorrelated attributes (named latent components) from the multidimensional space of system variables. That is, the PCA algorithm uses a linear combination of the original variables to construct the new variables while maintaining the maximum variance information [Fadhel 2019a]. The PCA represents the new variables in two subspaces [Fadhel 2019a]. The first, named the main subspace or the "representation" subspace. The second named complementary or residual subspace in which noises and outliers are rejected.

The search for the PCA, in the multidimensional space of the M variables, is carried out through a decomposition problem of the eigenvectors of the data covariance matrix. The first principal components, spanning the principal subspace, are given by the first q dominant eigenvectors of the data covariance matrix. Those dominant eigenvectors are associated with the highest q eigenvalues. The last unretained eigenvectors $(M - q)$ define the residual subspace. In the representation subspace containing the most significant variations, the eigenvectors are named load

vectors and the projection of the data onto these load vectors is named principal component scores [Fadhel 2019a]. The set of the load vector and the component score is named the search direction. The percentage of variance contained in each principal component (PC) is expressed by its corresponding eigenvalue. Additionally, each PC is aligned in a direction corresponding to the largest variation in the data, starting with the first PC. Therefore, the principal components are ordered from the most energized associated with the highest eigenvalue, to the least energized associated with the lowest eigenvalue. Finally, the main subspace is covered by the majority of powered PCs, while the residual is covered by the remaining PCs [Fadhel 2019a].

Without loss of generality and to simplify the notation, the dimensions of the matrix $\mathbb{F}_*$ will henceforth be expressed as $N \times M$, where $N = n_P$ and $M = ((L + 1) \times n_F)$. The matrix $\mathbb{F}_*$ will be also named the matrix $\mathbf{X}$. Let us consider that we want to select the features with the highest variance from the matrix $\mathbf{X}$, where x_j, $j = 1, \ldots, N$ is the j^{th} variable. First, we must center (mean zero) and reduce (unit variance) the variables for each observation $n_i, i = 1, \ldots, M$, of x_j as follows:

$$x_{i,c}n_i = \frac{x_j(n_i) - (\bar{x_j})}{\sigma_j}, \tag{6.16}$$

where $(x_j)_c$ is the centered and reduced variable, (x_j) and (σ_j) are respectively the mean value and the standard deviation of (x_j). With these values it is possible to form the new centered and reduced data matrix $\mathbf{X_c}$ of dimensions $(N \times M)$. Each row $\mathbf{X_c}(j, .)$ of the matrix $\mathbf{X_c}$ provides the status information (healthy or snail trail) of the n_P panels PV, centered and reduced. Once $\mathbf{X_c}$ is obtained, the covariance matrix is calculated as follows:

$$\mathbf{C} = \frac{1}{N-1}\mathbf{X_c}^T\mathbf{X_c}, \tag{6.17}$$

where $\mathbf{X_c}^T$ denotes the $\mathbf{X_c}$ transposed. The quality of the obtained representation depends on the latent components retained in the main representation space. Let us denote it by $\mathbf{P}$ as the column matrix of the charge vectors or linear transformation matrix, which are arranged in descending order of their corresponding eigenvalues [Fadhel 2019a]. The principal component scores are obtained by projecting the original centered and reduced data into the new space generated with $\mathbf{P}$ obtaining the matrix $\mathbf{T}$ of the principal component scores of dimensions $(N \times M)$. That is, the linear transformation matrix P transforms $\mathbf{X_c}$ into a new matrix of latent components $\mathbf{T}$ as follows:

$$\mathbf{T} = \mathbf{PX}, \tag{6.18}$$

where, each column $\mathbf{T}(., i)$ of the matrix $\mathbf{T}$ provides a PC for the set of PV panels n_P.

6.5.2 Isometric mapping

Alternatively to PCA, dimensionality reduction or features selection can be performed using the Isomap. Isomap stands for isometric mapping. This method approaches dimensionality reduction as the problem of creating a transformation from high dimension to low dimension as a graph problem [Samko 2006]. Isomap extends the metric multidimensional scale (MDS) [Hout 2013] by incorporating the concept of geodesic distances imposed by a weighted graph [Bouttier 2003].

In the domain of graph theory, the distance between two vertices on a graph corresponds to the number of edges in a shortest path connecting them. This distance is also known as the geodesic distance [Bouttier 2003]. Isomap is intended to preserve pairwise geodesic distances between conformations in a graph, that is, in the lower dimension. The distances $d_{\mathbf{X}}(i,j)$ between all pairs i, j of N data points in the high-dimensional input space $\mathbf{X}$ are required as input to the Isomap algorithm, generally measured using the standard Euclidean distance. The algorithm outputs coordinate vectors $\mathbf{Y}_i$ in a (lower) d-dimensional Euclidean space $\mathbf{Y}$ that best represents the intrinsic geometry of the data. Dimensionality reduction or feature selection using Isomap is based on three steps:

First step. The Isomap estimates the neighborhood graph. To obtain it, it need to start determined each of its points which are neighbors in the manifold $\mathbf{M}$, based on the distances $d_{\mathbf{X}}(i,j)$ between pairs of input points i, j in the input space $\mathbf{X}$. Then, having these input points, the set of neighbors for each point is determined. To determine the neighbors, the K nearest neighbors (K-Isomap) can be used for all those within a fixed radius ϵ (ϵ-Isomap) [Samko 2006]. Neighborhood relationships are plotted as a weighted graph $\mathbf{G}$ over the data points, with weighted edges $d_{\mathbf{X}}(i,j)$ between neighboring points. As mentioned in [Samko 2006], if the neighbors were determined using the K-Isomap method, the vertices in the graph can have degree greater than K since the nearest K neighborhood relationship need not be symmetric [Samko 2006]

Second step. The Isomap computes the shortest path graph given the neighborhood graph. Isomap then estimates the geodesic distances $d_{\mathbf{M}}(i,j)$ between all pairs of points in the manifold by computing the shortest path lengths $d_{\mathbf{G}}(i,j)$ in $\mathbf{G}$. $d_{\mathbf{G}}(i,j) = d_{\mathbf{X}}(i,j)$ if i, j are joined by an edge, and $d_{\mathbf{G}}(i,j) = \infty$ in otherwise. Then, for each value of k_i, $i = 1,\ldots,N$, all entries $d_{\mathbf{G}}(i,j)$ are replaced by $min\{d_{\mathbf{G}}(i,j), d_{\mathbf{G}}(i,k) + d_{\mathbf{G}}(k,j)\}$. The array of final values $\mathbf{D}_{\mathbf{G}} = \{d_{\mathbf{G}}(i,j)\}$ will contain the lengths of the shortest paths between all pairs of points in $\mathbf{G}$.

Third and final step. The Isomap constructs the lower dimensional embedding using classical MDS to the graph distance matrix $\mathbf{D}_{\mathbf{G}} = \{d_{\mathbf{G}}(i,j)\}$, constructing an embedding of the data in a d-dimensional Euclidean space that best preserves the estimated intrinsic geometry of the manifold. The coordinate vectors Y_i for points in Y are chosen to minimize the following cost function:

$$E = |\tau(\mathbf{D}_{\mathbf{G}}) - \tau(\mathbf{D}_{\mathbf{Y}})|_{L^2}, \tag{6.19}$$

where $\mathbf{D}_{\mathbf{Y}}$ denotes the matrix of Euclidean distances $\{d_{\mathbf{y}}(i,j) = |y_i - y_j|\}$ and $|A|l^2$

is the matrix norm. The τ operator converts distances to inner products, which uniquely characterize the geometry of the data in a way that supports efficient [Tenenbaum 2000b] optimization.

As it can be seen in the Isomap dimensionality reduction steps, only the free parameter is the neighborhood factor K or ϵ depending on the method used. The success of the Isomap method transformation lies in choosing an appropriate value for these two parameters. Generally these parameters are selected manually [Tenenbaum 2000b].

In order to provide an idea of the performance of some well-known supervised and unsupervised machine learning algorithms, Annex A.2 presents a series of tests with algorithms such as k-means and Random Forest (RF). These algorithms are tested using the feature extraction and transformation techniques proposed in this chapter. Taking into account the limitations that those algorithms demonstrated, the proposal of advanced machine learning algorithms adapted to this problem but that can be easily extrapolated to other domains makes sense. The advanced machine learning algorithms proposed in this thesis are described in Chapters 7-9.

6.6 Discussion and Conclusions

The set of feature extraction, selection and transformation techniques presented in this chapter demonstrated that they are capable of extracting small details from the signals, increasing the richness of the predictor matrix that can be used for the detection of snail trail faults and broken glass. In the case of fine faults such as the snail trail, in Figure 6.4, in the detail coefficients the panels with snail trail begin to stand out. This is extremely important since it is the first time that a fault detection study is carried out using artificial intelligence aimed at detecting this type of fault. In addition, the proposed signal decomposition and dimensionality reduction and selection methods based on correlation and variance managed to reduce the number of features from the original features matrix by 40%. This reduction of features while maintaining the pertinent information is key to guaranteeing accuracy in fault detection, reducing the computational calculation time. The reduction of computation is also a key aspect considering that the objective of all the approaches proposed in Chapters 7-9 is to embark them in the new PV systems data acquisition system presented in Chapter 5. On the other hand, regarding the PCA and Isomap methods are methods that when transforming to other spaces are capable of highlighting features that were not visible in the original space and even depending on the number of latent components selected, the number of features can be considerably reduced while also maintaining the information relevant.

Chapter 7

An Ensemble Based Diagnosis Algorithm (EB-diag)

Contents

7.1 Approach description 198
7.2 Dataset 199
7.3 Selected features for fault detection 200
7.4 EB-diag composition 202
7.4.1 k-Nearest-Neighbor 202
7.4.2 Support Vector Machines 203
7.4.3 Decision Trees 205
7.4.4 Majority voting 207
7.5 Discussion and Conclusions 208

Performance, safety and reliability of photovoltaic plants are strongly linked to the ability to detect abnormal loss of power production and faults as soon as they appear. For these reasons, a major objective in this field is to develop intelligent Fault Detection and Isolation (FDI) paradigms that may greatly benefit from Ensemble Learning (EL). Until now, these techniques applied on sustainable energy sources as high photovoltaic PV power plants proved to be too complex. Nevertheless, recent advances of the scientific community make these techniques more applicable and can hence ensure high-performance operation of PV systems. The technique proposed, in this chapter, combines several learning models, namely the Support Vector Machine (SVM), K-Nearest Neighbor (kNN) and Decision Trees (DT), instead of using a single learning model. The combined model is aimed at detecting classical faults but its own distinguishing property over existing models is its capacity at detecting faults whose electrical features are similar to that of a healthy panel. In the proposed methodology, first, a predictor matrix is built by extracting time-frequency characteristics (using wavelet decomposition) and statistics from the string PV current signal. Then, due to the high dimension of the matrix of predictors, two feature selection and dimensionality reduction algorithms (PCA and Isomap) are used. Finally, the reduced predictor matrix is constituted between the ensemble learning algorithms. This method is validated with a real PV string of 8 panels (4 healthy and 4 with snail trail).

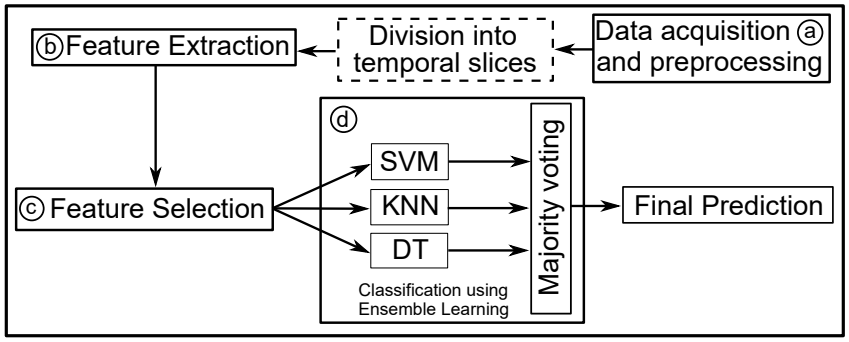

Figure 7.1: Description of stages of the proposed approach. a) Data acquisition and preprocessing. b) Feature extraction. c) Feature selection. d) Fault detection and classification based on Ensemble Learning.

7.1 Approach description

It is important to mention that as demonstrated in Chapter 2, few studies have focused on the detection of fine faults such as snail trail, due to the high similarity of the electrical signal of this fault compared to the electrical signal of a healthy panel. It is for this reason that as a contribution to the detection of snail trail faults, this chapter presents a new approach that takes advantage of multiple learning algorithms, named Ensemble Learning (EL). EL is a new approache assembling properties of different techniques with good trade-offs [Sagi 2018]. The main idea of this proposed method is to combine several models in a meta-algorithm combining most properties of each technique in the aim to improve the classification results of any of the base fault detection techniques. To combine the results of the diagnosis, currently there exists multiple options such as average, weighted average, majority voting and weighted majority voting.

A summary of the proposed methodology is presented in Figure 7.1.

As shown in Figure 7.1, it is first necessary to capture the panel string current and perform the respective pre-processing. This stage is named *Data acquisition and pre-processing*. Once this data acquisition stage is accomplished, in order to fully explore and then exploit the richness of the effect of any change in the electrical current signal of each PV panel strongly connected with any change on weather conditions in particular in level and quality of irradiation at different times of the day, the signals are divided into 4 time slices named: Morning, Midday, Afternoon and Evening as proposed in [Sepúlveda Oviedo 2022]. Then, the third stage is concerned with *Feature extraction*, working under the assumption that some faults may be visible in the time domain and others in the frequency domain, a multiresolution signal decomposition based on the discrete wavelet transform over each time slice is used to analyses them simultaneously. As a product of this decomposition, a set of detail and approximation coefficients are obtained on which the extraction of statistical characteristics is performed. The features obtained from the 4 time slices are then put together in a matrix of predictors noted F^*. This choice allows improvements on

the performance of diagnosis algorithms. The fourth stage named *Feature selection* is performed. In that stage each corresponding matrix F^* is subjected to a feature selection process based on a space transformation (Isomap) [Tenenbaum 2000b] and Principal component analysis (PCA) [Esmael 2012]. This reduction of dimensionality allows to compress data representing an important quantity of information with a major part not pertinent or is redundant to obtain at the end only a significantly smaller number, keeping only the pertinent information. Finally, the stage *Fault detection and classification based on Ensemble Learning* is carried out. In that stage, the new reduced matrix F can be used as input to a set of machine learning algorithms (K Nearest Neighbors (KNN), Support-vector machine SVM and Decision tree learning) that detect and classify PV system faults. To reach good results, the 3 algorithms are combined based on the method named "majority voting" in the approach named ensemble learning (EL).

The objective of the approach, presented in this chapter, is to improve the results of fault detection and classification in PV systems even in cases where there is a reduced number of PV panels (2 or more panels). In addition, improving the computational time taken by conventional fault detection systems while increasing their accuracy are other objectives of this approach. The photovoltaic platform used in this investigation has 16 PV modules. These modules are divided into 8 for training and 8 for testing. The classifiers used in this chapter are trained using 8 modules, 4 in a healthy state and 4 with the snail trail fault. The methodology is validated to date from a series of 8 panels, different from those of the training, of which 4 panels are healthy panels and 4 panels show traces of discoloration (snail tracks or snail trails). Both the training signals and the signals for testing and validation were captured for a period of one day, in each of the seasons of the year. Information needed is only one variable. The choice is to measure the current of each panel obtained with a Tigo optimizer and its associated automatized data acquisition able to capture and store data. The objective is to evaluate and then validate the proposed approach with the help of an existing data acquisition obtained with a commercial product including its available data acquisition described in the next section.

In order to fully explore and then exploit the richness of the effect of any change in the electrical current signal of each PV panel strongly connected with any change on weather conditions in particular in level and quality of irradiation at different times of the day, the signals are divided into 4 time slices named: Morning, Midday, Afternoon and Evening as proposed in [Sepúlveda Oviedo 2022].

Each of the stages of the methodology is explained in detail below.

7.2 Dataset

The fault detection presented in this chapter is performed on a photovoltaic plant located in the Adream building of the LAAS-CNRS laboratory in Toulouse, France. The PV plant is made up of 16 PV modules with reference *SLK60P6L*, with the

capacity to generate power between 205 and 240 Wp. Each of the modules is instrumented with a commercial data acquisition system provided by the company Tigo[1].

This Tigo platform is capable of capturing the voltage and current of each module. However, for this study only the electrical current signal of each PV module is used. These 8 current signals build a current matrix named $F*$ for each time slice. Tigo's data acquisition system captures current signals with a sampling time of one minute. The signals used in this chapter are captured in the year 2020 on August 6 between 7:00 a.m. and 8:00 p.m., November 6 between 7:45 a.m. and 5:15 p.m., February 6 between 8:00 a.m. and 6:00 p.m. and finally on May 6 between 7:00 a.m. and 8:00 p.m. These dates were carefully selected approximately in the middle of each of the seasons of the year, to measure the robustness of the proposed approach. The data began to be captured every day as soon as the PV panel began to produce. The data capture in each day is finished once the panel stops producing. The signals captured on each day of each season are shown in Figure 7.2.

In the same Figure 7.2, it is possible to observe the 4 time slices (Morning, Midday, Afternoon and Evening). The electrical signals framed in the orange color in Figure 7.2 correspond to the healthy PV panels. Other signals in red color correspond to PV panels with snail trail faults. This type of default represents corrosion of the sheet of the encapsulation surface [Li 2021c]. Although at the beginning this fault does not cause a severe or critical reduction in the performance of the photovoltaic panels, with time, if the panels continue to be exposed to the same conditions of solar radiation, the fault can evolve producing cracks or microcracks in the PV cells that can even completely stop the production of the PV system [Kim 2016, Koentges 2014]. As it can be seen in Figure 7.2, the behavior of photovoltaic panels with a snail trail (Red color) is very similar to that of healthy photovoltaic panels (orange) in all scenarios (Summer, Fall, Winter and Spring). Once the data is captured, the feature extraction is performed.

7.3 Selected features for fault detection

The feature extraction and selection used in the approach proposed in this chapter is based on some of the algorithms explained in Chapter 6. First, feature extraction is performed using multi-resolution signal decomposition with 3 decomposition levels. For the selection of the mother wavelet, in this chapter, all wavelet families were tested and the *mother wavelet* Daubechies38 (db38) due to its computational speed and good decomposition result. Then the extraction of statistical features *Mean* μ, Skewness (F1), Kurtosis (F2), Power spectral density (F6) and Entropy (F7), explained in Section 6.3.2, is carried out. Finally, to reduce the dimensionality of the matrix or select the appropriate features, the PCA and Isomap methods, explained in Sections 6.5.1 and respectively, are used independently. In the case

[1]To obtain the description of this system, visit here

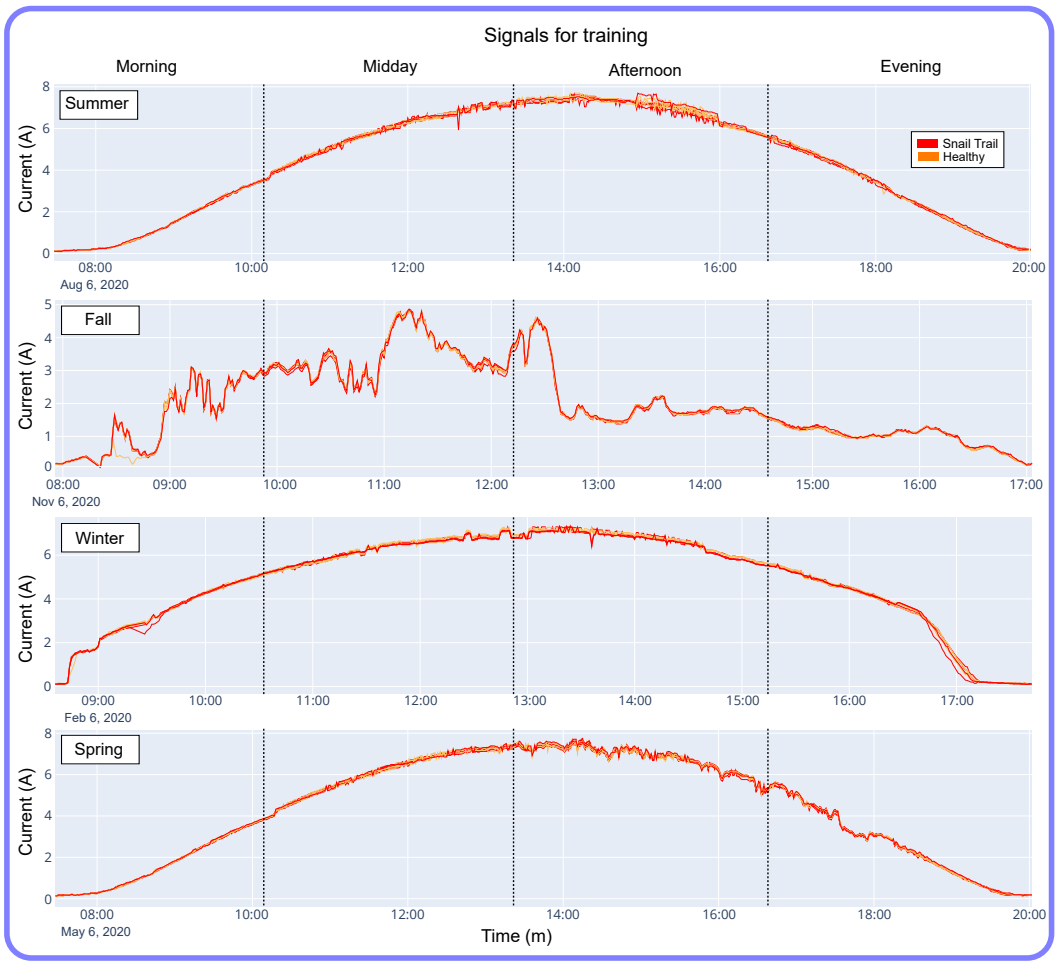

Figure 7.2: Electric current signals from 8 photovoltaic modules used in the training of the proposed methodology. The signals are captured during a full day in the 4 seasons of the year. for different health states: healthy (orange) and snail trails (red). The data is captured with a frequency of one minute. The 4 time slices proposed [Sepúlveda Oviedo 2022] and adopted in this chapter are represented using dotted lines.

study, applying feature selection to $\mathbb{M}_{\mathbb{F},*}$ obtained after feature extraction, where $* \in \{\ morning, midday, afternoon, evening\ \}$, the matrices $M_{F,*}^{PCA}$ and $M_{F,*}^{ISOMAP}$ are obtained respectively, as explained in the Section 6.5.

7.4 EB-diag composition

In this section, the main details of machine learning techniques used in this work are described. The classification methods are applied in parallel to the reduced matrices resulting from the PCA and Isomap methods $M_{F,*}^{PCA}$ and $M_{F,*}^{ISOMAP}$ and to the matrix without the feature selection $\mathbb{M}_{\mathbb{F},*}$. Performing the classification on the matrices $M_{F,*}^{PCA}$ and $M_{F,*}^{ISOMAP}$ significantly reduces the calculation time, since the high dimensionality of the features is reduced with a minimum loss of information. A description of the three methods that constitute the Ensemble Learning method is presented below.

7.4.1 k-Nearest-Neighbor

The non-parametric algorithm K-Nearest-Neighbor (kNN) is one of the most used models for classification thanks to its features and simplicity [Zhang 2007]. kNN finds the nearest neighbors for a given sample based on some distance metric of interest [Wang 2020]. To determine the class kNN, it is considered that the samples of known class are $x = [x_1 x_2, \ldots, x_k]$ and those of the data to be classified are $y = [y_1, \ldots, y_k]$. According to [Dhibi 2021], the distance between the two samples x and y is defined as:

$$d(x, y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}, \tag{7.1}$$

Then the kNN assigns the element to a class in which the distance of formulation (7.1) is minimum. The only free parameter to fit is the number of k nearest neighbors in the training sample space. The proper choice of k has a significant impact on the diagnosis performance of the kNN algorithm. In the simplest case where k = 1, the object-oriented class is the nearest neighbor class of the object. On one hand, as the value of k increases, the model can tolerate more noise. That is, a large k reduces the impact of variance caused by random error, but risks missing a small but important pattern. On the other hand, a small value of k makes the model more sensitive to noise. The key to choose an appropriate value of k is to strike a balance between overfitting and underfitting [Zhang 2014]. Overfitting occurs when the model learns complex details and noise from the training data to the point where it detracts from its effectiveness on new data [Zhang 2014]. On the other hand, when a machine learning model is underfitting, it does not learn correctly from the training data and has serious difficulties classifying new data [Zhang 2014]. Figure 7.3 shows an example of the K-Nearest-Neighbor (kNN) classification.

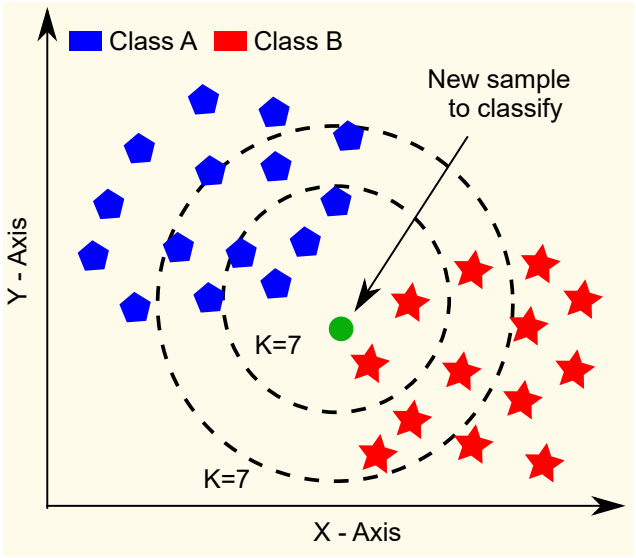

Figure 7.3: Example of classifying a new sample using K-Nearest-Neighbor (kNN).

This chapter uses the Euclidean distance, however there are other interesting metrics such as the Hamming distance and the Manhattan distance [Ruan 2017].

7.4.2 Support Vector Machines

The support vector machine (SVM) is one of the most powerful, complex and widely applied classification algorithms for fault diagnosis [Cervantes 2020] [Natarajan 2020]. SVM uses, unlike passive learning methods, an objective function obtained from the training data to make the classification decisions. The main idea of SVM is to map the input space training data into a higher dimensional feature space through a mapping function and then apply linear SVM on this space. In general, the SVM classifier seeks to find an optimal separating hyperplane as a decision plane, maximizing the margin between two classes. As an example to facilitate the understanding of the SVM algorithm, a two-dimensional plane is analyzed.

In this two-dimensional case, the SVM finds a straight line, also named the optimal classification surface or hyperplane H. This is the line that separates the samples into two types of classes. In this method there are also so-called support vectors. Each support vector corresponds to the closest sample to the hyperplane H [Cervantes 2020]. Figure 7.4 illustrates the hyperplane H and the support vectors of SVM.

In Figure 7.4 the blue dots and the red squares represent two different sample types. H_1 and H_2 are the two parallel hyperplanes which are the support vectors of the two sides. As it can be seen in Figure 7.4 there is also the concept of margin. A margin, sometimes named a class interval, represents the distance between the hyperplanes H_1 and H_2.

To formally understand the SVM, suppose we have two classes of balanced

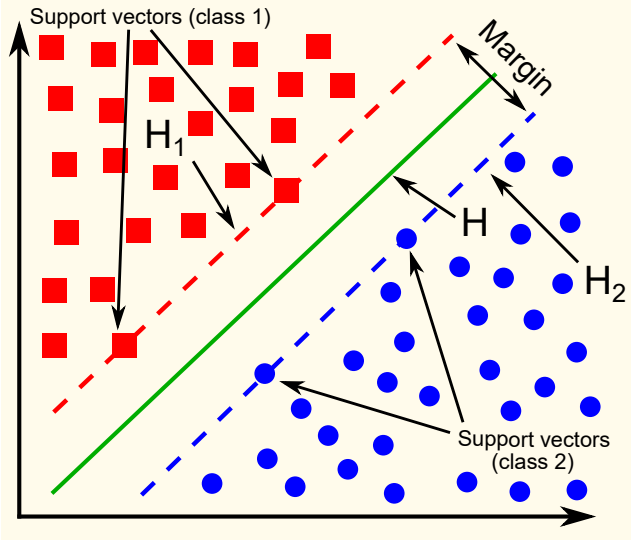

Figure 7.4: Example of classification hyperplane representation of SVM algorithm. H_1 and H_2 correspond to the hyperplanes of classes 1 and 2 respectively. H corresponds to the optimal hyperplane.

data $\mathbf{x} = [x_1, \ldots, x_n]$ and $\mathbf{y} = [y_1, \ldots, y_n]$, $n = 1, \ldots, N$, where N is the total number of samples in each class. A pair of samples is defined as (x_i, y_i). The optimal hyperplane divides linear separable samples (x_i, y_i) into two categories and is formally defined as:

$$H = (\mathbf{W}^T\mathbf{x}) + b = 0, \tag{7.2}$$

where $\mathbf{W}$ is a vector perpendicular to the hyperplane that represents the weight vector and determines the direction of the hyperplane. b denotes the bias vector or displacement term, which determines the distance between the hyperplane and the origin. The distance from any point in the sample space to the hyper-plane can be written as:

$$\mathbf{Y} = \frac{|\mathbf{W}^T\mathbf{x} + b|}{|\mathbf{W}|}, \tag{7.3}$$

The sum of the distance between the two heterogeneous support vectors and the hyper-plane is:

$$\mathbf{Y} = \frac{2}{|\mathbf{W}|}, \tag{7.4}$$

To find the maximum margin, that is, to find $\mathbf{W}$ and b the following constrained optimization problem must be solved:

$$\min_{\mathbf{W},b} \frac{1}{2}|\mathbf{W}|^2, \quad s.t \quad y_i(\mathbf{W}^T\mathbf{x_i} + b) \geq 1, i = 1, \ldots, n, \tag{7.5}$$

The basic SVM model assumes that the training samples are linearly separable in the sample space. This means that there is a hyperplane to divide the samples into different categories. However, it is often difficult to find the proper kernel function for Equation (7.5) in the real application [Cervantes 2020]. The Kernel Function can be seen as a method used to take data as input and transform it into another required form [Chen 2014]. As a solution to this problem support vector machines are allowed to make errors in some samples. This introduces the concept of "soft margin". Vector machines compatible with the "soft margin" are defined according to [Chapelle 2001] as:

$$\min_{\mathbf{W},b,\xi} \frac{1}{2}|\mathbf{W}|^2 + C\sum_{i=1}^{n}\xi_i, \quad s.t \;\; y_i(\mathbf{W}^T\mathbf{x_i} + b) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \;\; i = 1, \ldots, n, \quad (7.6)$$

where ξ_i. is the distance of $\mathbf{x_i}$ from the corresponding class's margin if $\mathbf{x_i}$ is on the wrong side of the margin, otherwise zero. As it can see Equation (7.5) required all samples to satisfy the constraint, which is named "hard margin" while Equation (7.6) While the margin is maximized, the number of samples that do not satisfy the constraint must be minimized. For this, Lagrangian multipliers are introduced in Equation (7.6), obtaining:

$$\mathbf{L}(\mathbf{W}, b, \alpha, \xi, \mu) = \frac{1}{2}|W|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - \xi_i - y_i(\mathbf{W}^T\mathbf{x_i} + b)) - \sum_{i=1}^{n}\mu_i\xi_i, \quad (7.7)$$

where $\alpha_i \geq 0$, $\mu_i \geq 0$ is the Lagrange multiplier. Solving the Equation 7.7 the optimal classification function is obtained as follows:

$$f(x) = sign\,(\mathbf{W}^T\mathbf{x}) + b) = \left(\sum_{i=1}^{n}\alpha_i\mu_i y_i K(x_i, x) + b\right), \quad (7.8)$$

where $\mathbf{x}$ denotes the input vector to be classified. K is the Kernel function and where $\alpha_i \geq 0$, $\mu_i \geq 0$ and $\xi_i \geq 0$.

7.4.3 Decision Trees

Tree-based ML techniques are among the most widely used nonlinear models in many applications, where Random Forest (RF) and Decision Tree (DT) are the most popular having in some cases an accuracy greater than that of neural networks [Lundberg 2020]. The DT model uses two types of nodes, which are the decision node and the leaf node. Decision nodes have multiple branches and are used to make any decision, while leaf nodes are the result of these decisions [Mahesh 2019]. An illustration of these nodes is presented in Figure 7.5.

DTs are successive models where a numerical feature is compared to a threshold value at each test. In general, conceptual rules are much easier to construct than numerical weights for a neural network of connections between nodes. DTs are mainly used for clustering and data mining purposes [Anuradha 2014].

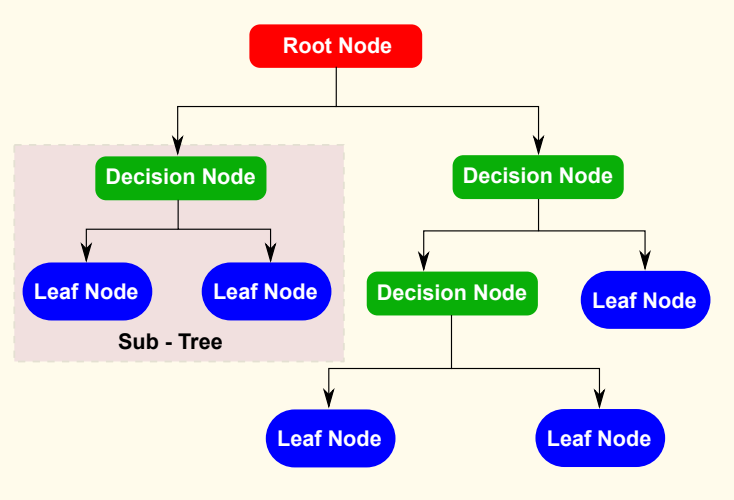

Figure 7.5: Structure of the Decision Tree (DT) classifier. The DT classifier is composed of two types of nodes. The first type corresponds to the decision node and the second to the leaf node.

The Decision Tree model is also a positive learning model, more exactly, it is defined as a tree-structured classification model including common algorithms such as the Iterative Dichotomies 3 (ID3), Successor of ID3 (C4.5), Automatic Chi-Square Interaction Detector (CHAID), Classification and Regression Tree (CART), Generalized Unbiased Interaction Detection and Estimation (GUIDE), Multivariate Adaptive Regression Splines (MARS), Classification Rule with Unbiased Interaction Selection and Estimation (CRUISE), Conditional Inference Trees (CTREE) or Unbiased and Efficient Statistical Tree (QUEST)[Jiao 2020].

In the methodology proposed, algorithm C4.5 is selected due to its results in detecting faults in PV systems [Benkercha 2018]. The C4.5 algorithm extracts the conditional entropy in Equation (7.9) and the information using the entropy in Equation (7.10) of the sample as follows:

$$H(X) = H(p) = -\sum_{i} P_i log_2 P_i, \tag{7.9}$$

$$H(Y|X) = \sum_{i} P_i H(Y|X = x_i), \tag{7.10}$$

where P_i is the ratio of the sample number of the subset and the i-th attribute value. That is, P_i is the probability that the sample X belongs to the category i. Resulting in the information gain equation as follows:

$$G(D, A) = H(D) - H(D|A), \tag{7.11}$$

where D and A are the whole sample set and the specific sample set respectively. The C4.5 classifier uses the information gain ratio to establish the information entropy of D over A.

To carry out ensemble learning it is necessary to use a final label selection strategy. The final label selection strategy is explained below.

7.4.4 Majority voting

The three classifiers described above (kNN, SVM, and DT) can be considered as "weak learners" that can be integrated to create an improved learner by training them to work together [Zhang 2020]. This study uses the principle of majority voting (MV). In this principle, the prediction results of three diagnosis models are compared to determine the final class designations. Weighted majority voting, relative majority voting and absolute majority voting can be used to carry out this vote. For example, suppose the category or class, to which the panels belong, to be predicted is $Y = \{y_1, \ldots, y_{n_P}\}$, where for each predicted sample x, the predicted results of the 3 weak learners (kNN, SVM and DT) are $(h_1(x), h_2(x), h_3(x))$. The weighted majority voting method is based on multiplying the votes of the weak classifiers by a weight w_i, $i = 1, \ldots, N_{wl}$, where N_{wl} is the number of weak learners. Then, the multiplication products of each class are added together and that result is used to predict the class with the highest value as the final class. The final label is named $H(x)$ and is defined as follows:

$$H(x) = y_{argmax_j} \sum_{i=0}^{N_{wl}} w_i h_i^j(x), \tag{7.12}$$

where j is the number of categories. Alternatively, there is the method named relative majority voting. This method selects the category with the highest number of votes among the results predicted from the sample x by the N_{wl} weak learners. The final category c_j with the most votes is chosen. In the event that two classes have the same number of votes, the final category is randomly selected between the two classes. Relative majority voting is defined as follows:

$$H(x) = y_{argmax_j} \sum_{i=0}^{N_{wl}} h_i^j(x), \tag{7.13}$$

In the same way as the relative majority voting method, the absolute majority voting method can only generate the final label if the highest voting rate of a certain category exceeds 50%, otherwise, it refuses to issue a prediction. Relative majority voting is adopted for this investigation due to its interesting results in fault detection [Zhang 2020] and the non-reliance on weights w_i being assigned arbitrarily. In the Relative majority voting method, the healthy and snail trail labels are used. An example of classification of a current signal of a Snail Trail panel is presented in Figure 7.6

The classification results for the $N_{wl} = 3$ weak learners (kNN, SVM and DT) and the EL method are presented and illustrated through the current signals of 8 PV modules different from those used for training. The current signals are captured under the same conditions as the signals presented in Figure 7.2.

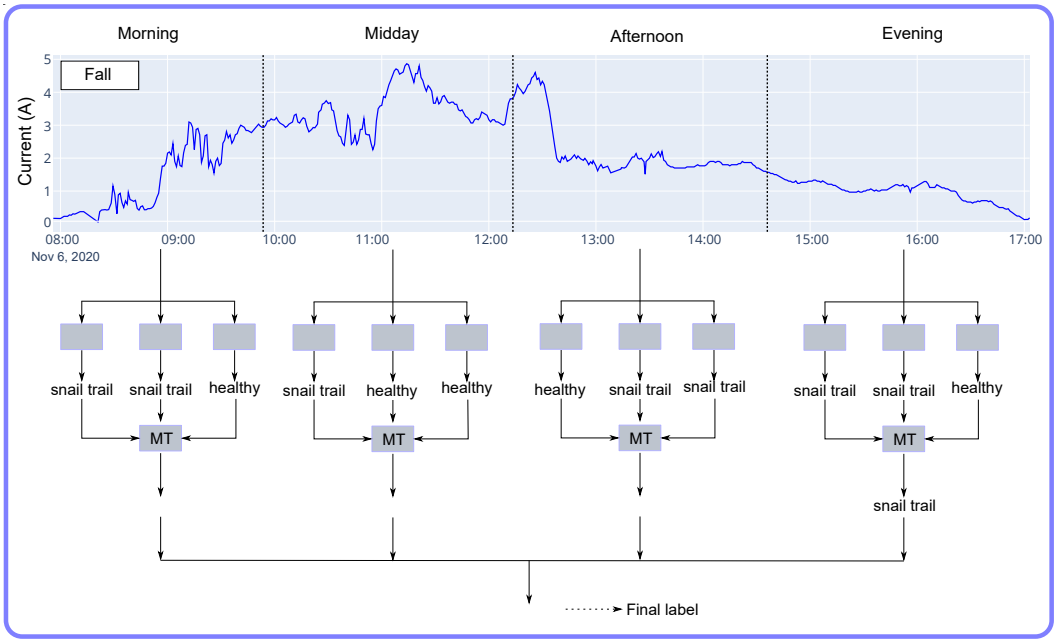

Figure 7.6: Example of fault classification of a PV panel with snail trail using EL based on Majority voting.

7.5 Discussion and Conclusions

The selected characteristics with significant variance (obtained with the dimensionality reducers) are a priori those that can be useful to solve problems of detection and classification of the health status of photovoltaic panels. These features are processed by the feature selection algorithms and then processed by the classification methods. All algorithms (kNN, SVM, DT and EL) are trained and tested with the same photovoltaic panels. Then, the algorithms are tested with the signals presented in Figure 7.7

To evaluate the degree of predictions of the classification algorithms, number of panels correctly classified, the F value and the confusion matrix are used. The F_{value} metric does not take into account true negatives (TN). For example, in a classification example with two classes (class 1 and class 2), a true negative is generated when a sample (in this case, a PV panel) that does not belong to a class for example 1 is effectively classified in class 2. The F_{value} is contained between 0 and 1, with 1 being the best performance and 0 being the worst. The F_{value} is defined as follows:

$$F_{value} = 2 * \frac{pr * re}{pr + re}, \tag{7.14}$$

where the term pr, represents the precision that can be seen as the cost of false positives and is defined as follows:

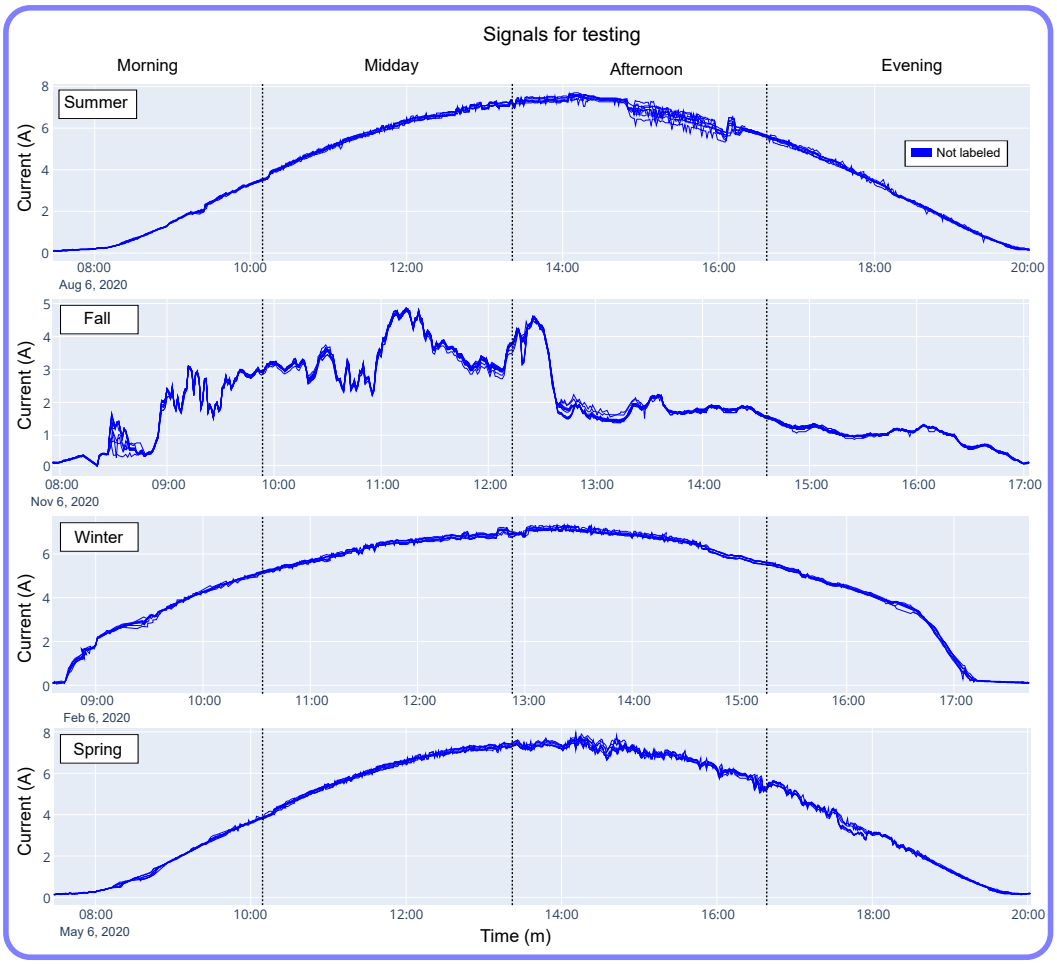

Figure 7.7: Electric current signals from 8 photovoltaic modules used in the testing of the proposed methodology. The signals were captured during a full day in the 4 seasons of the year. The data is captured with a frequency of one minute. The 4 time slices proposed [Sepúlveda Oviedo 2022] and adopted in this chapter are represented using dotted lines.

$$pr = \frac{True_{Pos}}{(True_{Pos} + False_{Pos})}, \tag{7.15}$$

The term re represents the recall, this recall is the estimate of the number of panels correctly classified based on the total number of panels belonging to the class. The recall is defined as follows:

$$re = \frac{True_{Pos}}{(True_{Pos} + False_{Neg})}, \tag{7.16}$$

Tables 7.1-7.4 present the results of the classification methods, for each season of the year, as a function of F_{value}. The values reported in Tables 7.1-7.4 are divided by each time slice. In the first scenario (without approach), the extraction of statistical features is performed directly on the current signal. That is, no signal decomposition and no dimensionality reduction (NSD_NDR). In the second scenario, signal decomposition is performed using MSD, statistical feature extraction, and dimensionality reduction using PCA and Isomap.

Season	Temporal Slice	Methodology		kNN	SVM	DT	EL
Summer	Morning	Without Approach	NSD_NDR	0,62	0,62	0,5	**0,71**
		New Approach	PCA	0,73	0,7	0,53	**0,83**
			Isomap	0,7	0,71	0,54	**0,8**
	Midday	Without Approach	NSD_NDR	0,63	0,65	0,53	**0,72**
		New Approach	PCA	0,63	0,71	0,54	**0,78**
			Isomap	0,66	0,65	0,54	**0,75**
	Afternoon	Without Approach	NSD_NDR	0,62	0,64	0,54	**0,72**
		New Approach	PCA	0,63	0,71	0,59	**0,76**
			Isomap	0,68	0,64	0,55	**0,74**
	Evening	Without Approach	NSD_NDR	0,63	0,63	0,51	**0,7**
		New Approach	PCA	0,67	0,66	0,53	**0,71**
			Isomap	0,67	0,67	0,53	**0,7**

Table 7.1: Fault detection and classification results (F_{value}) for signals captured in Summer. In Without Approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed.

As it can be seen in Tables 7.1-7.4, the performance of the classifiers increases with the use of the method proposed in this work. In a complementary way, only on the classification results using the proposed methodology with the PCA method, the confusion matrix is used. The confusion matrix is only used over the PCA method, because in all the scenarios presented in the Tables 7.1-7.4, the PCA method outperforms the Isomap method. The confusion matrix is a widely known tool that allows visualizing the performance of a supervised learning algorithm or classification algorithm [Mahesh 2019]. In this matrix, each column represents the number of predictions of each class, while each row represents the instances in the actual class. This allows to see what types of successes and errors our model is

Season	Temporal Slice	Methodology		kNN	SVM	DT	EL
Fall	Morning	Without Approach	NSD_NDR	0,63	0,64	0,5	**0,72**
		New Approach	PCA	0,67	0,73	0,53	**0,81**
			Isomap	0,63	0,63	0,5	**0,76**
	Midday	Without Approach	NSD_NDR	0,62	0,64	0,54	**0,72**
		New Approach	PCA	0,66	0,67	0,6	**0,81**
			Isomap	0,67	0,71	0,51	**0,8**
	Afternoon	Without Approach	NSD_NDR	0,62	0,65	0,53	**0,7**
		New Approach	PCA	0,68	0,64	0,59	**0,81**
			Isomap	0,62	0,65	0,54	**0,76**
	Evening	Without Approach	NSD_NDR	0,62	0,62	0,5	**0,7**
		New Approach	PCA	0,7	0,69	0,57	**0,81**
			Isomap	0,67	0,7	0,5	**0,79**

Table 7.2: Fault detection and classification results (F_{value}) for signals captured in Fall. In Without Approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed.

having when going through the learning process with the current data as a function of the time of each PV panel of the string. Figure 7.8 shows the results of the classification algorithms for each season of the year, after dimensionality reduction using PCA. The results are presented in the form of a confusion matrix where 0 is the healthy class and 1 is the class of the panels with snail trail. In this chapter, it is considered that if at least in a time slice the sample is classified as faulty, the final label is assigned as a faulty panel.

As it can be seen in Tables 7.1-7.4 and in Figure 7.8 the algorithm EL is capable of clearly discriminating between the 2 types of panels (healthy and snail trail), reducing the information to be processed to the essential, eliminating redundant or irrelevant information. In addition, it can be seen how the Ensemble learning approach proposed in this work is much superior to that of the kNN, SVM and DT algorithms. In addition, the algorithm for the use of time slices includes the analysis of the evolution of the faults detected over time.

The various contributions highlighted above make both the proposed approach effective to detect the faults of PV systems and is likely to reduce maintenance costs significantly. Also, the ensemble learning approach (EL) is generic and can be easily extrapolated to other diagnosis problems in other domains. Finally, the approach proposed in this chapter is easily integrated with devices such as inverters that capture current measurements of strings, panels, or arrays, etc. as a function of time.

The approach proposed in this chapter is aimed at making a significant contribution to the preventive maintenance of PV systems. An improvement in the preventive maintenance of the plants translates into an increase in the guarantee of continuous production of these PV systems. This becomes critical when taking into account that PV systems distribute around 2% of the total energy consumption in

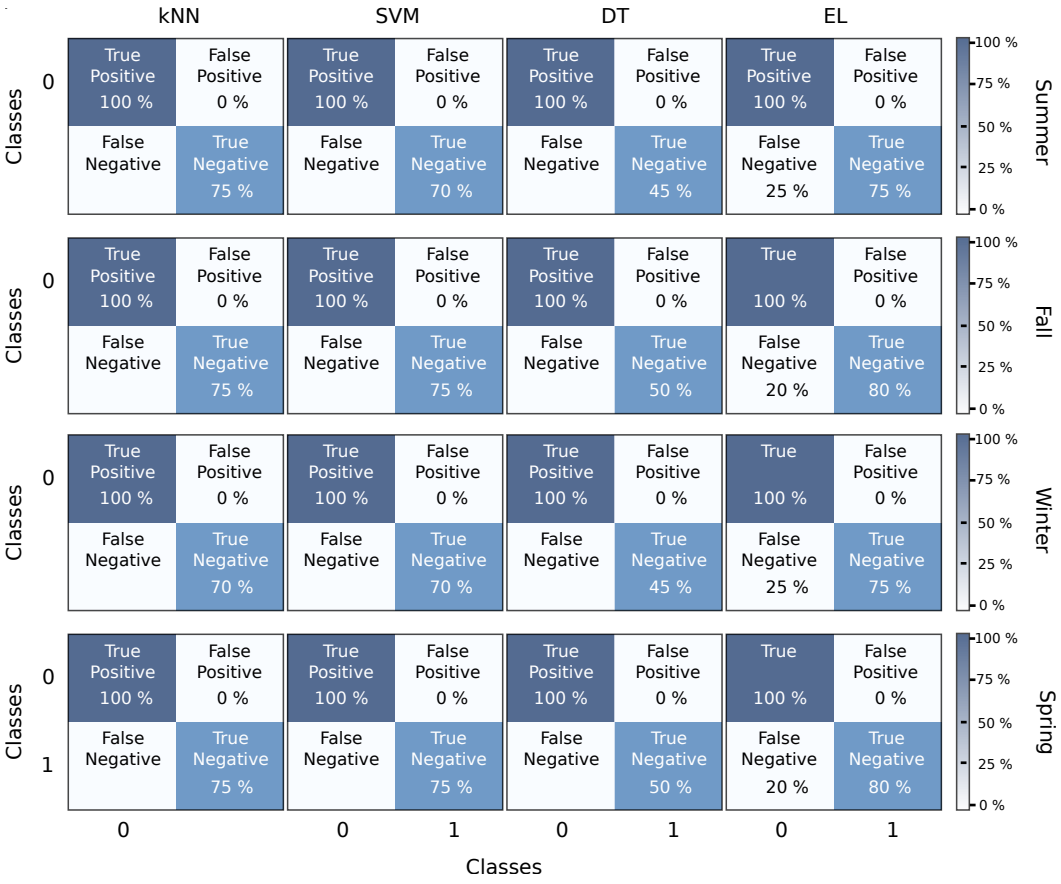

Figure 7.8: Confusion matrix of the results of the classification algorithms for each season of the year, after dimensionality reduction using PCA. The class 0 corresponds to healthy panels and the class 1 corresponds to panels with a snail trail.

Season	Temporal Slice	Methodology		kNN	SVM	DT	EL
Winter	Morning	Without Approach	NSD_NDR	0,63	0,64	0,54	**0,71**
		New Approach	PCA	0,7	0,75	0,54	**0,77**
			Isomap	0,67	0,65	0,57	**0,75**
	Midday	Without Approach	NSD_NDR	0,63	0,63	0,54	**0,72**
		New Approach	PCA	0,73	0,74	0,54	**0,74**
			Isomap	0,65	0,67	0,5	**0,72**
	Afternoon	Without Approach	NSD_NDR	0,64	0,65	0,51	**0,71**
		New Approach	PCA	0,64	0,74	0,57	**0,74**
			Isomap	0,71	0,68	0,6	**0,73**
	Evening	Without Approach	NSD_NDR	0,61	0,64	0,5	**0,7**
		New Approach	PCA	0,69	0,72	0,59	**0,85**
			Isomap	0,72	0,67	0,54	**0,74**

Table 7.3: Fault detection and classification results (F_{value}) for signals captured in Winter. In Without Approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed.

the world and present losses of more or less 18.9% per year due to the occurrence of faults. Also, this type of chapter is vital considering that the growth of photovoltaic energy is expected to continue during the coming decades, and it is even estimated that by 2050 photovoltaic energy will supply around 11 % of global electricity generation and reduce 2.3 Gigatonnes (Gt) of CO2 emissions per year. Similarly, it is important to highlight that this work proposes and develops an approach based on machine learning that only needs a set of MPP current signals over time.

The approach proposed in this chapter uses only the current signal of the panels, a reduced number of samples, as well as a reduced number of features that greatly reduce the costs of data collection, data storage and computation time. Furthermore, the diagnosis process proposed here proved to be computationally simple and efficient. This approach is validated using a real string of 8 PV modules and its efficiency is validated by separating two different health scenarios: healthy and snail trail. Another interesting aspect is that this approach is able to detect this type of fault even in time slices such as (*Morning* and *Afternoon*) where the irradiation is lower and therefore it is more difficult to diagnose faults. These results on those two time slices are tested on different days of different months to check its generality. In all cases, a result consistent with that presented in this chapter is obtained.

Another interesting aspect is that this approach is capable of detecting this type of fault even in time slices such as *Morning* and *Afternoon*, (see Tables 7.1-7.4) where the irradiation is lower and therefore it is more difficult to diagnose faults. These results demonstrate its high potential to classify or discriminate panels with faults whose power reduction is weak, but which may be the cause of other severe faults, even under low irradiation conditions. The analysis by time slices is another interesting aspect of this approach, since it considers that the detection of a fault in a time interval can become a serious fault later on or disappear simply causing

Season	Temporal Slice	Methodology		kNN	SVM	DT	EL
Spring	Morning	Without Approach	NSD_NDR	0,61	0,62	0,5	**0,7**
		New Approach	PCA	0,66	0,67	0,61	**0,81**
			Isomap	0,63	0,65	0,61	**0,76**
	Midday	Without Approach	NSD_NDR	0,63	0,63	0,55	**0,71**
		New Approach	PCA	0,65	0,72	0,58	**0,79**
			Isomap	0,7	0,65	0,6	**0,71**
	Afternoon	Without Approach	NSD_NDR	0,62	0,63	0,53	**0,72**
		New Approach	PCA	0,69	0,64	0,59	**0,81**
			Isomap	0,63	0,67	0,6	**0,76**
	Evening	Without Approach	NSD_NDR	0,63	0,64	0,55	**0,71**
		New Approach	PCA	0,63	0,71	0,62	**0,83**
			Isomap	0,67	0,69	0,57	**0,79**

Table 7.4: Fault detection and classification results (F_{value}) for signals captured in Spring. In Without Approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed.

a slight loss of performance. Therefore, this approach provides a fault evolution monitoring tool that directly contributes to preventive and corrective maintenance of large PV plants. This approach succeeded in detecting snail trail type faults with great precision, which until today, can only be detected by regularly visiting the photovoltaic plant, which is extremely more expensive.

Finally, because this approach does not require a high computational capacity, it can be easily integrated as an embedded system in photovoltaic inverters, or data acquisition systems or other similar data acquisition systems from other domains. The data storage space required for the predictor matrix is reduced, but since this approach can be extrapolated to larger time slices according to the desired scale, this system can be easily integrated with databases on local servers, converting the control and/or data acquisition systems into powerful fault detection and location tools.

Chapter 8

A Serial Diagnosis Algorithm (Serial-diag)

Contents

8.1 Approach description 216
8.2 Dataset 217
8.3 DTW Hierarchical clustering 218
8.3.1 Dynamic Time Warping 218
8.3.2 Hierarchical Clustering 219
8.4 Selected features for fault detection 221
8.5 Diagnosis of PV panels 221
8.5.1 PLS Regression model 221
8.5.2 Diagnosis of the health status of PV panels 222
8.5.3 PLS-LDA classification method 224
8.6 Discussion and Conclusions 227

Diagnosis aims at predicting the health status of components and systems. In photovoltaic (PV) systems, it is vital to guarantee energy production and extend the useful life of PV power plants. Multiple diagnosis algorithms are proposed for this purpose in the literature. The accuracy of these algorithms depends directly on the quality of the data with which they are adjusted or trained, i.e., the features. In this chapter, an innovative approach for diagnosis in PV systems is proposed, which includes a feature selection stage. This approach first discriminates severely affected PV panels using basic electrical features. In a second stage, it discriminates the other faulty panels using more elaborated time-frequency features and selecting the most relevant features through correlation and variance analysis. Finally, the approach diagnoses the health status of PV panels using a nonlinear regression method named partial least squares. This later is then combined to linear discriminant analysis and compared. The approach is validated with real current data from a PV plant composed of 12 PV panels with a power between 205 and $240Wp$ in three health states (broken glass, healthy, big snail snails). The results obtained show that the proposed approach efficiently diagnoses the three health states. It determines the level of degradation of the panels, which indicates priorities to corrective and predictive maintenance actions. Furthermore, it is cost-effective since it uses

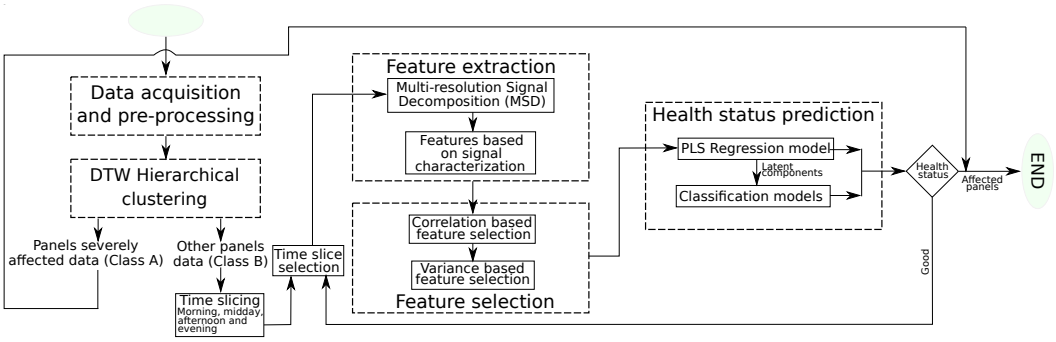

Figure 8.1: The five stages of the proposed approach. *i)* Data acquisition and preprocessing; *ii)* DTW Hierarchical clustering; *iii)* Feature extraction; *iv)* Feature selection; and *v)* Health status diagnosis.

only electrical measurements that are already available in standard PV data acquisition systems. Above all, the approach is generic and it can be easily extrapolated to other diagnosis problems in other domains.

8.1 Approach description

It is important to remember that, as explained in Chapter 2 and Chapter 3, in recent years multiple fault detection techniques are proposed. However, it is interesting to note that in these works:

- No special attention is paid to the process of feature extraction for training
- These approaches have not been tested on faulty PV panels whose fault signature is similar to that of healthy panels
- No special attention is paid to fault detection under low irradiation conditions such as at the beginning and end of the day.
- Reduced computational time
- Variety of faults detected

As a contribution to solving the issues mentioned above, this chapter presents a new approach for health status diagnosis in PV systems. Figure 8.1 illustrates the five stages of the proposal.

As shown in Figure Figure 8.1, it is first necessary to capture the panel string current and perform the respective pre-processing. This stage is named *Data acquisition and pre-processing*. Once this data acquisition stage is accomplished, the second stage named *DTW Hierarchical clustering* is performed. This stage applies the *Hierarchical Clustering* (HC) [Nielsen 2016] to the time series issued from the captured signals, for which the time series similarity index of *Dynamic Time Warping* (DTW) [Wang 2019, Jeong 2011], explained in Section 8.3.1, is used as

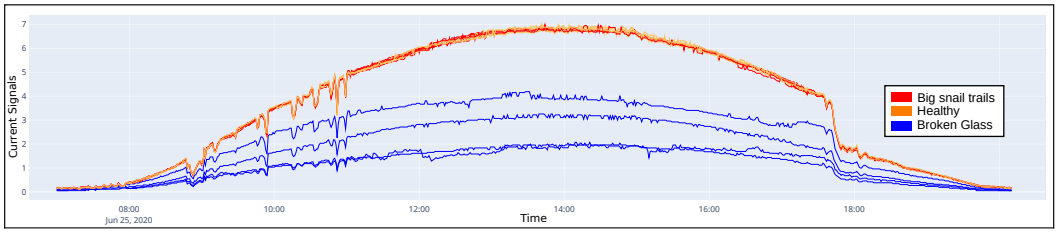

Figure 8.2: Behavior of the current over one day for different health statuses: healthy (yellow), broken glass (blue), and big snail trails (red) for a period of 13 hours every minute.

distance. This stage performs a coarse grain discrimination, aiming to separate the PV panels in two groups, those whose production is heavily affected (cluster A) and the others (cluster B). The third stage is concerned with *feature extraction*. It is intended to be carried out only on cluster B to achieve refined discrimination. This stage leverages signal decomposition with the Multi-resolution signal decomposition based on *Discrete Wavelet Transform* (DWT) [Yi 2017c, Cesar 2017], explained in Section 6.3.1, to generate a set of features. Then a Features based on signal characterization is carried out, using the features $F1$-$F5$ exposed in the Section 6.3.2. The fourth stage named *feature selection* uses the Two-stage cascading dimensionality selection and reduction method proposed in Section 6.4 based on the correlation and variance analysis to select the appropriate features. Finally, the fifth stage performs the health status diagnosis of the PV system by two methods. It first uses the *Partial Least Squares* (PLS) algorithm as a diagnosis method based on a regression model. Then, this stage uses the PLS latent components, obtained as a product of the dimensionality reduction of PLS, as input to the *Linear Discriminant Analysis* (LDA) algorithm to evaluate the results of the diagnosis with the PLS algorithm and to perform an alternative diagnosis of the health status of the PV panels.

8.2 Dataset

Similar to the algorithm in Chapter 7, the approach proposed in this chapter is conducted using only the data from the current signal. The difference is that the number of panels decreases from 16 to 12 and the number of health states (labels of samples) is increased to 3 (healthy, snail trail, broken glass). These 12 panels build a database of 12 current signals captured in parallel with a sampling time of one minute for 13 hours from 7:00 a.m. to 8:00 p.m. on June 25, 2020. For a PV panel PV_i, the data takes the form of a time series denoted by $I_{i\{1:n_I\}} = \{i_{i,t_1}, \ldots, i_{i,t_{n_I}}\}$, where n_I is the number of samples of the $i-th$ time series that has a sampling period of one minute and $t_i, i = 1..n_I$, is the date of the sample. The analysis is carried out in a time window of one day. However, it is possible to use the same methodology on different time slices. Figure 8.2 presents the PV panel current behaviors over one day for different health statuses after data cleaning.

The blue color corresponds to the PV panels with a broken glass fault, the

yellow color corresponds to the healthy PV panels and the red color to the big snail trail fault. The big snail trail represents corrosion of the sheet of the encapsulation surface and although it does not significantly decrease the performance of the PV panels, it can be the cause of fractures or micro cracks in the modules that reduce the production of a PV panel. As shown in Figure 8.2, the behavior of the PV panels with a big snail trail is very similar to that of healthy PV panels.

8.3 DTW Hierarchical clustering

In this stage, Hierarchical clustering (HC) is used to construct the two clusters A and B allowing to separate the panels severely affected data from the other panels data (cf. Figure 8.1) based on the similarity of the current time series $I_{i\{1:n_I\}}$ of the different PV panels PV_i, $i = 1, \ldots, n_P$. The time series similarity index is taken as the Dynamic time warping (DTW) index due to its well-known performance [Lines 2015, Li 2021d].

In the following subsections, HC and DTW are presented for generic time series that are then instantiated to the current times series $I_{i\{1:n_I\}}$ of each PV panel PV_i, $i = 1, \ldots, n$ of our case study.

8.3.1 Dynamic Time Warping

DTW is a well-known technique that is based on the principle of dynamic programming to deform two temporal sequences in a non-linear way and find optimal alignments between them [Jun 2011, Tanaka 2016]. To measure the similarity between two time series $S_{\{1:\eta_s\}}$ and $T_{\{1:\eta_t\}}$ the matrix of distances D of dimensions $(\eta_s \times \eta_t)$ is built. Each entry $d(i, j)$ corresponds to a local distance between S and T given by the Euclidean distance between s_i, $i = 1, \ldots, \eta_s$ and t_j, $j = 1, \ldots, \eta_t$.

A valid warping path $W_k = \{w_{k,1}, \ldots, w_{k,\eta_{W_k}}\}$, where η_{W_k} is the number of elements of the path W_k in matrix D, is defined using the above distances and satisfying the three following constraints:

1. Endpoint constraints: $w_{k,1} = d(1, 1)$ and $w_{k,\eta_{W_k}} = d(\eta_s, \eta_t)$.

2. Monotonicity constraint: If $w_{k,\alpha+1} = d(i, j)$ and $w_{k,\alpha} = d(i^{'}, j^{'})$, then $i \geq i^{'}$ and $j \geq j^{'}$, $\forall \alpha = 1, \ldots, \eta_{W_k}$

3. Continuity constraint: If $w_{k,\alpha+1} = d(i, j)$ and $w_{k,\alpha} = d(i^{'}, j^{'})$, then $i \leq i^{'} + 1$ and $j \leq j^{'} + 1$, $\forall \alpha = 1, \ldots, \eta_{W_k}$

Let us define $\mathbb{W}$ as the set of valid warping paths and $W_k^{\oplus}$ as the sum of elements of a valid warping path W_k, i.e., $W_k^{\oplus} = \sum_{p=1}^{\eta_{W_k}} w_{k,p}$. Therefore, the DTW (S,T) distance is given by the minimum warping path among all valid paths in D:

$$\mathrm{DTW}(S, T) = \min_{W_k \in \mathbb{W}} W_k^{\oplus}. \tag{8.1}$$

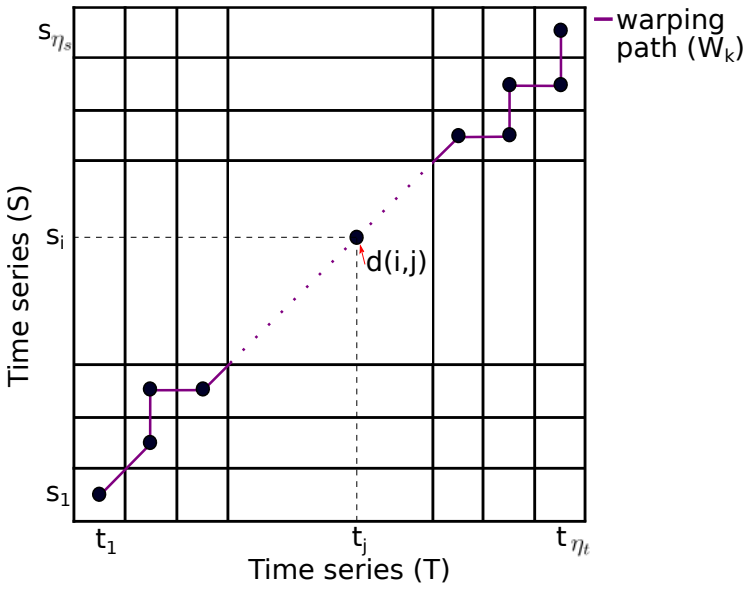

Figure 8.3: Example of warping path in the distance matrix **D**. Each entry $d(i, j)$ represents a local distance between the time series S and T given by the euclidean distance between each point s_i, and t_j.

A more detailed description of DTW is presented in [Li 2021d, Sammour 2019]. Figure 8.3 illustrates the principle of DTW.

The results of DTW are used as input to a hierarchical clustering algorithm.

8.3.2 Hierarchical Clustering

Agglomerative hierarchical clustering (AHC) is a well-known method that allows several individuals to be grouped into clusters according to the degree of similarity between the individuals. For this, the algorithm uses a degree of similarity between individuals and groups, and between groups [Tanaka 2016]. Then in each iteration, the groups with the shortest distance are merged into a single cluster [Badr 2016, Aminikhanghahi 2017] from bottom to top in the hierarchical grouping. This process continues until reaching the final condition [Rani 2012, Saleh 2009]. The result of the clustering is generally presented in the form of a tree called dendrogram [Sammour 2019]. The final clustering of the AHC depends on the level at which the dendrogram is cut [Nielsen 2016].

This algorithm is applied to the time series of the current $I_{i\{1:n_I\}}$ of each PV panel PV_i, $i = 1, \ldots, n_P$. The degree of similarity is given by the DTW. The result of the hierarchical clustering on the current signals of the PV panels is presented in Figure 8.4.

As shown in Figure 8.4, the PV panels are grouped into two large clusters A (green color) and B (red color). Since the group of PV panels from cluster A is easily discriminable, the detailed analysis of the third stage is applied only on the PV panels of cluster B. In order to analyze in detail, the behavior of the PV panels

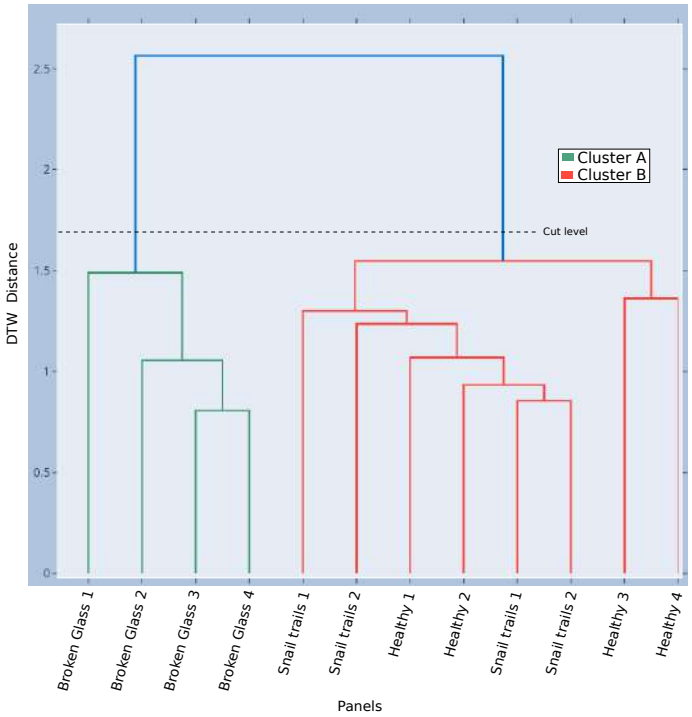

Figure 8.4: Dendrogram of the agglomerative hierarchical clustering of current signals. Cluster A in green groups the severely affected panels (broken glass). Cluster B of color B groups the healthy panels and those with big snail trails.

of cluster B under the different irradiation conditions of the day, the signals are divided into 4 slices called: morning, midday, afternoon and evening.

The feature extraction is carried out on each of these slices. In this article, it is assumed that if the classes are discriminated in at least one of the slices, the algorithm is efficient and it is possible to detect anomalies between the PV panels of group B. Next, the features selected for the diagnosis of the panels are presented.

8.4 Selected features for fault detection

The feature extraction and selection used in the approach proposed in this chapter is based on some of the algorithms explained in Chapter 6. This stage is performed only on the signals from the panels of cluster B. First, feature extraction is performed using multi-resolution signal decomposition with 3 decomposition levels. As in the ensemble learning approach of Chapter 6, the *mother wave* Daubechies38 (db38) is selected as the basis for the decomposition of multiresolution signal decomposition. Then the extraction of statistical features Skewness (F1), Kurtosis (F2), Variance (F3), $P - P_{value}$ (F4) and Energy (F5), explained in Section 6.3.2, is carried out. Finally, to reduce the dimensionality of the matrix, the two-stage cascading dimensionality selection and reduction method from Section 6.4 is used. In the case study, applying feature selection to $\mathbb{F}_*$, where $* \in \{$ *morning*, *midday*, $afternoon$, *evening* $\}$, leads to four reduced feature matrices F_* of 13, 12, 11, and 16 dimensions respectively are obtained.

8.5 Diagnosis of PV panels

The selected features aim to solve four classification problems of the health status of the PV panels. Each of these problems can be formulated as a diagnosis problem based on a regression model or a classification problem where the response variable is the label, the predictors being the features obtained in Section 6.4.

The PLS algorithm provides very interesting results over other conventional methods when the objective is class diagnosis [Liu 2007]. In addition, PLS defines latent components that can be subsequently used as predictors in a classification problem, providing an alternative method to diagnosis or a validation method of the results of the PLS based diagnosis. In this sense, the PLS algorithm can be seen as a dimension reduction method that is coupled with a regression model. It performs dimensionality reduction and classification based on regression simultaneously [Boulesteix 2006].

8.5.1 PLS Regression model

The *PLS* algorithm is based on the iterative nonlinear partial least squares algorithm (NIPALS) [Wold 1966, Wold 1982] adapted to reduce the dimensionality in ill-conditioned over-determined regression problems [Liu 2007]. Assume the matrix of predictors to be given by a centralized and normalized matrix F of dimension

$(n_B \times \eta_c^{\oplus})$, and the matrix of targets or response variables be given by a matrix Y of dimension $(n_B \times q)$. The PLS algorithm is based on the decomposition of Y and F into latent components T such that:

$$Y = TQ^T + U, \tag{8.2}$$

$$F = TP^T + E, \tag{8.3}$$

where, P and Q are matrices of coefficients, of dimensions $(\eta_c^{\oplus} \times \eta_{PLS})$ and $(q \times \eta_{PLS})$ respectively, that show how the latent components are related to F and Y. E and U are matrices of random errors of dimensions $(n_B \times \eta_c^{\oplus})$ and $(n_B \times q)$ respectively. Finally, T is a $(n_B \times \eta_{PLS})$ matrix giving the uncorrelated latent or PLS components of n_B observations. T can be seen as a linear transformation of F given by Equation (8.4).

$$T = FK, \tag{8.4}$$

where K is a $(\eta_c^{\oplus} \times \eta_{PLS})$ matrix of weights. The columns of T and K are denoted as $T(.,h) = (t_{1,h}, \ldots, t_{n_B,h})^T$ and $K(.,h) = (k_{1,h}, \ldots, k_{\eta_c^{\oplus},h})^T$, $h = 1, \ldots, \eta_{PLS}$. The rows of F are denoted as $F(j,.) = (f_{j,1}, \ldots, f_{j,\eta_c^{\oplus}})$, $j = 1, \ldots, n_B$. Based on Equation (8.4), each term $t_{j,h}$ of $T(.,h)$ is calculated according to:

$$t_{j,h} = (f_{j,1}, \ldots, f_{j,\eta_c^{\oplus}}) * (k_{1,h}, \ldots, k_{\eta_c^{\oplus},h})^T = \sum_{i=1}^{\eta_c^{\oplus}} f_{j,i} k_{i,h}, \tag{8.5}$$

where each element $k_{i,h}$, $i = 1, \ldots, \eta_c^{\oplus}$, corresponds to the normalized covariance of the response variable with each predictor given by:

$$k_{i,h} = \frac{Cov(f_{j,i}, y_j)}{\sqrt{\sum_{i=1}^{\eta_c^{\oplus}} Cov^2(f_{j,i}, y_j)}}, \tag{8.6}$$

Once T is constructed, the matrix Q^T is obtained as the least squares solution of the equation (8.2). Then, the regression model is defined according to:

$$Y = FB + U, \tag{8.7}$$

Where, B is a $(n_B \times q)$ matrix of regression coefficients defined according to:

$$B = KQ^T, \tag{8.8}$$

8.5.2 Diagnosis of the health status of PV panels

In the case study presented in this chapter, the response variables Y are categorical. In other words, each response variable y_i, $i = 1, \ldots, n_B$, of the matrix Y takes only one of the possible n_B unordered values. For example, in our case, each categorical

variable y_i takes the value of $y_i = 2$ (big snail trails), $y_i = 3$ (healthy) or $y_i = 0$ otherwise.

In the proposed approach, we first use the non-linear PLS algorithm as a dimensionality reducer. In [Dai 2006], the PLS and other dimensionality reduction algorithms are analyzed. Particularly in categorical scenarios, dimensionality reduction using PLS shows results similar to PCA [Boulesteix 2006] with high diagnosis accuracy [Man 2004, Huang 2005]. The set of components that are obtained as a result of dimensionality reduction using PLS is named the set of PLS latent components. These PLS latent components are used for the diagnosis based on the regression model of Equation (8.7). The PLS is fitted with 60% of the data and tested with the remaining 40% of the data.

In order to evaluate the accuracy of the regression model, the complementary metrics Root Mean Squared Error ($RMSE$) and R-Squared or Coefficient of determination metrics (R^2) are used. The $RMSE$ measures the standard deviation between the predicted values and the actual values of the observation [Pham 2019]. A number close to zero implies a high precision of the model. The $RMSE$ for n_B samples is defined as:

$$RMSE = \sqrt{\frac{1}{n_B}\sum_{i=1}^{n_B}(y_i - \hat{y}_i)^2}, \tag{8.9}$$

where y_i are observed values and $\hat{y}_i$ are the fitted values of the response variable Y for the ith case. The $RMSE$ does not provide information about the explained component of the regression fit [Ostertagová 2012]. Because of this, the metric R^2 is used in a complementary way. R^2 measures the percentage of variation in the response variable Y explained by the predictors F [Ostertagová 2012]. The value of R^2 ranges from 0 to 1, where 1 corresponds to the best diagnosis or prediction and 0 corresponds to a poor diagnosis or prediction. The R^2 metric for n_B samples is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n_B}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_B}(y_i - \bar{y})^2}, \tag{8.10}$$

where $\bar{y} = \sum_{i=1}^{n_B} y_i$ represents the mean value of the response variable Y. Similarly, the confusion matrix is used as a tool for evaluating the performance of the PLS algorithm. The confusion matrix represents a count of the number of accurately classified negative and positive samples represented as True Negative (TN) and True Positive (TN) respectively. Also, it represents the number of real negative samples classified as positive stands for False Positive (FP) and the number of real positive samples classified as negative stands for False Negative (FN) [Kulkarni 2020].

The results of the diagnosis of health status for all matrices F_*, where $* \in \{$ $morning$, $midday$, $afternoon$, $evening$ $\}$, are reported in Figure 8.5.

As can be seen in Figure 8.5, the PLS algorithm is able to correctly predict 7 of the 8 PV panels of cluster B in the 4 time slices. In the $Midday$ time slice, it is possible to observe how the PLS algorithm classifies a Big Snail Trail panel as a new

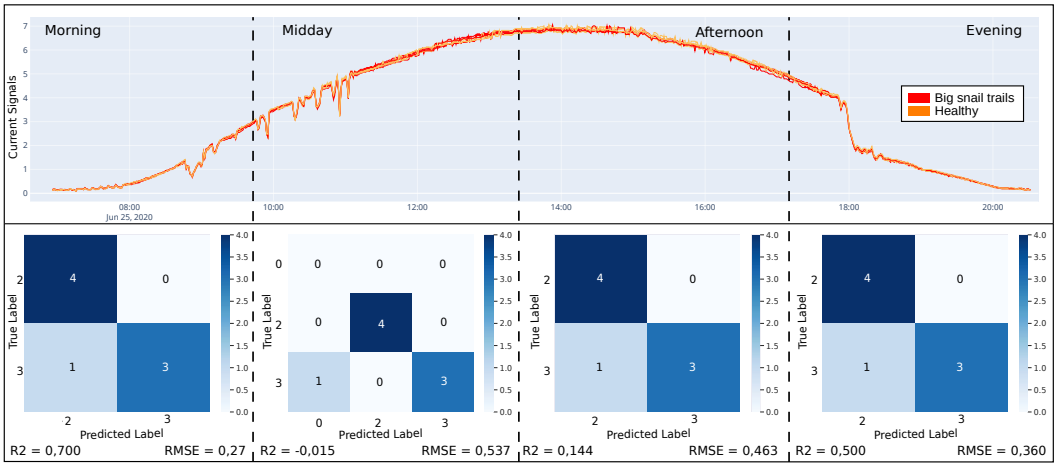

Figure 8.5: Diagnosis of the PV panels of cluster B with PLS. Diagnosis accuracy with R^2 and $RMSE$ metrics for the four time slices *morning*, *midday*, $afternoon$, and *evening* with PLS.

different class (label 0). Furthermore, the performance of the diagnosis of the PLS method on the time slices $Midday$ and $Afternoon$ is related to the similarity of the current signals between the PV panels PV_i, $i = 1, \ldots, n$ when solar irradiation is at its highest value. In the same Figure 8.5, analyzing the value of R^2 and $RMSE$, in the $Morning$ and $Evening$ slices, it is possible to observe that the model can explain the 70% and 50%, respectively, of what is happening in the actual data. While in the $Midday$ and $afternoon$ slices, the model reaches a maximum of 14% of the data. The performance of the PLS algorithm is strongly affected by the number of individuals who are used to fit the model.

8.5.3 PLS-LDA classification method

Alternatively, a health status classification method that uses the PLS latent components (given by T) as input of a classical classification method is proposed. The use of PLS as a dimension reducer for classification problems is studied in [Liu 2007, Nguyen 2002, Boulesteix 2004]. This method allows to classify the health status and to validate the health status results generated with the PLS diagnosis of Section 8.5.2.

The classification algorithm is selected to be Linear Discriminant Analysis (LDA) due to the interesting results reported when it is used with the PLS dimensionality reduction [Tang 2014, Boulesteix 2004]. In addition, this algorithm has already been used in fault detection in PV systems [Fadhel 2018]. The LDA algorithm projects the original data matrix T (predictors) from a high-dimensional space into a new low-dimensional space that makes within-class scatter as small as possible and between-class scatter as large as possible.

Given a number of classes G, the LDA determines the center class φ_{C_g}, $g = 1, \ldots, G$, for each class C_g according to:

$$\varphi_{C_g} = \frac{1}{n_e} \sum_{i=1}^{n_e} e_i, \tag{8.11}$$

where n_e is the number of elements e_i in class C_g. Then, the LDA algorithm computes the within-class S_W and the between-class S_B scatters. The S_W is calculated according to:

$$S_W = \sum_{g=1}^{G} S_{C_g}, \tag{8.12}$$

where S_{C_g} is defined as:

$$S_{C_g} = \sum_{i=1}^{n_e} (e_i - \varphi_{C_g})(e_i - \varphi_{C_g})^T, \tag{8.13}$$

The between-class scatter S_B is calculated according to the expression:

$$S_B = \sum_{i=1}^{G} (\varphi_g - \varphi)(\varphi_g - \varphi)^T, \tag{8.14}$$

where, φ is the mean value of all data in matrix T. Finally, the LDA finds a linear projection v that discriminates as much as possible the set of classes of the data. This projection is obtained by maximizing the expression:

$$J(v) = \frac{v^T S_w v}{v^T S_B v}, \tag{8.15}$$

The discriminant axes of v have as eigenvalues $\lambda_1, \ldots, \lambda_{\eta_{PLS}}$ and correspond to the decomposition of the matrix $S_w S_B^{-1}$. This decomposition into eigenvalues defines the projection space of the original data of the matrix T. To evaluate the degree of correct diagnosis (ability to identify positive and negative samples) the confusion matrix and the F_{value}, defined in Chapter 7, are used.

The LDA algorithm is trained and tested with the same PV panels that fit the model of Section 8.5.2. The total number of components generated in dimensionality reduction using PLS are used. The classification results using the PLS-LDA method, together with the F_{value} and the confusion matrix for each time slice are presented in Figure 8.6.

As shown in Figure 8.6, with the exception of the *Midday* time slice, in the other time slices the PLS-LDA method classifies the PV panels in the same classes as using the PLS algorithm. In the *Midday* time slice the different class (label 0) generated by the PLS algorithm is removed. As a summary, Table 8.1 presents the final diagnosis accuracy for the time slices *morning*, *midday*, *afternoon*, and *evening* of the PLS-LDA and PLS methods.

As seen in Table 8.1, the PLS-LDA method classifies the four time slices with an F_{value} of 0.875 (high precision) compared to the diagnosis accuracy presented by the PLS method that does not give homogeneous results for all the time slices (see

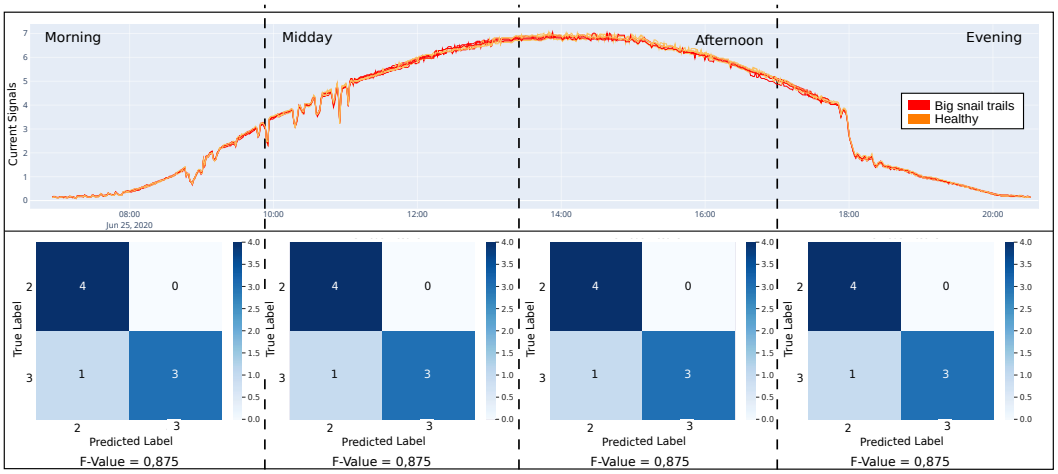

Figure 8.6: diagnosis of the PV panels of cluster B with $PLS-LDA$. Diagnosis accuracy with $F-Value$ metric for the four time slices for the time slices *morning*, *midday*, $afternoon$, and *evening* with $PLS-LDA$.

Time slice	**PLS**		**PLS-LDA**
	R^2	**RMSE**	F_{value}
Morning	0,700	0,274	0,875
Midday	-0,015	0,537	0,875
Afternoon	0,144	0,463	0,875
Evening	0,500	0,360	0,875

Table 8.1: Diagnosis accuracy for the four time slices with the PLS and $PLS-LDA$ methods.

Midday line in Table 8.1. Let us recall that in this chapter it is considered that if it is possible to discriminate healthy PV panels from another set of PV panels in at least one time slice, then it is possible to establish which PV panels are faulty with the available data.

8.6 Discussion and Conclusions

The approach presented in this chapter responds to current energy concerns regarding the guarantee of continuous energy production in photovoltaic systems. These systems distribute approximately 2% of the energy consumed in the world [Pillai 2019a] and present annual losses of around 18.9% of power due to the presence of faults [Firth 2010].

This chapter proposes and develops a health state diagnosis dedicated to photovoltaic systems. The method is based on a set of features all extracted from the MPP current signal. This approach is tested with a string of 12 photovoltaic panels and validated for efficiency by separating three different health scenarios: healthy, big snail trail, and broken glass.

To summarize, the approach uses, in a first stage, a simple hierarchical clustering based on Dynamic Time Warping, to group the PV panels into two clusters A and B, where cluster A contains the severely affected PV panels and group B contains the others. At this early stage, the method clearly discriminates between healthy and broken glass types, which points at priority predictive maintenance actions and reduces overall costs consequently. In a second stage, the use of a set of in-depth time-frequency features allows for a more precise approach to detect tiny faults and shows its ability to discriminate weakly affected panels from healthy panels.

The second stage is validated by advantageously identifying photovoltaic panels with big snail trail faults despite the difficulty of discriminating them from healthy panels. This represents a clear contribution with respect to previous works such as [Garoudja 2017a] that fails to detect faults whose behavior is highly similar to that of healthy panels. It is also important to highlight that our method has the clear advantage to require very simple data acquisition. Indeed, only the MPP current is required. Nowadays, this type of detection can only be achieved by regularly visiting the PV plant, which is extremely expensive.

A further advantage is that the approach proposed in this chapter only requires a reduced number of individuals of each class, which reduces the cost of data acquisition and storage.

Another interesting point is that faults that occur under low irradiation (*Morning* and *Evening*) are generally the most difficult to diagnose, however, the proposed method presents the best performance in these situations.

Another contribution is to base the diagnosis process on four time slices of the day. The detection of a fault in a time slice may grow into a serious fault later or vanish simply inducing a slight loss of performance. The method hence provides information about specific time points of the day that should be monitored.

Therefore, this diagnosis by time slices allows analyzing the impact and evolution of faults over time. Let us note that different time slices could be used to increase resolution in diagnosing faults such as arc faults [Wang 2013], partial shadowing [Kumar 2018], LL-faults [Dadhich 2019] that occur with low levels of irradiation.

Referring to time aspects, it should also be noted that multiresolution signal decomposition is extremely efficient at detecting the exact time a signal changes as well as the type and extent of the change [Misiti 2013]. This provides an advantage over the Fourier transform because if the fault manifests faster than the sampling window of the Fourier analysis, like it is the case of arc faults, it is very likely that they go completely undetected.

The various contributions highlighted above make the proposed method an effective method for monitoring PV systems and likely to significantly reduce maintenance costs.

Interestingly, the method that is proposed is based on generic algorithms that could be applied to PV array faults that are not considered in this chapter, and also to other applications of the energy sector. This is considered in our future work. It is also envisaged to make the measurements of the electrical quantities, including the current, at a higher frequency than that used in the tests of this chapter in order to check whether the diagnosis is thereby improved.

Chapter 9

An adaptive Diagnosis algorithm (Adaptive-diag)

Contents

9.0.1 General Scheme of operation 230
9.0.2 Normalization 233
9.0.3 Model based on knowledge 238
9.0.4 Maintenance priority 241
9.1 Discussion and Conclusions 241

As has been mentioned on multiple occasions in previous reports, the objective of diagnosis is to predict or identify the health status of the PV system components. In Article [Sepúlveda Oviedo 2022], in addition to the detection of fine faults, it was shown that the detection of various types of faults is possible, with efficient machine learning algorithms in computational terms and that they can also perform the analysis of the evolution of a fault in the process. weather. However, although in [Sepúlveda Oviedo 2022] the objective of diagnosis is achieved, there are still multiple aspects to be taken into account to overcome. First, the algorithm has been conceived with data from only one type of solar panel. For this reason, it is necessary to retrain it when a new PV plant has to be diagnosed. Second, it is not able to determine a complete maintenance priority. Furthermore, it is not able to self-adapt depending on the aging of the PV plant. For such reasons a new diagnostic approach is presented in this section.

This new adaptive machine learning approach called Adaptive-diag is integrated into the versatile new Solar Vitality data acquisition system. The first characteristic of this methodology is that it uses data normalization oriented to photovoltaic systems. This normalization allows the usability of the diagnostic model on PV systems of different technologies, topologies and installation characteristics (inclination, installation age, degradation rate among others). In addition, this approach combines supervised and unsupervised learning, as well as learning model and data-based This approach uses a complex PV power prediction model, the techniques described in [Sepúlveda Oviedo 2022], together with the normalization presented in this section, to detect, locate and identify faults in photovoltaic systems This approach uses wind speed, ambient temperature, irradiance, data sheet information, and age of the PV plant to automatically generate a healthy reference string or

panel group. With this information, this approach not only is capable of detecting faults in the photovoltaic system, but also automatically generates a report of maintenance priority for faulty panels. This system is also evolutionary, since as faults are detected and new samples are classified, the internal database grows. Once new clusters are detected, the system trains itself and updates the fault diagnosis model. The full description of this new approach is presented below.

9.0.1 General Scheme of operation

The diagnostic system, shown in Figure 9.1, needs three types of information to perform the diagnosis.

First, it collects the technical information of the PV plant, this information comes from the field reports of the PV plant and its data sheet. Aspects compiled include installation age, panel modifications, technology, topology, panel behavior under standard conditions, and more.

Second, the platform uses the "Diagnosis Box" to assess the electrical behavior of the photovoltaic plant. The behavior of the installation is represented only by the acquisition of current and voltage as a function of time.

Finally, weather behavior is recorded by measuring ambient temperature, wind speed and irradiation using the weather station. Following the analysis carried out by the Diagnosis Box, a report is obtained which contains 4 results:

- Healthy Strings
- Faulty Strings
- Detected Faults
- Maintenance priority

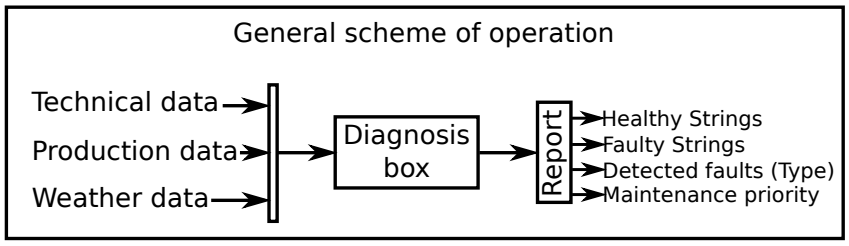

Figure 9.1: General Scheme of operation.

The diagnosis of the plant is carried out at the end of the day after having collected all the electrical and meteorological data of the day. The diagnostic platform has two main processes.

9.0.1.1 Offline operation scheme (server)

The first process takes place offline (see Figure 9.2) and is responsible for training the AI system. In this process, the DataBase D database consisting of data from healthy and faulty PV strings is used to build a machine learning model. This will then be used, together with a knowledge-based model derived from the technical data, to carry out the diagnosis of new PV installations in the online phase.

In this process, the signals are first divided into 4 time slices and on each of these slices, the characteristics (features) of the signal are extracted as illustrated by Figure 9.2 and a model is built by supervised machine learning. The frequency of execution of this process is variable. This process must be performed when new labeled data, i.e. which corresponds to an identified fault, is present in sufficient quantity in the database D or when it is observed that the behavior of a known cluster (corresponding to a fault) deviates from the actual behavior observed for this fault in the field. Database D is fed by the online process as explained below.

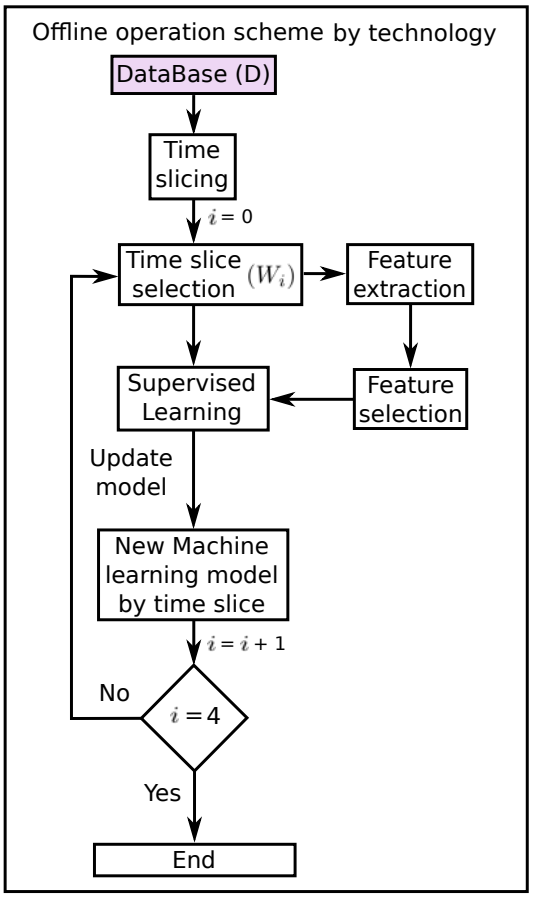

Figure 9.2: Offline operation scheme (server).

The most relevant elements of Figure 9.2 are explained below.

Time slicing and time slice selection In order to analyze in detail the behavior of the PV panels in DataBase (D) under the different irradiation conditions of the day, the signals are divided into 4 slices called: *morning, midday, afternoon and evening.*

The feature extraction is carried out on each of these slices. In this chapter, it is assumed that if the classes are discriminated in at least one of the slices, the algorithm is efficient and it is possible to detect anomalies between the PV panels of DataBase (D). In order to explain and illustrate our approach, feature extraction is explained and illustrated using the midday slice as an example.

Feature extraction This stage is based on Multiresolution Signal Decomposition, followed by the extraction of statistical features as proposed in [Ahmad 2018, Kurukuru 2020, Haque 2019, Dadhich 2019] and presented in Chapter 6.

Multi-resolution signal decomposition , Feature extraction and selection based on signal characterization sections are described in the Chapter 6.

Supervised Learning

9.0.1.2 Online operation scheme (box)

The second process of the diagnostic platform takes place online. This option is used to diagnose new installations.

In this process, as shown in Figure 9.3, sensor data and technical information (reports, datasheet, etc.) are collected. With the technical information, a first cluster of healthy panels is built (healthy model based on knowledge). Next, we consider the data from electrical and meteorological sensors acquired during a full day. This data is normalized using a set of equations to place all strings on the same scale for comparison (normalization).

If the data acquired by the sensors comes from a chain whose health is known and healthy, then the chain is stored in the cluster of healthy chains. Otherwise, all normalized signals in the chain are first tested with the knowledge-based model (check model from knowledge). If they match healthy behavior, the string is also stored in the cluster of healthy strings.

Otherwise, the chain's normalized signals are sent to the maintenance priority calculation stage on one side and on the other side they are tested by time slices with the model resulting from offline learning. (classification using the trained machine learning model). If the fault is recognized, the signals are sent to the DataBase D database. If the fault is not recognized, an expert intervenes to identify the fault before sending it to the DataBase D database.

To determine the maintenance priority, a residual calculation is carried out between the signals of the chain and the centroid of the cluster made up of the data of the healthy chains.

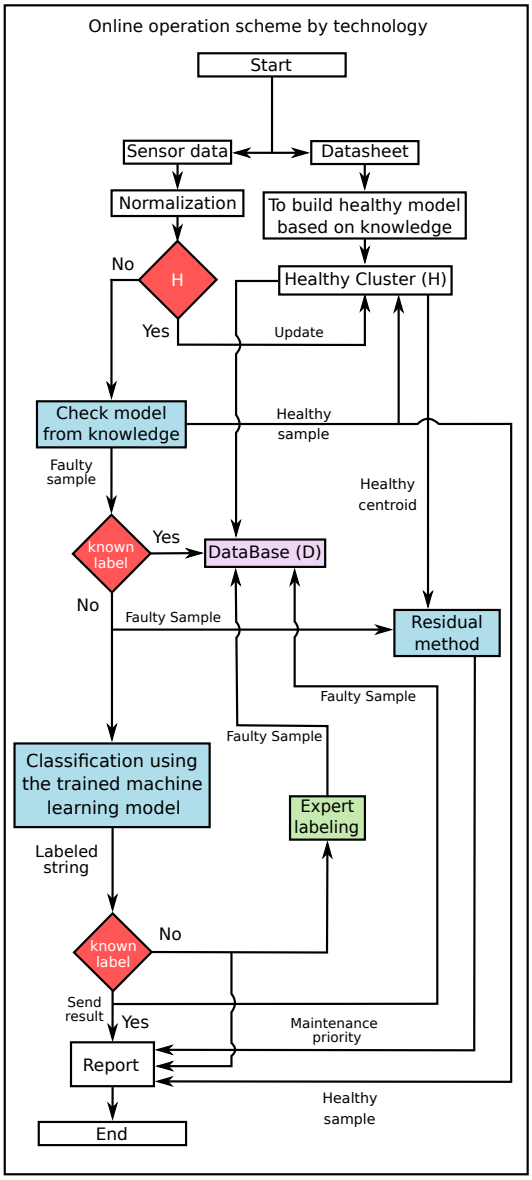

Figure 9.3: Offline operation scheme (server).

The most relevant elements of Figure 9.3 are explained below.

9.0.2 Normalization

A global study [Felder 2017] of 1.9 million modules installed in different climates found that climatic conditions have a strong impact on the appearance of faults due to material interactions. These effects are more pronounced in hot arid climates than in tropical and temperate climates. This difference in the occurrence of faults is directly linked to the rate of degradation, which in turn is linked to weather conditions [Ndiaye 2013]. For this reason, to compare PV systems that are under different climatic conditions, or with different configuration conditions or

technology, a data normalization stage is proposed.

Normalizing the performance ratio of PV systems allows systems with different nominal performances and orientations to be compared with each other [Herteleer 2017]. To this end, multiple approaches are proposed. Some documents propose the use of *Final Yield* to compare two photovoltaic installations. This measure normalizes the energy produced with respect to the size of the system. This parameter is strongly linked to the location and type of installation, and allows comparing the production of similar photovoltaic installations with different sizes but located in a specific geographical region [Santiago 2017]. This quantity represents the time that the photovoltaic array would need to operate at its nominal power P_{STC} to generate the same energy [Haeberlin 2003]. The Final Yield is calculated on its discrete over a time interval τ as follows:

$$Y_{f,\tau} = \frac{1}{P_{STC}} \int_0^\tau P(t)\,\mathrm{d}t = \frac{1}{P_{STC}} \sum_{t=0}^{\tau} P(t) = \frac{E_{AC}}{P_{STC}}, \tag{9.1}$$

where E_{AC} is the AC energy of the inverter, determined from the monitoring of the AC power output of the inverter P_{AC}, in the time interval τ. In the same way, to quantify and compare the performance of photovoltaic systems, the Performance Ratio (PR) metric is widely used [Woyte 2014]. The dimensionless metric PR is calculated in the range of days, and occasionally shorter periods, for example 5 minute intervals are recommended in [Woyte 2014]. PR represents the relationship between the Final Yield $Y_{f,\tau}$ and the Reference $Y_{r,\tau}$, over a time interval τ as follows:

$$PR_\tau = \frac{Y_{f,\tau}}{Y_{r,\tau}}, \tag{9.2}$$

where the reference $Y_{r,\tau}$ is defined as follows:

$$Y_{f,\tau} = \frac{1}{G_{STC}} \int_0^\tau G(t)\,\mathrm{d}t = \frac{1}{G_{STC}} \sum_{t=0}^{\tau} G(t) = \frac{E_G}{G_{STC}}, \tag{9.3}$$

where E_G is the sloped irradiance, determined by multiplying the in-plane irradiance value, $G(W/m^2)$, by the monitoring time interval τ. $G_{STC} = 1000W/m^2$ is the irradiance under standard STC conditions. This magnitude represents the time that the photovoltaic field must receive radiation with a value of $G_{STC} = 1000W/m^2$ to generate an energy E_G [Haeberlin 2003, IEC 2017b, Malvoni 2017, Trillo-Montero 2014]. Its value depends on the location, orientation and inclination of the photovoltaic system and the weather conditions [Santiago 2017]. A temperature-corrected form of PR [IEA 2014, Dierauf 2013] can be calculated by using Equations (9.4) and (9.5) defined as:

$$T_{corr} = 1 + \gamma(T_{cell} - T_{STC}) = 1 + \gamma \triangle T_{STC}, \tag{9.4}$$

$$P^* = \frac{P(T_{cell})}{T_{corr}}, \tag{9.5}$$

and then replacing P by P^* in Equation (9.1), with P the power (or efficiency) that shows a temperature dependence, as widely documented [Skoplaki 2009a]. This same idea has been extended in [Leloux 2012] including and correcting for the effects of temperature. In a complementary way, the work of Sandia (multiprogram laboratory operated by Sandia Corporation) proposes the use of the AC efficiency of the regular system (that is, not normalized) [King 2004a] and an extension for the DC side is presented in [Copper 2013a]. A modification of the model presented in [King 2004a] is carried out in [Huld 2010, Huld 2011]. They propose a module-level model of the relative instantaneous efficiency η_{rel} and its hourly and annual equivalent. In [Herteleer 2017] the normalized efficiency of a photovoltaic module or system is proposed as a metric of photovoltaic performance. That metric is used for monitoring and analysis purposes and can be implemented on time scales ranging from seconds to days and more. In the same document [Herteleer 2017], a modification to the temperature-corrected form of the power classification model proposed in [Huld 2010, Huld 2011] for photovoltaic modules is proposed. This modification is applicable not only to individual photovoltaic modules, but also to photovoltaic arrays and systems on both the AC and DC sides. In [Walker 2020] an adjustment to the Performance Ratio is proposed, using a constant annual degradation of 6%.

All these elaborate models have taken place because the MPP signals of the photovoltaic arrays (photovoltaic voltage (VMPP) and photovoltaic current (IMPP)) vary strongly throughout a year. Specifically, these changes are closely linked to the dependence of PV systems on environmental conditions such as: *i)* solar radiation; *ii)* temperature; and *iii)* wind speed. This study works under the hypothesis that reference modules and photovoltaic arrays degrade in a similar way or at insignificant rates. This hypothesis is the result of the analysis of the studies of optical degradation of the surface of a photovoltaic array and the reference modules under similar work environments carried out in [Meyer 2004, Cueto 2010]. In those studies the comparison is made with exposure to ultraviolet rays, thermal stress and humidity, similar. This chapter is based on the fact that the unavoidable degradation rate of the photovoltaic system is between 0.6% and 1% per year [Jordan 2016]. For specific fault detection calculations, this chapter adopts the value 0.6% year as an assumption on which to normalize the performance of PV systems.

Other more precise approaches, but which require a larger number of sensors, are based on the normalization of the electrical and meteorological variables of the PV systems. The fault detection approach presented in this chapter works under this hypothesis. For this, new normalized parameters are proposed as a modification of the set of empirical equations presented in [Zhao 2015b].

Furthermore, this chapter works under the hypothesis that the change of the VOC and ISC conditions caused by the cell temperature difference associated with the short-circuit and open-circuit conditions is negligible as presented in [Bharti 2009].

Performing the normalization of parameters such as voltage and current is highly advantageous, since they would remain constant, even when the PV modules degrade uniformly. The objective of this normalization is that the panels that are analyzed can be compared under the same scale.

For this reason, one of the contributions of this section lies in the construction and formal presentation of a set of normalization equations for photovoltaic data aimed at fault diagnosis. This set of equations not only improves data visualization but also the accuracy of fault detection algorithms. The normalized variables are: Voltage, Current, Irradiation, Ambient temperature, and wind speed.

9.0.2.1 Voltage

For the normalized voltage $V_{String,norm}$, the following parameters are used:

- V_{string}: String tension
- V_{STC}: Voltage under STC conditions
- n_b: Number of panels in the string
- d_{n_V}: Natural degradation
- α: Age of the PV plant (in days)
- β: Annual degradation rate

$$d_{n_V} = 1 + \alpha * \frac{-\beta}{365.24} \tag{9.6}$$

$$V_{String,norm} = \frac{V_{string}}{V_{STC} * n_b} * d_{n_V} \tag{9.7}$$

The result of the normalized voltage $V_{String,norm}$ should vary between 0 and 2.

9.0.2.2 Current

For the normalized current, the following parameters are used:

- I_{string}: String current
- I_{STC}: Current under STC conditions
- d_{n_I}: Natural degradation
- α: Age of the PV plant (in days)
- β: Annual degradation rate

$$d_{n_I} = 1 + \alpha * \frac{-\beta}{365.24} \tag{9.8}$$

$$I_{String,norm} = \frac{I_{string}}{I_{STC} * d_{n_I}} \tag{9.9}$$

Current values must be between 0 and 2 (values may slightly exceed 1 if Irr > 1000 W/m^2)

9.0.2.3 Irradiation

This chapter works under the hypothesis that the irradiation values G_{poa} vary between 0 and 1500 W/m^2.

$$G_{poa,norm} = \frac{G_{poa}}{G_{STC}} \tag{9.10}$$

where $G_{STC} = 1000$

$$G_{poa,norm} = G/G_{STC} \tag{9.11}$$

The normalized irradiation values $G_{poa,norm}$ must vary between 0 and 1.5 W/m^2.

9.0.2.4 Ambient Temperature

The adopted hypothesis assumes a average temperature in France of 12°C and a variation between $average - 40$°C and $average + 40$°C, where between -28°C and $+52$°C.

$$T_{amb,norm} = \frac{T_{amb,celcius} + 28}{80} \tag{9.12}$$

Ambient temperature $T_{amb,norm}$ can be used together with irradiance $G_{poa,norm}$ to estimate the temperature of the solar cell [Zhao 2015b].

$$T_{cell} = T_amb + \frac{NOCT - 20°C}{800W/m^2} * G_T \tag{9.13}$$

where the nominal operating cell temperature (NOCT) is chosen as 50°C [Skoplaki 2009b]

9.0.2.5 Wind speed

Finally, for the normalized wind speed, a variation between 0 and 10 m/s is proposed.

$$WS_{nomr} = \frac{WS_{m/s}}{10} \tag{9.14}$$

The knowledge-based model is explained below.

9.0.3 Model based on knowledge

As mentioned in [Berghout 2021a], due to the long lifespan of PV panels, and associated computational costs, such as that of in-memory storage, it is difficult to collect the necessary patterns similar to degradation. As mentioned in Eder et al. [Eder 2018], accelerated tests can be an alternative solution for reconstruction of data-driven models. However, data-driven samples of artificially aged PV panels suffer from the lack of important descriptive patterns related to deterioration or damage processes. In addition, some samples, such as I-V/P-V or thermographic images, are generally difficult or impossible to label, even for ML developers. Therefore, knowledge-driven models are used in this case to fill the gaps in uncompleted lists of unlabeled samples [Berghout 2021c]. Two main types of ML models can be found in this type of learning, namely, generative models [Theis 2016] and domain adaptation learning by considering the domain to be invariant [Zhao 2019a, Baktashmotlagh 2013]. Generative models are ML tools capable of generating new examples or preliminary hypotheses using training data. These new examples or preliminary hypotheses are used either to improve the representation of the features or to provide the necessary information that is assumed to be hidden in the original feature space. Similarly, domain adaptation learning by considering the domain to be invariant is an alternative solution for adjusting the data distribution in the target domain, once similar complete data are available in the source domain. Mathematical formulations of the loss erm of generative models are relatively similar to domain invariant learning when feeding a discriminative model [Bai 2019, Song 2017].

In recent ML modeling, specifically for condition monitoring, Generative Adversarial Networks (GANs) and Transfer Learning (TL) have been among the commonly used types of generative models and domain invariant adaptation learning approaches [Kusiak 2019, Serin 2020]. GANs represent a new effective generative adversarial learning theory specific to data augmentation. GAN is a ML technique developed by Goodfellow in 2015 [Goodfellow 2014], in which the main idea is to train a generative model, such as a deep network, to generate real examples from fake data in a form of "minimum of two players game" approach. Unlike traditional generative models that try to extract features, GANs are trained in a supervised manner by associating a discriminator to classify these representations to only the two preceding categories of fake/not fake. By comparison, TL can be applied to any learning algorithm by moving learning parameters from different distributions of the source domain to the target domain, and minimizing a common and full loss function of the entire contributed domains in the adaptation process [Weiss 2016, Long 2014]. Knowledge-driven models have also been investigated according to the two discussed data acquisition methods.

To address knowledge-driven modeling using data acquired from ordinary sensors, a set of recent algorithms for PV condition monitoring are discussed in this review. For instance, Lu et al. [Lu 2019a], proposed a hybrid deep TL algorithm adaptable to several domain distributions using a CNN for DC arc faults (i.e., can

be caused by short-circuit or ground faults) diagnosis. First, the algorithm attempts to learn representative examples from the learning samples in the source domain data. Then, a dummy generation process of new samples in the target domain is followed by the TL process using GANs. A total of 25,000 samples were collected for a real PV system (see Lu et al. [Lu 2019a]) during normal healthy operating conditions. In addition, 5000 arc fault samples were used to construct the source domain dataset. Accordingly, 20% of the randomly selected samples from healthy operating states were reserved for the validation process. Three types of arcing faults at the start, middle, and end of the PV string were considered. Lu et al. [Lu 2021a], in a work similar to their previous study (i.e., Lu et al. [Lu 2019a]), extended their experiments using almost the same training tools and frameworks by involving three additional datasets.

In the context of knowledge-driven image acquisition, a number of studies can be noted. Tang et al. [Tang 2020], in the study of a prediction problem using a limited number of electroluminescence images, augmented their data by combining GANs and traditional image processing techniques. Then, generated examples for data augmentation reasons were fed into a CNN fault detection algorithm of PV modules. Five types of PV cell degradation fault (i.e., micro-cracks in polycrystalline silicon, micro-cracks in monocrystalline silicon, finger interruptions in monocrystalline silicon, finger interruptions in polycrystalline silicon, and breaks) were studied. Akram et al. [Akram 2020] also examined a TL-based approach to train a CNN for PV faults. However, a more complicated study was involved in this case, in which fault classification in two different datasets was considered. An electroluminescence image dataset was used to train the CNN in the source domain and infrared image datasets were used for training in the target domain. The infrared images enabled manual labeling of the degradation faults with eight types of faults, namely, failed cell interconnection, cell cracking, cracks isolating cell parts, failed/resistive soldering bonds, localized shunting in cells, high current density at bus bars, breakage of module glass, and failed cells in outdoor infrared images. It should be noted that the use of knowledge-driven models has been lacking in PV fault detection. As a result, the attention of scientists has moved toward traditional and deep learning techniques in this field. According to [Berghout 2021a], this type of knowledge-guided paradigms are especially useful in cases like the following:

- Cases where the test samples are subject to a higher level of variation, or their data distribution is different from the distribution of the data used for training;
- Training data is incomplete or many labels are missing;
- Data is incomplete and subject to many outliers.

Taking into account all these previously mentioned aspects, another alternative to generate models based on knowledge is the use of equations that define the behavior of a PV system [Dhoke 2019, Herteleer 2017, Huang 2020, Huld 2011,

Huld 2010]. In this case, multiple researchers have worked on the development and tests of equations that have shown a prediction of the power of PV systems quite close to that of a PV system. This approach is adopted in this work because it provides great precision and is also easily adaptable to real data from PV plants. The model proposed for the present is a variant of the model presented in [King , King 2004b] proposed in [Chianese 2003, Kenny 2003, Huld 2008]. The model applied for the PV power P_i of each PV string PV_i, $i = 1, \ldots, n_s$, where n_s is the number of strings, in this work has the following form:

$$P_i = G_{poa,norm} * (P_{STC} + a + b + c + d + e + f) \tag{9.15}$$

$$a = k_1 * ln(G_{poa,norm})$$

$$b = k_2 * (ln(G_{poa,norm}))^2$$

$$c = k_3 * T$$

$$d = k_4 * T * ln(G_{poa,norm})$$

$$e = k_5 * T * (ln(G_{poa,norm}))^2$$

$$f = k_6 * (T)^2$$

where the normalized in-plane irradiance and module temperatures are given by

$$T = T_{module} - T_{STC} \tag{9.16}$$

where $T_{STC} = 25°C$. The main difference from the original model is that the terms for current and voltage at maximum power point (MPP) have been multiplied together to a single expression for the module power at MPP. In this way the expression for the module power is linear in the empirical coefficients P_{STC} and k_1–k_6, and it is possible to fit the model to data that contain only the measured power at given G and T. Another way of expressing this is in terms of the relative conversion efficiency, defined as:

$$\eta_i = P_i/(P_{STC} * G_{poa,norm}) \tag{9.17}$$

The relative efficiency is the ratio of the module efficiency under given conditions of G and T to the efficiency at STC. The empirical coefficients P_{STC} and k_1–k_6 must be found by fitting the function to measured data (indoor or outdoor). The fit is done by a least-square procedure. Some of the typical values reported in the

Coefficient	c-Si	CIS	CdTe
k1	-0.017237	-0.005554	-0.046689
k2	-0.040465	-0.038724	-0.072844
k3	-0.004702	-0.003723	-0.002262
k4	0.000149	-0.000905	0.000276
k5	0.000170	-0.001256	0.000159
k6	0.000005	0.000001	-0.000006

Table 9.1: Typical values of the coefficients of Equation (9.15).

literature are presented in Table 9.1.

Finally, this chapter explains how the maintenance priority calculation works.

9.0.4 Maintenance priority

Once the theoretical power is calculated and added to the cluster of healthy panels along with the actual healthy panels that have been verified in the field, the centroid of that cluster is calculated. Then, the string detected as faulty is compared with that centroid to determine the magnitude of the difference from the reference (healthy cluster centroid). The difference of each string i, also called residual r_i, is defined as:

$$r_i = |P_i - P_c| \,, \forall \; i \in [1, n_s], \tag{9.18}$$

where P_c is the power of the centroid of the healthy strings and n_s is the number of strings analyzed. Once the residuals of all the strings have been calculated, they are arranged in descending order to determine those whose r_i is the highest.

9.1 Discussion and Conclusions

With new emerging solar cell technologies like the ones we introduced in Section 2, and increases in financial incentives from governments, the global solar PV industry is growing exponentially as seen in Figure 2.4. This growth has been accompanied by multiple studies focused on guaranteeing the continuous and optimal production of PV systems. Most studies are devoted to fault detection and diagnosis. However, in order to detect a wide range of faults, it is necessary to have high-performance monitoring systems that generate enormous amounts of data (Big Data). It is there, where the use of AI and more specifically of machine learning makes sense, extracting behaviors that are difficult to detect with other conventional methods. As presented in Section 2, faults can appear in any of the components, during the installation and/or operation of the PV system.

This chapter focuses only on the study of the faults that occur on the DC side after the installation of the PV system. However, considering that it is possible to capture multiple variables in a power generation system in the form of time series, these algorithms may be easily extrapolated to the AC side.

The objective of this study is to provide a set of tools and knowledge to improve the efficiency and reliability of PV systems. It is important to highlight that the detection of a large number of faults is really a challenge when taking into account the strong relationship that exists between the production of the PV system and changes in weather conditions or automatic corrective actions by inverters or optimizers, different fault locations, mismatches between PV modules, among others [Ahmad 2018].

Having clear the context and the difficulty involved in making this type of detection, one of the first innovative aspects of this chapter is the application of advanced feature extraction and selection techniques in conjunction with machine learning algorithms to detect fine faults in PV systems. This approach is motivated by the objective of proposing diagnosis methods aimed at preventive maintenance of PV systems, but that can also be used at other scales or extrapolated to other types of systems such as storage systems, microgrids, among others.

As the study is oriented to preventive maintenance of PV systems, it is necessary to detect faults from their occurrence, or faults whose observable electrical signature is similar to that of healthy panels. In this way, this chapter makes a great effort at the signal processing level to implement a transformation that extracts not only the behavior in the time domain, but also manages to extract changes in the frequency domain. This type of time/frequency analysis can provide vital elements for class discrimination.

It is necessary to highlight that this chapter not only proposes a machine learning algorithm for fault detection in embedded PV systems. In addition, it proposes a whole context of fault diagnosis that includes a new PV system data acquisition system, a versatile mobile weather station, and machine learning algorithms that, due to their computational efficiency and rapid response characteristics, can be embedded in other types of devices such as inverters, optimizers, etc.

A deep investigation about the current monitoring systems of PV systems and their limitations was carried out. The result of this research allowed a series of adaptations in Solar Vitality that make it viable and effective for small, medium and large-scale photovoltaic plants, without compromising the desired performance. Among the critical parameters of Solar Vitality, it must be taken into account that it must be guaranteed that the acquisition of all the data sent by Arduino does not exceed the sensor data acquisition time. In addition, it must be ensured that the portable power supply can supply the necessary current to avoid data loss or corrupt data on the Raspberry. When data acquisition is performed at high frequencies such as milliseconds or less, problems such as drift start to become apparent and must be addressed to avoid false fault diagnosis results. In general, Solar Vitality and the meteorological station proposed in this chapter proved to be able to efficiently monitor PV plants. In addition, due to the use of the Raspberry Pi board, it is possible to ship different machine learning algorithms that work in parallel coded in high-level languages such as Python. The platform was put into operation in different PV plants, demonstrating high performance and continuity in operation. Solar Vitality also demonstrated that it is efficient in terms of storing large amounts

of data thanks to the format transformations it performs on the captured signals. It also demonstrated great versatility and easy parameterization to be adapted to different topologies of PV plants. However, based on the data captured with the Tigo data acquisition system and the Solar Vitality prototype together with the meteorological station proposed in this section, it is necessary to mention that not only because high quality data acquisition and sampling rate can ensure fine fault detection.

On the other hand, this chapter not only proposes an operational machine learning algorithm, but also demonstrates that it can be embedded in a real system, with very good performance since it is tested in different PV plants. Another of the great advantages of the approach presented in this chapter is that it does not need to cut off the production of the PV plant in order to carry out fault detection. This is the big difference with the wide number of approaches proposed for fault detection in PV systems using the I(V) characteristic curve.

Another of the problems found in the literature is that many of the approaches are trained with data based on simulation. The problem with these simulation models is that on many occasions they fail to faithfully represent the real behavior of a PV plant. This chapter overcomes this limitation since all the work is based on real data. Even the algorithm in Section 9, which contains a part of model-based learning, only uses that information as the base, but then the cluster created with that model-based data is dynamically adapted based on the real data captured in the PV plant.

Another interesting aspect of the chapter is that the analysis by time slices has managed to demonstrate that it is not only efficient for fault detection, but also for analyzing their evolution, or the discrimination between temporary and permanent faults. That is, if a fault appears in a single time slice, it is possible that it could be a non-permanent fault. If the fault appears in two time slices, the panel or PV string could be put under supervision, while if it appears in more than two, it probably has a permanent fault.

In addition, the system not only detects faults and allows evolution analysis, but the approach presented in this chapter can also send a maintenance priority report, exponentially reducing the decision-making time for the replacement of faulty panels.

Conclusions and Perspectives

With new emerging solar cell technologies like the ones we introduced in Chapter 2, and increases in financial incentives from governments, the global solar PV industry is growing exponentially as seen in Figure 2.4. This growth has been accompanied by multiple studies focused on guaranteeing the continuous and optimal production of PV systems. Most studies are devoted to fault detection and diagnosis. However, in order to detect a wide range of faults, it is necessary to have high-performance monitoring systems that generate enormous amounts of data (Big Data). It is there, where the use of AI and more specifically of machine learning makes sense, extracting behaviors that are difficult to detect with other conventional methods. As presented in Chapter 2, faults can appear in any of the components, during the installation and/or operation of the PV system.

This thesis focuses only on the study of the faults that occur on the DC side after the installation of the PV system. However, considering that it is possible to capture multiple variables in a power generation system in the form of time series, these algorithms may be easily extrapolated to the AC side.

The objective of this study is to provide a set of tools and knowledge to improve the efficiency and reliability of PV systems. It is important to highlight that the detection of a large number of faults is really a challenge when taking into account the strong relationship that exists between the production of the PV system and changes in weather conditions or automatic corrective actions by inverters or optimizers, different fault locations, mismatches between PV modules, among others [Ahmad 2018].

Having clear the context and the difficulty involved in making this type of detection, one of the first innovative aspects of this thesis is the application of advanced feature extraction and selection techniques in conjunction with machine learning algorithms to detect fine faults in PV systems. This approach is motivated by the objective of proposing diagnosis methods aimed at preventive maintenance of PV systems, but that can also be used at other scales or extrapolated to other types of systems such as storage systems, microgrids, among others.

As the thesis is oriented to preventive maintenance of PV systems, it is necessary to detect faults from their occurrence, or faults whose observable electrical signature is similar to that of healthy panels. In this way, this thesis makes a great effort at the signal processing level to implement a transformation that extracts not only the behavior in the time domain, but also manages to extract changes in the frequency domain. This type of time/frequency analysis can provide vital elements for class discrimination.

In order to position this research and to know the current limitations in the area of diagnosis of PV systems, this thesis presents for the first time an extremely deep and complete study of a large number of articles that builds a state of the art on the subject of interest. This is possible due to two innovative methodologies

proposed for the systematic analysis of information. Otherwise, carrying out this analysis with conventional methods such as expert analysis would not have been possible.

It is necessary to highlight that this thesis not only proposes a set of machine learning algorithms for fault detection in embedded PV systems. In addition, it proposes a whole context of fault diagnosis that includes a new PV system data acquisition system, a versatile mobile weather station, and machine learning algorithms that, due to their computational efficiency and rapid response characteristics, can be embedded in other types of devices such as inverters, optimizers, etc.

In the same way, it is important to emphasize that two of the three algorithms for detecting faults based on machine learning proposed in this thesis (see Chapters 7-8), are designed to work with the minimum number of variables (electrical current) to respect the economic limitations of industry. In addition, this feature makes these algorithms easily implementable, cost-effective, and accurate in installations that are not only high power but also residential.

On the other hand, this thesis not only proposes a set of operational machine learning algorithms, but also demonstrates that they can be embedded in a real system, with very good performance since they are tested in different PV plants. Another of the great advantages of all the approaches presented in this thesis is that they do not need to cut off the production of the PV plant in order to carry out fault detection. This is the big difference with the wide number of approaches proposed for fault detection in PV systems using the I(V) characteristic curve.

Another of the problems found in the literature is that many of the approaches are trained with data based on simulation. The problem with these simulation models is that on many occasions they fail to faithfully represent the real behavior of a PV plant. This thesis overcomes this limitation since all the work is based on real data. Even the algorithm in Chapter 9, which contains a part of model-based learning, only uses that information as the base, but then the cluster created with that model-based data is dynamically adapted based on the real data captured in the PV plant.

Another interesting aspect of the thesis is that the analysis by time slices has managed to demonstrate that it is not only efficient for fault detection, but also for analyzing their evolution, or the discrimination between temporary and permanent faults. That is, if a fault appears in a single time slice, it is possible that it could be a non-permanent fault. If the fault appears in two time slices, the panel or PV string could be put under supervision, while if it appears in more than two, it probably has a permanent fault. In addition, the system not only detects faults and allows evolution analysis, but the approach presented in Chapters 8 and 9 are also capable of sending a maintenance priority report, exponentially reducing decision-making time for the replacement of panels.

With all the above, the list of the main contributions of this thesis are presented below.

Main contributions of the thesis

To address the issues discussed above, this research presents contributions with respect to ten aspects:

1. **A review of the state of the art in fault detection in PV systems** that includes conventional detection methods and advanced methods based on Machine Learning. In this scientific context, this research is led with two novel methodologies for computational and systematic analysis of the literature. These new approaches can be easily extrapolated based on bibliometrics and topic modeling and cover more articles to have a more precise idea of the current state of the art. In addition, this type of review not only presents relevant articles, but also analyzes aspects such as: i) existing collaborative work relationships between countries, authors, scientific institutions and the most successful machine learning algorithms in the area depending on the type of learning (supervised, unsupervised, reinforcement, semi-supervised) and the families of the master algorithm of machine learning [Domingos 2015]. This allows identifying, according to the conditions of the problem, the most suitable algorithms for fault detection. Finally, this analysis determines interesting research topics and challenges related to fault detection in these systems.

2. **A formal dictionary of faults** that contains four types of identified faults sources: external causes, material interaction, component aging or caused by other faults (cause-effect circle). In turn, within this dictionary a new multilevel classification of system faults is proposed according to the type of fault, the component where it occurs (cell, module, arrangement, protection system or junction box), whether structural, electrical , caused by abnormal increases in temperature (hot spot), bad connections or shadow (due to obstacles or dirt). Each fault is exposed with its due explanation and graphic illustration. This dictionary also includes the aspects of frequency of occurrence and impact in terms of human safety and loss of energy.

3. **A novel platform for data acquisition and monitoring named Solar-Vitality** in PV systems aimed at diagnosing faults in PV strings. This platform contains two embedded systems in charge of: i) a versatile new photovoltaic monitoring system that captures the current and voltage at the PV string level; and ii) automatic fault detection.

4. **A portable weather station** adaptable to different configurations of PV plants. This weather station captures climatological variables such as ambient temperature, wind speed, and irradiation.

5. **A contribution to signal processing and analysis for fault feature extraction and selection** that includes a set of transformation operations performed on the fault signals as a guide to increase the richness of the signals

analyzed by the machine learning algorithms and therefore improve the ability to discriminate between classes.

6. **An ensemble learning algorithm named EB-diag** able to detect snail trace faults in PV modules. EB-diag combines several learning models, rather than using a single learning model. Also, this approach takes advantage of the feature extraction and selection techniques from Point 5, greatly improving detection accuracy. The results of this approach demonstrate that it is capable of classifying healthy PV modules and those with snail tracks/trails efficiently and cost-effectively, since it uses only the electrical current signal of the modules obtained from standard PV data acquisition systems. Furthermore, the approach is generic and can be easily extrapolated to other diagnostic problems in other domains.

7. **A new hybrid Machine Learning algorithm named Serial-diag** for fault detection in PV systems. These approaches are even capable of detecting and diagnosing faults such as the snail trail whose behavior is similar to that of a healthy panel. This algorithm is also tested to detect panels with broken glass, managing to classify them efficiently. In addition, this proposed approach proved to be very fast in computational terms because, thanks to the proposed combination of unsupervised and supervised learning, the heaviest calculation is only performed on a part of the faulty panels.

8. **An efficient normalization method for data from PV plants** that makes an important contribution. This type of approach not only makes it possible to compare PV strings with different numbers of PV panels, but also with different temperatures, irradiations, wind speeds, and even technologies. This approach also includes the degradation factor of the PV plant.

9. **A PV power prediction model adjusted to real data** that uses the following variables: ambient temperature, wind speed, irradiation, STC power, number of panels connected in series, plant installation date and annual degradation rate. Due to all these parameters, the proposed model is capable of estimating PV production at the PV module or PV string level. The estimation results of the module are compared with real data from a PV plant, obtaining a high level of coincidence with the real data. This model is also used to generate a data augmentation strategy, and generation of synthetic faults, as a solution to problems of insufficient amount of data or unbalanced data.

10. **A novel machine learning approach integrated to the new data acquisition system Solar Vitality** embedded in the new and versatile data acquisition system Solar Vitality. This approach combines supervised and unsupervised learning, as well as model and data-based learning. This approach uses data from the modeling of point 9, the techniques described in

point 5, together with the normalization of point 8, to detect, locate and identify faults in PV systems. This approach is able to use wind speed, ambient temperature, irradiation, datasheet information and age of the PV plant to automatically generate a cluster of healthy reference panels or strings. With this information, this approach is not only capable of detecting faults in the PV system but also automatically generating a maintenance priority report for faulty panels. This system is also evolutionary, since as errors are detected and new samples are classified, the internal database grows. Once new clusters are detected, the system trains itself and updates the fault diagnosis model.

Perspectives

This thesis leaves the door open for multiple future works because it addresses two large areas: hardware and software.

Hardware

Diagnosis-oriented Data Acquisition (Solar Vitality)

In future work the dimensions of Solar Vitality should be reduced. For this, a market analysis of different components that maintain signal quality but reduce energy consumption and dimensions must be carried out. This can also improve the weak point of Solar Vitality, which is the cost associated with high-performance sensors. In addition, the microcontrollers used in the latest version of Solar Vitality could be replaced by ones with lower power consumption. Also a touch screen could be added to the platform, to reduce its reliance on a computer to set up the system and start it. The inclusion of the screen also allows adding a new functionality that is real-time supervision.

In addition, this would make it easier to parameterize the system before startup. The two data acquisition cards and data processing cards could be replaced by a single card with the embedded system so that prototype size and power consumption could be reduced. If possible, a custom electronic card that allows integrating the voltage dividers should be made. This would increase the robustness of the platform considering that it is a portable prototype. Another study on panels failing at different data capture speeds should be done to determine if a high speed ADC is really necessary or could be changed to one with lesser features. This would reduce the price, and also the amount of data to be stored.

Weather station

Other wireless communication protocols should be explored to avoid cabling between the weather station and the new PV system monitoring platform. This will also facilitate the coupling of the station with the PV system. It would be interesting to examine the possibility of adding humidity sensors that are related to the

accelerated degradation of the PV modules and the appearance of other faults as demonstrated in Chapter 2. Special connectors must be put on the terminals of sensor cables. the weather station that goes directly to an electrical connection box that speeds up the connection process and avoids wrong connections of the sensors.

Software

Data storage

The system could be complemented by moving to relational tables that allow having primary and secondary keys to avoid confusion between the data. An automatic backup system could be implemented to avoid data loss. The data could no longer be stored on a local phpmyadmin server and instead go to online platforms such as those provided by Amazon or others.

Data pre-processing

The kalman estimator system that eliminates noise in the signals captured by the Arduino should be verified with different field tests. One of the most important tests is the withdrawal and appearance of the measurement source to adapt the coefficients adequately with the response times of the sensors and avoid hiding faults.

Diagnosis system

The approaches presented in Chapters 7 and 8 could be improved using the data normalization presented in Chapter 9. In addition, they could be modified to perform a multivariable classification to include aspects such as humidity, wind speed, the irradiation and ambient temperature. This, taking into account the excellent results obtained in Chapters 7 and 8. In addition, all the parameters of Chapter 9 for the model could be improved by including an optimization algorithm (heuristic or Meta-heuristic).

The fault database to improve training must continue to be built in the field, ideally with PV plants that have different configurations, technologies and types of panels. This would exponentially increase the robustness of the system. In addition, it would show whether it is necessary to further modify the data normalization proposed in the thesis.

Another idea would be to explore the possibility of coupling the system with an unmanned vehicle system (at least one) that captures images in order to effectively locate power plant faults that have a thermal signature at the panel level. Tests should be carried out to ship the algorithms proposed in Chapters 6-8 in devices such as inverters or optimizers to test their usability and accuracy.

Furthermore, a new coupled system dedicated to MPPT point tracking based on machine learning techniques could be easily coupled with the systems already presented in this thesis.

Maintenance priority reporting system

For the maintenance reporting system, a system coded in Visual Basic could automatically generate reports for the client, sending a series of change recommendations based on the detected faults.

Appendix A

Annexes

A.1 Examples of signals captured with the monitoring platform

For this first data capture test, the information from the temperature and irradiation sensors is captured with a frequency of one second and then compared to the measurements taken by the Adream building monitoring system. In this experiment, a pyranometer was also used to compare the measurement with the reference cell. The reference cell and the pyranometer were secured under the same slope of a solar panel located on the terrace of the Adream building as shown in Figure A.1. The data capture in the arduino is made every second, while the data of the Adream building is collected every 15 seconds.

Figure A.1: External connection of the sensors: Irradiance (yellow), Temperature (blue), Current (red) and Electrical power supply (white arrow)

Two comparison experiments are performed under different irradiation conditions. In addition, the captured current was also compared with the data recorded by the building's monitoring system. The temperature could not be compared because the building's temperature sensor had some problems and is not calibrated. The results are presented in Figures A.2 - A.3.

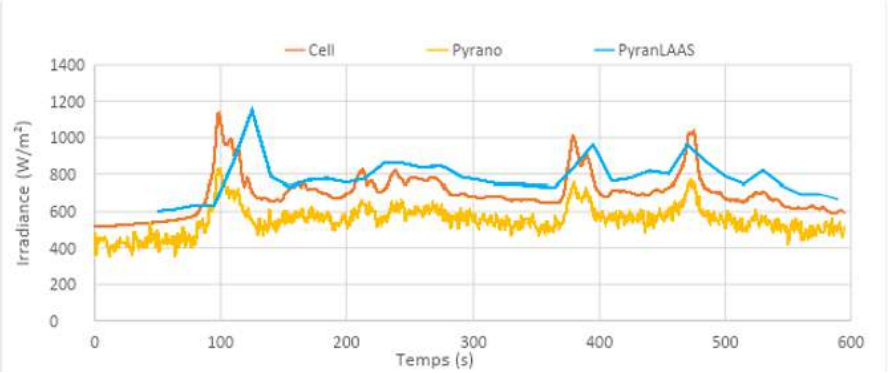

Figure A.2: Irradiation comparison.

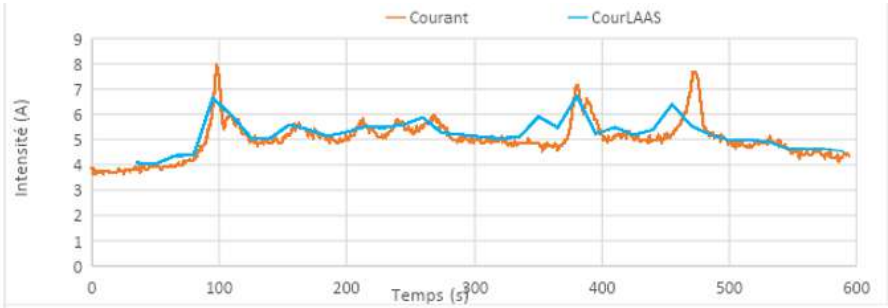

Figure A.3: Current comparison.

A second scenario was performed with other environmental conditions to verify if the behavior followed the same pattern. The results of the second scenario are presented in Figures A.4 - A.5.

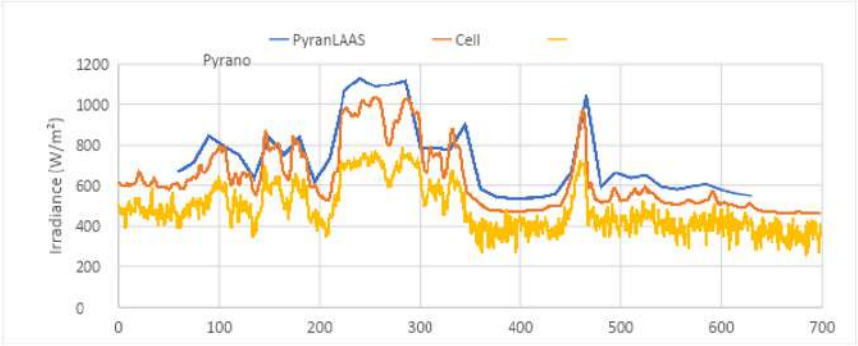

Figure A.4: Irradiation comparison.

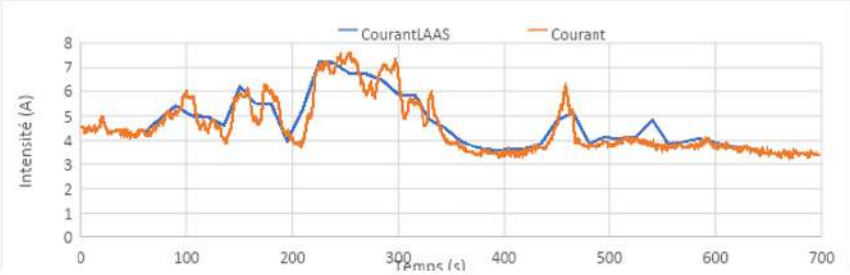

Figure A.5: Current comparison.

As can be seen in Figures A.2 and A.4, the reference cell again has the highest precision. The difference in measurement may be related to the place where the cell is fixed with respect to the sensors of the Adream building. Regarding Figures A.3 and A.5, the current transducer has a behavior similar to that observed by the supervision platform. In the following test the monitoring platform is connected to an isolated test panel. The interior of the first version of the monitoring platform is shown in Figure A.6.

Figure A.6: Data acquisition platform

As can be seen in Figure A.6, the acquisition platform contains the temperature sensor, current transducer, fuse holder, irradiance and temperature sensors and the Arduino Mega. Figure A.7 shows the front view of the solar panel with the photodiode (yellow circle at the top left) and the reference cell (yellow circle at the bottom left).

Figure A.7: Positioning of sensors on the panel

Two tests are performed and a hall effect current sensor was added. The first test captures the data each second and in the second test data is captured each 100 milliseconds. The results are presented in Figure A.8 andA.9

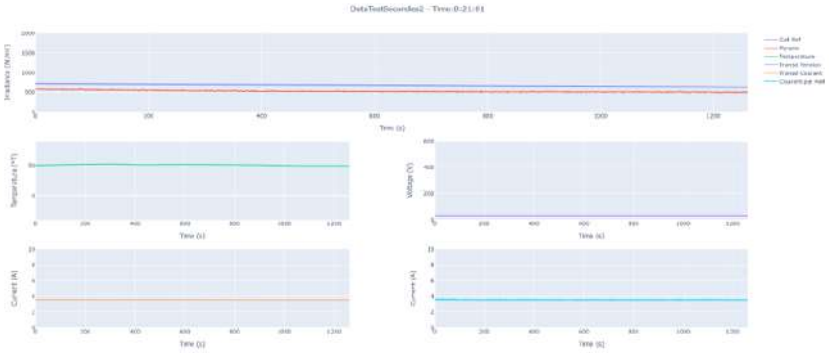

Figure A.8: Acquisition every second.

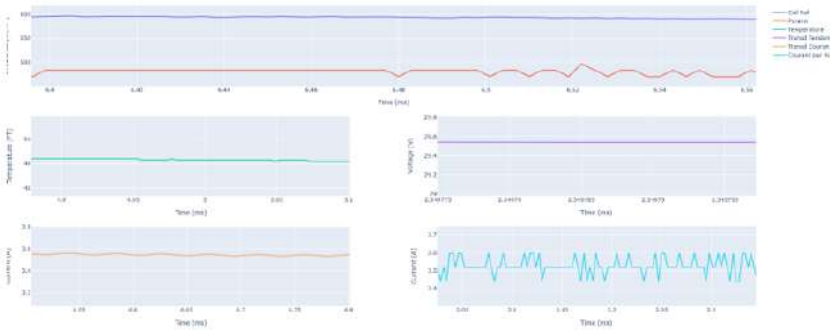

Figure A.9: Acquisition every 100 milliseconds.

As can be seen, the greatest oscillation with respect to current is obtained with the Hall effect sensor, so it is again recommended to use only the current transducer

due to its high precision. A third test is performed capturing the data every 55 milliseconds for 10 minutes. The results are presented in Figure A.9.

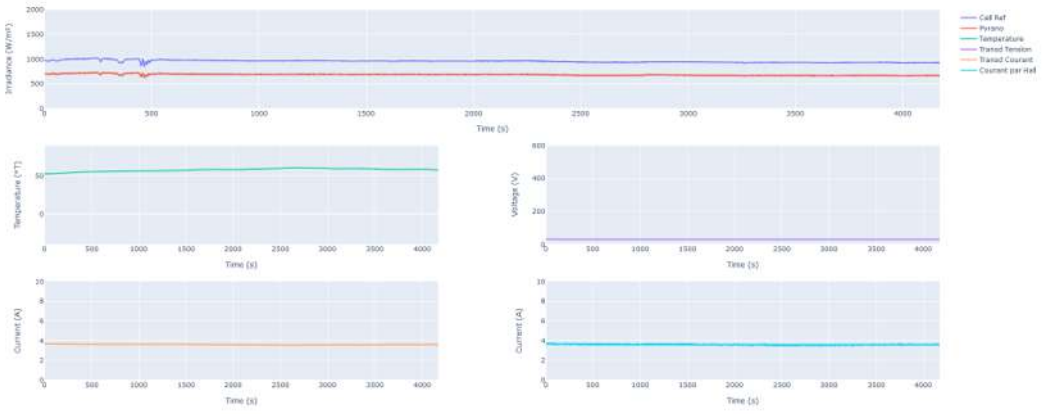

Figure A.10: Test of measurements made every 55 milliseconds.

As can be seen in Figure A.9, the behavior of the system every 55 microseconds is equivalent to the behavior every second. The only known limitation so far comes from the number of measurements per second that the arduino MEGA's ADCs can take.

A.2 Test of conventional machine learning algorithms.

For the examples presented here, only the data from the current signal, that make up a current database of 12 panels, with signals captured in parallel with a sampling time of one minute for 13 hours from 7:00 a.m. to 8:00 p.m. on June 25, 2020. The objective of demonstrating the results with these algorithms is to highlight the limitations of the actual acquisition platforms on the market for fault detection and classification of Broken Glass and Snail Trail faults. First, DTW is used in conjunction with K-means [Chen 2020a] to cluster the current signals. Then, other feature extraction and transformation methods, from Sections 6.3 and 6.4, are used as input to the Random Forest supervised learning method.

A.2.1 K-means Clustering

The K-means algorithm is an unsupervised classification algorithm that tries to divide N data objects into $k \geq 2$ partitions or groups, where each would have an object (mean) as its group center. That is, it groups them so that the objects in one group are similar to each other and different from those in other groups. The center of each group represents all the objects ($I_{i\{1:n_I\}}$ of each PV panel PV_i) within that group. Like hierarchical clustering, its result depends on the parameter k assigned. Then the algorithm assigns the rest of the objects to the appropriate groups and recalculates new centers. This process is done until the algorithm considers the cluster centers to be stable [Niennattrakul 2007].

To group the signals, this algorithm calculates the similarity between the time series and the centroid of its group. As a similarity metric, the DTW is used. There are several works that use the k-means algorithm for time series classification [Huang 2016, Niennattrakul 2007, Soheily-Khah 2016, Jang 2011]. In the following paragraphs, the clustering problem related to K-means is formalized.

Let $N = \{I_{1\{1:n_I\}}, \ldots, I_{n\{1:n_I\}}\}$, be the set of $n = 12$ time series of the current to be clustered by a similarity criterion. For a k-partition, $k \geq 2$, $P = \{c_1, \ldots, c_k\}$, is the set of cluster of N, let $U = \{u_1, \ldots, u_k\}$ be the set of centroids of P and $W = \{w_{1,1}, \ldots, w_{n,c_k}\}$ be the matrix of weights of dimensions $(n \times k)$, where each row $W(z,.)$, $z = 1, \ldots, n$, denotes the belonging of a current signal $I_{z\{1:n_I\}}$ with all k clusters and each column $W(.,q)$, $q = 1, \ldots, k$ denotes the belonging of all n current signals to a cluster c_q. Therefore, the clustering problem can be formulated as an optimization problem [Selim 1984] which is described as follows:

$$P : minimize \ z(W,U) = \sum_{z=1}^{n} \sum_{q=1}^{k} w_{z,q} d(x_z, u_q), \tag{A.1}$$

$$subject \ to \ \sum_{z=1}^{n} w_{z,q} = 1, for \ z = 1, \ldots, n,$$

$$w_{z,q} = 0 \ or \ 1, \ for \ z = 1, \ldots, n, \ and \ q = 1, \ldots, k,$$

where $w_{z,q} = 1$ implies object $I_{z\{1:n_I\}}$ belongs to cluster c_q and $d(x_z, u_q)$ denotes the DTW between $I_{z\{1:n_I\}}$ and c_q for $z = 1, \ldots, n$, and $q = 1, \ldots, k$. Figure A.11 presents the results of the clustering with K-means applied to the time series of the current $I_{i\{1:n_I\}}$ of each PV panel PV_i, $i = 1, \ldots, n$. In Figure A.11, Figure A.11a represents the original signals. Figure A.11b represents the centroids found by k-means algorithm. Finally, Figure A.11c represents an overlay of the original signals and the centroids.

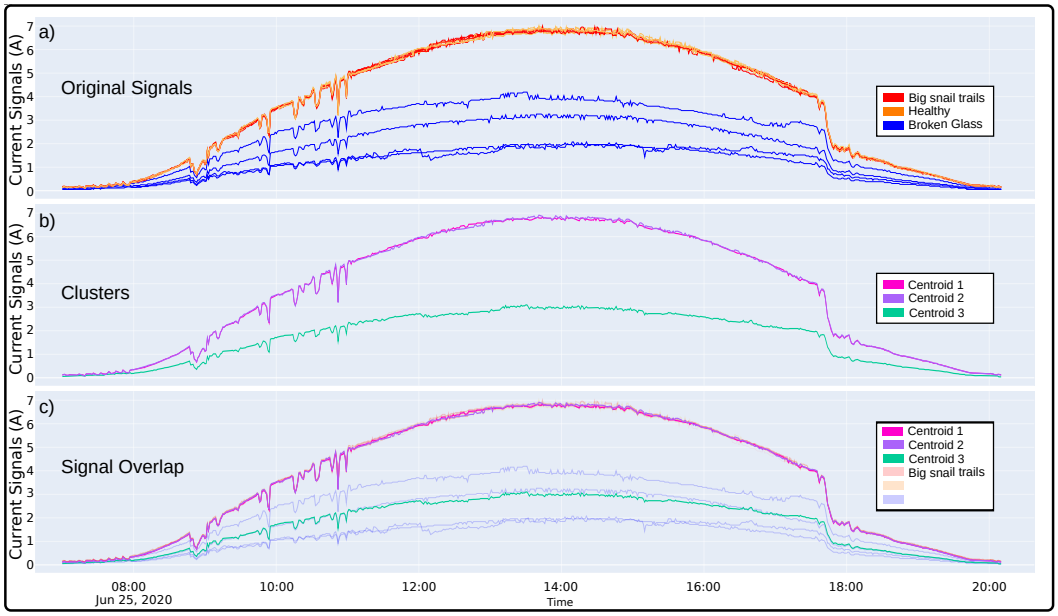

Figure A.11: K-means clustering applied to the time series of the current $I_{i\{1:n_I\}}$ of each PV panel PV_i, $i = 1, \ldots, n$. a) the original signals. b) the centroids found using the k-means algorithm. c) An overlay of the original signals and the centroids

As can be seen in the Figure A.11b, the algorithm is able to determine three centroids. The first and second centroids that group the healthy PV panels and Snail Trails PV panels. The third centroid groups the panels with the Broken Glass fault. It is evident that both in the Hierarchical Clustering algorithm and in the K-means clustering the group of photovoltaic panels of cluster A (Broken Glass) is easily discriminable. Nevertheless, as can be seen in Section A.2.1, unsupervised learning algorithms are really limited to detect the Snail Trail fault as it requires really fine detection. In order to improve the detection of this type of fine faults, other feature extraction and transformation methods, from Sections 6.3 and 6.4, are used as input for a supervised classification algorithm named Random Forest (RF).

A.2.2 Random Forest (RF)

The Rf model, having several integrated classifiers, usually improves diagnosis performance than the use of individual classifiers [Zhang 2020]. RF is an ensemble learning method for classification that builds multiple decision trees during the training phase and generates the final class by majority voting [Ho 1998]. This approach represents an improvement in the accuracy of fault detection compared to the use of a single decision tree. This approach proposes the use of only the decision tree as weak learners, and the strong learner Random Forest model is integrated using the bagging algorithm [Badarna 2019]. This bagging method randomly selects samples from the original training set of samples to generate a training set for each member of the set by random band playback [Zhang 2020]. The bagging

method allows each subtree to be trained on different parts of the same training set, with the goal of reducing the variance [Ho 1995]. It has been mentioned in the literature that RF mitigates the overfitting of DT during training and therefore generally outperforms DT [Ali 2012].

In the domain of fault detection and classification in PV systems, RF has been used to analyze the characteristics extracted from the I-V curves, allowing the detection of faults such as partial shading, open circuit, short circuit and degradation faults in an array PV [Hu 2017a]. In [Chen 2018c] it is used for classification of partial shading, Line to Line LF, PV chain open circuit and degradation fault. In [Heinrich 2020], the soiling rate is detected and analyzed using RF on the signals of V_{MPP}, I_{MPP} and T_m. However, in this thesis the RF is used for the first time for the detection of Snail Trail faults.

For this approach two predictor matrices M_F are evaluated and it will be used only to detect the panels with the Snail Trail fault due to the high limitation of the HC and k-means algorithms. The first M_F (Without combined feature extraction approach) is constructed only by extracting the 7 statistical characteristics exposed in Section 6.3.2. The second M_F, named new approach (combined feature extraction approach), is built using first the signal decomposition of Section 6.3.1 and then on each component of the reconstructed signal, the extraction of the 7 statistical features of the Section 6.3.2 is carried out.

To evaluate the degree of correct predictions (ability to identify positive and negative samples) the confusion matrix and the F_{value} are used. The F_{value} metric does not take into account the true negatives (TN), for this reason, in cases of unbalanced classes it improves the perception of the performance of the algorithm [Ferri 2009]. The F_{value} ranges from 0 to 1, where 1 indicates the best performance and 0 the worst. The F_{value} is defined as:

$$F_{value} = 2 * \frac{precision * recall}{precision + recall} \tag{A.2}$$

where the precision, $precision = TP/(TP + FP)$, allows us to measure the cost of false positives. The recall, $recall = TP/(TP + FN)$, allows estimating the number of individuals correctly classified as true positives compared to the total number of elements belonging to the class.

The confusion matrix is a widely known tool that allows visualizing the performance of a supervised learning algorithm or classification algorithm [Demir 2022]. In this matrix, each column represents the number of predictions of each class, while each row represents the instances in the actual class. This allows to see what types of successes and errors our model is having when going through the learning process with the current data as a function of the time of each PV panel of the string.

In this RF approach, to ensure the performance of strong learners, the differences with weak students should be as large as possible [Zhang 2020]. The methodology of this approach starts with randomly selected n samples with reproduction as training samples on all n samples for each tree in the random forest. This process is done to increase the chance that the samples in each tree are different, while the samples in

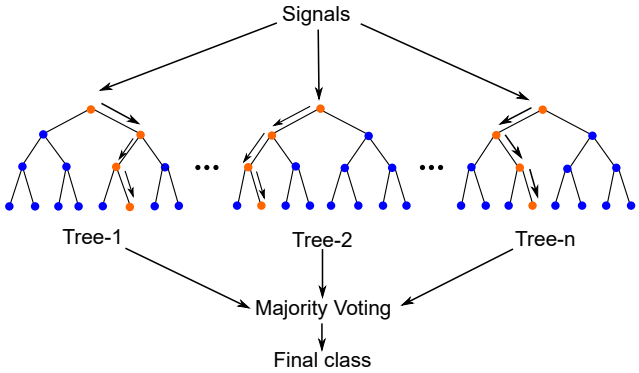

Figure A.12: Example of classifying a new sample using Random forest (RF) model

Approach	Without approach	New approach
	Reference	Combined feature extraction
RF	0.55	**0,73**
Decision Tree	0,47	0,52

Table A.1: Fault detection and classification results with Random forest (RF) model.

the same tree are repeatable. As a second measure to guarantee that the samples of each tree are different, from the characteristics M (number of columns of the matrix M_F), only k characteristics are selected for each tree. The most common way to calculate k is by $log_2 M$ or $\sqrt{M}$.

Because a sample can be entered into different decision trees, it is likely to fall into different categories due to differences between the trees. It is there that the final label is assigned using the votes from each decision tree. Like the ensemble learning algorithm, the RF model, by combining multiple Decision Tree algorithms, instead of using a single learning model, considerably increases detection accuracy. This approach presents contributions such as: 1) It reduces the calculation time necessary to detect and classify faults in PV systems compared to the Decision Trees algorithm; 2) It also achieves classification using only the MPP current sensor (no additional sensors required); and 3) Despite the small number of individuals (current signals from the panels), it manages to classify all healthy panels with a low cost of learning, data acquisition and storage.

As can be seen from Figure A.13 and Table A.1, Random Forests outperform decision trees. However, it is important to mention that it can be observed that the RF performance can be strongly affected in scenarios with a small number or a low-dimensional data set (Without approach). When the predictor matrix used as input for the RF is of very low dimension, it is possible that in some scenarios the RF performance is lower than that of a single decision tree [Zhang 2020].

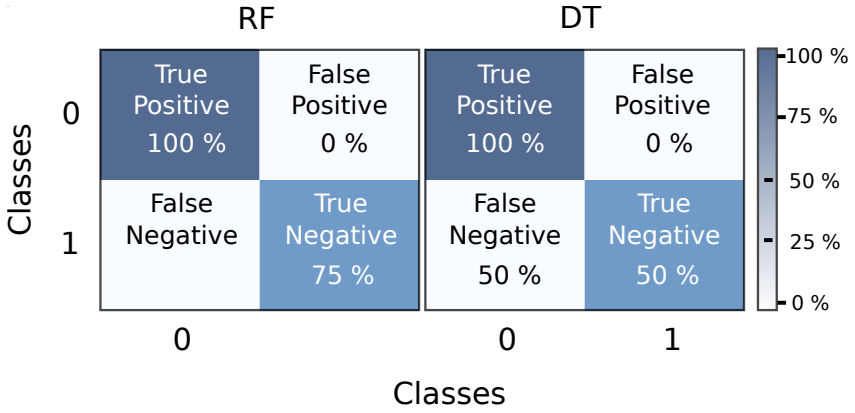

Figure A.13: Confusion matrix of the results of the RF model. The class 0 corresponds to healthy panels and the class 1 corresponds to panels with a Snail Trail.

A.2.3 Discussion and Conclusions

As can be seen, some of the conventional and most representative methods of machine learning, both unsupervised and supervised, have serious problems in detecting Snail Trail faults. As mentioned in Chapter 2 the Snail Trail fault does not significantly decrease the performance of solar panels (emulates healthy behavior) and is therefore difficult to detect at the electrical signal level. However, as stated above, it is a fault that can originate from microcracks and severe cracks to corrosion and hot spots. Therefore, its early detection is vital.

It is interesting to note that for the detection of faults whose electrical signature is different from that of a healthy panel, unsupervised detection algorithms such as hierarchical clustering and k-means represent a great opportunity with low computational cost since they do not require the multiple feature extraction. A disadvantage of the two unsupervised methods is that the number of desired classes must be defined a priori. For example, as mentioned in [Nielsen 2016] the final result of the HC depends on the level at which the bunches are cut. This characteristic of the HC could be seen as an advantage, if what is desired is to differentiate the level of affectation between panels of the same fault, since, as can be seen in Figure 8.4, if the cut-off level is increased, it is possible to determine sub-clusters linked to the level fault impact. This aspect is vital to establish a priority in preventive maintenance. To test the robustness of the HC and k-means algorithms, both are tested with different time windows, always obtaining the same result even with 3-minute windows. This aspect is vital to establish a priority in preventive maintenance.

Regarding the RF approach, it is possible to notice how the increase in the number of features also increases the precision in the detection and classification of faults. In addition, it can be seen that although RF has limitations for Snail Trail detection, it is capable of detecting 3 out of 4 panels with the Snail Trail fault despite the small number of individuals (samples) and the high similarity between the samples of each class. One of the advantages of all the methods presented in

this section is that they manage to group the signals using a single variable, which is the panel current. This can translate into a significant reduction in the number of sensors for fault diagnosis whose electrical signature is different from that of a healthy panel. In addition, it is not necessary to cut PV production to carry out the diagnosis and they work with a reduced number of individuals of each class (panels per class), which avoids the need for large numbers of individuals to train the diagnosis system.

Considering that Snail Trail faults are currently detected with regular staff visits to PV plants, these approaches really present an important contribution to automatic fault detection. In general, the diagnosis process proposed here using the extraction and transformation of signatures together with the supervised machine learning algorithms, although it continues with limitations regarding the complete detection of the panels with Snail Trail, is a step towards understanding the existing limitations. For the reasons stated above a new fault detection algorithm is exposed in Chapters 7-9.

Bibliography

[. 2019] ., S. and Zinger, D. *Review on Methods of Fault Diagnosis in Photovoltaic System Applications.* Journal of Engineering Science and Technology Review, vol. 12, pages 53–66, 10 2019. [Online]. Available: http://dx.doi.org/10.25103/jestr.125.07. (Cited in pages 92 and 111.)

[Abbas 2022] Abbas, A. N., Chasparis, G. and Kelleher, J. D. *Interpretable Hidden Markov Model-Based Deep Reinforcement Learning Hierarchical Framework for Predictive Maintenance of Turbofan Engines.* 2022. [Online]. Available: `https://arxiv.org/abs/2206.13433`. (Cited in page 123.)

[AbdulMawjood 2018a] AbdulMawjood, K., Refaat, S. S. and Morsi, W. G. *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), pages 1–8, 2018. (Cited in pages 56, 57, 59, 83, 84, 85, 86, 88, and 89.)

[AbdulMawjood 2018b] AbdulMawjood, K., Refaat, S. S. and Morsi, W. G. *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), pages 1–8, Doha, 2018. IEEE. (Cited in pages 69, 70, 71, and 77.)

[Abubakar 2021] Abubakar, A., Almeida, C. F. M. and Gemignani, M. *Review of Artificial Intelligence-Based Failure Detection and Diagnosis Methods for Solar Photovoltaic Systems.* Machines, vol. 9, no. 12, 2021. (Cited in page 92.)

[Adinoyi 2013] Adinoyi, M. J. and Said, S. A. *Effect of dust accumulation on the power outputs of solar photovoltaic modules.* Renewable Energy, vol. 60, pages 633–636, 2013. (Cited in page 65.)

[Adothu 2019] Adothu, B., Bhatt, P., Chattopadhyay, S., Zele, S., Oderkerk, J., Sagar, H., Costa, F. R. and Mallick, S. *Newly developed thermoplastic polyolefin encapsulant–A potential candidate for crystalline silicon photovoltaic modules encapsulation.* Solar Energy, vol. 194, pages 581–588, 2019. (Cited in page 61.)

[Afrasiabi 2019] Afrasiabi, M., Mohammadi, M., Rastegar, M. and Kargarian, A. *Multi-agent microgrid energy management based on deep learning forecaster.* Energy, vol. 186, page 115873, 2019. (Cited in page 118.)

[Afrasiabi 2021] Afrasiabi, M., Mohammadi, M., Rastegar, M. and Afrasiabi, S. *Advanced Deep Learning Approach for Probabilistic Wind Speed Forecasting.*

IEEE Transactions on Industrial Informatics, vol. 17, no. 1, pages 720–727, 2021. [Online]. Available: http://dx.doi.org/10.1109/TII.2020.3004436. (Cited in page 116.)

[Afrasiabi 2022a] Afrasiabi, S., Afrasiabi, M., Jarrahi, M. A. and Mohammadi, M. *Fault Location and Faulty Line Selection in Transmission Networks: Application of Improved Gated Recurrent Unit.* IEEE Systems Journal, vol. 16, no. 3, pages 5056–5066, 2022. [Online]. Available: http://dx.doi.org/10.1109/JSYST.2022.3172406. (Cited in page 116.)

[Afrasiabi 2022b] Afrasiabi, S., Allahmoradi, S., Salimi, M., Liang, X. and Chung, C. *Machine Learning-Based Condition Monitoring of Solar Photovoltaic Systems: A Review.* In 2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pages 49–54, 2022. (Cited in pages 92, 114, and 119.)

[Aghaei 2015] Aghaei, M., Gandelli, A., Grimaccia, F., Leva, S. and Zich, R. E. *IR real-time analyses for PV system monitoring by digital image processing techniques.* In 2015 International Conference on Event-based Control, Communication, and Signal Processing (EBCCSP), pages 1–6, 2015. (Cited in pages vi and 66.)

[Aghaei 2016] Aghaei, M., Leva, S. and Grimaccia, F. *PV power plant inspection by image mosaicing techniques for IR real-time images.* In 2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC), pages 3100–3105, 2016. (Cited in page 141.)

[Aghaei 2022] Aghaei, M., Fairbrother, A., Gok, A., Ahmad, S., Kazim, S., Lobato, K., Oreski, G., Reinders, A., Schmitz, J., Theelen, M., Yilmaz, P. and Kettle, J. *Review of degradation and failure phenomena in photovoltaic modules.* Renewable and Sustainable Energy Reviews, vol. 159, page 112160, 2022. (Cited in page 56.)

[Ahmad 2018] Ahmad, S., Hasan, N., Bharath Kurukuru, V. S., Ali Khan, M. and Haque, A. *Fault Classification for Single Phase Photovoltaic Systems using Machine Learning Techniques.* In 8th IEEE India International Conference on Power Electronics (IICPE), pages 1–6, 2018. (Cited in pages 30, 183, 184, 186, 232, 242, and 245.)

[Ahmadi 2011] Ahmadi, H. and Khaksar, Z. Power spectral density technique for fault diagnosis of an electromotor, volume 165 of *Hruschka E.R., Watada J., do Carmo Nicoletti M. (eds) Integrated Computing Technology. INTECH 2011. Communications in Computer and Information Science.* Springer, Berlin, Heidelberg, 2011. (Cited in page 144.)

[Ahmadipour 2018a] Ahmadipour, M., Hizam, H., Lutfi Othman, M. and Amran Mohd Radzi, M. *An Anti-Islanding Protection Technique Using a Wavelet*

Packet Transform and a Probabilistic Neural Network. Energies, vol. 11, no. 10, 2018. (Cited in page 183.)

[Ahmadipour 2018b] Ahmadipour, M., Hizam, H., Othman, M. L., Radzi, M. A. M. and Murthy, A. S. *Islanding detection technique using Slantlet Transform and Ridgelet Probabilistic Neural Network in grid-connected photovoltaic system.* Applied Energy, vol. 231, pages 645–659, 2018. (Cited in page 183.)

[A.h.mohamed 2015] A.h.mohamed and A.m.nassar. *New Algorithm for Fault Diagnosis of Photovoltaic Energy Systems.* International Journal of Computer Applications, vol. 114, no. 9, pages 26–31, March 2015. (Cited in pages 122 and 123.)

[Akbari 2021] Akbari, A., Ebrahimi, J., Jafarian, Y. and Bakhshai, A. *A Multilevel Inverter Topology With an Improved Reliability and a Reduced Number of Components.* IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. PP, pages 1–1, 06 2021. [Online]. Available: http://dx.doi.org/10.1109/JESTPE.2021.3089867. (Cited in page 54.)

[Akinlolu 2020] Akinlolu, M. and Haupt, T. C. *A Bibliometric review of trends in construction safety Technology Research.* Proceedings of International Structural Engineering and Construction, vol. 7, no. 2, 2020. (Cited in page 96.)

[Akram 2015] Akram, M. N. and Lotfifard, S. *Modeling and Health Monitoring of DC Side of Photovoltaic Array.* IEEE Transactions on Sustainable Energy, vol. 6, no. 4, pages 1245–1253, 2015. (Cited in pages 115 and 180.)

[Akram 2019] Akram, M. W., Li, G., Jin, Y., Chen, X., Zhu, C., Zhao, X., Khaliq, A., Faheem, M. and Ahmad, A. *CNN based automatic detection of photovoltaic cell defects in electroluminescence images.* Energy, vol. 189, page 116319, 2019. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0360544219320146`. (Cited in page 117.)

[Akram 2020] Akram, M. W., Li, G., Jin, Y., Chen, X., Zhu, C. and Ahmad, A. *Automatic detection of photovoltaic module defects in infrared images with isolated and develop-model transfer deep learning.* Solar Energy, vol. 198, pages 175–186, 2020. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0038092X20300621`. (Cited in pages 117 and 239.)

[Al Ahmar 2010] Al Ahmar, E., Choqueuse, V., Benbouzid, M. E. H., Amirat, Y., El Assad, J., Karam, R. and Farah, S. *Advanced signal processing techniques for fault detection and diagnosis in a wind turbine induction generator drive train: A comparative study.* In 2010 IEEE Energy Conversion Congress and Exposition, pages 3576–3581, 2010. (Cited in page 144.)

[Al Mamun 2022] Al Mamun, M. A., Azad, M. A. K., Al Mamun, M. A. and Boyle, M. *Review of flipped learning in engineering education: Scientific mapping and research horizon.* Education and Information Technologies, vol. 27, no. 1, pages 1261–1286, 2022. (Cited in page 96.)

[Alam 2013a] Alam, M. K., Khan, F., Johnson, J. and Flicker, J. *PV ground-fault detection using spread spectrum time domain reflectometry (SSTDR).* In 2013 IEEE Energy Conversion Congress and Exposition, pages 1015–102, 2013. (Cited in page 144.)

[Alam 2013b] Alam, M. K., Khan, F. H., Johnson, J. and Flicker, J. *PV faults: Overview, modeling, prevention and detection techniques.* In 2013 IEEE 14th Workshop on Control and Modeling for Power Electronics (COMPEL), pages 1–7, 2013. (Cited in pages 69, 70, and 77.)

[Alam 2013c] Alam, M. K., Khan, F. H., Johnson, J. and Flicker, J. *PV faults: Overview, modeling, prevention and detection techniques.* In 2013 IEEE 14th Workshop on Control and Modeling for Power Electronics (COMPEL), pages 1–7, Salt Lake City, UT, USA, 2013. IEEE. (Cited in page 180.)

[Alam 2015a] Alam, M. K., Khan, F., Johnson, J. and Flicker, J. *A Comprehensive Review of Catastrophic Faults in PV Arrays: Types, Detection, and Mitigation Techniques.* IEEE Journal of Photovoltaics, vol. 5, no. 3, pages 982–997, 2015. (Cited in pages vi, 30, 69, 70, and 71.)

[Alam 2015b] Alam, M. K., Khan, F., Johnson, J. and Flicker, J. *A Comprehensive Review of Catastrophic Faults in PV Arrays: Types, Detection, and Mitigation Techniques.* IEEE Journal of Photovoltaics, vol. 5, no. 3, pages 982–997, 2015. (Cited in page 180.)

[Aletras 2013] Aletras, N. and Stevenson, M. *Evaluating Topic Coherence Using Distributional Semantics.* pages 13–22, 03 2013. (Cited in page 101.)

[Ali 2012] Ali, J., Khan, R., Ahmad, N. and Maqsood, I. *Random Forests and Decision Trees.* International Journal of Computer Science Issues(IJCSI), vol. 9, 09 2012. (Cited in page 260.)

[Ali 2017] Ali, M. H., Rabhi, A., Hajjaji, A. E. and Tina, G. M. *Real Time Fault Detection in Photovoltaic Systems.* Energy Procedia, vol. 111, pages 914–923, 2017. 8th International Conference on Sustainability in Energy and Buildings, SEB-16, 11-13 September 2016, Turin, Italy. (Cited in page 87.)

[Ali 2020] Ali, M. U., Khan, H. F., Masud, M., Kallu, K. D. and Zafar, A. *A machine learning framework to identify the hotspot in photovoltaic module using infrared thermography.* Solar Energy, vol. 208, pages 643–651, 2020. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0038092X20308665`. (Cited in page 119.)

[alias Balamurugan 2011] alias Balamurugan, A., Rajaram, R., Pramala, S., Rajalakshmi, S., Jeyendran, C. and Dinesh Surya Prakash, J. *NB+: An improved Naïve Bayesian algorithm.* Knowledge-Based Systems, vol. 24, no. 5, pages 563–569, 2011. (Cited in page 112.)

[Almeida 2018] Almeida, F. *Canvas Framework for Performing Systematic Reviews Analysis.* Multidisciplinary Journal for Education, Social and Technological Sciences, vol. 5, no. 1, page 65–85, Mar. 2018. (Cited in page 93.)

[Alrifaey 2022] Alrifaey, M., Lim, W. H., Ang, C. K., Natarajan, E., Solihin, M. I., Juhari, M. R. M. and Tiang, S. S. *Hybrid Deep Learning Model for Fault Detection and Classification of Grid-Connected Photovoltaic System.* IEEE Access, vol. 10, pages 13852–13869, 2022. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2022.3140287. (Cited in page 118.)

[Alter 2000] Alter, O., Brown, P. and D., B. *Singular value decomposition for genome-wide expression data processing and modeling.* Proc Natl Acad Sci U S A, vol. 97, no. 18, 2000. (Cited in page 192.)

[Alves dos Reis Benatto 2020] Alves dos Reis Benatto, G., Mantel, C., Spataru, S., Santamaria Lancia, A. A., Riedel, N., Thorsteinsson, S., Poulsen, P. B., Parikh, H., Forchhammer, S. and Sera, D. *Drone-Based Daylight Electroluminescence Imaging of PV Modules.* IEEE Journal of Photovoltaics, vol. 10, no. 3, pages 872–877, 2020. [Online]. Available: http://dx.doi.org/10.1109/JPHOTOV.2020.2978068. (Cited in page 85.)

[Ameur 2021] Ameur, A., Berrada, A., Loudiyi, K. and Adomatis, R. *Chapter 6 - Performance and energetic modeling of hybrid PV systems coupled with battery energy storage.* In Berrada, A. and El Mrabet, R., editors, Hybrid Energy System Models, pages 195–238. Academic Press, 2021. (Cited in page 42.)

[Aminikhanghahi 2017] Aminikhanghahi, S. and Cook, D. *A survey of methods for time series change point detection.* Knowl Inf Syst 51, page 339–367, 2017. (Cited in page 219.)

[Andreoni-López 2012] Andreoni-López, M. E., Galdeano Mantiñan, F. J. and Molina, M. G. *Implementation of wireless remote monitoring and control of solar photovoltaic (PV) system.* In 2012 Sixth IEEE/PES Transmission and Distribution: Latin America Conference and Exposition (T&D-LA), pages 1–6, 2012. (Cited in page 146.)

[Andrianajaina 2017] Andrianajaina, T., Sambatra, E., Andrianirina, C., Razafimahefa, T. D. and Heraud, N. *Modeling, analysis and comparison of shading effects in a photovoltaic array using different configurations.* 07 2017. (Cited in pages v and 53.)

[Anil Kumar 2016] Anil Kumar, T. C., Singh, G. and Naikan, V. N. A. *Effectiveness of vibration and current monitoring in detecting broken rotor bar and bearing faults in an induction motor.* In 2016 IEEE 6th International Conference on Power Systems (ICPS), pages 1–5, 2016. (Cited in page 144.)

[Ansari 2021] Ansari, S., Ayob, A., Lipu, M. S. H., Saad, M. H. M. and Hussain, A. *A Review of Monitoring Technologies for Solar PV Systems Using Data Processing Modules and Transmission Protocols: Progress, Challenges and Prospects.* Sustainability, vol. 13, no. 15, 2021. (Cited in page 137.)

[Antonanzas 2016] Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., de Pison, F. M. and Antonanzas-Torres, F. *Review of photovoltaic power forecasting.* Solar Energy, vol. 136, pages 78–111, 2016. (Cited in pages 106, 108, 109, and 125.)

[Anuradha 2014] Anuradha and Gupta, G. *A self explanatory review of decision tree classifiers.* In International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), pages 1–7, 2014. (Cited in page 205.)

[Anwari 2011] Anwari, M., Dom, M. M. and Rashid, M. *Small Scale PV Monitoring System Software Design.* Energy Procedia, vol. 12, pages 586–592, 2011. (Cited in page 146.)

[Appiah 2019a] Appiah, A., Zhang, X., Ayawli, B. and Kyeremeh, F. *Review and Performance Evaluation of Photovoltaic Array Fault Detection and Diagnosis Techniques.* International Journal of Photoenergy, vol. 2019, pages 1–19, 02 2019. [Online]. Available: http://dx.doi.org/10.1155/2019/6953530. (Cited in page 68.)

[Appiah 2019b] Appiah, A. Y., Zhang, X., Ayawli, B. B. K. and Kyeremeh, F. *Long Short-Term Memory Networks Based Automatic Feature Extraction for Photovoltaic Array Fault Diagnosis.* IEEE Access, vol. 7, pages 30089–30101, 2019. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2019.2902949. (Cited in page 118.)

[Araneo 2009] Araneo, R., Lammens, S., Grossi, M. and Bertone, S. *EMC Issues in High-Power Grid-Connected Photovoltaic Plants.* IEEE Transactions on Electromagnetic Compatibility, vol. 51, no. 3, pages 639–648, 2009. (Cited in page 29.)

[Ard 2022] *Arduino Playground*, 2022. Accessed: 2022-09-30. (Cited in page 163.)

[Arduino 2020] Arduino. *Arduino Due*, 2020. (Cited in page 163.)

[Arunkumar 2019] Arunkumar, K. and Manjunath, D. T. *A brief review/survey of vibration signal analysis in time domain.* SSRG International Journal of Electronics and Communication Engineering, vol. 3, no. 3, pages 12–55, 2019. (Cited in page 186.)

[Asadpour 2020] Asadpour, R., Sulas-Kern, D. B., Johnston, S., Meydbray, J. and Alam, M. A. *Dark Lock-in Thermography Identifies Solder Bond Failure as the Root Cause of Series Resistance Increase in Fielded Solar Modules.* IEEE Journal of Photovoltaics, vol. 10, no. 5, pages 1409–1416, 2020. (Cited in page 85.)

[Asif 2021] Asif, M., Tariq, M., Sarwar, A., Hussan, M. R., Ahmad, S., Mihet-Popa, L. and Shah Noor Mohamed, A. *A Robust Multilevel Inverter Topology for Operation under Fault Conditions.* Electronics, vol. 10, no. 24, 2021. (Cited in page 54.)

[Atl 2022] *Atlas de radiación solar*, 2022. Accessed: 2022-09-30. (Cited in pages 18 and 156.)

[Ayaz 2014] Ayaz, E. *A Review Study on Mathematical Methods for Fault Detection Problems in Induction Motors.* Balkan Journal of Electrical and Computer Engineering, vol. 2, no. 3, pages 156 – 165, 2014. [Online]. Available: `https://dergipark.org.tr/tr/download/article-file/39723`. (Cited in page 144.)

[Azad 2022] Azad, A. and Parvin, S. *Bibliometric analysis of photovoltaic thermal (PV/T) system: From citation mapping to research agenda.* Energy Reports, vol. 8, pages 2699–2711, 2022. (Cited in page 96.)

[Aziz 2020] Aziz, F., Ul Haq, A., Ahmad, S., Mahmoud, Y., Jalal, M. and Ali, U. *A Novel Convolutional Neural Network-Based Approach for Fault Classification in Photovoltaic Arrays.* IEEE Access, vol. 8, pages 41889–41904, 2020. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2020.2977116. (Cited in pages 116 and 117.)

[Bachmann 2012] Bachmann, J., Buerhop-Lutz, C., Steim, R., Schilinsky, P., Hauch, J. A., Zeira, E. and Christoh J., B. *Highly sensitive non-contact shunt detection of organic photovoltaic modules.* Solar Energy Materials and Solar Cells, vol. 101, pages 176–179, 2012. (Cited in page 85.)

[Badarna 2019] Badarna, M. and Shimshoni, I. *Selective sampling for trees and forests.* Neurocomputing, vol. 358, pages 93–108, 2019. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0925231219306617`. (Cited in page 259.)

[Badr 2016] Badr, H., Zaitchik, B. and Dezfuli, A. *A tool for hierarchical climate regionalization.* Earth Sci Inform 8, page 949–958, 2016. (Cited in page 219.)

[Badr 2021] Badr, M. M., Hamad, M. S., Abdel-Khalik, A. S., Hamdy, R. A., Ahmed, S. and Hamdan, E. *Fault Identification of Photovoltaic Array Based on Machine Learning Classifiers.* IEEE Access, vol. 9, pages 159113–159132, 2021. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2021.3130889. (Cited in page 115.)

[Baghaee 2020] Baghaee, H. R., Mlakić, D., Nikolovski, S. and Dragicević, T. *Support Vector Machine-Based Islanding and Grid Fault Detection in Active Distribution Networks.* IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 8, no. 3, pages 2385–2403, 2020. (Cited in page 119.)

[Baharin 2014] Baharin, K. A., Rahman, H. A., Hassan, M. Y. and Gan, C. K. *Hourly irradiance forecasting in Malaysia using support vector machine.* In 2014 IEEE Conference on Energy Conversion (CENCON), pages 185–190, 2014. (Cited in page 112.)

[Bai 2019] Bai, W., Quan, C. and Luo, Z.-W. *Improving Generative and Discriminative Modelling Performance by Implementing Learning Constraints in Encapsulated Variational Autoencoders.* Applied Sciences, vol. 9, no. 12, 2019. (Cited in page 238.)

[Bai 2020] Bai, Y., Li, H. and Liu, Y. *Visualizing research trends and research theme evolution in E-learning field: 1999-2018.* Scientometrics, vol. 126, 11 2020. (Cited in page 97.)

[Bakker 2007] Bakker, B. *Reinforcement learning by backpropagation through an LSTM model/critic.* In 2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning, pages 127–134, 2007. (Cited in page 123.)

[Baktashmotlagh 2013] Baktashmotlagh, M., Harandi, M. T., Lovell, B. C. and Salzmann, M. *Unsupervised Domain Adaptation by Domain Invariant Projection.* In 2013 IEEE International Conference on Computer Vision, pages 769–776, 2013. (Cited in page 238.)

[Banerjee 2022] Banerjee, A., Singh, D., Sahana, S. and Nath, I. *Chapter 3 - Impacts of metaheuristic and swarm intelligence approach in optimization.* In Mishra, S., Tripathy, H. K., Mallick, P. K., Sangaiah, A. K. and Chae, G.-S., editors, Cognitive Big Data Intelligence with a Metaheuristic Approach, Cognitive Data Science in Sustainable Computing, pages 71–99. Academic Press, 2022. (Cited in page 126.)

[Barja-Martinez 2021] Barja-Martinez, S., AragÃ¼Ã©s-PeÃ±alba, M., Ãngrid MunnÃ©-Collado, Lloret-Gallego, P., Bullich-MassaguÃ©, E. and Villafafila-Robles, R. *Artificial intelligence techniques for enabling Big Data services in distribution networks: A review.* Renewable and Sustainable Energy Reviews, vol. 150, page 111459, 2021. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1364032121007413`. (Cited in page 111.)

[Basnet 2020] Basnet, B., Chun, H. and Bang, J. *An Intelligent Fault Detection Model for Fault Detection in Photovoltaic Systems.* Journal of Sensors, no. 6960328, pages 1–11, 2020. (Cited in page 191.)

[Bastani 2019] Bastani, K., Namavari, H. and Shaffer, J. *Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints.* Expert Systems with Applications, vol. 127, pages 256–271, 2019. (Cited in page 98.)

[Bastidas-Rodriguez 2013] Bastidas-Rodriguez, J., Petrone, G., Ramos-Paja, C. and Spagnuolo, G. *Photovoltaic modules diagnostic: An overview.* In IECON 2013 - 39th Annual Conference of the IEEE Industrial Electronics Society, pages 96–101, 2013. (Cited in page 59.)

[Bay 2008] Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L. *Speeded-Up Robust Features (SURF).* Computer Vision and Image Understanding, vol. 110, no. 3, pages 346–359, 2008. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1077314207001555`. Similarity Matching in Computer Vision and Multimedia. (Cited in page 119.)

[Bayrak 2013] Bayrak, G. and Cebeci, M. *Monitoring a grid connected PV power generation system with labview.* In 2013 International Conference on Renewable Energy Research and Applications (ICRERA), pages 562–567, 2013. (Cited in page 139.)

[Bayram 2017] Bayram, D. and Şeker, S. *Redundancy-Based Predictive Fault Detection on Electric Motors by Stationary Wavelet Transform.* IEEE Transactions on Industry Applications, vol. 53, no. 3, pages 2997–3004, 2017. (Cited in page 183.)

[Belaout 2018a] Belaout, A., Krim, F., Mellit, A., Talbi, B. and Arabi, A. *Multiclass adaptive neuro-fuzzy classifier and feature selection techniques for photovoltaic array fault detection and classification.* Renewable Energy, vol. 127, pages 548–558, 2018. (Cited in page 183.)

[Belaout 2018b] Belaout, A., Krim, F., Mellit, A., Talbi, B. and Arabi, A. *Multiclass adaptive neuro-fuzzy classifier and feature selection techniques for photovoltaic array fault detection and classification.* Renewable Energy, vol. 127, pages 548–558, 2018. (Cited in page 184.)

[Bellini 2006] Bellini, A., Concari, C., Franceschini, G., Tassoni, C. and Toscani, A. *Vibrations, currents and stray flux signals to asses induction motors rotor conditions.* In IECON 2006 - 32nd Annual Conference on IEEE Industrial Electronics, pages 4963–4968, 2006. (Cited in page 144.)

[Ben Belghith 2014] Ben Belghith, O. and Sbita, L. *Remote GSM module monitoring and Photovoltaic system control.* In 2014 First International Conference on Green Energy ICGE 2014, pages 188–192, 2014. (Cited in pages 137 and 138.)

[Benghanem 1998a] Benghanem, M. and Maafi, A. *Data acquisition system for photovoltaic systems performance monitoring.* IEEE Transactions on In-

strumentation and Measurement, vol. 47, no. 1, pages 30–33, 1998. [Online]. Available: http://dx.doi.org/10.1109/19.728784. (Cited in page 146.)

[Benghanem 1998b] Benghanem, M. and Maafi, A. *Performance of stand-alone photovoltaic systems using measured meteorological data for Algiers.* Renewable Energy, vol. 13, no. 4, pages 495–504, 1998. (Cited in page 145.)

[Benghanem 2009a] Benghanem, M. *Low cost management for photovoltaic systems in isolated site with new IV characterization model proposed.* Energy Conversion and Management, vol. 50, no. 3, pages 748–755, 2009. (Cited in page 145.)

[Benghanem 2009b] Benghanem, M. *Measurement of meteorological data based on wireless data acquisition system monitoring.* Applied Energy, vol. 86, no. 12, pages 2651–2660, 2009. (Cited in page 145.)

[Benghanem 2010] Benghanem, M. *RETRACTED: A low cost wireless data acquisition system for weather station monitoring.* Renewable Energy, vol. 35, no. 4, pages 862–872, 2010. (Cited in page 145.)

[Benkercha 2018] Benkercha, R. and Moulahoum, S. *Fault detection and diagnosis based on C4.5 decision tree algorithm for grid connected PV system.* Solar Energy, vol. 173, pages 610–634, 2018. (Cited in pages 113 and 206.)

[Benninger 2020] Benninger, M., Hofmann, M. and Liebschner, M. *Anomaly detection by comparing photovoltaic systems with machine learning methods.* In NEIS 2020; Conference on Sustainable Energy Supply and Energy Storage Systems, pages 1–6, 2020. (Cited in page 128.)

[Berasategi Arostegi 2013] Berasategi Arostegi, A. *New Optimized Electrical Architectures of a Photovoltaic Generators with High Conversion Efficiency.* Theses, Université Paul Sabatier - Toulouse III, June 2013. (Cited in pages 50 and 52.)

[Berghout 2021a] Berghout, T., Benbouzid, M., Bentrcia, T., Ma, X., Djurović, S. and Mouss, L.-H. *Machine Learning-Based Condition Monitoring for PV Systems: State of the Art and Future Prospects.* Energies, vol. 14, no. 19, 2021. (Cited in pages 116, 238, and 239.)

[Berghout 2021b] Berghout, T., Benbouzid, M., Ma, X., Djurović, S. and Mouss, L.-H. *Machine Learning for Photovoltaic Systems Condition Monitoring: A Review.* In IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society, pages 1–5, 2021. (Cited in page 92.)

[Berghout 2021c] Berghout, T., Benbouzid, M. and Mouss, L.-H. *Leveraging Label Information in a Knowledge-Driven Approach for Rolling-Element Bearings Remaining Useful Life Prediction.* Energies, vol. 14, no. 8, 2021. (Cited in page 238.)

[Berry 2019] Berry, M., Mohamed, A. and Yap, B. Supervised and unsupervised learning for data science. Springer, Cham, Switzerland, 2019. (Cited in page 112.)

[Bharath KurukuruF 2020] Bharath KurukuruF, V. S., Blaabjerg, F., Khan, M. A. and Haque, A. *A Novel Fault Classification Approach for Photovoltaic Systems.* Energies, vol. 13, no. 2, page 308, 2020. [Online]. Available: http://dx.doi.org/https://doi.org/10.3390/en13020308. (Cited in page 144.)

[Bharti 2009] Bharti, R., Kuitche, J. and TamizhMani, M. G. *Nominal Operating Cell Temperature (NOCT): Effects of module size, loading and solar spectrum.* In 2009 34th IEEE Photovoltaic Specialists Conference (PVSC), pages 001657–001662, 2009. (Cited in page 235.)

[Bjurström 2011] Bjurström, A. and Polk, M. *Climate change and interdisciplinarity: A co-citation analysis of IPCC Third Assessment Report.* Scientometrics, vol. 87, pages 525–550, 06 2011. (Cited in page 95.)

[Blaesser 1995] Blaesser, G. and Munro, D. *Guidelines for the Assessment of Photovoltaic Plants Document A Photovoltaic System Monitoring.* Technical Report, Joint Research Centre, 1995. (Cited in pages 133, 157, and 158.)

[Blaesser 1997] Blaesser, G. *PV system measurements and monitoring the European experience.* Solar Energy Materials and Solar Cells, vol. 47, no. 1, pages 167–176, 1997. (Cited in pages 18, 145, and 156.)

[Blei 2001] Blei, D., Ng, A. and Jordan, M. *Latent Dirichlet Allocation.* volume 3, pages 601–608, 01 2001. (Cited in page 99.)

[Blei 2003] Blei, D. M., Ng, A. Y. and Jordan, M. I. *Latent dirichlet allocation.* Journal of machine Learning research, vol. 3, no. Jan, pages 993–1022, 2003. (Cited in page 99.)

[Blei 2007] Blei, D. M. and Lafferty, J. D. *A correlated topic model of Science.* The Annals of Applied Statistics, vol. 1, no. 1, pages 17 – 35, 2007. (Cited in page 99.)

[Borg 2005] Borg, I. and Groenen, P. Modern multidimensional scaling. Springer, Berlin, 2 edition, 2005. (Cited in page 97.)

[Boulesteix 2004] Boulesteix, A. *PLS Dimension Reduction for Classification with Microarray Data.* Statistical Applications in Genetics and Molecular Biology, vol. 3, pages 1 – 30, 2004. (Cited in page 224.)

[Boulesteix 2006] Boulesteix, A.-L. and Strimmer, K. *Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.* Briefings in Bioinformatics, vol. 8, no. 1, pages 32–44, 2006. (Cited in pages 221 and 223.)

[Bouma 2009] Bouma, G. *Normalized (pointwise) mutual information in collocation extraction*. 2009. (Cited in page 101.)

[Bouttier 2003] Bouttier, J., Di Francesco, P. and Guitter, E. *Geodesic distance in planar graphs*. Nuclear Physics B, vol. 663, no. 3, pages 535–567, 2003. (Cited in page 194.)

[Bouzidi 2012] Bouzidi, K., Chegaar, M. and Aillerie, M. *Solar Cells Parameters Evaluation from Dark I-V Characteristics*. Energy Procedia, vol. 18, pages 1601–1610, 2012. Terragreen 2012: Clean Energy Solutions for Sustainable Environment (CESSE). (Cited in page 50.)

[Bower 1994] Bower, W. and Wiles, J. *Analysis of grounded and ungrounded photovoltaic systems*. In Proceedings of 1994 IEEE 1st World Conference on Photovoltaic Energy Conversion - WCPEC (A Joint Conference of PVSC, PVSEC and PSEC), volume 1, pages 809–812 vol.1, 1994. (Cited in page 69.)

[Bracewell 1986] Bracewell, R. and Bracewell, R. The fourier transform and its applications. McGraw-Hill, 1986. (Cited in page 142.)

[Bressan 2013] Bressan, M., Dupé, V., Jammes, B., Talbert, T. and Alonso, C. *Monitoring and Analysis of Two Grid Connected PV Systems*. In PVSEC, pages 1–5, Paris, France, September 2013. (Cited in page 138.)

[Bressan 2014] Bressan, M. *Développement d'un outil de supervision et de contrôle pour une installation solaire photovoltaïque*. Theses, Université de Perpignan, June 2014. (Cited in pages 50 and 52.)

[Bressan 2016] Bressan, M., El Basri, Y., Galeano, A. and Alonso, C. *A shadow fault detection method based on the standard error analysis of I-V curves*. Renewable Energy, vol. 99, pages 1181–1190, 2016. (Cited in page 87.)

[Brooks 2011] Brooks, B. The bakersfield fire: A lesson in ground-fault protection. 2011. (Cited in pages 2 and 34.)

[Bun 2011a] Bun, L. *Détection et localisation de défauts dans un système photovoltaïque*. Theses, Université de Grenoble, November 2011. (Cited in pages 45, 51, 52, and 54.)

[Bun 2011b] Bun, L. *Détection et localisation de défauts pour un système PV*. Theses, Université de Grenoble, November 2011. (Cited in pages 4 and 35.)

[Buško 2011] Buško, V. Psychological testing theory, pages 1138–1139. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. (Cited in page 189.)

[Calderón 2020] Calderón, A., Barreneche, C., Hernández-Valle, K., Galindo, E., Segarra, M. and Fernández, A. I. *Where is Thermal Energy Storage (TES) research going? – A bibliometric analysis*. Solar Energy, vol. 200, pages 37–50, 2020. (Cited in pages 95 and 96.)

[Cao 2020] Cao, Y., Dong, Y., Cao, Y., Yang, J. and Yang, M. Y. *Two-stream convolutional neural network for non-destructive subsurface defect detection via similarity comparison of lock-in thermography signals.* NDT & E International, vol. 112, page 102246, 2020. (Cited in page 85.)

[Caprioglio 2020] Caprioglio, P., Wolff, C. M., Sandberg, O. J., Armin, A., Rech, B., Albrecht, S., Neher, D. and Stolterfoht, M. *On the Origin of the Ideality Factor in Perovskite Solar Cells.* Advanced Energy Materials, vol. 10, no. 27, page 2000502, 2020. (Cited in page 50.)

[Carvalho 2019] Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. d. P., Basto, J. P. and Alcalá, S. G. S. *A systematic literature review of machine learning methods applied to predictive maintenance.* Computers & Industrial Engineering, vol. 137, page 106024, 2019. (Cited in page 180.)

[CEC 1997] CEC. *Guidelines for the Assessment of Photovoltaic Plants, Document B, Analysis and Presentation of Monitoring Data.* Technical Report, Commission of the European Communities, 1997. (Cited in pages 3 and 35.)

[Cervantes 2020] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. and Lopez, A. *A comprehensive survey on support vector machine classification: Applications, challenges and trends.* Neurocomputing, vol. 408, pages 189–215, 2020. (Cited in pages 203 and 205.)

[Cesar 2017] Cesar, T. M., Pimentel, S. P., Marra, E. G. and Alvarenga, B. P. *Wavelet Transform analysis for grid-connected photovoltaic systems.* In 6th International Conference on Clean Electrical Power (ICCEP), pages 1–6, 2017. (Cited in page 217.)

[Chamarthi 2020] Chamarthi, P. K., Agarwal, V. and Al-Durra, A. *A New 1-, Seventeen Level Inverter Topology With Less Number of Power Devices for Renewable Energy Application.* Frontiers in Energy Research, vol. 8, 2020. (Cited in page 54.)

[Chamberlin 2011] Chamberlin, C. E., Rocheleau, M. A., Marshall, M. W., Reis, A. M., Coleman, N. T. and Lehman, P. A. *Comparison of PV module performance before and after 11 and 20 years of field exposure.* In 2011 37th IEEE Photovoltaic Specialists Conference, pages 000101–000105, 2011. (Cited in page 29.)

[Chang 2015] Chang, M., Chen, C., Hsueh, C. H., Hsieh, W. J., Yen, E., Ho, K. L., Chuang, H. P., Lee, C. Y. and Chen, H. *The reliability investigation of PV junction box based on 1GW worldwide field database.* 2015 IEEE 42nd Photovoltaic Specialist Conference (PVSC), pages 1–4, 2015. (Cited in pages vi and 72.)

[Chao 2008] Chao, K.-H., Ho, S.-H. and Wang, M.-H. *Modeling and fault diagnosis of a photovoltaic system.* Electric Power Systems Research, vol. 78, no. 1, pages 97–105, 2008. (Cited in pages 87 and 89.)

[Chao 2017] Chao, K.-H. and Chen, C.-T. *A remote supervision fault diagnosis meter for photovoltaic power generation systems.* Measurement, vol. 104, pages 93–104, 2017. (Cited in page 147.)

[Chapelle 2001] Chapelle, O., Vapnik, V., Bousquet, O. and Mukherjee, S. *Choosing Multiple Parameters for Support Vector Machines.* Machine Learning, vol. 46, 05 2001. (Cited in page 205.)

[Chen 2003] Chen, C. Mapping scientific frontiers. Springer, Berlin, 2003. (Cited in page 97.)

[Chen 2010] Chen, B., Zhu, L., Kifer, D. and Lee, D. *What Is an Opinion About? Exploring Political Standpoints Using Opinion Scoring Model.* Proceedings of the AAAI Conference on Artificial Intelligence, vol. 24, no. 1, pages 1007–1012, Jul. 2010. (Cited in page 99.)

[Chen 2014] Chen, F., Tang, B., Song, T. and Li, L. *Multi-fault diagnosis study on roller bearing based on multi-kernel support vector machine with chaotic particle swarm optimization.* Measurement, vol. 47, pages 576–590, 2014. (Cited in pages 188 and 205.)

[Chen 2017] Chen, Z., Wu, L., Cheng, S., Lin, P., Wu, Y. and Lin, W. *Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and I-V characteristics.* Applied Energy, vol. 204, pages 912–931, 2017. (Cited in pages 115 and 141.)

[Chen 2018a] Chen, L., Li, S. and Wang, X. *Quickest Fault Detection in Photovoltaic Systems.* IEEE Transactions on Smart Grid, vol. 9, no. 3, pages 1835–1847, 2018. (Cited in pages 68, 70, and 122.)

[Chen 2018b] Chen, Z., Han, F., Wu, L., Yu, J., Cheng, S., Lin, P. and Chen, H. *Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents.* Energy Conversion and Management, vol. 178, pages 250–264, 2018. (Cited in pages 29 and 113.)

[Chen 2018c] Chen, Z., Han, F., Wu, L., Yu, J., Cheng, S., Lin, P. and Chen, H. *Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents.* Energy Conversion and Management, vol. 178, pages 250–264, 2018. (Cited in page 260.)

[Chen 2019] Chen, J., Jin, Y., Akram, M., Li, K. and Chen, E. *Novel multi-convolutional neural network fusion approach for smile recognition.* Multimedia Tools and Applications, vol. 78, 06 2019. [Online]. Available: http://dx.doi.org/10.1007/s11042-018-6945-x. (Cited in page 117.)

[Chen 2020a] Chen, C.-H., Lin, W.-Y. and Lee, M.-Y. *The Applications of K-means Clustering and Dynamic Time Warping Average in Seismocardiography Template Generation.* In 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 1000–1007, 2020. (Cited in page 257.)

[Chen 2020b] Chen, X., Zou, D., Cheng, G. and Xie, H. *Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education.* Computers & Education, vol. 151, page 103855, 2020. (Cited in page 98.)

[Chianese 2003] Chianese, D., Friesen, G., Cereghetti, N., Realini, A., Bura, E., Rezzonico, S. and Bernasconi, A. *Final Report 2000–2003: qualità a e resa energetica di moduli ed impianti fotovoltaici—centrale LEEE-TISO.* Technical Report, OFEN, 2003. (Cited in page 240.)

[Chikh 2015] Chikh, A. and Chandra, A. *An Optimal Maximum Power Point Tracking Algorithm for PV Systems With Climatic Parameters Estimation.* IEEE Transactions on Sustainable Energy, vol. 6, no. 2, pages 644–652, 2015. [Online]. Available: http://dx.doi.org/10.1109/TSTE.2015.2403845. (Cited in page 141.)

[Chine 2015] Chine, W., Mellit, A., Pavan, A. M. and Lughi, V. *Fault diagnosis in photovoltaic arrays.* In 2015 International Conference on Clean Electrical Power (ICCEP), pages 67–72, 2015. (Cited in page 87.)

[Chine 2016a] Chine, W., Mellit, A., Lughi, V., Malek, A., Sulligoi, G. and Massi Pavan, A. *A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks.* Renewable Energy, vol. 90, pages 501–512, 2016. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0960148116300362`. (Cited in page 113.)

[Chine 2016b] Chine, W., Mellit, A., Lughi, V., Malek, A., Sulligoi, G. and Pavan, A. M. *A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks.* Renewable Energy, vol. 90, pages 501–512, 2016. (Cited in page 114.)

[Chouay 2021] Chouay, Y. and Ouassaid, M. *A Multi-stage SVM Based Diagnosis Technique for Photovoltaic PV Systems.* In Advances in Robotics, Automation and Data Analytics, volume 1350, pages 183–193. Springer International Publishing, Cham, 2021. (Cited in pages 119 and 180.)

[Chouder 2010] Chouder, A. and Silvestre, S. *Automatic supervision and fault detection of PV systems based on power losses analysis.* Energy Conversion and Management, vol. 51, no. 10, pages 1929–1937, 2010. (Cited in page 88.)

[Church 1990] Church, K. W. and Hanks, P. *Word Association Norms, Mutual Information, and Lexicography.* Computational Linguistics, vol. 16, no. 1, pages 22–29, 1990. (Cited in page 101.)

[Cid Pastor 2006a] Cid Pastor, A. *Conception et réalisation de modules photovoltaïques électroniques.* Theses, INSA de Toulouse, September 2006. (Cited in page 45.)

[Cid Pastor 2006b] Cid Pastor, A. *Conception et réalisation de modules photovoltaïques électroniques.* Theses, INSA de Toulouse, September 2006. (Cited in page 46.)

[Circutor 2022] Circutor. *PhotoScada.* 2022. [Online]. Available: `http://docs.circutor.com/docs/CT_SolPhotoSCADA_FR.pdf`. (Cited in pages 4 and 35.)

[Copper 2013a] Copper, J., Bruce, A., Spooner, T., Calais, M., Pryor, T. and Watt, M. *Australian technical guidelines for monitoring and analysing photovoltaic systems.* Technical Report, Australian PV institute, 2013. (Cited in page 235.)

[Copper 2013b] Copper, J., Bruce, A., Spooner, T., Calais, M., Pryor, T. and Watt, M. *Australian Technical Guidelines for Monitoring and Analysing Photovoltaic Systems.* Technical Report, 11 2013. (Cited in page 140.)

[Coronato 2020] Coronato, A., Naeem, M., De Pietro, G. and Paragliola, G. *Reinforcement learning for intelligent healthcare applications: A survey.* Artificial Intelligence in Medicine, vol. 109, page 101964, 2020. (Cited in page 122.)

[Correia 2021] Correia, I., Lameirinhas, R., Fernandes, C. and Torres, J. P. *Comparative Study of Copper Indium Gallium Selenide (CIGS) Solar Cell With Other Solar Technologies.* Sustainable Energy Fuels, vol. 5, 01 2021. [Online]. Available: http://dx.doi.org/10.1039/D0SE01717E. (Cited in page 46.)

[Costa 2015] Costa, F. B., Souza, B. A., Brito, N. S. D., Silva, J. A. C. B. and Santos, W. C. *Real-Time Detection of Transients Induced by High-Impedance Faults Based on the Boundary Wavelet Transform.* IEEE Transactions on Industry Applications, vol. 51, no. 6, pages 5312–5323, 2015. (Cited in page 183.)

[Cotterell 2012] Cotterell, M. Installation guidelines: Electrical, pages 819–834. 01 2012. (Cited in page 69.)

[Cristaldi 2010] Cristaldi, L., Faifer, M., Ferrero, A. and Nechifor, A. *On-line monitoring of the efficiency of photo-voltaic panels for optimizing maintenance scheduling.* In 2010 IEEE Instrumentation & Measurement Technology Conference Proceedings, pages 954–959, 2010. (Cited in pages 137 and 139.)

[Cristaldi 2015] Cristaldi, L., Faifer, M., Lazzaroni, M., Khalil, M. M. A. F., Catelani, M. and Ciani, L. *Diagnostic architecture: A procedure based on the analysis of the failure causes applied to photovoltaic plants.* Measurement, vol. 67, pages 99–107, 2015. (Cited in pages 60, 64, 72, and 133.)

[Cristaldi 2016] Cristaldi, L., Faifer, M. and Lazzaroni, M. *A cooperative monitoring and diagnostic architecture for PV systems.* In 2016 IEEE Sensors Applications Symposium (SAS), pages 1–6, 2016. (Cited in page 133.)

[Cross 2018] Cross, B. *Chapter III-1-C - PV System Monitoring.* In Kalogirou, S. A., editor, McEvoy's Handbook of Photovoltaics (Third Edition), pages 1183–1191. Academic Press, third edition edition, 2018. (Cited in page 132.)

[Cubukcu 2020] Cubukcu, M. and Akanalci, A. *Real-time inspection and determination methods of faults on photovoltaic power systems by thermal imaging in Turkey.* Renewable Energy, vol. 147, pages 1231–1238, 2020. (Cited in pages 83 and 84.)

[Cueto 2010] Cueto, J. and Rummel, S. *Degradation of Photovoltaic Modules Under High Voltage Stress in the Field: Preprint.* Proceedings of SPIE - The International Society for Optical Engineering, vol. 7773, 08 2010. [Online]. Available: http://dx.doi.org/10.1117/12.861226. (Cited in page 235.)

[Cusido 2008] Cusido, J., Romeral, L., Ortega, J. A., Rosero, J. A. and GarcÍa-Garcia Espinosa, A. *Fault Detection in Induction Machines Using Power Spectral Density in Wavelet Decomposition.* IEEE Transactions on Industrial Electronics, vol. 55, no. 2, pages 633–643, 2008. [Online]. Available: http://dx.doi.org/10.1109/TIE.2007.911960. (Cited in page 144.)

[d. J. Rangel-Magdaleno 2009] d. J. Rangel-Magdaleno, J., d. J. Romero-Troncoso, R., Osornio-Rios, R. A., Cabal-Yepez, E. and Contreras-Medina, L. M. *Novel Methodology for Online Half-Broken-Bar Detection on Induction Motors.* IEEE Transactions on Instrumentation and Measurement, vol. 58, no. 5, pages 1690–1698, 2009. [Online]. Available: http://dx.doi.org/10.1109/TIM.2009.2012932. (Cited in page 144.)

[Dadhich 2019] Dadhich, K., Kurukuru, V. S. B., Khan, M. A. and Haque, A. *Fault Identification Algorithm for Grid Connected Photovoltaic Systems using Machine Learning Techniques.* In International Conference on Power Electronics, Control and Automation (ICPECA), pages 1–6, 2019. (Cited in pages 23, 183, 184, 186, 228, and 232.)

[Dai 2006] Dai, J., Lieu, L. H. and Rocke, D. M. *Dimension Reduction for Classification with Gene Expression Microarray Data.* Statistical Applications in Genetics and Molecular Biology, vol. 5, 2006. (Cited in page 223.)

[Dai 2021] Dai, S., Wang, D., Li, W., Zhou, Q., Tian, G. and Dong, H. *Fault Diagnosis of Data-Driven Photovoltaic Power Generation System Based on Deep*

Reinforcement Learning. Mathematical Problems in Engineering, vol. 2021, pages 1–10, 11 2021. (Cited in page 122.)

[Dalal 2005] Dalal, N. and Triggs, B. *Histograms of oriented gradients for human detection*. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 886–893 vol. 1, 2005. (Cited in page 119.)

[Daliento 2016] Daliento, S., Di Napoli, F., Guerriero, P. and d'Alessandro, V. *A modified bypass circuit for improved hot spot reliability of solar panels subject to partial shading*. Solar Energy, vol. 134, pages 211–218, 2016. (Cited in page 87.)

[Data 2019] Data, G. Global solar photovoltaic capacity expected to exceed 1,500gw by 2030, says globaldata. 2019. (Cited in pages 1 and 33.)

[Daubechies 1990] Daubechies, I. *The wavelet transform, time-frequency localization and signal analysis*. IEEE Transactions on Information Theory, vol. 36, no. 5, pages 961–1005, 1990. (Cited in page 183.)

[Daud 2010] Daud, A., Li, J., Zhou, L. and Muhammad, F. *Latent Dirichlet allocation (LDA) and topic modeling: models, applications, future challenges, a survey*. Frontiers of Computer Science in China, vol. 4, 06 2010. (Cited in page 99.)

[Davidson 2001] Davidson, G., Wylie, B. and Boyack, K. *Cluster stability and the use of noise in interpretation of clustering*. In IEEE Symposium on Information Visualization, 2001. INFOVIS 2001., pages 23–30, 2001. (Cited in page 97.)

[De Nooy 2018] De Nooy, W., Mrvar, A. and Batagelj, V. Exploratory social network analysis with pajek. Structural Analysis in the Social Sciences. Cambridge University Press, 3 edition, 2018. (Cited in page 97.)

[de Oliveira 2018] de Oliveira, M. C. C., Diniz Cardoso, A. S. A., Viana, M. M. and de Freitas Cunha Lins, V. *The causes and effects of degradation of encapsulant ethylene vinyl acetate copolymer (EVA) in crystalline silicon photovoltaic modules: A review*. Renewable and Sustainable Energy Reviews, vol. 81, pages 2299–2317, 2018. (Cited in pages vi and 64.)

[Delgosha 2021] Delgosha, M. S., Hajiheydari, N. and Talafidaryani, M. *Discovering IoT implications in business and management: A computational thematic analysis*. Technovation, page 102236, 2021. (Cited in pages 98 and 100.)

[Demant 2014] Demant, M., Oswald, M., Welschehold, T., Nold, S., Bartsch, S., Schoenfelder, S. and Rein, S. *Micro-Cracks in Silicon Wafers and Solar Cells: Detection and Rating of Mechanical Strength and Electrical Quality*. 09 2014. (Cited in page 119.)

[Demant 2016] Demant, M., Welschehold, T., Oswald, M., Bartsch, S., Brox, T., Schoenfelder, S. and Rein, S. *Microcracks in Silicon Wafers I: Inline Detection and Implications of Crack Morphology on Wafer Strength.* IEEE Journal of Photovoltaics, vol. 6, no. 1, pages 126–135, 2016. [Online]. Available: http://dx.doi.org/10.1109/JPHOTOV.2015.2494692. (Cited in page 119.)

[Demir 2022] Demir, F. *14 - Deep autoencoder-based automated brain tumor detection from MRI data.* In Bajaj, V. and Sinha, G., editors, Artificial Intelligence-Based Brain-Computer Interface, pages 317–351. Academic Press, 2022. (Cited in page 260.)

[Demirhan 2018] Demirhan, H. and Renwick, Z. *Missing value imputation for short to mid-term horizontal solar irradiance data.* Applied Energy, vol. 225, pages 998–1012, 2018. (Cited in page 141.)

[Demirtas 2008] Demirtas, M., Sefa, I., Irmak, E. and Colak, I. *Low-cost and high sensitive microcontroller based data acquisition system for renewable energy sources.* In 2008 International Symposium on Power Electronics, Electrical Drives, Automation and Motion, pages 196–199, 2008. (Cited in pages 145 and 146.)

[Deshkar 2015] Deshkar, S. N., Dhale, S. B., Mukherjee, J. S., Babu, T. S. and Rajasekar, N. *Solar PV array reconfiguration under partial shading conditions for maximum power extraction using genetic algorithm.* Renewable and Sustainable Energy Reviews, vol. 43, pages 102–110, 2015. (Cited in page 68.)

[Devaraju 2015] Devaraju, J., Suhas, K., Mohana, H. and Patil, V. A. *Wireless Portable Microcontroller based Weather Monitoring Station.* Measurement, vol. 76, pages 189–200, 2015. (Cited in page 147.)

[Dey 2019] Dey, S., Roy, S. S., Samanta, K., Modak, S. and Chatterjee, S. *Autocorrelation Based Feature Extraction for Bearing Fault Detection in Induction Motors.* In 2019 International Conference on Electrical, Electronics and Computer Engineering (UPCON), pages 1–5, 2019. (Cited in page 144.)

[Dhanraj 2021] Dhanraj, J. A., Mostafaeipour, A., Velmurugan, K., Techato, K., Chaurasiya, P. K., Solomon, J. M., Gopalan, A. and Phoungthong, K. *An Effective Evaluation on Fault Detection in Solar Panels.* Energies, vol. 14, no. 22, 2021. (Cited in page 67.)

[Dhere 2012] Dhere, N. G. and Shiradkar, N. S. *Fire hazard and other safety concerns of photovoltaic systems.* Journal of Photonics for Energy, vol. 2, no. 1, pages 1 – 14, 2012. (Cited in page 29.)

[Dhibi 2020] Dhibi, K., Fezai, R., Mansouri, M., Trabelsi, M., Kouadri, A., Bouzara, K., Nounou, H. and Nounou, M. *Reduced Kernel Random Forest*

Technique for Fault Detection and Classification in Grid-Tied PV Systems. IEEE Journal of Photovoltaics, vol. 10, no. 6, pages 1864–1871, 2020. (Cited in pages 113 and 127.)

[Dhibi 2021] Dhibi, K., Mansouri, M., Bouzrara, K., Nounou, H. and Nounou, M. *An Enhanced Ensemble Learning-Based Fault Detection and Diagnosis for Grid-Connected PV Systems.* IEEE Access, vol. 9, pages 155622–155633, 2021. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2021.3128749. (Cited in pages 115 and 202.)

[Dhimish 2016] Dhimish, M. and Holmes, V. *Fault detection algorithm for grid-connected photovoltaic plants.* Solar Energy, vol. 137, pages 236–245, 2016. (Cited in page 88.)

[Dhimish 2018a] Dhimish, M. Fault detection and performance analysis of photovoltaic installations. March 2018. (Cited in pages 3 and 34.)

[Dhimish 2018b] Dhimish, M., Holmes, V., Mehrdadi, B. and Dales, M. *Comparing Mamdani Sugeno fuzzy logic and RBF ANN network for PV fault detection.* Renewable Energy, vol. 117, pages 257–274, 2018. (Cited in page 125.)

[Dhoke 2019] Dhoke, A., Sharma, R. and Saha, T. K. *An approach for fault detection and location in solar PV systems.* Solar Energy, vol. 194, pages 197–208, 2019. (Cited in page 240.)

[Dierauf 2013] Dierauf, T., Growitz, A., Kurtz, S., Becerra Cruz, J. L., Riley, E. and Hansen, E. *Weather-Corrected Performance Ratio.* Technical Report, National Renewable Energy Lab.(NREL), Golden, CO, USA, 2013. (Cited in page 234.)

[Dietrich 2008] Dietrich, S., Pander, M., Ebert, M. and Bagdahn, J. *Mechanical Assessment of Large Photovoltaic Modules by Test and Finite Element Analysis.* 09 2008. (Cited in page 56.)

[Dirnberger 2015] Dirnberger, D., Blackburn, G., MÃ¼ller, B. and Reise, C. *On the impact of solar spectral irradiance on the yield of different PV technologies.* Solar Energy Materials and Solar Cells, vol. 132, pages 431–442, 2015. (Cited in pages v, 47, and 48.)

[Domingos 2015] Domingos, P. The master algorithm: How the quest for the ultimate learning machine will remake our world. The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books, New York, NY, US, 2015. Pages: xxi, 329. (Cited in pages 8, 106, 108, 109, and 247.)

[Dominković 2022] Dominković, D., Weinand, J., Scheller, F., D'Andrea, M. and McKenna, R. *Reviewing two decades of energy system analysis with bibliometrics.* Renewable and Sustainable Energy Reviews, vol. 153, page 111749, 2022. (Cited in page 96.)

[Dong 2012] Dong, B., Xu, G., Luo, X., Cai, Y. and Gao, W. *A bibliometric analysis of solar power research from 1991 to 2010.* Scientometrics, vol. 93, no. 3, pages 1101–1117, 2012. (Cited in page 96.)

[Donoho 2003] Donoho, D. L. and Grimes, C. *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data.* Proceedings of the National Academy of Sciences, vol. 100, no. 10, pages 5591–5596, 2003. (Cited in page 192.)

[Dorise 2022] Dorise, A., Travé-Massuyès, L., Subias, A. and Alonso, C. *Dyd^2: Dynamic Double anomaly Detection.* In IFAC Safeprocess 2022 :11th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes, Pafos, Cyprus, June 2022. IFAC. (Cited in page 127.)

[Dou 2013] Dou, L., You, J., Hong, Z., Xu, Z., Li, G., Street, R. A. and Yang, Y. *25th anniversary article: a decade of organic/polymeric photovoltaic research.* Advanced materials (Deerfield Beach, Fla.), vol. 25, no. 46, page 6642â€"6671, December 2013. (Cited in page 46.)

[Dreidy 2013] Dreidy, M., Alsayid, B., Alsadi, S. and Jallad, J. *Partial Shading of PV System Simulation with Experimental Results.* Smart Grid and Renewable Energy, vol. 04, 09 2013. [Online]. Available: http://dx.doi.org/10.4236/sgre.2013.46049. (Cited in page 66.)

[Drews 2004] Drews, A., Betcke, J., Lorenz, E., Heinemann, D., Toggweiler, P., Stettler, S., Rasmussen, J., van Sark, W., Heilscher, G., Schneider, M., Wiemken, E., Heydenreich, W. and Beyer, H. G. *Intelligent performance check of PV system operation based on satellite data.* 06 2004. (Cited in page 138.)

[Drews 2007] Drews, A., de Keizer, A., Beyer, H., Lorenz, E., Betcke, J., van Sark, W., Heydenreich, W., Wiemken, E., Stettler, S., Toggweiler, P., Bofinger, S., Schneider, M., Heilscher, G. and Heinemann, D. *Monitoring and remote failure detection of grid-connected PV systems based on satellite observations.* Solar Energy, vol. 81, no. 4, pages 548–564, 2007. (Cited in pages 88 and 138.)

[Dross 2017] Dross, F. and Meydbray, J. *PV Module Reliability Scorecard Report.* Technical Report, DNV.GL, 2017. (Cited in pages v, 58, and 62.)

[Du 2014] Du, H., Li, N., Brown, M. A., Peng, Y. and Shuai, Y. *A bibliographic analysis of recent solar energy literatures: The expansion and evolution of*

a research field. Renewable Energy, vol. 66, no. C, pages 696–706, 2014. (Cited in page 96.)

[Duerr 2016] Duerr, I., Bierbaum, J., Metzger, J., Richter, J. and Philipp, D. *Silver Grid Finger Corrosion on Snail Track affected PV Modules – Investigation on Degradation Products and Mechanisms.* Energy Procedia, vol. 98, pages 74–85, 2016. (Cited in page 180.)

[Duhamel 1990] Duhamel, P. and Vetterli, M. *Fast fourier transforms: A tutorial review and a state of the art.* Signal Processing, vol. 19, no. 4, pages 259 – 299, 1990. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/016516849090158U`. (Cited in page 142.)

[Duman 2021] Duman, S., Li, J., Wu, L. and Yorukeren, N. *Symbiotic Organisms Search Algorithm-Based Security-Constrained AC–DC OPF Regarding Uncertainty of Wind, PV and PEV Systems.* Soft Comput., vol. 25, no. 14, page 9389–9426, jul 2021. (Cited in page 71.)

[Dumas 1982] Dumas, L. N. and Shumka, A. *Photovoltaic Module Reliability Improvement through Application Testing and Failure Analysis.* IEEE Transactions on Reliability, vol. R-31, no. 3, pages 228–234, 1982. [Online]. Available: http://dx.doi.org/10.1109/TR.1982.5221325. (Cited in page 62.)

[Dunlop 2006] Dunlop, E. D. and Halton, D. *The performance of crystalline silicon photovoltaic solar modules after 22 years of continuous outdoor exposure.* Progress in Photovoltaics: Research and Applications, vol. 14, no. 1, pages 53–64, 2006. (Cited in page 62.)

[Dunn 2012] Dunn, L., Gostein, M. and Emery, K. *Comparison of pyranometers vs. PV reference cells for evaluation of PV array performance.* In 2012 38th IEEE Photovoltaic Specialists Conference, pages 002899–002904, 2012. (Cited in page 135.)

[Díaz-Dorado 2010] Díaz-Dorado, E., Suárez-García, A., Carrillo, C. and Cidrás, J. *Influence of the shadows in photovoltaic systems with different configurations of bypass diodes.* In SPEEDAM 2010, pages 134–139, 2010. (Cited in page 51.)

[Díaz 2007] Díaz, P., Egido, M. and Nieuwenhout, F. *Dependability analysis of stand-alone photovoltaic systems.* Progress in Photovoltaics: Research and Applications, vol. 15, no. 3, pages 245–264, 2007. (Cited in pages 3 and 35.)

[Ebeid 2016] Ebeid, I. and Arango, J. *Mallet vs GenSim: Topic Modeling Evaluation Report*, 04 2016. (Cited in page 101.)

[Ebner 2010] Ebner, R., Zamini, S. and ÃšjvÃ¡ri, G. *Defect Analysis in Different Photovoltaic Modules Using Electroluminescence (EL) and Infrared (IR)-Thermography.* pages 333–336, 01 2010. (Cited in page 85.)

[Eck 2009] Eck, N. J. v. and Waltman, L. *How to normalize cooccurrence data? An analysis of some well-known similarity measures.* Journal of the American Society for Information Science and Technology, vol. 60, no. 8, pages 1635–1651, 2009. (Cited in page 98.)

[Eder 2018] Eder, G. C., Voronko, Y., Hirschl, C., Ebner, R., ÃšjvÃ¡ri, G. and MÃ¼hleisen, W. *Non-Destructive Failure Detection and Visualization of Artificially and Naturally Aged PV Modules.* Energies, vol. 11, no. 5, 2018. (Cited in page 238.)

[Eke 2012] Eke, R., Sertap Kavasoglu, A. and Kavasoglu, N. *Design and implementation of a low-cost multi-channel temperature measurement system for photovoltaic modules.* Measurement, vol. 45, no. 6, pages 1499–1509, 2012. (Cited in page 146.)

[El Basri 2015] El Basri, Y., Bressan, M., Seguier, L., Alawadhi, H. and Alonso, C. *A proposed graphical electrical signatures supervision method to study PV module failures.* Solar Energy, vol. 116, pages 247–256, 2015. (Cited in page 87.)

[El Bouchikhi 2013] El Bouchikhi, E. H., Choqueuse, V. and Benbouzid, M. E. H. *A parametric spectral estimator for faults detection in induction machines.* In IECON 2013 - 39th Annual Conference of the IEEE Industrial Electronics Society, pages 7358–7363, 2013. (Cited in page 144.)

[Eler 2015] Eler, D. M., Batista Martins Teixeira, J., Macanha, P. A. and Garcia, R. E. *Simplified Stress and Simplified Silhouette Coefficient to a Faster Quality Evaluation of Multidimensional Projection Techniques and Feature Spaces.* In 2015 19th International Conference on Information Visualisation, pages 133–139, 2015. (Cited in page 105.)

[energy agency 2016] energy agency, I. Snapshot of global photovoltaic markets. 2016. (Cited in pages 1 and 33.)

[EPIA 2011] EPIA. *Solar generation 6,.* Technical Report, European Photovoltaic Industry Association, 2011. (Cited in pages v and 46.)

[Eskandari 2020a] Eskandari, A., Milimonfared, J. and Aghaei, M. *Line-line fault detection and classification for photovoltaic systems using ensemble learning model based on I-V characteristics.* Solar Energy, vol. 211, pages 354–365, 2020. (Cited in page 115.)

[Eskandari 2020b] Eskandari, A., Milimonfared, J., Aghaei, M. and Reinders, A. H. *Autonomous Monitoring of Line-to-Line Faults in Photovoltaic Systems by Feature Selection and Parameter Optimization of Support Vector Machine Using Genetic Algorithms.* Applied Sciences, vol. 10, no. 16, 2020. [Online]. Available: http://dx.doi.org/10.3390/app10165527. (Cited in page 119.)

[Eskandari 2021] Eskandari, A., Milimonfared, J. and Aghaei, M. *Fault Detection and Classification for Photovoltaic Systems Based on Hierarchical Classification and Machine Learning Technique.* IEEE Transactions on Industrial Electronics, vol. 68, no. 12, pages 12750–12759, 2021. [Online]. Available: http://dx.doi.org/10.1109/TIE.2020.3047066. (Cited in pages 112 and 116.)

[Esmael 2012] Esmael, B., Arnaout, A., Fruhwirth, R. and Thonhauser, G. *A Statistical Feature-Based Approach for Operations Recognition in Drilling Time Series.* International Journal of Computer Information Systems and Industrial Management Applications, vol. 4, no. 6, pages 100–108, 2012. (Cited in pages 186, 187, 188, 189, 191, and 199.)

[Etemadi 2008] Etemadi, A. and Sanaye-Pasand, M. *High-impedance fault detection using multi-resolution signal decomposition and adaptive neural fuzzy inference system.* IET Generation, Transmission and Distribution, vol. 2, no. 1, page 110 – 118, 2008. (Cited in page 184.)

[Europe 2016] Europe, S. *Analytical monitoring of grid-connected photovoltaic systems.* Technical Report, SolarPower Europe, 2016. (Cited in page 140.)

[Fadhel 2018] Fadhel, S., Migan, A., Delpha, C., Diallo, D., Bahri, I., Trabelsi, M. and Mimouni, M. *Data-driven approach for isolated PV shading fault diagnosis based on experimental I-V curves analysis.* pages 927–932, 2018. (Cited in pages 13, 29, 52, 56, 77, 90, 181, and 224.)

[Fadhel 2019a] Fadhel, S., Delpha, C., Diallo, D., Bahri, I., Migan, A., Trabelsi, M. and Mimouni, M. *PV shading fault detection and classification based on I-V curve using principal component analysis: Application to isolated PV system.* Solar Energy, vol. 179, pages 1–10, 2019. (Cited in pages 87, 192, and 193.)

[Fadhel 2019b] Fadhel, S. *Efficacité énergétique et surveillance d'un microgrid à courant continu alimenté par des panneaux photovoltaïques.* PhD thesis, Université Paris-Sud, 2019. (Cited in pages 83 and 85.)

[Falvo 2015] Falvo, M. and Capparella, S. *Safety issues in PV systems: Design choices for a secure fault detection and for preventing fire risk.* Case Studies in Fire Safety, vol. 3, pages 1–16, 2015. (Cited in pages 56 and 69.)

[Famili 1997] Famili, A., Shen, W.-M., Weber, R. and Simoudis, E. *Data preprocessing and intelligent data analysis.* Intelligent Data Analysis, vol. 1, no. 1, pages 3–23, 1997. (Cited in page 141.)

[Fan 2019] Fan, L., Zhang, F., Fan, H. and Zhang, C. *Brief review of image denoising techniques.* Visual Computing for Industry, Biomedicine, and Art, vol. 2, 12 2019. [Online]. Available: http://dx.doi.org/10.1186/s42492-019-0016-7. (Cited in page 141.)

[Fatama 2019] Fatama, A.-Z., Haque, A. and Khan, M. A. *A Multi Feature Based Islanding Classification Technique for Distributed Generation Systems.* In International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pages 160–166, 2019. (Cited in page 183.)

[Faziludeen 2013] Faziludeen, S. and Sabiq, P. V. *ECG beat classification using wavelets and SVM.* In 2013 IEEE Conference on Information Communication Technologies, pages 815–818, 2013. (Cited in page 143.)

[Feedgy 2022] Feedgy. *feedgy analytics.* 2022. [Online]. Available: `https://www.feedgy.solar/`. (Cited in pages 4 and 35.)

[Felder 2017] Felder, T., Hu, H., Gambogi, W., Choudhury, K. R., MacMaster, S., Lles, L. G. and Trout, T. J. *Field study and analysis of backsheet degradation in 450MW+ PV installations.* Technical Report, International Energy Agency IEA, 2017. (Cited in page 233.)

[Feldman 2022] Feldman, D., Dummit, K., Zuboy, J., Heeter, K., J. Xu and Margolis, R. *Winter 2021/2022 Solar Industry Update.* Technical Report, National Renewable Energy Lab.(NREL), Golden, CO, USA, 2022. (Cited in pages 1 and 33.)

[Ferencz 2018] Ferencz, K. and Domokos, J. *IoT Sensor Data Acquisition and Storage System Using Raspberry Pi and Apache Cassandra.* In 2018 International IEEE Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE), pages 000143–000146, 2018. (Cited in page 164.)

[Ferri 2009] Ferri, C., Hernández-Orallo, J. and Modroiu, R. *An experimental comparison of performance measures for classification.* Pattern Recognition Letters, vol. 30, no. 1, pages 27–38, 2009. (Cited in page 260.)

[Fezzani 2015] Fezzani, A., Mahammed, I. H., Drid, S. and Chrifi-alaoui, L. *Modeling and analysis of the photovoltaic array faults.* In 2015 3rd International Conference on Control, Engineering & Information Technology (CEIT), pages 1–9, 2015. (Cited in page 87.)

[Firth 2010] Firth, S., Lomas, K. and Rees, S. J. *A simple model of PV system performance and its use in fault detection.* Sol Energy, vol. 84, no. 4, page 624–635, 2010. (Cited in page 227.)

[Fitrianto 2019a] Fitrianto, M. I., Wahjono, E., Anggriawan, D. O., Prasetyono, E., Mubarok, R. H. and Tjahjono, A. *Identification and Protection of Series DC Arc Fault for Photovoltaic Systems Based on Fast Fourier Transform.* In 2019 International Electronics Symposium (IES), pages 159–163, 2019. (Cited in page 142.)

[Fitrianto 2019b] Fitrianto, M. I., Wahjono, E., Anggriawan, D. O., Prasetyono, E., Mubarok, R. H. and Tjahjono, A. *Identification and Protection of Series*

DC Arc Fault for Photovoltaic Systems Based on Fast Fourier Transform. In 2019 International Electronics Symposium (IES), pages 159–163, 2019. (Cited in page 142.)

[Flicker 2013] Flicker, J. and Johnson, J. *Photovoltaic ground fault and blind spot electrical simulations.* Technical Report, Sandia Corporation, 2013. (Cited in page 89.)

[Flicker 2015] Flicker, J., Johnson, J., Albers, M. and Ball, G. *Recommendations for isolation monitor ground fault detectors on residential and utility-scale PV systems.* In 2015 IEEE 42nd Photovoltaic Specialist Conference (PVSC), pages 1–6, 2015. (Cited in page 69.)

[Flicker 2016] Flicker, J. and Johnson, J. *Photovoltaic ground fault detection recommendations for array safety and operation.* Solar Energy, vol. 140, pages 34–50, 12 2016. [Online]. Available: http://dx.doi.org/10.1016/j.solener.2016.10.017. (Cited in pages vi and 72.)

[Foltz 1996] Foltz, P. *Latent Semantic Analysis for Text-Based Research.* Behavior Research Methods, vol. 28, pages 197–202, 02 1996. (Cited in page 99.)

[Fong 2015] Fong, S. *Using Hierarchical Time Series Clustering Algorithm and Wavelet Classifier for Biometric Voice Classification.* Journal of Biomedicine and Biotechnology, vol. 1, pages 1 – 12, 2015. [Online]. Available: http://dx.doi.org/https://doi.org/10.1155/2012/215019. (Cited in page 143.)

[Forero 2006] Forero, N., HernÃ¡ndez, J. and Gordillo, G. *Development of a monitoring system for a PV solar plant.* Energy Conversion and Management, vol. 47, no. 15, pages 2329–2336, 2006. (Cited in page 146.)

[Forman 1982] Forman, S. *Performance of Experimental Terrestrial Photovoltaic Modules.* IEEE Transactions on Reliability, vol. R-31, no. 3, pages 235–245, 1982. [Online]. Available: http://dx.doi.org/10.1109/TR.1982.5221326. (Cited in page 69.)

[Friesen 2018] Friesen, G., Herrmann, W., Belluardo, G. and Herteleer, B. *Photovoltaic module energy yield measurements.* Technical Report, International Energy Agency IEA, 2018. (Cited in pages 134 and 135.)

[Fruchterman 1991] Fruchterman, T. M. J. and Reingold, E. M. *Graph drawing by force-directed placement.* Software: Practice and Experience, vol. 21, no. 11, pages 1129–1164, 1991. (Cited in page 97.)

[Fucheng 2015] Fucheng, S., Wenyuan, S. and Hailong, S. *Gearbox fault diagnosis based on autocorrelation and HHT.* Vibroengineering PROCEDIA, vol. 5, no. 1, pages 270 –274, 2015. (Cited in page 144.)

[Fuentes 2014] Fuentes, M., Vivar, M., Burgos, J., Aguilera, J. and Vacas, J. *Design of an accurate, low-cost autonomous data logger for PV system monitoring using Arduino™ that complies with IEC standards.* Solar Energy Materials and Solar Cells, vol. 130, pages 529–543, 2014. (Cited in pages 18, 137, 156, and 157.)

[Gallardo-Saavedra 2019] Gallardo-Saavedra, S., Hernández-Callejo, L. and Duque-Pérez, O. *Quantitative failure rates and modes analysis in photovoltaic plants.* Energy, vol. 183, pages 825–836, 2019. (Cited in page 56.)

[García-Gutiérrez 2019] García-Gutiérrez, L. A. *Développement d'un contrôle actif tolérant aux défaillances appliqué aux systèmes PV.* PhD thesis, 2019. (Cited in page 56.)

[Garg 1993] Garg, K. C., Sharma, P. and Sharma, L. *Bradford's law in relation to the evolution of a field. A case study of solar power research.* Scientometrics, vol. 27, no. 2, pages 145–156, 1993. (Cited in page 96.)

[Garoudja 2017a] Garoudja, E., Chouder, A., Kara, K. and Silvestre, S. *An enhanced machine learning based approach for failures detection and diagnosis of PV systems.* Energy Conversion and Management, vol. 151, pages 496–513, 2017. (Cited in pages 23 and 227.)

[Garoudja 2017b] Garoudja, E., Chouder, A., Kara, K. and Silvestre, S. *An enhanced machine learning based approach for failures detection and diagnosis of PV systems.* Energy Conversion and Management, vol. 151, pages 496–513, 2017. (Cited in page 114.)

[Garrido-Alzar 1997] Garrido-Alzar, C. *Algorithm for extraction of solar cell parameters from I–V curve using double exponential model.* Renewable Energy, vol. 10, no. 2, pages 125–128, 1997. (Cited in page 144.)

[Gethers 2010] Gethers, M. and Poshyvanyk, D. *Using Relational Topic Models to capture coupling among classes in object-oriented software systems.* In 2010 IEEE International Conference on Software Maintenance, pages 1–10, 2010. (Cited in page 99.)

[Gokmen 2012] Gokmen, N., Karatepe, E., Celik, B. and Silvestre, S. *Simple diagnostic approach for determining of faulted PV modules in string based PV arrays.* Solar Energy, vol. 86, no. 11, pages 3364–3377, 2012. (Cited in page 67.)

[Goodfellow 2014] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. *Generative Adversarial Networks*, 2014. (Cited in page 238.)

[Goyal 2020] Goyal, D., Choudhary, A., Pabla, B. and Dhami, S. *Support vector machines based non-contact fault diagnosis system for bearings.* J Intell Manuf 31, page 1275–1289, 2020. (Cited in page 186.)

[Graps 1995] Graps, A. *An introduction to wavelets.* IEEE Computational Science and Engineering, vol. 2, no. 2, pages 50–61, 1995. [Online]. Available: http://dx.doi.org/10.1109/99.388960. (Cited in page 143.)

[Greco 2020] Greco, A., Pironti, C., Saggese, A., Vento, M. and Vigilante, V. *A deep learning based approach for detecting panels in photovoltaic plants.* pages 1–7, 2020. (Cited in page 117.)

[Green 2022] Green, M. A., Dunlop, E. D., Hohl-Ebinger, J., Yoshita, M., Kopidakis, N., Bothe, K., Hinken, D., Rauer, M. and Hao, X. *Solar cell efficiency tables (Version 60).* Progress in Photovoltaics: Research and Applications, vol. 30, no. 7, pages 687–701, 2022. (Cited in page 47.)

[Greene 2015] Greene, D. and Cross, J. P. *Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis*, 2015. (Cited in page 99.)

[Gritli 2012] Gritli, Y., Di Tommaso, A. O., Filippetti, F., Miceli, R., Rossi, C. and Chatti, A. *Investigation of motor current signature and vibration analysis for diagnosing rotor broken bars in double cage induction motors.* In International Symposium on Power Electronics Power Electronics, Electrical Drives, Automation and Motion, pages 1360–1365, 2012. (Cited in page 144.)

[Gritli 2013] Gritli, Y., Di Tommaso, A. O., Miceli, R., Filippetti, F. and Rossi, C. *Vibration signature analysis for rotor broken bar diagnosis in double cage induction motor drives.* In 4th International Conference on Power Engineering, Energy and Electrical Drives, pages 1814–1820, 2013. (Cited in page 144.)

[Guerriero 2016] Guerriero, P., Cuozzo, G. and Daliento, S. *Health diagnostics of PV panels by means of single cell analysis of thermographic images.* In 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC), pages 1–6, 2016. (Cited in page 84.)

[Guo 2020] Guo, X., Na, Z., Ma, D., Lu, Y. and Luo, X. *Fault diagnosis of photovoltaic system based on machine learning model fusion.* IOP Conference Series: Earth and Environmental Science, vol. 467, no. 1, page 012073, mar 2020. (Cited in page 116.)

[Gupta 2020] Gupta, R., Tanwar, S., Tyagi, S. and Kumar, N. *Machine Learning Models for Secure Data Analytics: A taxonomy and threat model.* Computer Communications, vol. 153, pages 406–440, 2020. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0140366419318493`. (Cited in page 110.)

[Gómez-Luna 2013] Gómez-Luna, E., Silva, D. and Aponte, G. *Selection of a mother wavelet for frequency analysis of transient electrical signals using WPD.* Ingeniare. Revista chilena de ingenierÃa, vol. 21, pages 262 – 270, 08 2013. (Cited in page 143.)

[Hac 2022] *Cooking Hacks by Libelium*, 2022. Accessed: 2022-09-30. (Cited in page 162.)

[Hachana 2016] Hachana, O., Tina, G. M. and Hemsas, K. E. *PV array fault DiagnosticTechnique for BIPV systems.* Energy and Buildings, vol. 126, pages 263–274, 2016. (Cited in page 87.)

[Hadj Arab 1989] Hadj Arab, A. Modélisation et simulation d'un système photovoltaïque de faible puissance. Master's thesis, HCR, 1989. (Cited in pages 50 and 51.)

[Hadke 2021] Hadke, S., Mishra, R. and Werulkar, A. *A Bibliometric analysis of Different Maximum Power Point Tracking Methods for Photovoltaic Systems.* International Journal of Trend in Scientific Research and Development, vol. 5, no. 4, pages 1353–1357, 2021. (Cited in page 96.)

[Haeberlin 2003] Haeberlin, H. *Normalized Representation of Energy and Power for Analysis of Performance and On-line Error Detection in PV-Systems.* 2003. (Cited in page 234.)

[Haeberlin 2007] Haeberlin, H. and Real, M. *Arc Detector for Remote Detection of Dangerous Arcs on the DC Side of PV Plants.* 2007. (Cited in page 142.)

[Haje Obeid 2017] Haje Obeid, N., Battiston, A., Boileau, T. and Nahid-Mobarakeh, B. *Early Intermittent Interturn Fault Detection and Localization for a Permanent Magnet Synchronous Motor of Electrical Vehicles Using Wavelet Transform.* IEEE Transactions on Transportation Electrification, vol. 3, no. 3, pages 694–702, 2017. (Cited in page 183.)

[Hajji 2021] Hajji, M., Harkat, M.-F., Kouadri, A., Abodayeh, K., Mansouri, M., Nounou, H. and Nounou, M. *Multivariate feature extraction based supervised machine learning for fault detection and diagnosis in photovoltaic systems.* European Journal of Control, vol. 59, pages 313–321, 2021. (Cited in pages 113 and 191.)

[Han 2014] Han, J., Choi, C.-s., Park, W.-k., Lee, I. and Kim, S.-h. *PLC-based photovoltaic system management for smart home energy management system.* IEEE Transactions on Consumer Electronics, vol. 60, no. 2, pages 184–189, 2014. (Cited in page 137.)

[Han 2019] Han, F., Chen, Z., Wu, L., Long, C., Yu, J., Lin, P. and Cheng, S. *An intelligent fault diagnosis method for PV arrays based on an improved rotation forest algorithm.* Energy Procedia, vol. 158, pages 6132–6138, 2019.

[Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876610219305211. Innovative Solutions for Energy Transitions. (Cited in page 113.)

[Haque 2019] Haque, A., Bharath, K. V. S., Khan, M. A., Khan, I. and Jaffery, Z. A. *Fault diagnosis of Photovoltaic Modules.* Energy Science & Engineering, vol. 7, no. 3, pages 622–644, 2019. (Cited in pages 30, 183, 184, 186, 191, and 232.)

[Hare 2016] Hare, J., Shi, X., Gupta, S. and Bazzi, A. *Fault diagnostics in smart micro-grids: A survey.* Renewable and Sustainable Energy Reviews, vol. 60, pages 1114–1124, 2016. (Cited in page 180.)

[Hariharan 2016a] Hariharan, R., Chakkarapani, M. and Ilango, G. S. *Challenges in the detection of line-line faults in PV arrays due to partial shading.* In 2016 International Conference on Energy Efficient Technologies for Sustainability (ICEETS), pages 23–27, 2016. (Cited in pages 2, 13, 29, 34, and 90.)

[Hariharan 2016b] Hariharan, R., Chakkarapani, M. and Ilango, G. S. *Challenges in the detection of line-line faults in PV arrays due to partial shading.* In 2016 International Conference on Energy Efficient Technologies for Sustainability (ICEETS), pages 23–27, 2016. (Cited in page 29.)

[Harrou 2019a] Harrou, F., Taghezouit, B. and Sun, Y. *Improved k NN-Based Monitoring Schemes for Detecting Faults in PV Systems.* IEEE Journal of Photovoltaics, vol. PP, pages 1–11, 03 2019. [Online]. Available: http://dx.doi.org/10.1109/JPHOTOV.2019.2896652. (Cited in page 120.)

[Harrou 2019b] Harrou, F., Taghezouit, B. and Sun, Y. *Robust and flexible strategy for fault detection in grid-connected photovoltaic systems.* Energy Conversion and Management, vol. 180, pages 1153–1166, 2019. (Cited in page 183.)

[Harrou 2021] Harrou, F., Saidi, A., Sun, Y. and Khadraoui, S. *Monitoring of Photovoltaic Systems Using Improved Kernel-Based Learning Schemes.* IEEE Journal of Photovoltaics, vol. 11, no. 3, pages 806–818, 2021. (Cited in page 119.)

[Haykin 1999] Haykin, S. Neural networks: A comprehensive foundation. Prentice Hall, 1999. (Cited in page 113.)

[He 2021] He, W., Yin, D., Zhang, K., Zhang, X. and Zheng, J. *Fault Detection and Diagnosis Method of Distributed Photovoltaic Array Based on Fine-Tuning Naive Bayesian Model.* Energies, vol. 14, no. 14, pages 1–17, 2021. (Cited in pages 112 and 180.)

[Heidarbeigi 2008] Heidarbeigi, K., Ahmadi, H. and Omid, M. *Fault diagnosis of Massey Ferguson gearbox using Power Spectral Density.* In 2008 18th

International Conference on Electrical Machines, pages 1–4, 2008. (Cited in page 144.)

[Heil 1989] Heil, C. E. and Walnut, D. F. *Continuous and Discrete Wavelet Transforms.* SIAM Review, vol. 31, no. 4, pages 628–666, 1989. (Cited in page 143.)

[Heilscher 2020] Heilscher, G., Reindl, T., Zhan, Y. and Idlbi, B. *Communication and control for high PV penetration under smart grid environment.* Technical Report, International Energy Agency IEA, 2020. (Cited in page 137.)

[Heinrich 2020] Heinrich, M., Meunier, S., Samé, A., Quéval, L., Darga, A., Oukhellou, L. and Multon, B. *Detection of cleaning interventions on photovoltaic modules with machine learning.* Applied Energy, vol. 263, page 114642, 2020. (Cited in page 260.)

[Hernández 2009] Hernández, J. C. and Vidal, P. G. *Guidelines for Protection Against Electric Shock in PV Generators.* IEEE Transactions on Energy Conversion, vol. 24, no. 1, pages 274–282, 2009. [Online]. Available: http://dx.doi.org/10.1109/TEC.2008.2008865. (Cited in page 89.)

[Herteleer 2017] Herteleer, B., Huyck, B., Catthoor, F., Driesen, J. and Cappelle, J. *Normalised efficiency of photovoltaic systems: Going beyond the performance ratio.* Solar Energy, vol. 157, pages 408–418, 2017. (Cited in pages 234, 235, and 240.)

[Hinton 2002] Hinton, G. E. and Roweis, S. *Stochastic Neighbor Embedding.* In Becker, S., Thrun, S. and Obermayer, K., editors, Advances in Neural Information Processing Systems, volume 15. MIT Press, 2002. (Cited in page 103.)

[Hinton 2006] Hinton, G. E., Osindero, S. and Teh, Y.-W. *A Fast Learning Algorithm for Deep Belief Nets.* Neural Comput., vol. 18, no. 7, page 1527–1554, jul 2006. (Cited in page 116.)

[Ho 1995] Ho, T. K. *Random decision forests.* In Proceedings of 3rd International Conference on Document Analysis and Recognition, volume 1, pages 278–282 vol.1, 1995. (Cited in page 260.)

[Ho 1998] Ho, T. K. *The random subspace method for constructing decision forests.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pages 832–844, 1998. [Online]. Available: http://dx.doi.org/10.1109/34.709601. (Cited in page 259.)

[Hofmann 1999] Hofmann, T. *Probabilistic Latent Semantic Indexing.* In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, page 50–57, New

York, NY, USA, 1999. Association for Computing Machinery. (Cited in page 99.)

[Hogarth 1977] Hogarth, R. M. Methods for aggregating opinions, pages 231–255. Springer Netherlands, Dordrecht, 1977. (Cited in page 189.)

[Hong 2022a] Hong, Y.-Y. and Pula, R. A. *Detection and classification of faults in photovoltaic arrays using a 3D convolutional neural network.* Energy, vol. 246, page 123391, 2022. (Cited in page 117.)

[Hong 2022b] Hong, Y.-Y. and Pula, R. A. *Methods of photovoltaic fault detection and classification: A review.* Energy Reports, vol. 8, pages 5898–5929, 2022. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S2352484722008022`. (Cited in pages 84, 85, 86, 87, 88, 92, 132, 139, and 141.)

[Hout 2013] Hout, M. C., Papesh, M. H. and Goldinger, S. D. *Multidimensional scaling.* WIREs Cognitive Science, vol. 4, no. 1, pages 93–103, 2013. (Cited in page 194.)

[Hu 2014] Hu, Y., Cao, W., Ma, J., Finney, S. J. and Li, D. *Identifying PV Module Mismatch Faults by a Thermography-Based Temperature Distribution Analysis.* IEEE Transactions on Device and Materials Reliability, vol. 14, no. 4, pages 951–960, 2014. [Online]. Available: http://dx.doi.org/10.1109/TDMR.2014.2348195. (Cited in page 65.)

[Hu 2015] Hu, Y., Zhang, J., Cao, W., Wu, J., Tian, G. Y., Finney, S. J. and Kirtley, J. L. *Online Two-Section PV Array Fault Diagnosis With Optimized Voltage Sensor Locations.* IEEE Transactions on Industrial Electronics, vol. 62, no. 11, pages 7237–7246, 2015. (Cited in page 87.)

[Hu 2017a] Hu, L., Ye, J., Chang, S., Li, H. and Chen, H. *A Novel Fault Diagnostic Technique for Photovoltaic Systems Based on Cascaded Forest.* In Proceedings of the Workshop on Smart Internet of Things, SmartIoT '17, New York, NY, USA, 2017. Association for Computing Machinery. (Cited in page 260.)

[Hu 2017b] Hu, Y., Gunapati, V. Y., Zhao, P., Gordon, D., Wheeler, N. R., Hossain, M. A., Peshek, T. J., Bruckman, L. S., Zhang, G.-Q. and French, R. H. *A Nonrelational Data Warehouse for the Analysis of Field and Laboratory Data From Multiple Heterogeneous Photovoltaic Test Sites.* IEEE Journal of Photovoltaics, vol. 7, no. 1, pages 230–236, 2017. (Cited in page 128.)

[Hua 2009] Hua, J., Lin, X., Xu, L., Li, J. and Ouyang, M. *Bluetooth wireless monitoring, diagnosis and calibration interface for control system of fuel cell bus in Olympic demonstration.* Journal of Power Sources, vol. 186, no. 2, pages 478–484, January 2009. (Cited in page 138.)

[Huang 2005] Huang, X., Pan, W., Grindle, S., Han, X., Chen, Y., Park, S., Miller, L. and Hall, J. *A comparative study of discriminating human heart failure etiology using gene expression profiles.* BMC Bioinformatics, vol. 6, no. 205, pages 1065–1084, 2005. (Cited in page 223.)

[Huang 2008] Huang, A. *Similarity measures for text document clustering.* Proceedings of the 6th New Zealand Computer Science Research Student Conference, 01 2008. (Cited in page 97.)

[Huang 2016] Huang, X., Ye, Y., Xiong, L., Lau, R. Y. K., Jiang, N. and Wang, S. *Time series k-means: A new k-means type smooth subspace clustering for time series data.* Inf. Sci., vol. 367-368, pages 1–13, 2016. (Cited in page 258.)

[Huang 2018] Huang, C., Wang, L., Yeung, R. S.-C., Zhang, Z., Chung, H. S.-H. and Bensoussan, A. *A Prediction Model-Guided Jaya Algorithm for the PV System Maximum Power Point Tracking.* IEEE Transactions on Sustainable Energy, vol. 9, no. 1, pages 45–55, 2018. (Cited in page 29.)

[Huang 2019] Huang, C., Du, J., Nie, B., Yu, R., Xiong, W. and Zeng, Q. *Feature Selection Method Based on Partial Least Squares and Analysis of Traditional Chinese Medicine Data.* Computational and Mathematical Methods in Medicine, pages 1–12, 2019. (Cited in page 192.)

[Huang 2020] Huang, J.-M., Wai, R.-J. and Yang, G.-J. *Design of Hybrid Artificial Bee Colony Algorithm and Semi-Supervised Extreme Learning Machine for PV Fault Diagnoses by Considering Dust Impact.* IEEE Transactions on Power Electronics, vol. 35, no. 7, pages 7086–7099, 2020. [Online]. Available: http://dx.doi.org/10.1109/TPEL.2019.2956812. (Cited in pages 115 and 240.)

[Hui 2017] Hui, K. H., Ooi, C. S., Lim, M. H., Leong, M. S. and Al-Obaidi, S. M. *An improved wrapper-based feature selection method for machinery fault diagnosis.* PLOS ONE, vol. 12, no. 12, pages 1–10, 2017. (Cited in page 186.)

[Huld 2008] Huld, T., Šúri, M. and Dunlop, E. *Geographical variation of the conversion efficieny of crystalline silicon photovoltaic modules in Europe.* Progress in Photovoltaics: Research and Applications, vol. 16, pages 595 – 607, 07 2008. [Online]. Available: http://dx.doi.org/10.1002/pip.846. (Cited in page 240.)

[Huld 2010] Huld, T., Gottschalg, R., Beyer, H. G. and Topič, M. *Mapping the performance of PV modules, effects of module type and data averaging.* Solar Energy, vol. 84, no. 2, pages 324–338, 2010. (Cited in pages 235 and 240.)

[Huld 2011] Huld, T., Friesen, G., Skoczek, A., Kenny, R. P., Sample, T., Field, M. and Dunlop, E. D. *A power-rating model for crystalline silicon PV modules.*

Solar Energy Materials and Solar Cells, vol. 95, no. 12, pages 3359–3369, 2011. (Cited in pages 235 and 240.)

[Hund 1995] Hund, T. D. and King, D. L. *Analysis techniques used on field degraded photovoltaic modules.* vol. 1, no. 1, 9 1995. (Cited in pages 84 and 85.)

[Hyvärinen 2000] Hyvärinen, A. and Oja, E. *Independent component analysis: algorithms and applications.* Neural Networks, vol. 13, no. 4, pages 411–430, 2000. (Cited in page 192.)

[Ibrahim 2020] Ibrahim, M. S., Dong, W. and Yang, Q. *Machine learning driven smart electric power systems: Current trends and new perspectives.* Applied Energy, vol. 272, page 115237, 2020. (Cited in page 110.)

[IEA 2007a] IEA. *Cost and Performance Trends in grid-connected photovoltaik systems and case studies 2007.* Technical Report, International Energy Agency IEA, 2007. (Cited in pages 42 and 46.)

[IEA 2007b] IEA. *Solar PV roadmap targets.* Technical Report, International Energy Agency IEA, 2007. (Cited in page 21.)

[IEA 2007c] IEA. *Technology Roadmap Solar photovoltaic energy.* Technical Report, International Energy Agency IEA, 2007. (Cited in page 21.)

[IEA 2014] IEA. *Analysis of Long-Term Performance of PV Systems*, 2014. (Cited in page 234.)

[IEC 1990] IEC. *International Electrotechnical Vocabulary Chapter 191.* Technical Report, International Standard IEC 61724, 1990. (Cited in page 55.)

[IEC 1994a] IEC. *Photovoltaic Devices. Requirements for Reference Solar Modules.* Technical Report, International Standard IEC 61724, 1994. (Cited in page 160.)

[IEC 1994b] IEC. *Requirements for reference solar modules.* Technical Report, International Standard IEC 61724, 1994. (Cited in page 135.)

[IEC 1998] IEC. *Photovoltaic System Performance Monitoring—Guidelines for Measurement, Data Exchange and Analysis.* Technical Report, International Standard IEC 61724, 1998. (Cited in pages vii, xi, 3, 17, 35, 133, 134, 135, 136, 137, 139, 140, 141, 155, 157, 158, 160, 164, 166, 172, 173, 174, and 177.)

[IEC 2005a] IEC. *Crystalline Silicon Terrestrial Photovoltaic Modules. Design Qualification and Type Approval.* Technical Report, International Standard IEC 61724, 2005. (Cited in pages 157 and 160.)

[IEC 2005b] IEC. *Photovoltaic solar energy conversion systems.* Technical Report, International Standard IEC 61215, 2005. (Cited in page 83.)

[IEC 2011] IEC. *Photovoltaic devices - Part 5: Determination of the equivalent cell temperature (ECT) of photovoltaic (PV) devices by the opencircuit voltage method.* Technical Report, International Standard IEC 61724, 2011. (Cited in page 136.)

[IEC 2015] IEC. *Requirements for photovoltaic reference devices.* Technical Report, International Standard IEC 61724, 2015. (Cited in page 135.)

[IEC 2016a] *Photovoltaic (PV) arrays - Design requirements," International Standard.* Standard, International Organization for Standardization, 2016. (Cited in page 59.)

[IEC 2016b] *Photovoltaic (PV) module safety qualification - Part 1: Requirements for construction.* Standard, International Organization for Standardization, 2016. (Cited in pages vi, 74, 77, and 78.)

[IEC 2017a] *Requirements for special installations or locations Solar photovoltaic (PV) power supply systems.* Standard, International Organization for Standardization, 2017. (Cited in page 59.)

[IEC 2017b] IEC. *Photovoltaic (PV) systems - Requirements for testing, documentation and maintenance - Part 3: Photovoltaic modules and plants Outdoor infrared thermography.* Technical Report, International Standard IEC 61724, 2017. (Cited in pages 136 and 234.)

[IEEE 2003] IEEE. *IEEE Standard for Interconnecting Distributed Resources with Electric Power Systems.* IEEE Std 1547-2003, pages 1–28, 2003. [Online]. Available: http://dx.doi.org/10.1109/IEEESTD.2003.94285. (Cited in page 132.)

[IFC 2015] IFC. *Utility-Scale Solar Photovoltaic Power Plants: A Project Developer's Guide.* Technical Report, 2015. (Cited in page 132.)

[Ikhsan 2013] Ikhsan, M., Purwadi, A., Hariyanto, N., Heryana, N. and Haroen, Y. *Study of Renewable Energy Sources Capacity and Loading Using Data Logger for Sizing of Solar-wind Hybrid Power System.* Procedia Technology, vol. 11, pages 1048–1053, 2013. (Cited in page 145.)

[IOS 2017] IOS. *Marking and documentation requirements for Photovoltaic Modules.* Technical Report, International Organization for Standardization EN 50380 ed. 2, 2017. (Cited in page 62.)

[IRENA 2019] IRENA. *Total renewable energy*, 2019. (Cited in page 110.)

[Ismail 2016] Ismail, N., Nordin, F., Alkahtani, A. and ZAM., S. *Detection of the Source of the Incipient Faults Produced by Single Phase Inverter using Feed-Forward Back-Propagation Neural Network.* Indian Journal of Science and Technology, vol. 9, pages 1–9, 2016. (Cited in page 186.)

[ISO 2013] ISO. *THIN-FILM TERRESTRIAL PHOTOVOLTAIC (PV) MODULES — DESIGN QUALIFICATION AND TYPE APPROVAL.* Technical Report, International Standard ISO 16077, 2013. (Cited in pages vii and 159.)

[Iyer 2013] Iyer, K. L. V., Lu, X., Usama, Y., Ramakrishnan, V. and Kar, N. C. *A Twofold Daubechies-Wavelet-Based Module for Fault Detection and Voltage Regulation in SEIGs for Distributed Wind Power Generation.* IEEE Transactions on Industrial Electronics, vol. 60, no. 4, pages 1638–1651, 2013. (Cited in page 184.)

[Jadidi 2020] Jadidi, S., Badihi, H. and Zhang, Y. *Fault Diagnosis in Microgrids with Integration of Solar Photovoltaic Systems:A Review.* IFAC-PapersOnLine, vol. 53, no. 2, pages 12091–12096, 2020. (Cited in page 180.)

[Jang 2011] Jang, M., Han, M.-S., Kim, J.-H. and Yang, H.-S. Dynamic time warping-based k-means clustering for accelerometer-based handwriting recognition, volume 363, pages 21–26. 07 2011. (Cited in page 258.)

[Janh 2000] Janh, U. *Analysis of Photovoltaic Systems. Report IEA-PVPS T2-01.* Technical Report, International Energy Agency IEA, 2000. (Cited in pages 4 and 35.)

[Jaskie 2021] Jaskie, K., Martin, J. and Spanias, A. *PV Fault Detection Using Positive Unlabeled Learning.* Applied Sciences, vol. 11, no. 12, 2021. (Cited in page 121.)

[Jean 2015] Jean, J., Brown, P. R., Jaffe, R. L., Buonassisi, T. and Bulović, V. *Pathways for solar photovoltaics.* Energy Environ. Sci., vol. 8, pages 1200–1219, 2015. (Cited in page 29.)

[Jelodar 2019] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L. *Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey.* Multimedia Tools and Applications, vol. 78, no. 11, pages 15169–15211, 2019. (Cited in pages 98 and 99.)

[Jenatton 2010] Jenatton, R., Obozinski, G. and Bach, F. *Structured Sparse Principal Component Analysis.* In Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), pages 366–373, Chia Laguna Resort, Sardinia, Italy, 2010. (Cited in page 192.)

[Jenkins 2017] Jenkins, N. and Ekanayake, J. Renewable energy engineering. Cambridge University Press, 2017. (Cited in pages v, 48, and 49.)

[Jensen 2001] Jensen, A. and Cour-Harbo, A. Ripples in mathematics:the discrete wavelet transform, pages 1–246. Springer ed, 2001. (Cited in page 186.)

[Jeong 2011] Jeong, Y.-S., Jeong, M. K. and Omitaomu, O. A. *Weighted dynamic time warping for time series classification.* Pattern Recognition, vol. 44, no. 9, pages 2231–2240, 2011. (Cited in page 216.)

[Ji 2016] Ji, W. and Zhang, J. *Phase error evaluation technique based on Fourier transform for refractive index detection limit of microfluidic differential refractometer.* Optik, vol. 127, no. 19, pages 7973–7977, 2016. (Cited in page 183.)

[Ji 2017] Ji, D., Zhang, C., Lv, M., Ma, Y. and Guan, N. *Photovoltaic Array Fault Detection by Automatic Reconfiguration.* Energies, vol. 10, no. 5, 2017. (Cited in pages 65 and 67.)

[Jiang 2015] Jiang, L. L. and Maskell, D. L. *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), pages 1–8, 2015. (Cited in page 113.)

[Jiang 2016] Jiang, H., Qiang, M. and Lin, P. *A topic modeling based bibliometric exploration of hydropower research.* Renewable and Sustainable Energy Reviews, vol. 57, pages 226–237, 2016. (Cited in pages 98 and 99.)

[Jiao 2020] Jiao, S., Song, J. and Liu, B. *A Review of Decision Tree Classification Algorithms for Continuous Variables.* Journal of Physics: Conference Series, vol. 1651, page 012083, 11 2020. (Cited in page 206.)

[Johansson 2016] Johansson, J. and Forsell, C. *Evaluation of Parallel Coordinates: Overview, Categorization and Guidelines for Future Research.* IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, pages 579–588, 2016. (Cited in page 189.)

[Johnson 2011a] Johnson, J., Kuszmaul, S., Bower, W. and Schoenwald, D. *Using PV module and line frequency response data to create robust arc fault detectors.* 09 2011. (Cited in page 89.)

[Johnson 2011b] Johnson, J., Pahl, B., Luebke, C., Pier, T., Miller, T., Strauch, J., Kuszmaul, S. and Bower, W. *Photovoltaic DC Arc Fault Detector testing at Sandia National Laboratories.* In 2011 37th IEEE Photovoltaic Specialists Conference, pages 003614–003619, 2011. (Cited in page 89.)

[Johnson 2012a] Johnson, J. *Electrical and thermal finite element modeling of arc faults in photovoltaic bypass diodes.* 05 2012. (Cited in pages 70 and 77.)

[Johnson 2012b] Johnson, J. and Kang, J. *Arc-fault detector algorithm evaluation method utilizing prerecorded arcing signatures.* In 2012 38th IEEE Photovoltaic Specialists Conference, pages 001378–001382, 2012. (Cited in page 89.)

[Johnson 2012c] Johnson, J., Montoya, M., McCalmont, S., Katzir, G., Fuks, F., Earle, J., Fresquez, A., Gonzalez, S. and Granata, J. *Differentiating series and parallel photovoltaic arc-faults.* In 2012 38th IEEE Photovoltaic Specialists Conference, pages 000720–000726, 2012. (Cited in pages 70 and 77.)

[Jolliffe 2002] Jolliffe, I. Principal component analysis. Springer, New York, 2002. (Cited in page 191.)

[Jones 2015] Jones, C. B., Stein, J. S., Gonzalez, S. and King, B. H. *Photovoltaic system fault detection and diagnostics using Laterally Primed Adaptive Resonance Theory neural network.* In 2015 IEEE 42nd Photovoltaic Specialist Conference (PVSC), pages 1–6, 2015. (Cited in page 114.)

[Jordan 2016] Jordan, D. C., Kurtz, S. R., VanSant, K. and Newmiller, J. *Compendium of photovoltaic degradation rates.* Progress in Photovoltaics: Research and Applications, vol. 24, no. 7, pages 978–989, 2016. (Cited in page 235.)

[Jordan 2017] Jordan, D. C., Silverman, T. J., Wohlgemuth, J. H., Kurtz, S. R. and VanSant, K. T. *Photovoltaic failure and degradation modes.* Progress in Photovoltaics: Research and Applications, vol. 25, no. 4, pages 318–326, 2017. (Cited in pages xi, 55, 61, 74, and 78.)

[Jun 2011] Jun, B. *Fault detection using dynamic time warping (DTW) algorithm and discriminant analysis for swine wastewater treatment.* Journal of Hazardous Materials, vol. 185, no. 1, pages 262–268, 2011. (Cited in page 218.)

[Juxing 2020] Juxing, W., Ziyu, W., Junfeng, Q., Qi, Z. and Shuiqing, X. *A Novel Multichannel Online Denoising Method for The Three Phase Electrical Signals of PV System.* In 2020 Chinese Control And Decision Conference (CCDC), pages 5021–5026, 2020. (Cited in page 141.)

[Kalogirou 2013] Kalogirou, S. A., Agathokleous, R. and Panayiotou, G. *On-site PV characterization and the effect of soiling on their performance.* Energy, vol. 51, pages 439–446, 2013. (Cited in page 65.)

[Kamada 1989] Kamada, T. and Kawai, S. *An algorithm for drawing general undirected graphs.* Information Processing Letters, vol. 31, no. 1, pages 7–15, 1989. (Cited in page 97.)

[Kamunda 2007] Kamunda, C., Carelse, X. F., Mathuthu, M. and Makarau, A. *Design and construction of microclimate monitoring system.* Review of Scientific Instruments, vol. 78, no. 8, page 086104, 2007. (Cited in page 145.)

[Kaplani 2012] Kaplani, E. *Degradation effects in sc-Si PV modules subjected to natural and induced ageing after several years of field operation.* Journal of Engineering Science and Technology Review, vol. 5, pages 18–23, 12

2012. [Online]. Available: http://dx.doi.org/10.25103/jestr.054.04. (Cited in page 64.)

[Kaplanis 2011] Kaplanis, S. and Kaplani, E. *Energy performance and degradation over 20years performance of BP c-Si PV modules.* Simulation Modelling Practice and Theory, vol. 19, no. 4, pages 1201–1211, 2011. Sustainable Energy and Environmental Protection "SEEP2009". (Cited in pages 62 and 87.)

[Kapucu 2021] Kapucu, C. and Cubukcu, M. *A supervised ensemble learning method for fault diagnosis in photovoltaic strings.* Energy, vol. 227, page 120463, 2021. (Cited in page 116.)

[Karimi 2000] Karimi, M., Mokhtari, H. and Iravani, M. R. *Wavelet based on-line disturbance detection for power quality applications.* IEEE Transactions on Power Delivery, vol. 15, no. 4, pages 1212–1220, 2000. [Online]. Available: http://dx.doi.org/10.1109/61.891505. (Cited in page 143.)

[Karimi 2019] Karimi, A. M., Fada, J. S., Hossain, M. A., Yang, S., Peshek, T. J., Braid, J. L. and French, R. H. *Automated Pipeline for Photovoltaic Module Electroluminescence Image Processing and Degradation Feature Classification.* IEEE Journal of Photovoltaics, vol. 9, no. 5, pages 1324–1335, 2019. [Online]. Available: http://dx.doi.org/10.1109/JPHOTOV.2019.2920732. (Cited in page 142.)

[Karimi 2020] Karimi, A. M., Fada, J. S., Parrilla, N. A., Pierce, B. G., Koyutürk, M., French, R. H. and Braid, J. L. *Generalized and Mechanistic PV Module Performance Prediction From Computer Vision and Machine Learning on Electroluminescence Images.* IEEE Journal of Photovoltaics, vol. 10, no. 3, pages 878–887, 2020. [Online]. Available: http://dx.doi.org/10.1109/JPHOTOV.2020.2973448. (Cited in page 141.)

[Karmacharya 2018] Karmacharya, I. M. and Gokaraju, R. *Fault Location in Ungrounded Photovoltaic System Using Wavelets and ANN.* IEEE Transactions on Power Delivery, vol. 33, no. 2, pages 549–559, 2018. [Online]. Available: http://dx.doi.org/10.1109/TPWRD.2017.2721903. (Cited in page 69.)

[Kashyap 2003] Kashyap, K. and Shenoy, U. *Classification of power system faults using wavelet transforms and probabilistic neural networks.* In International Symposium on Circuits and Systems (ISCAS), pages 1–4, 2003. (Cited in page 184.)

[Kato 2011] Kato, K. *PVRessQ! - A Research Activity on Reliability of PV System from an user's viewpoint in Japan,Real-time inspection and determination methods of faults on photovoltaic power systems by thermal imaging in Turkey.* Proceedings of SPIE - The International Society for Optical Engineering, vol. 8112, pages 18–, 09 2011. [Online]. Available: http://dx.doi.org/10.1117/12.896135. (Cited in page 83.)

[Kato 2015] Kato, L. *PV module failures observed in the field- solder bond and bypass diode failures.* Technical Report, International Energy Agency IEA, 2015. (Cited in page 71.)

[Kaushika 2003] Kaushika, N. and Gautam, N. *Energy yield simulations of interconnected solar PV arrays.* IEEE Transactions on Energy Conversion, vol. 18, no. 1, pages 127–134, 2003. [Online]. Available: http://dx.doi.org/10.1109/TEC.2002.805204. (Cited in page 53.)

[Kaushika 2007] Kaushika, N. and Rai, A. K. *An investigation of mismatch losses in solar photovoltaic cell networks.* Energy, vol. 32, no. 5, pages 755 – 759, 2007. (Cited in page 73.)

[Keiser 2005] Keiser, J. and Utzinger, J. *Trends in the core literature on tropical medicine: a bibliometric analysis from 1952-2002.* Scientometrics, vol. 62, no. 3, pages 351–365, March 2005. (Cited in page 95.)

[Kempe 2006] Kempe, M. D. *Modeling of rates of moisture ingress into photovoltaic modules.* Solar Energy Materials and Solar Cells, vol. 90, no. 16, pages 2720–2738, 2006. (Cited in page 62.)

[Kempe 2007] Kempe, M. D., Jorgensen, G. J., Terwilliger, K. M., McMahon, T. J., Kennedy, C. E. and Borek, T. T. *Acetic acid production and glass transition concerns with ethylene-vinyl acetate used in photovoltaic devices.* Solar Energy Materials and Solar Cells, vol. 91, no. 4, pages 315–329, 2007. (Cited in page 59.)

[Kenny 2003] Kenny, R., Friesen, G., Chianese, D., Bernasconi, A. and Dunlop, E. *Energy rating of PV modules: comparison of methods and approach.* In 3rd World Conference onPhotovoltaic Energy Conversion, 2003. Proceedings of, volume 2, pages 2015–2018 Vol.2, 2003. (Cited in page 240.)

[Khoshnami 2018] Khoshnami, A. and Sadeghkhani, I. *Two-stage power–based fault detection scheme for photovoltaic systems.* Solar Energy, vol. 176, pages 10–21, 2018. (Cited in page 192.)

[Kim 2016] Kim, N., Hwang, K.-J., Kim, D., Lee, J. H., Jeong, S. and Jeong, D. H. *Analysis and reproduction of snail trails on silver grid lines in crystalline silicon photovoltaic modules.* Solar Energy, vol. 124, pages 153–162, 2016. (Cited in pages 36, 61, 82, 180, and 200.)

[Kim 2021] Kim, J., Rabelo, M., Padi, S. P., Yousuf, H., Cho, E.-C. and Yi, J. *A Review of the Degradation of Photovoltaic Modules for Life Expectancy.* Energies, vol. 14, no. 14, 2021. (Cited in pages vi and 64.)

[King] King, D. L., Kratochvil, J. A. and Boyson, W. E. *Field experience with a new performance characterization procedure for photovoltaic arrays.* world

conference and exhibition on photovoltaic solar energy conversion. (Cited in page 240.)

[King 1999] King, D. L., Quintana, M. A., Kratochvil, J. A., Ellibee, D. E. and Hansen, B. R. *Photovoltaic module performance and durability following long-term field exposure*. AIP Conference Proceedings, vol. 462, no. 1, pages 565–571, 1999. (Cited in page 73.)

[King 2004a] King, D., Boyson, W. and Kratochvill, J. *Photovoltaic array performance model*. Technical Report, Sandia Corporation, 2004. (Cited in page 235.)

[King 2004b] King, D., Kratochvil, J. and Boyson, W. *Photovoltaic Array Performance Model*. PhD thesis, 01 2004. (Cited in page 240.)

[Kirchartz 2009] Kirchartz, T., Helbig, A., Reetz, W., Reuter, M., Werner, J. H. and Rau, U. *Reciprocity between electroluminescence and quantum efficiency used for the characterization of silicon solar cells*. Progress in Photovoltaics: Research and Applications, vol. 17, no. 6, pages 394–402, 2009. (Cited in pages 83 and 85.)

[Klavans 2006] Klavans, R. and Boyack, K. W. *Quantitative evaluation of large maps of science*. Scientometrics, vol. 68, pages 475–499, 2006. (Cited in page 97.)

[Koch 2016] Koch, S., Weber, T., Sobottka, C., Fladung, A., Clemens, P. and Berghold, J. *OUTDOOR ELECTROLUMINESCENCE IMAGING OF CRYSTALLINE PHOTOVOLTAIC MODULES: COMPARATIVE STUDY BETWEEN MANUAL GROUND-LEVEL INSPECTIONS AND DRONE-BASED AERIAL SURVEYS*. 06 2016. (Cited in page 85.)

[Koentges 2014] Koentges, M., Kurtz, S., Packard, C. E., Jahn, U., Berger, K. A., Kato, K., Friesen, T., Liu, H., Van Iseghem, M. and Wohlgemuth, J. *Review of failures of photovoltaic modules*. Technical Report, International Energy Agency IEA, 2014. (Cited in pages 180 and 200.)

[Köntges 2017] Köntges, M., Oreski, G., Magnus Herz, U. J., Hacke, P. and Weiss, K.-A. *Assessment of Photovoltaic Module Failures in the Field*. Technical Report, International energy agency photovoltaic power systems programme, 2017. (Cited in page 75.)

[Kopcsa 1998] Kopcsa, A. and Schiebel, E. *Science and technology mapping: A new iteration model for representing multidimensional relationships*. Journal of the American Society for Information Science, vol. 49, no. 1, pages 7–17, 1998. (Cited in page 97.)

[Koutroulis 2003] Koutroulis, E. and Kalaitzakis, K. *Development of an integrated data-acquisition system for renewable energy sources systems monitoring*.

Renewable Energy, vol. 28, no. 1, pages 139–152, 2003. (Cited in pages 145 and 146.)

[Koutroulis 2012] Koutroulis, E. and Blaabjerg, F. *A New Technique for Tracking the Global Maximum Power Point of PV Arrays Operating Under Partial-Shading Conditions.* IEEE Journal of Photovoltaics, vol. 2, no. 2, pages 184–190, 2012. [Online]. Available: http://dx.doi.org/10.1109/JPHOTOV.2012.2183578. (Cited in page 67.)

[Kratochvil 2004] Kratochvil, J. A., Boyson, W. E. and King, D. L. *Photovoltaic array performance model.* 8 2004. (Cited in page 136.)

[Kuhn 2018] Kuhn, K. D. *Using structural topic modeling to identify latent topics and trends in aviation incident reports.* Transportation Research Part C: Emerging Technologies, vol. 87, pages 105–122, 2018. (Cited in page 98.)

[Kuitche 2014] Kuitche, J. M., Pan, R. and TamizhMani (Mani), G. *Investigation of Dominant Failure Mode(s) for Field-Aged Crystalline Silicon PV Modules Under Desert Climatic Conditions.* IEEE Journal of Photovoltaics, vol. 4, no. 3, pages 814–826, 2014. [Online]. Available: http://dx.doi.org/10.1109/JPHOTOV.2014.2308720. (Cited in pages 57 and 78.)

[Kulkarni 2020] Kulkarni, A., Chong, D. and Batarseh, F. A. *5 - Foundations of data imbalance and solutions for a data democracy.* In Data Democracy, pages 83–106. Academic Press, 2020. (Cited in page 223.)

[Kumar 2018] Kumar, B. P., Ilango, G. S., Reddy, M. J. B. and Chilakapati, N. *Online Fault Detection and Diagnosis in Photovoltaic Systems Using Wavelet Packets.* IEEE Journal of Photovoltaics, vol. 8, no. 1, pages 257–265, 2018. (Cited in pages 23, 183, 184, and 228.)

[Kurtz 2013] Kurtz, S., Newmiller, J., Kimber, A., Flottemesch, R., Riley, E., Dierauf, T., McKee, J. and Krishnani, P. *Analysis of Photovoltaic System Energy Performance Evaluation Method.* 11 2013. (Cited in pages 133, 157, and 158.)

[Kurukuru 2020] Kurukuru, V. S. B., Blaabjerg, F., Khan, M. A. and Haque, A. *A Novel Fault Classification Approach for Photovoltaic Systems.* Energies, vol. 13, no. 2, 2020. (Cited in pages 183, 184, 186, 191, and 232.)

[Kusiak 2019] Kusiak, A. *Convolutional and generative adversarial neural networks in manufacturing.* International Journal of Production Research, vol. 58, pages 1–11, 09 2019. [Online]. Available: http://dx.doi.org/10.1080/00207543.2019.1662133. (Cited in page 238.)

[Köntges 2011a] Köntges, M., Kunze, I., Kajari-Schröder, S., Breitenmoser, X. and Bjørneklett, B. *The risk of power loss in crystalline silicon based photovoltaic*

modules due to micro-cracks. Solar Energy Materials and Solar Cells, vol. 95, no. 4, pages 1131–1137, 2011. (Cited in page 60.)

[Köntges 2011b] Köntges, M., Kajari-Schröder, S., Kunze, I. and Jahn, U. *Crack Statistic of Crystalline Silicon Photovoltaic Modules.* 09 2011. (Cited in page 56.)

[Köntges 2014a] Köntges, M., Kurtz, S., Packard, C., Jahn, U., Berger, K., Kato, K., Friesen, T., Liu, H. and Van Iseghem, M. *Performance and reliability of photovoltaic systems, subtask 3.2: review of failures of photovoltaic modules, international energy agency.* Technical Report, International energy agency photovoltaic power systems programme, 2014. (Cited in page 55.)

[Köntges 2014b] Köntges, M., Kurtz, S., Packard, C., Jahn, U., Berger, K., Kato, K., Friesen, T., Liu, H., Van Iseghem, M., Wohlgemuth, j., Miller, D., Kempe, M., Hacke, P., Reil, F., Bogdanski, N., Herrmann, W., Buerhop, C., Razongles, G. and Friesen, G. *Review of Failures of Photovoltaic Modules*, 01 2014. (Cited in pages vi, 36, 56, 60, 61, 63, 64, 65, 68, 71, 72, 73, 77, 78, and 83.)

[Köntges 2018] Köntges, M., Altmann, S., Heimberg, T., Jahn, U., Berger, K. and Pvps, I. *Mean Degradation Rates In Pv Systems For Various Kinds Of Pv Module Failures.* pages 1435–1443, 05 2018. (Cited in pages vi and 72.)

[Lambrou 1998] Lambrou, T., Kudumakis, P., Speller, R., Sandler, M. and Linney, A. *Classification of audio signals using statistical features on time and wavelet transform domains.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 6, pages 3621–3624 vol.6, 1998. (Cited in page 188.)

[Lebreton 2022] Lebreton, C., Kbidi, F., Alicalapa, F., Benne, M. and Damour, C. *PV Fault Diagnosis Method Based on Time Series Electrical Signal Analysis.* Engineering Proceedings, vol. 18, no. 1, 2022. (Cited in page 183.)

[Lecun 1998] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, vol. 86, no. 11, pages 2278–2324, 1998. [Online]. Available: http://dx.doi.org/10.1109/5.726791. (Cited in page 116.)

[Leloux 2012] Leloux, J., Narvarte, L. and Trebosc, D. *Review of the performance of residential PV systems in Belgium.* Renewable and Sustainable Energy Reviews, vol. 16, no. 1, pages 178–184, 2012. (Cited in page 235.)

[Leva 2019] Leva, S., Mussetta, M. and Ogliari, E. *PV Module Fault Diagnosis Based on Microconverters and Day-Ahead Forecast.* IEEE Transactions on Industrial Electronics, vol. 66, no. 5, pages 3928–3937, 2019. [Online]. Available: http://dx.doi.org/10.1109/TIE.2018.2879284. (Cited in pages 113 and 114.)

[Li 2012] Li, Z., Wang, Y., Zhou, D. and Wu, C. *An Intelligent Method for Fault Diagnosis in Photovoltaic Array.* In Xiao, T., Zhang, L. and Ma, S., editors, System Simulation and Scientific Computing, pages 10–16, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. (Cited in pages 113 and 114.)

[Li 2015] Li, H. and Xiao, D. *Fault diagnosis based on power spectral density basis transform.* Journal of Vibration and Control, vol. 21, no. 12, pages 2416–2433, 2015. [Online]. Available: `https://doi.org/10.1177/1077546313487242`. (Cited in page 144.)

[Li 2021a] Li, B., Delpha, C., Diallo, D. and Migan-Dubois, A. *Application of Artificial Neural Networks to photovoltaic fault detection and diagnosis: A review.* Renewable and Sustainable Energy Reviews, vol. 138, page 110512, 2021. (Cited in page 92.)

[Li 2021b] Li, B., Delpha, C., Diallo, D. and Migan-Dubois, A. *Application of Artificial Neural Networks to photovoltaic fault detection and diagnosis: A review.* Renewable and Sustainable Energy Reviews, vol. 138, page 110512, 2021. (Cited in page 180.)

[Li 2021c] Li, B. *Health monitoring of photovoltaic modules using electrical measurements.* Theses, Université Paris-Saclay, October 2021. (Cited in pages 56, 58, 60, 61, 62, 63, 77, 134, 135, 141, 143, 180, 181, 186, and 200.)

[Li 2021d] Li, H. *Time works well: Dynamic time warping based on time weighting for time series data mining.* Information Sciences, vol. 547, pages 592–608, 2021. (Cited in pages 218 and 219.)

[Licciardo 2018] Licciardo, G. D., Cappetta, C., Di Benedetto, L., Rubino, A. and Liguori, R. *Multiplier-Less Stream Processor for 2D Filtering in Visual Search Applications.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 1, pages 267–272, 2018. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2016.2603068. (Cited in page 110.)

[Lim 2021] Lim, M. K., Li, Y., Wang, C. and Tseng, M.-L. *A literature review of blockchain technology applications in supply chains: A comprehensive analysis of themes, methodologies and industries.* Computers & Industrial Engineering, vol. 154, page 107133, 2021. (Cited in page 93.)

[Lin 2015] Lin, H., Chen, Z., Wu, L., Lin, P. and Cheng, S. *On-line Monitoring and Fault Diagnosis of PV Array Based on BP Neural Network Optimized by Genetic Algorithm.* In Bikakis, A. and Zheng, X., editors, Multi-disciplinary Trends in Artificial Intelligence, pages 102–112, Cham, 2015. Springer International Publishing. (Cited in page 123.)

[Lin 2017] Lin, P., Lin, Y., Chen, Z.-C., Wu, L., Chen, L. and Cheng, S. *A Density Peak-Based Clustering Approach for Fault Diagnosis of Photovoltaic Arrays.*

International Journal of Photoenergy, vol. 2017, pages 1–14, 03 2017. (Cited in page 86.)

[Lindroos 2016] Lindroos, J. and Savin, H. *Review of light-induced degradation in crystalline silicon solar cells.* Solar Energy Materials and Solar Cells, vol. 147, pages 115–126, 2016. (Cited in page 62.)

[Lines 2015] Lines, J. and Bagnall, A. *Time series classification with ensembles of elastic distance measures.* Data Min Knowl Disc 29, page 565–592, 2015. (Cited in page 218.)

[Liu 2007] Liu, Y. and Rayens, W. *PLS and dimension reduction for classification.* Computational Statistics, vol. 22, page 189–208, 2007. (Cited in pages 221 and 224.)

[Liu 2017] Liu, Q., Zhao, Y., Zhang, Y., Kang, D., Lv, Q. and Shang, L. *Hierarchical context-aware anomaly diagnosis in large-scale PV systems using SCADA data.* In 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), pages 1025–1030, 2017. (Cited in page 128.)

[Liu 2021a] Liu, H., Yang, J., Ye, M., James, S. C., Tang, Z., Dong, J. and Xing, T. *Using t-distributed Stochastic Neighbor Embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data.* Journal of Hydrology, vol. 597, page 126146, 2021. (Cited in pages 103 and 104.)

[Liu 2021b] Liu, Y., Ding, K., Zhang, J., Li, Y., Yang, Z., Zheng, W. and Chen, X. *Fault diagnosis approach for photovoltaic array based on the stacked auto-encoder and clustering with I-V curves.* Energy Conversion and Management, vol. 245, page 114603, 2021. (Cited in pages 118 and 124.)

[Liu 2022] Liu, Y., Ding, K., Zhang, J., Lin, Y., Yang, Z., Chen, X., Li, Y. and Chen, X. *Intelligent fault diagnosis of photovoltaic array based on variable predictive models and Iâ€“V curves.* Solar Energy, vol. 237, pages 340–351, 2022. (Cited in page 114.)

[Livera 2019a] Livera, A., Theristis, M., Makrides, G. and Georghiou, G. E. *Recent advances in failure diagnosis techniques based on performance data analysis for grid-connected photovoltaic systems.* Renewable Energy, vol. 133, pages 126–143, 2019. (Cited in pages 82, 92, and 136.)

[Livera 2019b] Livera, A., Theristis, M., Makrides, G. and Georghiou, G. E. *Recent advances in failure diagnosis techniques based on performance data analysis for grid-connected photovoltaic systems.* Renewable Energy, vol. 133, pages 126–143, 2019. (Cited in pages 111, 134, 137, and 180.)

[Long 2014] Long, M., Wang, J., Ding, G., Pan, S. J. and Yu, P. S. *Adaptation Regularization: A General Framework for Transfer Learning.* IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 5, pages 1076–1089, 2014.

[Online]. Available: http://dx.doi.org/10.1109/TKDE.2013.111. (Cited in page 238.)

[Lu 2012] Lu, K. and Wolfram, D. *Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches.* Journal of the American Society for Information Science and Technology, vol. 63, no. 10, pages 1973–1986, 2012. (Cited in page 98.)

[Lu 2018] Lu, S., Phung, B. and Zhang, D. *A comprehensive review on DC arc faults and their diagnosis methods in photovoltaic systems.* Renewable and Sustainable Energy Reviews, vol. 89, pages 88–98, 2018. (Cited in page 180.)

[Lu 2019a] Lu, S., Sirojan, T., Phung, B. T., Zhang, D. and Ambikairajah, E. *DA-DCGAN: An Effective Methodology for DC Series Arc Fault Diagnosis in Photovoltaic Systems.* IEEE Access, vol. 7, pages 45831–45840, 2019. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2019.2909267. (Cited in pages 118, 125, 238, and 239.)

[Lu 2019b] Lu, X., Lin, P., Cheng, S., Lin, Y., Chen, Z., Wu, L. and Zheng, Q. *Fault diagnosis for photovoltaic array based on convolutional neural network and electrical time series graph.* Energy Conversion and Management, vol. 196, pages 950–965, 2019. (Cited in page 141.)

[Lu 2021a] Lu, S., Ma, R., Sirojan, T., Phung, B. and Zhang, D. *Lightweight transfer nets and adversarial data augmentation for photovoltaic series arc fault detection with limited fault data.* International Journal of Electrical Power Energy Systems, vol. 130, page 107035, 2021. (Cited in page 239.)

[Lu 2021b] Lu, S.-D., Wang, M.-H., Wei, S.-E., Liu, H.-D. and Wu, C.-C. *Photovoltaic Module Fault Detection Based on a Convolutional Neural Network.* Processes, vol. 9, no. 9, 2021. (Cited in page 86.)

[Lundberg 2020] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I. *From local explanations to global understanding with explainable AI for trees.* Nat Mach Intell, vol. 2, pages 56–67, 2020. (Cited in page 205.)

[López-Escalante 2016] López-Escalante, M., Caballero, L. J., Martín, F., Gabás, M., Cuevas, A. and Ramos-Barrado, J. *Polyolefin as PID-resistant encapsulant material in PV modules.* Solar Energy Materials and Solar Cells, vol. 144, pages 691–699, 2016. (Cited in page 68.)

[Maaløe 2020] Maaløe, L., Winther, O., Spataru, S. and Sera, D. *Condition Monitoring in Photovoltaic Systems by Semi-Supervised Machine Learning.* Energies, vol. 13, no. 3, 2020. (Cited in pages 112 and 121.)

[Madeti 2017a] Madeti, S. R. and Singh, S. N. *Monitoring system for photovoltaic plants: A review.* Renewable and Sustainable Energy Reviews, vol. 67, pages 1180–1207, 2017. (Cited in page 29.)

[Madeti 2017b] Madeti, S. R. and Singh, S. *A comprehensive study on different types of faults and detection techniques for solar photovoltaic system.* Solar Energy, vol. 158, pages 161–185, 2017. (Cited in pages 30, 56, 83, 84, 85, and 180.)

[Madeti 2017c] Madeti, S. R. and Singh, S. *A comprehensive study on different types of faults and detection techniques for solar photovoltaic system.* Solar Energy, vol. 158, pages 161–185, 2017. (Cited in page 92.)

[Madeti 2017d] Madeti, S. R. and Singh, S. *Monitoring system for photovoltaic plants: A review.* Renewable and Sustainable Energy Reviews, vol. 67, pages 1180–1207, 2017. (Cited in pages 133, 136, 137, and 138.)

[Maghami 2016] Maghami, M. R., Hizam, H., Gomes, C., Radzi, M. A., Rezadad, M. I. and Hajighorbani, S. *Power loss due to soiling on solar panel: A review.* Renewable and Sustainable Energy Reviews, vol. 59, pages 1307–1316, 2016. (Cited in page 65.)

[Mahesh 2019] Mahesh, B. *Machine Learning Algorithms -A Review.* International Journal of Science and Research, pages 381–386, 01 2019. (Cited in pages 205 and 210.)

[Mahjoubi 2012] Mahjoubi, A., Mechlouch, R. F. and Brahim, A. B. *Data Acquisition System for Photovoltaic Water Pumping System in the Desert of Tunisia.* Procedia Engineering, vol. 33, pages 268–277, 2012. (Cited in pages 145 and 146.)

[Makarskas 2021] Makarskas, V., Jurevičius, M., Zakis, J., Kilikevičius, A., Borodinas, S., Matijošius, J. and Kilikevičienė, K. *Investigation of the influence of hail mechanical impact parameters on photovoltaic modules.* Engineering Failure Analysis, vol. 124, page 105309, 2021. (Cited in page 56.)

[Mallat 2008] Mallat, S. A wavelet tour of signal processing 3rd ed., page 1 – 745. Academic Press, Cambridge, MA, USA, 2008. (Cited in page 184.)

[Malvoni 2017] Malvoni, M., Leggieri, A., Maggiotto, G., Congedo, P. and De Giorgi, M. *Long term performance, losses and efficiency analysis of a 960kWP photovoltaic system in the Mediterranean climate.* Energy Conversion and Management, vol. 145, pages 169–181, 2017. (Cited in page 234.)

[Man 2004] Man, M. Z., Dyson, G., Johnson, K. and Liao, B. *Evaluating Methods for Classifying Expression Data.* Journal of Biopharmaceutical Statistics, vol. 14, no. 4, pages 1065–1084, 2004. (Cited in page 223.)

[Manganiello 2015] Manganiello, P., Balato, M. and Vitelli, M. *A Survey on Mismatching and Aging of PV Modules: The Closed Loop.* IEEE Transactions on Industrial Electronics, vol. 62, no. 11, pages 7276–7286, 2015. [Online]. Available: http://dx.doi.org/10.1109/TIE.2015.2418731. (Cited in pages 57, 59, 60, and 62.)

[Manno 2021] Manno, D., Cipriani, G., Ciulla, G., Di Dio, V., Guarino, S. and Lo Brano, V. *Deep learning strategies for automatic fault diagnosis in photovoltaic systems by thermographic images.* Energy Conversion and Management, vol. 241, page 114315, 2021. (Cited in page 117.)

[Manohar 2017] Manohar, M. and Koley, E. *SVM based protection scheme for microgrid.* In 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), pages 429–432, 2017. (Cited in page 143.)

[Manohar 2019] Manohar, M. *Enhancing the reliability of protection scheme for PV integrated microgrid by discriminating between array faults and symmetrical line faults using sparse auto encoder.* IET Renewable Power Generation, vol. 13, pages 308–317(9), February 2019. (Cited in page 124.)

[Mao 2015] Mao, G., Liu, X., Du, H., Zuo, J. and Wang, L. *Way forward for alternative energy research: A bibliometric analysis during 1994–2013.* Renewable and Sustainable Energy Reviews, vol. 48, pages 276–286, 2015. (Cited in page 95.)

[Mao 2018] Mao, G., Huang, N., Chen, L. and Wang, H. *Research on biomass energy and environment from the past to the future: A bibliometric analysis.* Science of The Total Environment, vol. 635, pages 1081–1090, 2018. (Cited in page 96.)

[Maral 2004] Maral, G. Vsat networks. Wiley series in communication and distributed systems. Wiley, 2004. (Cited in page 144.)

[Martinez 2017] Martinez, J., Belahcen, A. and Muetze, A. *Analysis of the Vibration Magnitude of an Induction Motor With Different Numbers of Broken Bars.* IEEE Transactions on Industry Applications, vol. 53, no. 3, pages 2711–2720, 2017. [Online]. Available: http://dx.doi.org/10.1109/TIA.2017.2657478. (Cited in page 144.)

[Masmoudi 2016] Masmoudi, F., Ben Salem, F. and Derbel, N. *Single and double diode models for conventional mono-crystalline solar cell with extraction of internal parameters.* In 2016 13th International Multi-Conference on Systems, Signals & Devices (SSD), pages 720–728, 2016. (Cited in page 59.)

[Massa 2021] Massa, G., Pappalardo, D., Acunzo, R. and Oliviero, M. *An Overview on Artificial Intelligence Techniques for Advanced Operation and Mainte-*

nance of PhotoVoltaic Power Plants. In 2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), pages 734–740, 2021. (Cited in pages 92, 110, and 111.)

[Massi Pavan 2011] Massi Pavan, A., Mellit, A. and De Pieri, D. *The effect of soiling on energy production for large-scale photovoltaic plants.* Solar Energy, vol. 85, no. 5, pages 1128–1136, 2011. (Cited in page 65.)

[Massi Pavan 2014] Massi Pavan, A., Mellit, A., De Pieri, D. and Lughi, V. *A study on the mismatch effect due to the use of different photovoltaic modules classes in large-scale solar parks.* Progress in Photovoltaics: Research and Applications, vol. 22, no. 3, pages 332–345, 2014. (Cited in page 65.)

[Mbuli 2020] Mbuli, N., Mathonsi, M., Seitshiro, M. and Pretorius, J.-H. C. *Decomposition forecasting methods: A review of applications in power systems.* Energy Reports, vol. 6, pages 298 – 306, 2020. (Cited in page 143.)

[McCallum 2002] McCallum, A. K. Mallet: A machine learning for language toolkit. 2002. (Cited in page 101.)

[McCalmont 2013] McCalmont, S. *Low Cost Arc Fault Detection and Protection for PV Systems: January 30, 2012 - September 30, 2013.* 10 2013. (Cited in page 70.)

[McEvoy 2013] McEvoy, A., Castaner, L. and Markvart, T. Solar cells: Materials, manufacture and operation. Academic Press, 2 edition, 2013. (Cited in pages 49, 50, and 88.)

[Meg 2022] *Discovery kit with STM8L152C6 MCU*, 2022. Accessed: 2022-09-30. (Cited in pages 162 and 163.)

[Mehta 2017] Mehta, S., Azad, A. P., Chemmengath, S. A., Raykar, V. and Kalyanaraman, S. *DeepSolarEye: Power Loss Prediction and Weakly Supervised Soiling Localization via Fully Convolutional Networks for Solar Panels.* CoRR, vol. abs/1710.03811, 2017. (Cited in page 117.)

[Mekki 2016] Mekki, H., Mellit, A. and Salhi, H. *Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules.* Simulation Modelling Practice and Theory, vol. 67, pages 1–13, 2016. (Cited in page 113.)

[Mellit 2018a] Mellit, A., Tina, G. M. and Kalogirou, S. A. *Fault detection and diagnosis methods for photovoltaic systems: A review.* Renewable and Sustainable Energy Reviews, vol. 91, pages 1–17, 2018. (Cited in page 180.)

[Mellit 2018b] Mellit, A., Tina, G. and Kalogirou, S. *Fault detection and diagnosis methods for photovoltaic systems: A review.* Renewable and Sustainable

Energy Reviews, vol. 91, pages 1–17, 2018. (Cited in pages 69, 72, 82, 83, and 92.)

[Mellit 2021] Mellit, A. and Kalogirou, S. *Artificial intelligence and internet of things to improve efficacy of diagnosis and remote sensing of solar photovoltaic systems: Challenges, recommendations and future directions.* Renewable and Sustainable Energy Reviews, vol. 143, 2021. (Cited in page 180.)

[Meng 2018] Meng, Y., Zhang, Z., Yin, H. and Ma, T. *Automatic detection of particle size distribution by image analysis based on local adaptive canny edge detection and modified circular Hough transform.* Micron, vol. 106, pages 34–41, 2018. (Cited in page 141.)

[Meyer 2004] Meyer, E. and van Dyk, E. *Assessing the reliability and degradation of photovoltaic module performance parameters.* IEEE Transactions on Reliability, vol. 53, no. 1, pages 83–92, 2004. [Online]. Available: http://dx.doi.org/10.1109/TR.2004.824831. (Cited in pages 60 and 235.)

[Michael Mills-Price 2014] Michael Mills-Price, M., Rourke, M., and Kite, D. *Adaptive Control Strategies and Communications for utility Integration of Photovoltaic Solar Sites.* In Power and Energy Automation Conference, pages 1337–1341, 2014. (Cited in pages 137, 138, and 139.)

[Mikolajczyk 2005] Mikolajczyk, K. and Schmid, C. *A performance evaluation of local descriptors.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 10, pages 1615–1630, 2005. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2005.188. (Cited in page 119.)

[Mimno 2011] Mimno, D., Wallach, H. M., Talley, E., Leenders, M. and McCallum, A. *Optimizing Semantic Coherence in Topic Models.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, page 262–272, USA, 2011. Association for Computational Linguistics. (Cited in page 101.)

[Ministry of Housing 2017] Ministry of Housing, C. . L. G. Grenfell tower. 2017. (Cited in pages 2 and 34.)

[Misiti 2013] Misiti, M., Misiti, Y., Oppenheim, G. and Poggi, J. *Wavelet Toolbox - User's Guide*, 2013. (Cited in pages 23 and 228.)

[Miwa 2006a] Miwa, M., Yamanaka, S., Kawamura, H. and Ohno, H. *Diagnosis of a Power Output Lowering of PV Array with a (dI/dV)-V Characteristic.* Proceeding of IEEE 4th World Conference on Photovoltaic Energy Conversion, vol. 2, 05 2006. (Cited in page 87.)

[Miwa 2006b] Miwa, M., Yamanaka, S., Kawamura, H., Ohno, H. and Kawamura, H. *Diagnosis of a Power Output Lowering of PV Array with a (-dI/dV)-V*

Characteristic. In 2006 IEEE 4th World Conference on Photovoltaic Energy Conference, volume 2, pages 2442–2445, 2006. (Cited in page 87.)

[Molenbroek 1991] Molenbroek, E., Waddington, D. and Emery, K. *Hot spot susceptibility and testing of PV modules*. volume 1, pages 547 – 552 vol.1, 11 1991. (Cited in page 56.)

[Momeni 2020] Momeni, H., Sadoogi, N., Farrokhifar, M. and Gharibeh, H. F. *Fault Diagnosis in Photovoltaic Arrays Using GBSSL Method and Proposing a Fault Correction System*. IEEE Transactions on Industrial Informatics, vol. 16, no. 8, pages 5300–5308, 2020. (Cited in pages 120 and 121.)

[Moradi Sizkouhi 2021] Moradi Sizkouhi, A., Aghaei, M. and Esmailifar, S. M. *A deep convolutional encoder-decoder architecture for autonomous fault detection of PV plants using multi-copters*. Solar Energy, vol. 223, pages 217–228, 2021. (Cited in page 117.)

[Moreno 2020] Moreno, C., González, A., Olazagoitia, J. L. and Vinolas, J. *The Acquisition Rate and Soundness of a Low-Cost Data Acquisition System (LC-DAQ) for High Frequency Applications*. Sensors, vol. 20, no. 2, 2020. (Cited in page 164.)

[Morstatter 2018] Morstatter, F. and Liu, H. *In Search of Coherence and Consensus: Measuring the Interpretability of Statistical Topics*. Journal of Machine Learning Research, vol. 18, no. 169, pages 1–32, 2018. (Cited in page 101.)

[Moya-Anegón 2007] Moya-Anegón, S. G. F. d., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F. J. and Herrero-Solana, V. *Visualizing the marrow of science*. Journal of the American Society for Information Science and Technology, vol. 58, no. 14, pages 2167–2179, 2007. (Cited in page 97.)

[Muhammadsharif 2017] Muhammadsharif, F., Yahya, M., Hameed, S., Aziz, F., Sulaiman, K., Rasheed, M. and Ahmad, Z. *Employment of single-diode model to elucidate the variations in photovoltaic parameters under different electrical and thermal conditions*. PLOS ONE, vol. 12, page e0182925, 08 2017. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0182925. (Cited in page 50.)

[Mukaro 1998] Mukaro, R., Carelse, X. and Olumekor, L. *First performance analysis of a silicon-cell microcontroller-based solar radiation monitoring system*. Solar Energy, vol. 63, no. 5, pages 313–321, 1998. (Cited in pages 145 and 146.)

[Mukaro 1999] Mukaro, R. and Carelse, X. *A microcontroller-based data acquisition system for solar radiation and environmental monitoring*. IEEE Transactions on Instrumentation and Measurement, vol. 48, no. 6, pages 1232–1238,

1999. [Online]. Available: http://dx.doi.org/10.1109/19.816142. (Cited in page 145.)

[Mukaro 2008] Mukaro, R. and Tinarwo, D. *Performance evaluation of hot-box reflector solar cooker using a microcontroller-based measurement system.* International Journal of Energy Research, vol. 32, pages 1339 – 1348, 11 2008. [Online]. Available: http://dx.doi.org/10.1002/er.1441. (Cited in page 145.)

[Mukherjee 2017] Mukherjee, A., Routray, A. and Samanta, A. K. *Method for Online Detection of Arcing in Low-Voltage Distribution Systems.* IEEE Transactions on Power Delivery, vol. 32, no. 3, pages 1244–1252, 2017. [Online]. Available: http://dx.doi.org/10.1109/TPWRD.2015.2392385. (Cited in page 142.)

[Munoz 2011] Munoz, M., Alonso-García, M., Vela, N. and Chenlo, F. *Early degradation of silicon PV modules and guaranty conditions.* Solar Energy, vol. 85, no. 9, pages 2264–2274, 2011. (Cited in pages 61, 63, 64, and 73.)

[Murtadho 2020] Murtadho, M., Prasetyono, E. and Anggriawan, D. O. *Detection of Parallel Arc Fault on Photovoltaic System Based on Fast Fourier Transform.* In 2020 International Electronics Symposium (IES), pages 21–25, 2020. (Cited in page 142.)

[Murugesan 2020] Murugesan, D. N., Anitha, R. and Ganesan, M. *N-Semi Regular Graph-Based Fuzzy Semi-supervised Learning Approach for Fault Detection and Classification in Photovoltaic Arrays.* International Journal of Advanced Research in Engineering and Technology, vol. 11, no. 9, pages 887–896, 2020. (Cited in page 121.)

[Mustak 2021] Mustak, M., Salminen, J., PlÃ©, L. and Wirtz, J. *Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda.* Journal of Business Research, vol. 124, pages 389–404, 2021. (Cited in pages 98 and 99.)

[Nanopoulos 2001] Nanopoulos, A., Alcock, R. and Manolopoulos, Y. Feature-based classification of time-series data, page 49–61. Nova Science Publishers, Inc., USA, 2001. (Cited in page 186.)

[NAS 2022] *Atmospheric Science Data Center*, 2022. Accessed: 2022-09-30. (Cited in pages 18 and 156.)

[Natarajan 2020] Natarajan, K., Kumar B, P. and Kumar, V. *Fault Detection of Solar PV system using SVM and Thermal Image Processing.* International Journal of Renewable Energy Research, vol. 10, pages 967–977, 06 2020. (Cited in page 203.)

[Navid 2021a] Navid, Q., Hassan, A., Fardoun, A. A., Ramzan, R. and Alraeesi, A. *Fault Diagnostic Methodologies for Utility-Scale Photovoltaic Power Plants:*

A State of the Art Review. Sustainability, vol. 13, no. 4, page 1629, 2021. (Cited in pages 29 and 180.)

[Navid 2021b] Navid, Q., Hassan, A., Fardoun, A. A., Ramzan, R. and Alraeesi, A. *Fault Diagnostic Methodologies for Utility-Scale Photovoltaic Power Plants: A State of the Art Review.* Sustainability, vol. 13, no. 4, 2021. (Cited in page 92.)

[Ndiaye 2013] Ndiaye, A., Charki, A., Kobi, A., Kébé, C. M., Ndiaye, P. A. and Sambou, V. *Degradations of silicon photovoltaic modules: A literature review.* Solar Energy, vol. 96, pages 140–151, 2013. (Cited in pages 29, 64, 65, and 233.)

[Ng 2001] Ng, A., Jordan, M. I. and Weiss, Y. *On Spectral Clustering: Analysis and an algorithm.* In NIPS, 2001. (Cited in page 192.)

[Nguyen 2002] Nguyen, D. V. and Rocke, D. M. *Tumor classification by partial least squares using microarray gene expression data* . Bioinformatics, vol. 18, no. 1, pages 39–50, 2002. (Cited in page 224.)

[Nguyen 2015] Nguyen, X. H. *Matlab/Simulink Based Modeling to Study Effect of Partial Shadow on Solar Photovoltaic Array.* Environmental Systems Research, vol. 4, pages 1–10, 2015. (Cited in pages 65 and 66.)

[Niazi 2019] Niazi, K. A. K., Akhtar, W., Khan, H. A., Yang, Y. and Athar, S. *Hotspot diagnosis for solar photovoltaic modules using a Naive Bayes classifier.* Solar Energy, vol. 190, pages 34–43, 2019. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0038092X19307340`. (Cited in page 112.)

[Nielsen 2016] Nielsen, F. In introduction to hpc with mpi for data science, page 195–211. Springer International Publishing, 2016. (Cited in pages 20, 216, 219, and 262.)

[Nielsen 2019] Nielsen, M. W. and Börjesonb, L. *Gender diversity in the management field: Does it matter for research outcomes?* Research Policy, vol. 48, no. 7, pages 1617–1632, 2019. (Cited in page 98.)

[Niennattrakul 2007] Niennattrakul, V. and Ratanamahatana, C. A. *On Clustering Multimedia Time Series Data Using K-Means and Dynamic Time Warping.* In 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07), pages 733–738, 2007. (Cited in pages 257 and 258.)

[Novoa 2015] Novoa, F. *Adhesion and Reliability of Solar Module Materials.* PhD thesis, Stanford University, 2015. (Cited in page 64.)

[NREL 2022] NREL, N. R. E. L. *Best Research-Cell Efficiency Chart*, 2022. (Cited in pages 46 and 47.)

[Nussbaumer 1981] Nussbaumer, H. *The fast Fourier transform.* In Fast Fourier Transform and Convolution Algorithms, pages 80–111. Springer, 1981. (Cited in page 142.)

[Okere 2020] Okere, A. and Iqbal, M. T. *A Review of Conventional Fault Detection Techniques in Solar PV Systems and a Proposal of Long Range (LoRa) Wireless Sensor Network for Module Level Monitoring and Fault Diagnosis in Large Solar PV Farms.* European Journal of Electrical Engineering and Computer Science, vol. 4, no. 6, 2020. (Cited in page 180.)

[Omazic 2019] Omazic, A., Oreski, G., Halwachs, M., Eder, G., Hirschl, C., Neumaier, L., Pinter, G. and Erceg, M. *Relation between degradation of polymeric components in crystalline silicon PV module and climatic conditions: A literature review.* Solar Energy Materials and Solar Cells, vol. 192, pages 123–133, 2019. (Cited in pages vi, 62, and 65.)

[Onal 2021] Onal, Y. and Turhal, U. C. *Discriminative common vector in sufficient data Case: A fault detection and classification application on photovoltaic arrays.* Engineering Science and Technology, an International Journal, vol. 24, no. 5, pages 1168–1179, 2021. (Cited in page 191.)

[Onar 2008] Onar, O., Uzunoglu, M. and Alam, M. *Modeling, control and simulation of an autonomous wind turbine/photovoltaic/fuel cell/ultra-capacitor hybrid power system.* Journal of Power Sources, vol. 185, no. 2, pages 1273–1283, 2008. (Cited in page 29.)

[Oreski 2010] Oreski, G. and Wallner, G. M. *Damp heat induced physical aging of PV encapsulation materials.* In 2010 12th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, pages 1–6, 2010. (Cited in page 62.)

[Ostertagová 2012] Ostertagová, E. *Modelling using Polynomial Regression.* Procedia Engineering, vol. 48, pages 500–506, 2012. (Cited in page 223.)

[Oviedo 2011] Oviedo, S. J., Quiroga, J. E. and Borrás, C. *Experimental evaluation of motor current signature and vibration analysis for rotor broken bars detection in an induction motor.* In 2011 International Conference on Power Engineering, Energy and Electrical Drives, pages 1–6, 2011. (Cited in page 144.)

[Packard 2012] Packard, C., Wohlgemuth, J. and Kurtz, S. *Development of a Visual Inspection Checklist for Evaluation of Fielded PV Module Condition.* Technical Report, National Renewable Energy Lab. (NREL), 2012. (Cited in page 83.)

[Pan 2019] Pan, W., Jian, L. and Liu, T. *Grey system theory trends from 1991 to 2018: a bibliometric analysis and visualization.* Scientometrics, vol. 121, no. 3, pages 1407–1434, December 2019. (Cited in page 97.)

[Pang 2010] Pang, C. and Kezunovic, M. *Fast Distance Relay Scheme for Detecting Symmetrical Fault During Power Swing.* IEEE Transactions on Power Delivery, vol. 25, no. 4, pages 2205–2212, 2010. (Cited in pages 143 and 184.)

[Papadakis 2005] Papadakis, K., Koutroulis, E. and Kalaitzakis, K. *A server database system for remote monitoring and operational evaluation of renewable energy sources plants.* Renewable Energy, vol. 30, no. 11, pages 1649–1669, 2005. (Cited in page 146.)

[Parida 2011] Parida, B., Iniyan, S. and Goic, R. *A review of solar photovoltaic technologies.* Renewable and Sustainable Energy Reviews, vol. 15, no. 3, pages 1625–1636, 2011. (Cited in page 29.)

[Park 2011] Park, N., Han, C., Hong, W. and Kim, D. *The effect of encapsulant delamination on electrical performance of PV module.* In 2011 37th IEEE Photovoltaic Specialists Conference, pages 001113–001115, 2011. (Cited in page 62.)

[Parretta 2005] Parretta, A., Bombace, M., Graditi, G. and Schioppo, R. *Optical degradation of long-term, field-aged c-Si photovoltaic modules.* Solar Energy Materials and Solar Cells, vol. 86, no. 3, pages 349–364, 2005. (Cited in page 61.)

[Patel 2008] Patel, H. and Agarwal, V. *MATLAB-Based Modeling to Study the Effects of Partial Shading on PV Array Characteristics.* IEEE Transactions on Energy Conversion, vol. 23, no. 1, pages 302–310, 2008. [Online]. Available: http://dx.doi.org/10.1109/TEC.2007.914308. (Cited in pages 65 and 66.)

[Patidar 2019] Patidar, S., Jenkins, D. P., Peacock, A. and Mccallum, P. *Time Series Decomposition Approach for Simulating Electricity Demand Profile.* In Building Simulation 2019, 16th IBPSA International International Conference, pages 1388–1395, 2019. (Cited in page 143.)

[PATRO 2015] PATRO, S. G. and Sahu, D.-K. K. *Normalization: A Preprocessing Stage.* IARJSET, 03 2015. [Online]. Available: http://dx.doi.org/10.17148/IARJSET.2015.2305. (Cited in page 141.)

[PAU 2022] *Feature engineering for machine learning enabled early prediction of battery lifetime.* Journal of Power Sources, vol. 527, page 231127, 2022. (Cited in page 179.)

[Pedersen 2019] Pedersen, E., Rao, S., Katoch, S., Jaskie, K., Spanias, A., Tepedelenlioglu, C. and Kyriakides, E. *PV Array Fault Detection using Radial Basis Networks.* In 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), pages 1–4, 2019. (Cited in page 183.)

[Pereira 2018] Pereira, J. and Silveira, M. *Unsupervised Anomaly Detection in Energy Time Series Data Using Variational Recurrent Autoencoders with Attention.* In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 1275–1282, 2018. (Cited in page 128.)

[Perez 2017] Perez, L. and Wang, J. *The Effectiveness of Data Augmentation in Image Classification using Deep Learning.* 12 2017. (Cited in page 142.)

[Pern 1997] Pern, F. J. *Ethylene-vinyl acetate (EVA) encapsulants for photovoltaic modules: Degradation and discoloration mechanisms and formulation modifications for improved photostability.* Die Angewandte Makromolekulare Chemie, vol. 252, no. 1, pages 195–216, 1997. (Cited in page 63.)

[Peters 1993] Peters, H. and van Raan, A. *Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling.* Research Policy, vol. 22, no. 1, pages 23–45, 1993. (Cited in page 98.)

[Pham 2019] Pham, H. *A New Criterion for Model Selection.* Mathematics, vol. 7, no. 12, 2019. (Cited in page 223.)

[Phua 2019] Phua, B., Hsiao, P.-C., Wang, X. and Lennon, A. *Towards a more reliable manufacturing future â€“ Automatic classification of failure modes during adhesion testing of silicon solar cells.* Solar Energy, vol. 178, pages 61–68, 2019. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0038092X18311770`. (Cited in page 119.)

[Picault 2010] Picault, D. *REDUCTION OF MISMATCH LOSSES IN GRID-CONNECTED PHOTOVOLTAIC SYSTEMS USING ALTERNATIVE TOPOLOGIES.* Theses, Institut National Polytechnique de Grenoble - INPG, October 2010. (Cited in page 53.)

[Pichler 2016] Pichler, K., Lughofer, E., Pichler, M., Buchegger, T., Klement, E. P. and Huschenbett, M. *Fault detection in reciprocating compressor valves under varying load conditions.* Mechanical Systems and Signal Processing, vol. 70-71, pages 104 – 119, 2016. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0888327015004008`. (Cited in page 144.)

[Pierdicca 2020] Pierdicca, R., Paolanti, M., Felicetti, A., Piccinini, F. and Zingaretti, P. *Automatic Faults Detection of Photovoltaic Farms: solAIr, a Deep Learning-Based System for Thermal Images.* Energies, vol. 13, no. 24, 2020. (Cited in page 116.)

[Pigueiras 2014] Pigueiras, E., Moretón, R. and Luque, I. *Dust effects on PV array performance: In-field observations with non-uniform patterns.* Progress in Photovoltaics: Research and Applications, vol. 22, 06 2014. [Online]. Available: http://dx.doi.org/10.1002/pip.2348. (Cited in page 65.)

[Pillai 2018a] Pillai, D. S. and Rajasekar, N. *A comprehensive review on protection challenges and fault diagnosis in PV systems.* Renewable and Sustainable Energy Reviews, vol. 91, pages 18–40, 2018. (Cited in pages 56, 58, 65, 66, 69, 70, and 92.)

[Pillai 2018b] Pillai, D. S. and Rajasekar, N. *A comprehensive review on protection challenges and fault diagnosis in PV systems.* Renewable and Sustainable Energy Reviews, vol. 91, pages 18–40, 2018. (Cited in pages 111 and 180.)

[Pillai 2019a] Pillai, D. S., Blaabjerg, F. and Rajasekar, N. *A Comparative Evaluation of Advanced Fault Detection Approaches for PV Systems.* IEEE Journal of Photovoltaics, vol. 9, no. 2, pages 513–527, 2019. (Cited in pages 21, 180, and 227.)

[Pillai 2019b] Pillai, D. S., Blaabjerg, F. and Rajasekar, N. *A Comparative Evaluation of Advanced Fault Detection Approaches for PV Systems.* IEEE Journal of Photovoltaics, vol. 9, no. 2, pages 513–527, 2019. [Online]. Available: http://dx.doi.org/10.1109/JPHOTOV.2019.2892189. (Cited in pages 67 and 92.)

[Pingel 2010] Pingel, S., Frank, O., Winkler, M., Daryan, S., Geipel, T., Hoehne, H. and Berghold, J. *Potential Induced Degradation of solar cells and panels.* In 2010 35th IEEE Photovoltaic Specialists Conference, pages 002817–002822, 2010. (Cited in page 68.)

[Pise 2008] Pise, N. N. and Kulkarni, P. *A Survey of Semi-Supervised Learning Methods.* In 2008 International Conference on Computational Intelligence and Security, volume 2, pages 30–34, 2008. (Cited in page 111.)

[Polo 2017] Polo, J., Fernandez-Neira, W. and Alonso-García, M. *On the use of reference modules as irradiance sensor for monitoring and modelling rooftop PV systems.* Renewable Energy, vol. 106, pages 186–191, 2017. (Cited in page 135.)

[Prema 2015] Prema, V. and Uma Rao, K. *Time series decomposition model for accurate wind speed forecast.* Prema and Rao Renewables, vol. 2, no. 18, pages 2 – 11, 2015. (Cited in page 143.)

[Purwadi 2011] Purwadi, A., Haroen, Y., Ali, F. Y., Heryana, N., Nurafiat, D. and Assegaf, A. *Prototype development of a Low Cost data logger for PV based LED Street Lighting System.* In Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, pages 1–5, 2011. (Cited in pages 145 and 146.)

[PVG 2022] *PVGIS Photovoltaic Geographical Information System*, 2022. Accessed: 2022-09-30. (Cited in pages 18 and 156.)

[Qadir 2019] Qadir, K. W., Zafar, Q., Ebrahim, N. A., Ahmad, Z., Sulaiman, K., Akram, R. and Nazeeruddin, M. K. *Methodical review of the literature referred to the dye-sensitized solar cells: Bibliometrics analysis and road mapping.* Chinese Physics B, vol. 28, no. 11, page 118401, October 2019. (Cited in page 96.)

[Qibin Zhao 2005] Qibin Zhao and Liqing Zhang. *ECG Feature Extraction and Classification Using Wavelet Transform and Support Vector Machines.* In 2005 International Conference on Neural Networks and Brain, volume 2, pages 1089–1092, 2005. (Cited in page 143.)

[Rabii 2003] Rabii, A., Jraidi, M. and Bouazzi, A. *Investigation of the Degradation in Field-Aged Photovoltaic Modules.* pages 2004 – 2006 Vol.2, 06 2003. (Cited in page 61.)

[Rabla 2013] Rabla, M., Tisserand, E., Schweitzer, P. and Lezama, J. *Arc Fault Analysis and Localisation by Cross-Correlation in 270 V DC.* In IEEE 59th Holm Conference on Electrical Contacts (HOLM), pages 1–6, 2013. (Cited in pages 2, 30, and 34.)

[Rafiee 2009] Rafiee, J. and Tse, P. *Use of autocorrelation of wavelet coefficients for fault diagnosis.* Mechanical Systems and Signal Processing, vol. 23, no. 5, pages 1554 – 1572, 2009. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0888327009000685`. (Cited in page 144.)

[Rajput 2019] Rajput, P., Malvoni, M., Kumar, N. M., Sastry, O. and Tiwari, G. *Risk priority number for understanding the severity of photovoltaic failure modes and their impacts on performance degradation.* Case Studies in Thermal Engineering, vol. 16, page 100563, 2019. (Cited in pages vi and 73.)

[Rakshikar 2015] Rakshikar, N. *Zotero: an ultimate citation management tool for researchers and academicians.* In UGC Sponsored Two Day's National Seminar on Emerging Trends in Library Technology, pages 1–6, 08 2015. (Cited in page 94.)

[Ramírez 2021] Ramírez, I. S., Chaparro, J. R. P. and Márquez, F. P. G. *Machine Learning techniques implemented in IoT platform for fault detection in photovoltaic panels.* In 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), pages 429–434, 2021. (Cited in page 92.)

[Rani 2012] Rani, S. and Sikka, G. *Article: Recent Techniques of Clustering of Time Series Data: A Survey.* International Journal of Computer Applications, vol. 52, no. 15, pages 1–9, 2012. (Cited in page 219.)

[Rani 2013] Rani, B. I., Ilango, G. S. and Nagamani, C. *Enhanced Power Generation From PV Array Under Partial Shading Conditions by Shade Dis-*

persion Using Su Do Ku Configuration. IEEE Transactions on Sustainable Energy, vol. 4, no. 3, pages 594–601, 2013. [Online]. Available: http://dx.doi.org/10.1109/TSTE.2012.2230033. (Cited in page 68.)

[Rao 2021] Rao, B. *The Role of Artificial Intelligence (AI) in Condition Monitoring and Diagnostic Engineering Management (COMADEM): A Literature Survey.* American Journal of Artificial Intelligence, vol. 5, page 17, 01 2021. [Online]. Available: http://dx.doi.org/10.11648/j.ajai.20210501.12. (Cited in page 116.)

[Ray 2016] Ray, P. K., Panigrahi, B. K., Rout, P. K., Mohanty, A. and Dubey, H. *Detection of Faults in Power System Using Wavelet Transform and Independent Component Analysis.* In First International Conference on Advancement of Computer Communication & Electrical Technology, pages 1–5, 2016. (Cited in page 184.)

[Ray 2018] Ray, P. K., Mohanty, A., Panigrahi, B. K. and Rout, P. K. *Modified wavelet transform based fault analysis in a solar photovoltaic system.* Optik, vol. 168, pages 754–763, 2018. (Cited in pages 29, 183, 184, 186, and 187.)

[Reddy 2021] Reddy, O. Y., Chatterjee, S. and Chakraborty, A. K. *Bilayered fault detection and classification scheme for low-voltage DC microgrid with weighted KNN and decision tree.* International Journal of Green Energy, vol. 0, no. 0, pages 1–11, 2021. (Cited in page 124.)

[Redmon 2015] Redmon, J., Divvala, S. K., Girshick, R. B. and Farhadi, A. *You Only Look Once: Unified, Real-Time Object Detection.* CoRR, vol. abs/1506.02640, 2015. (Cited in page 117.)

[Ren 2015] Ren, S., He, K., Girshick, R. B. and Sun, J. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.* CoRR, vol. abs/1506.01497, 2015. [Online]. Available: `http://arxiv.org/abs/1506.01497`. (Cited in page 118.)

[Reyes-Belmonte 2020] Reyes-Belmonte, M. A. *A Bibliometric Study on Integrated Solar Combined Cycles (ISCC), Trends and Future Based on Data Analytics Tools.* Sustainability, vol. 12, no. 19, page 8217, 2020. (Cited in page 96.)

[Rezgui 2014] Rezgui, W., Mouss, N. K., Mouss, L.-H., Mouss, M. D., Amirat, Y. and Benbouzid, M. *Faults modeling of the impedance and reversed polarity types within the PV generator operation.* In 3rd International Symposium on Environmental Friendly Energies and Applications (EFEA), pages 1–6, 2014. (Cited in page 71.)

[Rezk 2017] Rezk, H., Tyukhov, I., Al-Dhaifallah, M. and Tikhonov, A. *Performance of data acquisition system for monitoring PV system parameters.* Measurement, vol. 104, pages 204–211, 2017. (Cited in page 147.)

[Rip 1984] Rip, A. and Courtial, J. P. *Co-word maps of biotechnology: An example of cognitive scientometrics.* Scientometrics, vol. 6, pages 381–400, 1984. (Cited in page 98.)

[Riza Alvy Syafi'i 2018] Riza Alvy Syafi'i, M. H., Prasetyono, E., Khafidli, M. K., Anggriawan, D. O. and Tjahjono, A. *Real Time Series DC Arc Fault Detection Based on Fast Fourier Transform.* In 2018 International Electronics Symposium on Engineering Technology and Applications (IES-ETA), pages 25–30, 2018. (Cited in page 142.)

[Röder 2015] Röder, M., Both, A. and Hinneburg, A. *Exploring the Space of Topic Coherence Measures.* In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15, page 399–408, New York, NY, USA, 2015. Association for Computing Machinery. (Cited in page 101.)

[Rodrigues 2017] Rodrigues, S., Ramos, H. G. and Morgado-Dias, F. *Machine Learning in PV Fault Detection, Diagnostics and Prognostics: A Review.* In 2017 IEEE 44th Photovoltaic Specialist Conference (PVSC), pages 3178–3183, Washington, DC, 2017. IEEE. (Cited in pages 106, 108, 109, and 125.)

[Romero-Cadaval 2015] Romero-Cadaval, E., Francois, B., Malinowski, M. and Zhong, Q.-C. *Grid-Connected Photovoltaic Plants: An Alternative Energy Source, Replacing Conventional Sources.* IEEE Industrial Electronics Magazine, vol. 9, no. 1, pages 18–32, 2015. (Cited in page 29.)

[Ross 2008] Ross, D., Lim, J., Lin, R. and Yang, M. *Incremental Learning for Robust Visual Tracking.* Int J Comput Vis, vol. 77, page 125–141, 2008. (Cited in page 192.)

[Rousseeuw 1987] Rousseeuw, P. J. *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, vol. 20, pages 53–65, 1987. (Cited in page 105.)

[Roweis 2000] Roweis, S. T. and Saul, L. K. *Nonlinear Dimensionality Reduction by Locally Linear Embedding.* Science, vol. 290, no. 5500, pages 2323–2326, 2000. (Cited in page 192.)

[Ruan 2017] Ruan, Y., Xue, X., Liu, H., Tan, J. and Li, X. *Quantum Algorithm for K-Nearest Neighbors Classification Based on the Metric of Hamming Distance.* International Journal of Theoretical Physics, vol. 56, 11 2017. (Cited in page 203.)

[Ryu 2019] Ryu, S., Nguyen, D. C., Ha, N., Park, H. J., Ahn, Y., Park, J.-Y. and Lee, S. *Light Intensity-dependent Variation in Defect Contributions to Charge Transport and Recombination in a Planar MAPbI3 Perovskite Solar Cell.* Scientific Reports, vol. 9, page 19846, 12 2019. [Online]. Available: http://dx.doi.org/10.1038/s41598-019-56338-6. (Cited in page 50.)

[S 1993] S, D. and Bowling, D. *Field experience with photovoltaic systems: ten-year assessment.* Technical Report, EPRI, 1993. (Cited in page 55.)

[S 2021] S, N. V. and Sugumaran, V. *Fault diagnosis of visual faults in photovoltaic modules: A Review.* International Journal of Green Energy, vol. 18, no. 1, pages 37–50, 2021. (Cited in page 21.)

[S4E 2022] S4E. *EnergySoft.* 2022. [Online]. Available: `https://www.s4e.fr/solutions-pvsoft/`. (Cited in pages 4 and 35.)

[Sagi 2018] Sagi, O. and Rokach, L. *Ensemble learning: A survey.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, page e1249, 2018. (Cited in page 198.)

[Saheb 2022a] Saheb, T., Dehghani, M. and Saheb, T. *Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis.* Sustainable Computing: Informatics and Systems, vol. 35, page 100699, 2022. (Cited in page 98.)

[Saheb 2022b] Saheb, T. and Dehghani, M. *Artificial intelligence for Sustainable Energy: A Contextual Topic Modeling and Content Analysis.* ArXiv, vol. abs/2110.00828, 2022. (Cited in page 100.)

[Saleh 2009] Saleh, M., Othman, Z. and Saleh, M. S. *Characteristics of agent-based hierarchical diff-EDF schedulability over heterogeneous real-time Packet networks.* European journal of scientific research, vol. 27, pages 431–453, 2009. (Cited in page 219.)

[Samko 2006] Samko, O., Marshall, A. and Rosin, P. *Selection of the optimal parameter value for the Isomap algorithm.* Pattern Recognition Letters, vol. 27, no. 9, pages 968–979, 2006. (Cited in page 194.)

[Sammour 2019] Sammour, M., Othman, Z. A., Rus, A. M. M. and Mohamed, R. *Modified Dynamic Time Warping for Hierarchical Clustering.* International Journal on Advanced Science, Engineering and Information Technology, vol. 9, no. 5, pages 1481–1487, 2019. (Cited in page 219.)

[Sangeetha 2018] Sangeetha, N. and Anita, X. *Entropy based texture watermarking using discrete wavelet transform.* Optik, vol. 160, pages 380–388, 2018. (Cited in page 183.)

[Santiago 2017] Santiago, I., Trillo Montero, D., Luna RodrÃguez, J. J., Moreno Garcia, I. M. and Palacios Garcia, E. J. *Graphical Diagnosis of Performances in Photovoltaic Systems: A Case Study in Southern Spain.* Energies, vol. 10, no. 12, 2017. (Cited in page 234.)

[Sathik 2021] Sathik, J., Aleem, S. H. E. A., Shalchi Alishah, R., Almakhles, D., Bertilsson, K., Bhaskar, M. S., Fernandez Savier, G. and Dhandapani, K. *A*

Multilevel Inverter Topology Using Diode Half-Bridge Circuit with Reduced Power Component. Energies, vol. 14, no. 21, 2021. (Cited in page 54.)

[Saxena 2017] Saxena, A., Surana, S. L. and Saini, D. *Hybrid Approach of Additive and Multiplicative Decomposition Method for Electricity Price Forecasting.* Skit research journal, vol. 7, no. 1, pages 13 – 20, 2017. (Cited in page 143.)

[Scharf 1991] Scharf, L. and Demeure, C. Statistical signal processing: detection, estimation, and time series analysis. Prentice Hall, 1991. (Cited in page 144.)

[Schirone 1994] Schirone, L., Califano, F. P. and Pastena, M. *Fault detection in a photovoltaic plant by time domain reflectometry.* Progress in Photovoltaics: Research and Applications, vol. 2, no. 1, pages 35–44, 1994. (Cited in page 86.)

[Schirripa Spagnolo 2012] Schirripa Spagnolo, G., Del Vecchio, P., Makary, G., Papalillo, D. and Martocchia, A. *A review of IR thermography applied to PV systems.* In 2012 11th International Conference on Environment and Electrical Engineering, pages 879–884, 2012. (Cited in pages 64, 65, 71, and 72.)

[Schmidhuber 2015] Schmidhuber, J. *Deep learning in neural networks: An overview.* Neural Networks, vol. 61, pages 85–117, 2015. (Cited in page 118.)

[Schölkopf 1999] Schölkopf, B., Smola, A. J. and Müller, K.-R. Kernel principal component analysis, page 327–352. MIT Press, Cambridge, MA, USA, 1999. (Cited in page 192.)

[Schvaneveldt 1988] Schvaneveldt, R., Dearholt, D. and Durso, F. *Graph theoretic foundations of pathfinder networks.* Computers & Mathematics with Applications, vol. 15, no. 4, pages 337–345, 1988. (Cited in page 97.)

[Seaton 2021] Seaton, H. The construction technology handbook. John Wiley & Sons, Hoboken, NJ, USA, 2021. (Cited in page 123.)

[Selim 1984] Selim, S. Z. and Ismail, M. A. *K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-6, no. 1, pages 81–87, 1984. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.1984.4767478. (Cited in page 258.)

[Sepúlveda Oviedo 2021] Sepúlveda Oviedo, E. H., Travé-Massuyès, L., Subias, A., Alonso, C. and Pavlov, M. *Hierarchical clustering and dynamic time warping for fault detection in photovoltaic systems.* In X Congreso internacional Ingeniería Mecánica, Mecatrónica y Automatización, Bogotá, Colombia, May 2021. (Cited in pages 123, 127, and 180.)

[Sepúlveda Oviedo 2022] Sepúlveda Oviedo, E. H., Travé-Massuyès, L., Subias, A., Alonso, C. and Pavlov, M. *Feature extraction and health status prediction in*

PV systems. Advanced Engineering Informatics, vol. 53, page 101696, 2022. (Cited in pages viii, ix, 2, 13, 29, 34, 90, 125, 127, 180, 198, 199, 201, 209, and 229.)

[Serin 2020] Serin, G., Sener, B., Ozbayoglu, M. and Unver, H. O. *Review of tool condition monitoring in machining and opportunities for deep learning*. The International Journal of Advanced Manufacturing Technology, vol. 109, 07 2020. [Online]. Available: http://dx.doi.org/10.1007/s00170-020-05449-w. (Cited in page 238.)

[Seyedmahmoudian 2016] Seyedmahmoudian, M., Horan, B., Soon, T. K., Rahmani, R., Than Oo, A. M., Mekhilef, S. and Stojcevski, A. *State of the art artificial intelligence-based MPPT techniques for mitigating partial shading effects on PV systems – A review*. Renewable and Sustainable Energy Reviews, vol. 64, pages 435–455, 2016. (Cited in page 29.)

[Shahsavari 2018] Shahsavari, A. and Akbari, M. *Potential of solar energy in developing countries for reducing energy-related emissions*. Renewable and Sustainable Energy Reviews, vol. 90, pages 275–291, 2018. (Cited in page 29.)

[Shaik 2015] Shaik, A. G. and Pulipaka, R. R. V. *A new wavelet based fault detection, classification and location in transmission lines*. International Journal of Electrical Power & Energy Systems, vol. 64, pages 35–40, 2015. (Cited in page 184.)

[Shariff 2013] Shariff, F., Rahim, N. A. and Ping, H. W. *Photovoltaic remote monitoring system based on GSM*. In 2013 IEEE Conference on Clean Energy and Technology (CEAT), pages 379–383, 2013. (Cited in pages 137 and 138.)

[Shariff 2015] Shariff, F., Rahim, N. A. and Hew, W. P. *Zigbee-based data acquisition system for online monitoring of grid-connected photovoltaic system*. Expert Systems with Applications, vol. 42, no. 3, pages 1730–1742, 2015. (Cited in pages 136, 137, and 147.)

[Sharma 2013] Sharma, V. and Chandel, S. *Performance and degradation analysis for long term reliability of solar photovoltaic systems: A review*. Renewable and Sustainable Energy Reviews, vol. 27, pages 753–767, 2013. (Cited in page 64.)

[Sharma 2016] Sharma, A., Amarnath, M. and Kankar, P. *Feature extraction and fault severity classification in ball bearings*. Journal of Vibration and Control, vol. 22, no. 1, pages 176–192, 2016. (Cited in page 186.)

[Shen 2021] Shen, Y., Ji, L., Xie, Y., Huang, G., Li, X. and Huang, L. *Research landscape and hot topics of rooftop PV: A bibliometric and network analysis*. Energy and Buildings, vol. 251, page 111333, 2021. (Cited in pages 30, 94, and 96.)

[Shimakage 2011] Shimakage, T., Nishioka, K., Yamane, H., Nagura, M. and Kudo, M. *Development of fault detection system in PV system.* In 2011 IEEE 33rd International Telecommunications Energy Conference (INTELEC), pages 1–5, 2011. (Cited in page 88.)

[Shobha 2018] Shobha, G. and Rangaswamy, S. *Chapter 8 - Machine Learning.* In Gudivada, V. N. and Rao, C., editors, Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications, volume 38 of *Handbook of Statistics*, pages 197–228. Elsevier, 2018. (Cited in page 110.)

[Shorten 2019] Shorten, C. and Khoshgoftaar, T. *A survey on Image Data Augmentation for Deep Learning.* Journal of Big Data, vol. 6, 07 2019. [Online]. Available: http://dx.doi.org/10.1186/s40537-019-0197-0. (Cited in page 142.)

[Shtovba 2020] Shtovba, S. and Petrychko, M. *Jaccard index-Based Assessing the Similarity of Research Fields in Dimensions.* 01 2020. (Cited in page 97.)

[Silvestre 2009] Silvestre, S., Boronat, A. and Chouder, A. *Study of bypass diodes configuration on PV modules.* Applied Energy, vol. 86, no. 9, pages 1632–1640, 2009. (Cited in page 51.)

[Silvestre 2013] Silvestre, S., Chouder, A. and Karatepe, E. *Automatic fault detection in grid connected PV systems.* Solar Energy, vol. 94, pages 119–127, 2013. (Cited in page 88.)

[Simon 2010] Simon, M. and Meyer, E. L. *Detection and analysis of hot-spot formation in solar cells.* Solar Energy Materials and Solar Cells, vol. 94, no. 2, pages 106–113, 2010. (Cited in page 65.)

[Singh 2006] Singh, B. N. and Tiwari, A. K. *Optimal selection of wavelet basis function applied to ECG signal denoising.* Digital Signal Processing, vol. 16, no. 3, pages 275–287, 2006. (Cited in page 184.)

[Skoplaki 2009a] Skoplaki, E. and Palyvos, J. *On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations.* Solar Energy, vol. 83, no. 5, pages 614–624, 2009. (Cited in page 235.)

[Skoplaki 2009b] Skoplaki, E. and Palyvos, J. *Operating temperature of photovoltaic modules: A survey of pertinent correlations.* Renewable Energy, vol. 34, no. 1, pages 23–29, 2009. (Cited in page 237.)

[SMA 2022] SMA. *Sunny Tripower.* 2022. [Online]. Available: `https://www.sma.de/fr/produits/surveillance-controle.html`. (Cited in pages 4 and 35.)

[Soheily-Khah 2016] Soheily-Khah, S., Douzal-Chouakria, A. and Gaussier, E. *Generalized k-means-based clustering for temporal data under weighted and kernel time warp.* Pattern Recognition Letters, vol. 75, pages 63–69, 2016. (Cited in page 258.)

[Solórzano 2013] Solórzano, J. and Egido, M. *Automatic fault diagnosis in PV systems with distributed MPPT.* Energy Conversion and Management, vol. 76, pages 925–934, 2013. (Cited in pages 64, 65, 66, 72, and 88.)

[Song 2017] Song, Z., Sun, J. and Yu, J. *Object Tracking by a Combination of Discriminative Global and Generative Multi-Scale Local Models.* Information, vol. 8, no. 2, 2017. (Cited in page 238.)

[Spanias 2017] Spanias, A. S. *Solar energy management as an Internet of Things (IoT) application.* In 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA), pages 1–4, 2017. (Cited in page 125.)

[Spataru 2015] Spataru, S., Sera, D., Kerekes, T. and Teodorescu, R. *Diagnostic method for photovoltaic systems based on light I–V measurements.* Solar Energy, vol. 119, pages 29–44, 2015. (Cited in page 87.)

[SPE 2022] SPE. *Global Market Outlook For Solar Power 2022 - 2026.* Technical Report, SolarPower Europe, 2022. (Cited in pages v, 1, 2, 42, 43, 44, and 45.)

[Spooner 2008] Spooner, E. and Wilmot, N. *SAFETY ISSUES, ARCING AND FUSING IN PV ARRAYS.* 01 2008. (Cited in page 70.)

[Stauffer 2015] Stauffer, Y., Ferrario, D., Onillon, E. and Hutter, A. *Power monitoring based photovoltaic installation fault detection.* In 2015 International Conference on Renewable Energy Research and Applications (ICRERA), pages 199–202, 2015. (Cited in page 88.)

[Steed 2012] Steed, C. A., Shipman, G., Thornton, P., Ricciuto, D., Erickson, D. and Branstetter, M. *Practical Application of Parallel Coordinates for Climate Model Analysis.* Procedia Computer Science, vol. 9, pages 877–886, 2012. (Cited in page 189.)

[Stellbogen 1993] Stellbogen, D. *Use of PV circuit simulation for fault detection in PV array fields.* In Conference Record of the Twenty Third IEEE Photovoltaic Specialists Conference - 1993 (Cat. No.93CH3283-9), pages 1302–1307, 1993. (Cited in page 87.)

[Stettler 2005] Stettler, S., Toggweiler, P., Wiemken, E., Heidenreich, W., de Keizer, A. C., van Sark, W., Feige, S., M., Schneider, Heilscher, G., Lorenz, E., Drews, A. R., Heinemann, D. and Beyer, H.-G. *FAILURE DETECTION ROUTINE FOR GRID CONNECTED PV SYSTEMS AS PART OF THE PVSAT-2 PROJECT.* 2005. (Cited in page 66.)

[STM 2022] *Discovery kit with STM8L152C6 MCU*, 2022. Accessed: 2022-09-30. (Cited in page 162.)

[Stoffel 2012] Stoffel, T., Gotseff, P. and Sengupta, M. *Evaluation of Photodiode and Thermopile Pyranometers for Photovoltaic Applications.* Technical Report, National Renewable Energy Lab.(NREL), Golden, CO, USA, 2012. (Cited in page 134.)

[Strobl 2010] Strobl, C. and Meckler, P. *Arc Faults in Photovoltaic Systems.* In Proceedings of the 56th IEEE Holm Conference on Electrical Contacts (HOLM), pages 1–7, 2010. (Cited in pages 2, 30, and 34.)

[Strohkendl 2010] Strohkendl, K., Herrmann, W., Vaassen, W., Althaus, J. and Reil, F. *The Effect of Transportation Impacts and Dynamic Load Tests on the Mechanical and Electrical Behaviour of Crystalline PV Modules.* In 25th European Photovoltaic Solar Energy Conference and Exhibition / 5th World Conference on Photovoltaic Energy Conversion, page 3989–3992, 2010. (Cited in page 56.)

[Sudiharto 2017] Sudiharto, I., Tjahjono, A. and Anggriawan, D. O. *Harmonic load identification based on fast fourier transform and levenberg marquardt backpropagation.* Journal of Theoretical and Applied Information Technology, vol. 95, no. 5, pages 1080–1087, 2017. [Online]. Available: http://dx.doi.org/10.1109/TPWRD.2015.2392385. (Cited in page 142.)

[Suryanto Hasyim 1986] Suryanto Hasyim, E., Wenham, S. and Green, M. *Shadow tolerance of modules incorporating integral bypass diode solar cells.* Solar Cells, vol. 19, no. 2, pages 109–122, 1986. (Cited in page 51.)

[Syafaruddin 2009] Syafaruddin, Hiyama, T. and Karatepe, E. *Feasibility of Artificial Neural Network for Maximum Power Point Estimation of Non Crystalline-Si Photovoltaic Modules.* In 2009 15th International Conference on Intelligent System Applications to Power Systems, pages 1–6, 2009. (Cited in page 114.)

[Syafaruddin 2011] Syafaruddin, Karatepe, E. and Hiyama, T. *Controlling of artificial neural network for fault diagnosis of photovoltaic array.* In 2011 16th International Conference on Intelligent System Applications to Power Systems, pages 1–6, 2011. (Cited in pages 113 and 114.)

[Syed 2017] Syed, S. and Spruit, M. *Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation.* In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 165–174, 2017. (Cited in pages 101 and 102.)

[Syed 2020] Syed, M. A., Dhomane, G. A. and Nandankar, P. V. *A review of Single-Phase Inverter Topology for Grid Connected Small Distributed Renewable Energy Generation.* 2020. (Cited in page 54.)

[Sánchez-Friera 2011] Sánchez-Friera, P., Piliougine, M., Peláez, J., Carretero, J. and Sidrach de Cardona, M. *Analysis of degradation mechanisms of crystalline silicon PV modules after 12 years of operation in Southern Europe.* Progress in Photovoltaics: Research and Applications, vol. 19, no. 6, pages 658–666, 2011. (Cited in page 72.)

[Takashima 2006] Takashima, T., Yamaguchi, J., Otani, K., Kato, K. and Ishida, M. *Experimental Studies of Failure Detection Methods in PV Module Strings.* In 2006 IEEE 4th World Conference on Photovoltaic Energy Conference, volume 2, pages 2227–2230, 2006. (Cited in page 86.)

[Takashima 2009] Takashima, T., Yamaguchi, J., Otani, K., Oozeki, T., Kato, K. and Ishida, M. *Experimental studies of fault location in PV module strings.* Solar Energy Materials and Solar Cells, vol. 93, no. 6, pages 1079–1082, 2009. 17th International Photovoltaic Science and Engineering Conference. (Cited in page 86.)

[Tanaka 2016] Tanaka, Y. and Takahashi, M. *Dynamic time warping-based cluster analysis and support vector machine-based prediction of solar irradiance at multi-points in a wide area.* International Symposium on Stochastic Systems Theory and its Applications (ISCIE), pages 210–215, 2016. (Cited in pages 218 and 219.)

[Tang 2014] Tang, L., Peng, S., Bi, Y., Shan, P. and Hu, X. *A New Method Combining LDA and PLS for Dimension Reduction.* PLOS ONE, vol. 9, no. 5, pages 1–10, 2014. (Cited in page 224.)

[Tang 2020] Tang, W., Yang, Q., Xiong, K. and Yan, W. *Deep learning based automatic defect identification of photovoltaic module using electroluminescence images.* Solar Energy, vol. 201, pages 453–460, 2020. (Cited in page 239.)

[Tao 2020] Tao, C., Wang, X., Gao, F. and Wang, M. *Fault diagnosis of photovoltaic array based on deep belief network optimized by genetic algorithm.* Chinese Journal of Electrical Engineering, vol. 6, no. 3, pages 106–114, 2020. [Online]. Available: http://dx.doi.org/10.23919/CJEE.2020.000024. (Cited in page 118.)

[Tarabsheh 2011] Tarabsheh, A. and Etier, I. *Analysis of the Ideality Factor of a-Si:H Solar Cells.* Journal of Solar Energy Engineering, vol. 133, page 011012, 02 2011. [Online]. Available: http://dx.doi.org/10.1115/1.4003294. (Cited in page 50.)

[Teh 2006] Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. *Hierarchical Dirichlet Processes.* Journal of the American Statistical Association, vol. 101, no. 476, pages 1566–1581, 2006. (Cited in page 99.)

[Tejwani 2014] Tejwani, R., Kumar, G. and Solanki, C. *Remote Monitoring for Solar Photovoltaic Systems in Rural Application Using GSM Voice Channel.*

Energy Procedia, vol. 57, pages 1526–1535, 2014. 2013 ISES Solar World Congress. (Cited in pages 137 and 138.)

[Tenenbaum 2000a] Tenenbaum, J. B., de Silva, V. and Langford, J. C. *A global geometric framework for nonlinear dimensionality reduction.* Science, vol. 290, no. 5500, pages 2319–23, 2000. (Cited in page 191.)

[Tenenbaum 2000b] Tenenbaum, J. B., de Silva, V. and Langford, J. C. *A Global Geometric Framework for Nonlinear Dimensionality Reduction.* Science, vol. 290, no. 5500, pages 2319–2323, 2000. (Cited in pages 195 and 199.)

[Tex 2022] *LaunchPad development kit*, 2022. Accessed: 2022-09-30. (Cited in page 162.)

[Theis 2016] Theis, L., van den Oord, A. and Bethge, M. *A note on the evaluation of generative models.* In International Conference on Learning Representations, Apr 2016. (Cited in page 238.)

[Tijare 2022] Tijare, P. V. and Prathuri, J. R. *Correlation Between K-means Clustering and Topic Modeling Methods on Twitter Datasets.* In Khanna, K., Estrela, V. V. and Rodrigues, J. J. P. C., editors, Cyber Security and Digital Forensics, pages 459–477, Singapore, 2022. Springer Singapore. (Cited in page 99.)

[Tina 2014] Tina, G. M. and Grasso, A. D. *Remote monitoring system for standalone photovoltaic power plants: The case study of a PV-powered outdoor refrigerator.* Energy Conversion and Management, vol. 78, pages 862–871, 2014. (Cited in page 146.)

[Tipping 1999] Tipping, M. E. and Bishop, C. M. *Probabilistic Principal Component Analysis.* Journal of the Royal Statistical Society, vol. 61, no. 3, pages 611–622, 1999. (Cited in page 191.)

[Tranfield 2003] Tranfield, D., Denyer, D. and Smart, P. *Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review.* British Journal of Management, vol. 14, no. 3, pages 207–222, 2003. (Cited in page 95.)

[Triki-Lahiani 2018a] Triki-Lahiani, A., Bennani-Ben Abdelghani, A. and Slama-Belkhodja, I. *Fault detection and monitoring systems for photovoltaic installations: A review.* Renewable and Sustainable Energy Reviews, vol. 82, pages 2680–2692, 2018. (Cited in pages 58, 66, 92, 132, 133, 136, 138, and 139.)

[Triki-Lahiani 2018b] Triki-Lahiani, A., Bennani-Ben Abdelghani, A. and Slama-Belkhodja, I. *Fault detection and monitoring systems for photovoltaic installations: A review.* Renewable and Sustainable Energy Reviews, vol. 82, pages 2680–2692, 2018. (Cited in page 180.)

[Trillo-Montero 2014] Trillo-Montero, D., Santiago, I., Luna-Rodriguez, J. and Real-Calvo, R. *Development of a software application to evaluate the performance and energy losses of grid-connected photovoltaic systems.* Energy Conversion and Management, vol. 81, pages 144–159, 2014. (Cited in page 234.)

[Tsai 2015] Tsai, D.-M., Li, G.-N., Li, W.-C. and Chiu, W.-Y. *Defect detection in multi-crystal solar cells using clustering with uniformity measures.* Advanced Engineering Informatics, vol. 29, no. 3, pages 419–430, 2015. (Cited in pages 123 and 180.)

[Tsanakas 2016] Tsanakas, J. A., Ha, L. and Buerhop, C. *Faults and infrared thermographic diagnosis in operating c-Si photovoltaic modules: A review of research and future challenges.* Renewable and Sustainable Energy Reviews, vol. 62, pages 695–709, 2016. (Cited in page 180.)

[Turrado 2014] Turrado, C., Meizoso-López, M., Sánchez-Lasheras, F., Rodriguez-Gomez, B., Calvo-Rolle, J. and de Cos Juez, F. *Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions.* Sensors, vol. 14, pages 20382–20399, 11 2014. (Cited in page 141.)

[UL 2018] UL. *Photovoltaic (PV) DC Arc-Fault Circuit Protection.* Technical Report, Underwriters Laboratories Inc., 2018. (Cited in page 118.)

[Upadhyay 2014] Upadhyay, S. and Sharma, M. *A review on configurations, control and sizing methodologies of hybrid energy systems.* Renewable and Sustainable Energy Reviews, vol. 38, pages 47–63, 2014. (Cited in page 29.)

[Upadhyaya 2013] Upadhyaya, H. M., Senthilarasu, S., Hsu, M.-H. and Kumar, D. K. *Recent progress and the status of dye-sensitised solar cell (DSSC) technology with state-of-the-art conversion efficiencies.* Solar Energy Materials and Solar Cells, vol. 119, pages 291–295, 2013. (Cited in page 46.)

[Usman 2017] Usman, M., Ahmed, S., Ferzund, J., Mehmood, A. and Rehman, A. *Using PCA and Factor Analysis for Dimensionality Reduction of Bioinformatics Data.* International Journal of Advanced Computer Science and Applications, vol. 8, no. 5, 2017. (Cited in page 192.)

[UTE 2008] UTE. *INSTALLATIONS ELECTRIQUES A BASSE TENSION GUIDE PRATIQUE Installations photovoltaïques.* Technical Report, 'UNION TECHNIQUE DE L'ELECTRICITE C - 15-712, 2008. (Cited in pages 52 and 53.)

[van der Maaten 2008a] van der Maaten, L. and Hinton, G. *Visualizing Data using t-SNE.* Journal of Machine Learning Research, vol. 9, no. 86, pages 2579–2605, 2008. (Cited in pages 103 and 104.)

[van der Maaten 2008b] van der Maaten, L. and Hinton, G. *Visualizing Data using t-SNE.* Journal of Machine Learning Research, vol. 9, no. 86, pages 2579–2605, 2008. (Cited in page 192.)

[van Eck 2006] van Eck, N. J., Waltman, L., van den Berg, J. and Kaymak, U. *Visualizing the computational intelligence field [Application Notes].* IEEE Computational Intelligence Magazine, vol. 1, no. 4, pages 6–10, 2006. (Cited in pages 97 and 98.)

[Van Eck 2007a] Van Eck, N. J. and Waltman, L. *BIBLIOMETRIC MAPPING OF THE COMPUTATIONAL INTELLIGENCE FIELD.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 15, no. 05, pages 625–645, 2007. (Cited in pages 97 and 98.)

[van Eck 2007b] van Eck, N. J. and Waltman, L. *VOS: A New Method for Visualizing Similarities Between Objects.* In Decker, R. and Lenz, H. J., editors, Advances in Data Analysis, pages 299–306, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. (Cited in page 97.)

[van Eck 2010a] van Eck, N. J. and Waltman, L. *Software survey: VOSviewer, a computer program for bibliometric mapping.* Scientometrics, vol. 84, pages 523–538, 2010. (Cited in pages 96 and 98.)

[van Eck 2010b] van Eck, N. J., Waltman, L., Dekker, R. and van den Berg, J. *A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS.* Journal of the American Society for Information Science and Technology, vol. 61, no. 12, pages 2405–2416, 2010. (Cited in page 97.)

[Van Gompel 2022] Van Gompel, J., Spina, D. and Develder, C. *Satellite based fault diagnosis of photovoltaic systems using recurrent neural networks.* Applied Energy, vol. 305, page 117874, 2022. (Cited in page 118.)

[Vargas-Quesada 2007] Vargas-Quesada, B. and Moya-Anegon, F. Visualizing the structure of science. 01 2007. (Cited in page 97.)

[Veerasamy 2021] Veerasamy, V., Wahab, N. I. A., Othman, M. L., Padmanaban, S., Sekar, K., Ramachandran, R., Hizam, H., Vinayagam, A. and Islam, M. Z. *LSTM Recurrent Neural Network Classifier for High Impedance Fault Detection in Solar PV Integrated Power System.* IEEE Access, vol. 9, pages 32672–32687, 2021. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2021.3060800. (Cited in page 118.)

[Vergura 2009] Vergura, S., Acciani, G., Amoruso, V., Patrono, G. E. and Vacca, F. *Descriptive and Inferential Statistics for Supervising and Monitoring the Operation of PV Plants.* IEEE Transactions on Industrial Electronics, vol. 56, no. 11, pages 4456–4464, 2009. (Cited in page 186.)

[Vergura 2015] Vergura, S., Marino, F. and Carpentieri, M. *Processing infrared image of PV modules for defects classification.* In 2015 International Conference on Renewable Energy Research and Applications (ICRERA), pages 1337–1341, 2015. (Cited in page 84.)

[Verhoeven 1998] Verhoeven, B. *Utility aspects of grid connected photovoltaic power systems.* Technical Report, International Energy Agency IEA, 1998. (Cited in pages 53 and 72.)

[Vighetti 2010] Vighetti, S. *Systèmes photovoltaïques raccordés au réseau : Choix et dimensionnement des étages de conversion.* Theses, Institut National Polytechnique de Grenoble - INPG, September 2010. (Cited in pages 42, 51, and 54.)

[Villagrán 2017] Villagrán, V., Montecinos, A., Franco, C. and Muñoz, R. C. *Environmental monitoring network along a mountain valley using embedded controllers.* Measurement, vol. 106, pages 221–235, 2017. (Cited in page 147.)

[Vyas 2016] Vyas, S., Kumar, R. and Kavasseri, R. *Unsupervised learning in islanding studies: Applicability study for predictive detection in high solar PV penetration distribution feeders.* In 2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering (UPCON), pages 361–366, 2016. (Cited in page 125.)

[Wali 2018] Wali, S. and Areeb, M. *Development of Low-Cost DAQ for Power System Signals Using Arduino.* In 2018 IEEE 21st International Multi-Topic Conference (INMIC), pages 1–5, 2018. (Cited in page 164.)

[Walker 2001] Walker, G. *Evaluating MPPT converter topologies using a Matlab PV Model.* Journal of Electrical and Electronics Engineering, Australia, vol. 21, 01 2001. (Cited in page 59.)

[Walker 2020] Walker, A., Desai, J. and heimiller, D. *Performance of Photovoltaic Systems Recorded by Open Solar Performance and Reliability Clearinghouse (oSPARC).* Technical Report, National Renewable Energy Lab.(NREL), Golden, CO, USA, 2020. (Cited in page 235.)

[Wang 2013] Wang, Z. and Balog, R. S. *Arc fault and flash detection in DC photovoltaic arrays using wavelets.* In IEEE 39th Photovoltaic Specialists Conference (PVSC), pages 1619–1624, 2013. (Cited in pages 2, 23, 34, 143, 183, 184, and 228.)

[Wang 2014a] Wang, Z., McConnell, S., Balog, R. and Johnson, J. *Arc fault signal detection - Fourier transformation vs. wavelet decomposition techniques using synthesized data.* pages 1–6, 2014. (Cited in pages 2, 30, and 34.)

[Wang 2014b] Wang, Z., McConnell, S., Balog, R. S. and Johnson, J. *Arc fault signal detection - Fourier transformation vs. wavelet decomposition techniques*

using synthesized data. In IEEE 40th Photovoltaic Specialist Conference (PVSC), pages 3239–3244, 2014. (Cited in pages 183, 184, and 186.)

[Wang 2015] Wang, X., Zheng, Y., Zhao, Z. and Wang, J. *Bearing Fault Diagnosis Based on Statistical Locally Linear Embedding.* Sensors, vol. 15, no. 7, pages 16225–16247, 2015. (Cited in pages 191 and 192.)

[Wang 2016a] Wang, W., Liu, A. C.-F., Chung, H. S.-H., Lau, R. W.-H., Zhang, J. and Lo, A. W.-L. *Fault Diagnosis of Photovoltaic Panels Using Dynamic Current–Voltage Characteristics.* IEEE Transactions on Power Electronics, vol. 31, no. 2, pages 1588–1599, 2016. (Cited in page 87.)

[Wang 2016b] Wang, Z. and Balog, R. S. *Arc fault and flash detection in photovoltaic systems using wavelet transform and support vector machines.* In 2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC), pages 3275–3280, 2016. (Cited in page 143.)

[Wang 2018a] Wang, F., Yu, Y., Zhang, Z., Li, J., Zhen, Z. and Li, K. *Wavelet Decomposition and Convolutional LSTM Networks Based Improved Deep Learning Model for Solar Irradiance Forecasting.* Applied Sciences, vol. 8, no. 8, 2018. (Cited in pages 143 and 186.)

[Wang 2018b] Wang, J., Zhao, X., Guo, X. and Li, B. *Analyzing the research subjects and hot topics of power system reliability through the Web of Science from 1991 to 2015.* Renewable and Sustainable Energy Reviews, vol. 82, pages 700–713, 2018. (Cited in page 97.)

[Wang 2019] Wang, X., Yu, F., Pedrycz, W. and Yu, L. *Clustering of interval-valued time series of unequal length based on improved dynamic time warping.* Expert Systems with Applications, vol. 125, pages 293–304, 2019. (Cited in page 216.)

[Wang 2020] Wang, Y., Pan, Z. and Pan, Y. *A Training Data Set Cleaning Method by Classification Ability Ranking for the k -Nearest Neighbor Classifier.* IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 5, pages 1544–1556, 2020. [Online]. Available: http://dx.doi.org/10.1109/TNNLS.2019.2920864. (Cited in page 202.)

[Wang 2021a] Wang, J., Lim, M. K., Wang, C. and Tseng, M.-L. *The evolution of the Internet of Things (IoT) over the past 20 years.* Computers & Industrial Engineering, vol. 155, page 107174, 2021. (Cited in page 93.)

[Wang 2021b] Wang, J., Lim, M. K., Wang, C. and Tseng, M.-L. *The evolution of the Internet of Things (IoT) over the past 20Â years.* Computers & Industrial Engineering, vol. 155, page 107174, 2021. (Cited in page 97.)

[Was 2022] *WaspmoteU*, 2022. Accessed: 2022-09-30. (Cited in page 162.)

[Wei 2019] Wei, S., Li, X., Ding, S., Yang, Q. and Yan, W. *Hotspots Infrared detection of photovoltaic modules based on Hough line transformation and Faster-RCNN approach.* In 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), pages 1266–1271, 2019. (Cited in page 118.)

[Weiss 2016] Weiss, K., Khoshgoftaar, T. and Wang, D. *A survey of transfer learning.* Journal of Big Data, vol. 3, 05 2016. [Online]. Available: http://dx.doi.org/10.1186/s40537-016-0043-6. (Cited in page 238.)

[Wescoat 2020] Wescoat, E., Mears, L., Goodnough, J. and Sims, J. *Frequency Energy Analysis in Detecting Rolling Bearing Faults.* Procedia Manufacturing, vol. 48, pages 980 – 991, 2020. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S2351978920315894`. 48th SME North American Manufacturing Research Conference, NAMRC 48. (Cited in page 144.)

[Wichert 2001] Wichert, B., Dymond, M., Lawrance, W. and Friese, T. *Development of a test facility for photovoltaic-diesel hybrid energy systems.* Renewable Energy, vol. 22, no. 1, pages 311–319, 2001. (Cited in page 146.)

[Wohlgemuth 2011] Wohlgemuth, J. H. and Kurtz, S. *Using accelerated testing to predict module reliability.* In 2011 37th IEEE Photovoltaic Specialists Conference, pages 003601–003605, 2011. (Cited in page 29.)

[Wold 1966] Wold, H. *Estimation of principal components and related models by iterative least squares.* 1966. (Cited in page 221.)

[Wold 1982] Wold, H. *Soft modelling: The Basic Design and Some Extensions.* 1982. (Cited in page 221.)

[Woyte 2014] Woyte, A., Richter, M., Moser, D., Reich, N., Green, M., Mau, S. and Beyer, H. *O&M Best Practices Guidelines.* Technical Report, International Energy Agency IEA, 2014. (Cited in pages 140 and 234.)

[Wu 2017] Wu, Y., Chen, Z., Wu, L., Lin, P., Cheng, S. and Lu, P. *An Intelligent Fault Diagnosis Approach for PV Array Based on SA-RBF Kernel Extreme Learning Machine.* Energy Procedia, vol. 105, pages 1070–1076, 2017. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1876610217305039`. 8th International Conference on Applied Energy, ICAE2016, 8-11 October 2016, Beijing, China. (Cited in page 115.)

[Wuest 2016] Wuest, T., Weimer, D., Irgens, C. and Thoben, K.-D. *Machine learning in manufacturing: advantages, challenges, and applications.* Production & Manufacturing Research, vol. 4, no. 1, pages 23–45, 2016. (Cited in pages 14, 90, and 91.)

[Xi 2021] Xi, P., Lin, P., Lin, Y., Zhou, H., Cheng, S., Chen, Z. and Wu, L. *Online Fault Diagnosis for Photovoltaic Arrays Based on Fisher Discrimination Dictionary Learning for Sparse Representation.* IEEE Access, vol. 9, pages 30180–30192, 2021. (Cited in page 128.)

[Xia 2012] Xia, Z., Xia, S., Wan, L. and Cai, S. *Spectral Regression Based Fault Feature Extraction for Bearing Accelerometer Sensor Signals.* Sensors, vol. 12, no. 10, pages 13694–13719, 2012. (Cited in pages 186, 191, and 192.)

[Xia 2015] Xia, K., He, Z., Yuan, Y., Wang, Y. and Xu, P. *An arc fault detection system for the household photovoltaic inverter according to the DC bus currents.* In 2015 18th International Conference on Electrical Machines and Systems (ICEMS), pages 1687–1690, 2015. (Cited in page 70.)

[Xiang 2021] Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C. and Chen, X. *A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data.* Frontiers in Genetics, vol. 12, page 320, 2021. (Cited in page 105.)

[Xie 2020] Xie, Q., Zhang, X., Ding, Y. and Song, M. *Monolingual and multilingual topic analysis using LDA and BERT embeddings.* Journal of Informetrics, vol. 14, no. 3, page 101055, 2020. (Cited in page 99.)

[Xu 2014] Xu, X., Zuo, L. and Huang, Z. *Reinforcement learning algorithms with function approximation: Recent advances and applications.* Information Sciences, vol. 261, pages 1–31, 2014. (Cited in page 122.)

[Xu 2021] Xu, Y., Zhen, D., Gu, J. X., Rabeyee, K., Chu, F., Gu, F. and Ball, A. D. *Autocorrelated Envelopes for early fault detection of rolling bearings.* Mechanical Systems and Signal Processing, vol. 146, page 106990, 2021. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0888327020303769`. (Cited in page 144.)

[Yang 1998] Yang, X., Zhang, Q.-Y. and Han, Y. *Fault signal detection with autocorrelation method.* In Ye, S., editor, Automated Optical Inspection for Industry: Theory, Technology, and Applications II, volume 3558, pages 601 – 606. International Society for Optics and Photonics, SPIE, 1998. (Cited in page 144.)

[Yang 2010] Yang, H., Xu, W., Wang, H. and Narayanan, M. *Investigation of reverse current for crystalline silicon solar cells—New concept for a test standard about the reverse current.* In 2010 35th IEEE Photovoltaic Specialists Conference, pages 002806–002810, 2010. (Cited in page 65.)

[Yang 2022] Yang, N.-C. and Ismail, H. *Voting-Based Ensemble Learning Algorithm for Fault Detection in Photovoltaic Systems under Different Weather Conditions.* Mathematics, vol. 10, no. 2, 2022. [Online]. Available: `https://www.mdpi.com/2227-7390/10/2/285`. (Cited in pages 115 and 116.)

[Yatimi 2014] Yatimi, H. and Aroudam, E. *A Detailed Study and Modeling of Photovoltaic Module under Real Climatic Conditions.* International Journal of Electronics and Electrical Engineering, vol. 3, 01 2014. (Cited in page 50.)

[Yau 2014] Yau, C.-K., Porter, A., Newman, N. and Suominen, A. *Clustering scientific documents with topic modeling.* vol. 100, no. 3, pages 767–786, 2014. (Cited in page 98.)

[Yi 2017a] Yi, Z. and Etemadi, A. H. *Fault Detection for Photovoltaic Systems Based on Multi-Resolution Signal Decomposition and Fuzzy Inference Systems.* IEEE Transactions on Smart Grid, vol. 8, no. 3, pages 1274–1283, 2017. (Cited in pages 183 and 184.)

[Yi 2017b] Yi, Z. and Etemadi, A. H. *Fault Detection for Photovoltaic Systems Based on Multi-Resolution Signal Decomposition and Fuzzy Inference Systems.* IEEE Transactions on Smart Grid, vol. 8, no. 3, pages 1274–1283, 2017. [Online]. Available: http://dx.doi.org/10.1109/TSG.2016.2587244. (Cited in page 192.)

[Yi 2017c] Yi, Z. and Etemadi, A. H. *Line-to-Line Fault Detection for Photovoltaic Arrays Based on Multiresolution Signal Decomposition and Two-Stage Support Vector Machine.* IEEE Transactions on Industrial Electronics, vol. 64, no. 11, pages 8546–8556, 2017. (Cited in pages 2, 13, 29, 34, 90, 184, and 217.)

[Yi 2017d] Yi, Z. and Etemadi, A. H. *Line-to-Line Fault Detection for Photovoltaic Arrays Based on Multiresolution Signal Decomposition and Two-Stage Support Vector Machine.* IEEE Transactions on Industrial Electronics, vol. 64, no. 11, pages 8546–8556, 2017. [Online]. Available: http://dx.doi.org/10.1109/TIE.2017.2703681. (Cited in page 69.)

[Yoon 2019] Yoon, H.-J., Lee, D. and Hovakimyan, N. *Hidden Markov Model Estimation-Based Q-learning for Partially Observable Markov Decision Process.* In 2019 American Control Conference (ACC), pages 2366–2371, 2019. (Cited in page 123.)

[Yordanov 2013] Yordanov, G. H., Midtgård, O.-M. and Saetre, T. O. *Ideality factor behavior between the maximum power point and open circuit.* In 2013 IEEE 39th Photovoltaic Specialists Conference (PVSC), pages 0729–0733, 2013. (Cited in page 50.)

[Youssef 2017] Youssef, A., El-Telbany, M. and Zekry, A. *The role of artificial intelligence in photo-voltaic systems design and control: A review.* Renewable and Sustainable Energy Reviews, vol. 78, pages 72–79, 2017. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1364032117305555`. (Cited in page 112.)

[Yu 2020] Yu, D. and He, X. *A bibliometric study for DEA applied to energy efficiency: Trends and future challenges.* Applied Energy, vol. 268, no. C, 2020. (Cited in page 96.)

[Zajonc 1962] Zajonc, R. B. *A Note on Group Judgements and Group Size.* Human Relations, vol. 15, no. 2, pages 177–180, 1962. (Cited in page 189.)

[Zhang 2004] Zhang, Z. and Zha, H. *Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment.* SIAM Journal on Scientific Computing, vol. 26, no. 1, pages 313–338, 2004. (Cited in page 192.)

[Zhang 2007] Zhang, M.-L. and Zhou, Z.-H. *ML-KNN: A lazy learning approach to multi-label learning.* Pattern Recognition, vol. 40, no. 7, pages 2038–2048, 2007. (Cited in page 202.)

[Zhang 2010] Zhang, G., Xie, S. and Ho, Y.-S. *A bibliometric analysis of world volatile organic compounds research trends.* Scientometrics, vol. 83, pages 477–492, 05 2010. (Cited in page 95.)

[Zhang 2013] Zhang, Y., Liang, M., Li, C., and Hou, S. *A Joint Kurtosis-Based Adaptive Bandstop Filtering and Iterative Autocorrelation Approach to Bearing Fault Detection.* ASME. J, vol. 135, no. 5, pages 270 –274, 2013. (Cited in page 144.)

[Zhang 2014] Zhang, Z. *Too much covariates in a multivariable model may cause the problem of overfitting.* J Thorac Dis, vol. 6, pages E196–E197, 2014. (Cited in page 202.)

[Zhang 2015] Zhang, P., Yan, F. and Du, C. *A comprehensive analysis of energy management strategies for hybrid electric vehicles based on bibliometrics.* Renewable and Sustainable Energy Reviews, vol. 48, pages 88–104, 2015. (Cited in page 96.)

[Zhang 2017] Zhang, Y., Chen, M., Huang, D., Wu, D. and Li, Y. *iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization.* Future Generation Computer Systems, vol. 66, pages 30–35, 2017. (Cited in page 99.)

[Zhang 2020] Zhang, Z., Han, H., Cui, X. and Fan, Y. *Novel application of multi-model ensemble learning for fault diagnosis in refrigeration systems.* Applied Thermal Engineering, vol. 164, page 114516, 2020. (Cited in pages 207, 259, 260, and 261.)

[Zhang 2021a] Zhang, D. and Allagui, A. *Chapter 8 - Fundamentals and performance of solar photovoltaic systems.* In Assad, M. E. H. and Rosen, M. A., editors, Design and Performance Optimization of Renewable Energy Systems, pages 117–129. Academic Press, 2021. (Cited in page 42.)

[Zhang 2021b] Zhang, H., Wang, P., Gao, X., Qi, Y. and Gao, H. *Out-of-sample data visualization using bi-kernel t-SNE.* Information Visualization, vol. 20, no. 1, pages 20–34, 2021. (Cited in pages 103, 104, and 105.)

[Zhang 2021c] Zhang, L., Liu, P., Zhao, L., Wang, G., Zhang, W. and Liu, J. *Air quality predictions with a semi-supervised bidirectional LSTM neural network.* Atmospheric Pollution Research, vol. 12, no. 1, pages 328–339, 2021. (Cited in page 120.)

[Zhang 2021d] Zhang, W., Li, B., Xue, R., Wang, C. and Cao, W. *A systematic bibliometric review of clean energy transition: Implications for low-carbon development.* PLOS ONE, vol. 16, no. 12, page e0261091, 2021. (Cited in page 96.)

[Zhao 2000] Zhao, W., Song, Y. and Min, Y. *Wavelet analysis based scheme for fault detection and classification in underground power cable systems.* Electric Power Systems Research, vol. 53, no. 1, pages 23–30, 2000. (Cited in pages 143 and 184.)

[Zhao 2011a] Zhao, Y., Lehman, B., de Palma, J.-F., Mosesian, J. and Lyons, R. *Challenges to overcurrent protection devices under line-line faults in solar photovoltaic arrays.* In 2011 IEEE Energy Conversion Congress and Exposition, pages 20–27, 2011. (Cited in pages 68, 69, and 71.)

[Zhao 2011b] Zhao, Y., Lehman, B., de Palma, J.-F., Mosesian, J. and Lyons, R. *Fault analysis in solar PV arrays under: Low irradiance conditions and reverse connections.* In 2011 37th IEEE Photovoltaic Specialists Conference, pages 002000–002005, 2011. (Cited in page 68.)

[Zhao 2012] Zhao, Y., Yang, L., Lehman, B., de Palma, J.-F., Mosesian, J. and Lyons, R. *Decision tree-based fault detection and classification in solar photovoltaic arrays.* In Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC), pages 93–99, 2012. (Cited in pages 69, 70, and 113.)

[Zhao 2013a] Zhao, Y., de Palma, J.-F., Mosesian, J., Lyons, R. and Lehman, B. *Line–Line Fault Analysis and Protection Challenges in Solar Photovoltaic Arrays.* IEEE Transactions on Industrial Electronics, vol. 60, no. 9, pages 3784–3795, 2013. (Cited in pages 2, 13, 29, 34, 68, and 90.)

[Zhao 2013b] Zhao, Y., de Palma, J.-F., Mosesian, J., Lyons, R. and Lehman, B. *Line–Line Fault Analysis and Protection Challenges in Solar Photovoltaic Arrays.* IEEE Transactions on Industrial Electronics, vol. 60, no. 9, pages 3784–3795, 2013. [Online]. Available: http://dx.doi.org/10.1109/TIE.2012.2205355. (Cited in page 69.)

[Zhao 2015a] Zhao, Y. *Fault detection, classification and protection in solar photovoltaic arrays.* PhD thesis, Northeastern University, 2015. (Cited in pages 2, 34, 67, and 69.)

[Zhao 2015b] Zhao, Y., Ball, R., Mosesian, J., de Palma, J.-F. and Lehman, B. *Graph-Based Semi-supervised Learning for Fault Detection and Classification in Solar Photovoltaic Arrays.* IEEE Transactions on Power Electronics, vol. 30, no. 5, pages 2848–2858, 2015. (Cited in pages 69, 120, 235, and 237.)

[Zhao 2018] Zhao, X., Wang, S. and Wang, X. *Characteristics and Trends of Research on New Energy Vehicle Reliability Based on the Web of Science.* Sustainability, vol. 10, no. 10, page 3560, 2018. (Cited in page 94.)

[Zhao 2019a] Zhao, H., des Combes, R. T., Zhang, K. and Gordon, G. J. *On Learning Invariant Representation for Domain Adaptation.* CoRR, vol. abs/1901.09453, 2019. (Cited in page 238.)

[Zhao 2019b] Zhao, Y., Liu, Q., Li, D., Kang, D., Lv, Q. and Shang, L. *Hierarchical Anomaly Detection and Multimodal Classification in Large-Scale Photovoltaic Systems.* IEEE Transactions on Sustainable Energy, vol. 10, no. 3, pages 1351–1361, 2019. (Cited in pages 124 and 128.)

[Zhao 2020] Zhao, X., Guo, J., Nie, F., Chen, L., Li, Z. and Zhang, H. *Joint Principal Component and Discriminant Analysis for Dimensionality Reduction.* IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 2, pages 433–444, 2020. (Cited in page 191.)

[Zheng 2017] Zheng, H., Wang, R., Xu, W., Wang, Y. and Zhu, W. *Combining a HMM with a Genetic Algorithm for the Fault Diagnosis of Photovoltaic Inverters.* Journal of Power Electronics, vol. 17, no. 4, pages 1014–1026, 07 2017. (Cited in page 123.)

[Zhou 2007] Zhou, F., Guo, H.-C., Ho, Y.-S. and Wu, C.-Z. *Scientometric Analysis of Geostatistics Using Multivariate Methods.* Scientometrics, vol. 73, pages 265–279, 12 2007. (Cited in page 95.)

[Zhou 2014] Zhou, H., Chen, Q., Li, G., Luo, S., bing Song, T., Duan, H.-S., Hong, Z., You, J., Liu, Y. and Yang, Y. *Interface engineering of highly efficient perovskite solar cells.* Science, vol. 345, no. 6196, pages 542–546, 2014. (Cited in page 46.)

[Zhou 2022] Zhou, X., Zhou, M., Huang, D. and Cui, L. *A probabilistic model for co-occurrence analysis in bibliometrics.* Journal of Biomedical Informatics, vol. 128, page 104047, 2022. (Cited in page 97.)

[Zhu 2018] Zhu, H., Lu, L., Yao, J., Dai, S. and Hu, Y. *Fault diagnosis approach for photovoltaic arrays based on unsupervised sample clustering and probabilistic*

neural network model. Solar Energy, vol. 176, pages 395–405, 2018. (Cited in page 29.)

[Zimmermann 2013] Zimmermann, C. G., Nömayr, C., Kolb, M. and Rucki, A. *A mechanism of solar cell degradation in high intensity, high temperature space missions*. Progress in Photovoltaics: Research and Applications, vol. 21, no. 4, pages 420–435, 2013. (Cited in page 62.)

[Zitt 2000] Zitt, M., Bassecoulard, E. and Okubo, Y. *Shadows of the Past in International Cooperation: Collaboration Profiles of the Top Five Producers of Science*. Scientometrics, vol. 47, pages 627–657, 2000. (Cited in page 98.)

[Zupic 2015] Zupic, I. and Čater, T. *Bibliometric Methods in Management and Organization*. Organizational Research Methods, vol. 18, no. 3, pages 429–472, 2015. (Cited in page 96.)

Résumé : Le diagnostic des défauts est essentiel pour garantir la production continue des systèmes photovoltaïques. Parmi les approches de diagnostic, le diagnostic basé sur les données apprend un modèle de diagnostic à partir d'une base de données de situations normales et défectueuses. Cette approche nécessite un contrôle strict par un système d'acquisition de données pour constituer une base de données substantielle et un algorithme d'apprentissage automatique capable de discriminer non seulement les défauts qui causent des pertes critiques de production, mais également les défauts subtils dont les symptômes doivent être séparés du bruit.

L'objectif de cette thèse est le développement de méthodes de diagnostic de défauts pour les installations photovoltaïques qui puissent s'embarquer sur une plateforme matérielle incluant un système d'acquisition et de traitement de données en temps réel, respectant les contraintes industrielles et prenant en compte le compromis coût/bénéfice en productivité ou en disponibilité. Afin de positionner notre recherche et de connaître les limites actuelles dans le domaine du diagnostic des systèmes photovoltaïques, cette thèse présente d'abord une étude extrêmement approfondie et complète qui construit un état de l'art sur le sujet. Celle se base sur un très grand nombre d'articles en tirant profit de l'analyse bibliométrique et de la modélisation thématique.

Pour résoudre le problème posé et contribuer à un diagnostic efficace des défauts dans les systèmes photovoltaïques, cette thèse propose un cadre matériel/logiciel complet pour le diagnostic qui comprend une nouvelle plate-forme d'acquisition de données de système PV, une station météo mobile polyvalente et des algorithmes d'apprentissage automatique qui, en raison de leur efficacité de calcul et de leurs temps de réponse rapide, peuvent être intégrés dans un système temps-réel.

La plate-forme d'acquisition de données et le logiciel embarqué ont été testés sur plusieurs centrales photovoltaïques et se sont avérés efficaces pour détecter divers défauts critiques dans les panneaux photovoltaïques.

Mots clés : Détection de défauts, Diagnosis, Surveillance, Centrales photovoltaïques, Intelligence artificielle, Apprentissage automatique

Abstract: Fault diagnosis is vital to ensure the continued production of photovoltaic (PV) systems. Among diagnosis approaches, data-driven diagnosis learns a diagnosis model from a database of normal and faulty situations. This approach requires strict control by a data acquisition system to constitute a substantial database and a machine learning algorithm capable of discriminating not only the faults that cause critical production losses, but also subtle faults whose symptoms must be separated from noise.

The objective of this thesis is the development of fault diagnosis methods for photovoltaic installations embedded in a physical system for data acquisition, treatment and detection of faults in real time, respecting industrial limitations and taking into account the cost/benefit compromise in productivity or uptime. In order to position the research and to know the current limitations in the area of diagnosis of PV systems, this thesis first presents an extremely deep and complete study of a large number of articles that builds a state of the art on the subject of interest based on bibliometric analysis and topic modeling.

To address the problem at hand and as a contribution to effective fault diagnosis in photovoltaic systems, this thesis proposes a complete hardware/software framework for fault diagnosis that includes a new PV system data acquisition platform, a versatile mobile weather station, and machine learning algorithms that, due to their computational efficiency and rapid response characteristics, can be embedded in a real system.

The data acquisition platform and embedded software has been tested on several PV plants and it has proved successful in detecting various critical faults in PV panels.

Keywords: Fault detection, Diagnosis, Photovoltaic plants, Supervision, Artificial intelligence, Machine learning, Fault detection fault diagnosis, Data acquisition, Advanced monitoring