PVOP

Solution 3. Concept

Near 100% automatic AI-based system based on a family of algorithm for faults detection and diagnosis and optimisation of resources in PV plants.

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor CINEA can be held responsible for them.

List of Acronyms

AI	Artificial Intelligence
KPI	Key Performance Indicator
PV	Photovoltaic
O&M	Operation and Maintenance
FAD	Fault and Anomalies Diagnosis
PR	Performance Ratio
SVDD	Support Vector Data Description
AGL	Above Ground Level
EVA	Ethylene Vinyl Acetate
EL	Electroluminescence
LL	Logical Layer
ML	Machine Learning
SN	Serial Number
UHR	Ultra-High Resolution
UV	Ultra-Violet

List of Tables

List of Figures

Keywords list

- Artificial Intelligence
- Machine learning
- Algorithm development
- Data Analytics
- Aerial Imagery
- Fault detection
- AI/ML
- Multi-temporal
- Multi-spectral
- High-Resolution

Table of Contents

1. Executive summary

Solar photovoltaic technology has been experiencing exponential growth in recent years and is expected to continue in the coming years. This and other factors mean that its operation is facing new challenges that can only be addressed by digitization and the use of advanced information analysis and management techniques.

The present WP addresses one part of this problem: the automatic diagnosis of faults and anomalies. This automation has great advantages in the operation and maintenance of the PV plants:

- *Considerably reduces the effort of analysing the plant operation, making it more efficient and productive.*
- *Problems are solved more efficiently and quickly, reducing their associated damage.*
- *Plant operators do not waste time on tedious inspections and have prior knowledge of the problem they are facing, reducing unnecessary work and improving workplace safety.*
- *Multiple problems that cannot be located with traditional techniques can be solved, especially those related to predictive maintenance.*

The approach of this project to face this challenge is through the use of AI. This document presents a state of the art on the latest related developments, which serves as a starting point for defining the techniques to be used.

Task 4.1 concept can be found in Chapter A: Near 100% automatic AI-based system based on a family of algorithm for faults detection and diagnosis and optimisation of resources in PV plants.

As the main objective of this chapter, the algorithms to be developed and their development plan are described. These are divided into two main groups with different approaches: *i)* combination of heuristic techniques and exploratory AI algorithms and *ii)* algorithms based on Machine Learning techniques.

The chapter A also describes the data structures that the developed system will have. This involves the definition of the input and output information, the latter having a proposal for classification of fault diagnoses. This definition is key to enable an intercomparison between algorithms.

Finally, the evaluation and comparison methods for the developments are defined. These are the KPIs that allow the evaluation of the algorithms' output in comparison with a known data set. A classification and ranking method will also be proposed to easily see all the developments carried out and which are the best to help in the operation and maintenance of PV solar plants.

Chapter A:

Near 100% automatic AI-based system based on a family of algorithm for faults detection and diagnosis and optimisation of resources in PV plants.

2. Introduction

The photovoltaic market is experiencing exponential growth, with forecasts that this will continue in the coming years. This brings new challenges to the sector, which is increasingly finding it difficult to maintain large parks, in portfolios that continue to grow and in turn in an environment of cost reduction. Digitalization and the use of new technologies are key to maintaining this situation.

This WP focuses on the automated diagnosis of faults and anomalies in power plants based on IA techniques. These techniques make it possible to address most of the challenges presented here and will provide a complete diagnosis solution. The next sections describe the state of the market, the addressed problem in this WP, and the AI-powered algorithm approach.

2.1. The market

According to the International Energy Agency (IEA), more than 413 GW of solar photovoltaic power was installed worldwide in 2023, reaching a cumulative figure of 1,589 GW. This represented a market with a value level of €214 billion. Within renewable energies, solar photovoltaic energy is the leader with 50% of the total installed renewable energy capacity in 2023.

This growth in installed capacity worldwide has been accompanied by an increase in the proportion of electrical energy of photovoltaic origin. During 2023, it is estimated that 5.5% of all electrical energy consumed in the world originated from solar photovoltaic energy. The European average of solar energy penetration has already reached 12% with the following countries above this average: Spain (15%), the Netherlands (14.5%), Germany (12.5%) and Denmark (12.5%).

Within the photovoltaic sector, the submarket for systems for monitoring and optimizing the operation of photovoltaic plants has even greater growth expectations than that for the development of new installations, since their implementation affects both new systems and those already in operation. The strong growth of the market for systems for monitoring and automatic analysis of PV installations is because technology must face new challenges where this type of system is essential: increasingly larger plants, portfolios of PV plants spread throughout the world, massive amounts of data, rapid portfolio growth, etc. There are several companies operating in the sector of systems for monitoring and optimizing PV plants, but in general these systems allow the status of the plant to be viewed and transfer the native alarms generated by the equipment itself, but they do not have true autonomous systems for the detection and diagnosis of faults, which prevents many operating anomalies from being corrected or anticipated and, therefore, leads to significant energy losses, as well as increased O&M costs.

2.2. The problem

The situation of the sector and its expectations brings new challenges in the operation and management of assets. These tasks cannot be carried out without a strong digitalization of the sector and innovation in line with this objective that allows for great automation. In these aspects, these are the main challenges:

1. *Large number of elements of different nature: Plants are increasingly larger, which implies an increase in the elements to be monitored. At the same time, the sector continues to grow technologically, with the appearance of new types of equipment or manufacturers that need to be incorporated into these systems. This implies that the analysis systems must be able to* ***adapt to different types of information****.*
2. *Large amounts of information: These elements generate a large amount of data, increasing in quantity as the operational analyses become more detailed. Analysis systems must be able to* ***manage massive amounts of data*** *and even have the ability to* ***adapt to data availability****.*
3. *Geographical distribution: PV plant portfolios are geographically distributed throughout the world. This implies that there are many sources of information facing different operating situations, requiring analysis systems that have the capacity to* ***adapt to different behaviours****.*
4. *Rapid growth: New PV plants are being incorporated into portfolios very quickly, and plant parks continue to grow, making it difficult to manage these new assets. Therefore, analysis systems must be* ***automatic*** *and able to adapt to new plants, with the* ***ability to learn****.*
5. *Accuracy in results: We must not forget that we are in an industrial environment, specifically in electrical installations, with its respective risks. Analysis systems that deliver useful information to the end user must be* ***accurate and robust*** *so that they provide real value to O&M.*

These challenges can be addressed with extensive digitalisation and automatic analysis. One of the main capabilities this analysis should have is the diagnosis of faults and anomalies. Broadly speaking, this consists of extracting from the massive amount of information coming from the plants, any pattern that may indicate a behaviour that is not correct. This can be done with different techniques, with AI being a great candidate for these tasks. The aim of this specific WP is the development of Faults and Anomalies Diagnosis (FAD) algorithms based on IA that can reach the market. These allow the analysis of large amounts of operating data and considerably help in plant operation tasks, increasing the performance of PV plants and reducing time and costs.

This integration must be focused on helping in the operation of the plants, which involves answering questions related to their status:

- ***Can I trust what I see?*** *The facilities are large with thousands of sensors. These have come from different manufacturers and have been assembled in the construction of the plant. Time passes and not everything continues to operate as it should. The information is not always reliable, it must be verified, many times the problem is the origin of the information.*
- ***Something is not working well?*** *The different elements of the plant may not be working as they should, from poor configuration to part of them breaking. These problems are usually linked to production losses.*
- ***Is there going to be a problem?*** *There are abnormal behaviours or tendencies that can lead to a bigger problem. These do not present a problem at the present, so they are not linked to production problems, but they can end up in greater damage.*

The aim of the development of this WP is to generate a set of algorithms that provide answers to these questions. These will be related to the possible outcomes in Section 4.2.

2.3. The algorithmic approach

Automatic fault detection applied to photovoltaic systems has been a trending research topic in recent years. Classic models used for fault detection are electrical circuit simulation, statistical analysis, electrical signal approaches, predictive models compared to real models, and comparisons between measured and simulated energy yields. Other methods are based on AI techniques, specifically machine learning or deep learning, depending on the data sources and objectives. In view of the many companies announcing specific products for this, one might think that the solution is very advanced. However, most of them are limited to presenting graphic descriptions of the measured variables, analysing the overall performance of the photovoltaic plant by calculating the PR and estimating the most relevant losses (thermal, DC/AC conversion, etc.). This allows a rough analysis (with a margin of uncertainty around 3%) of how a plant meets the profitability expectations placed on it. Unfortunately, it leaves out other more advanced possibilities, such as early fault detection, which are precisely those that allow optimizing operating costs. In addition, faults that can be detected from the combination of different data sources. Currently, this integration is very limited, and it is necessary to resort to several tools, some for photovoltaic plant sensors, others for aerial image campaigns, others para manual works, etc.

Algorithms, including those based on artificial intelligence, can work with information of different nature, which in turn can be structured in many ways. This often makes it difficult to apply algorithms to the same problems, even if they are conceptually very similar. This project seeks to develop a family of algorithms that can solve the challenges described above and be a bridge for future developments in the sector. To achieve this goal, certain bases for the

generalization are established, as comparison and evaluation of these algorithms. On the other hand, algorithms can be of a very varied nature and their technology changes over time.

Current monitoring systems generally incorporate visualization platforms and a native alarm management system. In other words, they transfer the operating alarms given by the equipment itself to the central control platform. In some cases, they also present simple operating indicators (KPI). However, they do not incorporate fault detection systems, do not provide diagnoses of the incidents that occur, nor do they help in the planning of interventions. These tasks are left to the operators, in the case of central installations, or to the owner or maintainer in the case of distributed installations. The absence of advanced Fault and Anomalies Diagnosis (FAD) procedures, that allow the detection, diagnosis and prevention of faults to be automated directly, translates into a higher incidence of anomalies, a longer duration of these and an increase in the severity of their effects. Regarding the control of the operation of the installation, this translates into a decrease in production and, therefore, in the returns of the project.

In the field of maintenance, the objective of incorporating the results of this project is to improve the planning of preventive measures and replace corrective measures with preventive measures as far as possible. In other words, the objective is, on the one hand, to optimize planning, reduce O&M costs and maximize production and, on the other, to anticipate maintenance interventions before incidents occur, which allows extending the life of the equipment by avoiding serious incidents and reducing the cost of maintenance (generally, preventive maintenance is less expensive than corrective maintenance). It must be considered that O&M costs are one of the main charges of a PV plant throughout its useful life, since they represent a recurring outlay throughout the useful life and, therefore, any reduction in these results is an improvement in the profitability of PV projects.

This project proposes a FAD system composed of a family of algorithms capable of detecting and diagnosing faults in PV plants, with a higher level of robustness and capable of working with different data sources at the same time. Due to the breadth of the state of the art, this Working Package will focus on certain techniques that have emerged as promising in previous studies. The main ones that will be used in the project are listed below, separating those that make use of AI and those that do not:

1) Techniques without Artificial Intelligence

These techniques do not require advanced machine learning models or neural networks, but are based on statistical methods, heuristic rules or simpler algorithms. They are useful when data is limited or when the complexity of the system does not justify the use of AI.

a) *Heuristic rules: Based on prior knowledge of the system. These rules are programmed by PV experts and are used to detect common faults. A common technique is Threshold Analysis, predefined limits are set for key parameters (energy production, panel temperature, solar radiation, etc.). If the actual values exceed or fall below these thresholds, a fault is assumed.*
b) *Comparison with an Ideal Model: The performance of the system can be compared with that of an ideal or expected model without using AI. Any significant discrepancy can signal a fault.*

2) Artificial Intelligence Techniques

These techniques use advanced machine learning models to detect complex patterns in data that may not be evident with traditional methods:

a) *Recurrent Neural Networks (RNN): They learn complex patterns in the operating data of PV systems and can predict faults. Recurrent neural networks such as LSTM are very useful for sequential data (such as energy production over time). Example: A neural network that learns the pattern of behaviour of a PV system under normal conditions and detects anomalous deviations that precede faults.*
b) *Convolutional Neural Network (CNN): They were historically employed for image processing applications. The effectiveness of CNNs in fault detection comes from their ability to learn from large datasets, adapt to different fault conditions and improve classification accuracy over time. By using labelled datasets of known faults and normal operating conditions, CNNs can be trained to distinguish between different types of anomalies with high accuracy. This results in faster fault identification and reduced downtime in PV systems. CNNs are characterized by several essential components:*

    i. Convolutional layers: These perform convolution operations on the input data, allowing the network to detect local patterns and features.

    ii. Pooling layers: These layers decrease the dimensionality of the data, emphasizing the most critical features while reducing computational load.

    iii. Fully connected layers: These layers conduct high-level inference using the features extracted by earlier layers.

c) *Anomaly Detection Algorithms (One-Class SVM, Isolation Forest): These algorithms train models with normal operating data to identify outliers. Example: An algorithm that detects anomalies in energy production or inverter behaviour based on historical normal operating data.*
d) *Clustering algorithms (k-means, DBSCAN): Identify groups or clusters of similar performance and highlight anomalous behaviour that could indicate fault. These methods can be adapted to AI when combined with unsupervised learning techniques. Example: Grouping PV panels according to their performance and detecting a group that is operating significantly worse than the rest.*
e) *Transfer learning models or Deep Learning: These allow for the inclusion of an adaptation in the algorithms' own operation phase, making them more robust in integration into new plants.*

Most of these algorithms require a learning process, which will allow them to generate results from historical results. The training process is the stage where the algorithm will learn from the known input-output data. Figure 1 shows a schematic of the process. The input data is processed by the algorithm, its output is evaluated against a known output for that input, and the algorithm is tuned to improve its accuracy.

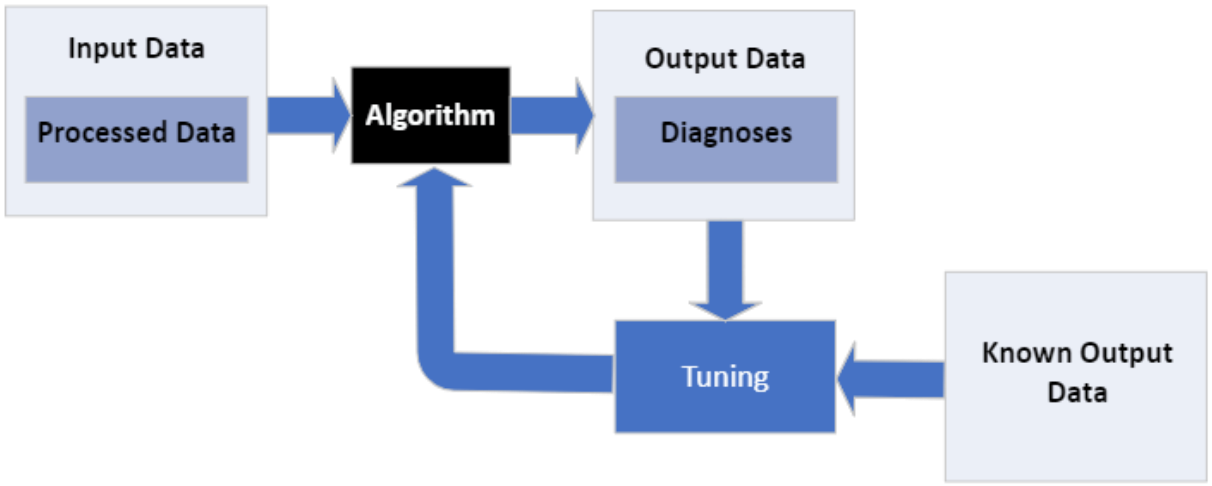

Figure 1. Supervised training process with known data sets.

Once the training process is finished, an evaluation stage usually begins with another set of known data. The algorithms that have been correctly adjusted will pass this evaluation stage and can be used in production. The algorithm can also continue its learning during the operation phase itself, following a scheme like that shown in Figure 2. It should be noted that in this phase, supervision of the learning and the quality of the data set may be difficult to ensure. The learning process can be reduced in this case, attending only to adjustments of some of its components that may allow a certain adaptation to the temporal evolution of the installation. These are cases of reinforcement learning or transfer learning.

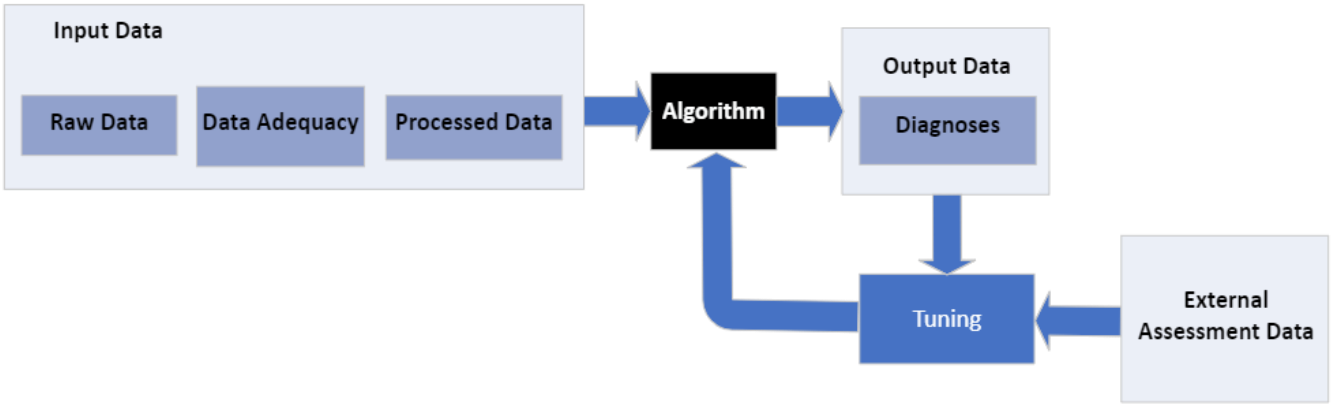

Figure 2. Supervised training process during operation stage.

3. State of the Art

The exponential growth of PV plants means that most of them have been built in the last decade. As they are relatively new installations, they usually have high levels of digitalization in their supervision. These generate a large amount of information that is difficult to manage with traditional techniques and the automation of their analysis is a growing topic of research in the sector [1][2].

During the operation of a plant, numerous faults and anomalies occur, which must be managed by the plant operation for its correct functioning. A fault is understood as any event that causes a loss in the plant's production. Anomalies are incorrect functioning, which may not be related to an energy loss, but which may be indicative of major problems in the future. Due to the large number of plants and their size, the detection of these faults and anomalies can be a complex task for the operation, which benefits considerably from the use of advanced monitoring techniques [3].

The classification of these faults and anomalies has been described in different studies in the literature, there are even standards that describe them, such as IEC 63019 aimed at defining the availability of a plant.

In general, a PV plant is composed of several elements in a hierarchical electrical structure: modules, string, string boxes, inverters and transformers. In parallel to this electrical structure, there would be other elements such as the tracking system or the operating condition sensors. There are multiple problems that can arise, and they are interrelated both hierarchically and unstructured [2].

Faults in PV systems can arise due to various causes, including environmental conditions, manufacturing defects, installation errors, and equipment degradation over time. Prompt detection and resolution of these faults are crucial for optimizing performance, ensuring safety, and extending the system's durability. To this end, various characterization methodologies have been developed to identify defects in PV systems, each with its own capabilities and complexities. Automatic diagnosis strategies for PV faults are generally classified into two main categories: visual inspection and data analysis.

Automated visual inspection is a powerful branch in the analysis of fault and anomaly operations in solar plants. These inspections should be carried out both before and after the modules are exposed to environmental or electrical factors [4]. However, in the following, this work will focus on the second branch, automatic data analysis, specifically in FAD systems.

Fault or anomaly detection from data analysis seeks to find these problems from the sensor data recorded by the different elements of the installation itself. These data are continuously recorded by the plant's monitoring and supervision systems and different algorithms can be applied to them to draw conclusions about their behaviour. The range of techniques that can be used is extensive, from deterministic techniques or those focused on statistical analysis to the introduction of automatic learning mechanisms or artificial intelligence. Although there is advanced and recent work in the field of analysis without the direct use of AI techniques, as is the case of [5] and [6], the aim of this project is to focus on analysis with AI.

In recent years, the development of AI algorithms applied to PV plant operation analysis has grown steadily. The handling of large amounts of data has led to extensive development of Machine Learning techniques that allow machines to learn and improve automatically from experience, without being explicitly programmed to perform specific tasks. Table 1 shows a list of recent academic work related with the use of ML in data analysis for PV plant faults and anomaly detection and diagnosis.

Ref.	Proposed techniques	Type of faults	Sensor data
[7]	- Artificial Neural Network - Stacking ensemble learning	- Open circuit diode - Dust - Shading - Shunted diode	- Solar irradiance - Air temperature - Cell temperature - PV output power - IV curve
[8]	- Graph Neural Network	*- Production faults*	*- DC current* *- AC current*
[9]	- Binary Firefly Algorithm - Naive Bayes-based machine learning	- Partial shading	- DC Voltage - DC Current - Irradiance - Temperature
[10]	- Support vector machine - Continuous Markov model	- Inverter faults - Degradation	- DC Voltages - DC Currents - Harmonics - Waveform - Alerts - Irradiance - Ambient temperature - Module temperature - Humidity - Inverter
[11]	- Multiple regression analysis - Support vector machine	- Dust - Open-circuit - Short-circuit	- Irradiance - Cell temperature - Ambient temperature - Inclinometer - Accelerometer - Strain gauge - DC voltage - DC current - AC grid
[12]	- Contextual information - Linear Regression - Multilayer Perceptron - Random Forest Regressor - Support Vector Regressor - Long Short-Term Memory	- Production faults	- Date - Time - AC and DC Power - Current - Voltage - Inverter generation - Horizontal irradiation

			- Tilt irradiation - Ambient temperature - Module temperature - Wind speed
[13]	- Lasso feature selection - Ensemble learning algorithm - Logistic Regression - Support Vector Machine - k-Nearest Neighbours - Genetic algorithm	- Open-circuit - Line-to-line	- DC Current - DC Voltage
[14]	- Data dimensionality reduction - Random Forest - Regression Trees - k-Nearest Neighbour	- Inverter - Current sensor - Grid anomaly - Partial shading - Open-circuit - MPPT/IPPT controller - Boost converter	- DC Current - DC Voltage
[15]	- Signal processing - Artificial neural network	- Open-circuit	- Irradiance - Temperature - Parametric variation - Irregular distribution of input irradiances
[16]	- Semi-supervised learning - Ensemble learning algorithm - Decision trees - k-Nearest neighbours - Support vector machines	- Arc fault - Line-to-line - Open-circuit - MPPT faults - Partial shading	- DC Voltage - DC Current - Solar irradiance
[17]	- Supervised decision tree - Labelling data	- Shading - Inverter thermal degradation - Fuse burnt - Site outage	- Solar irradiation - Rated capacity - Output power - DC Voltage - Current waveforms
[18]	- Salp Swarm Algorithm - Feature selection - k-Nearest neighbours - Discriminant analysis - Decision tree - Support vector machine	- Line-to-line - Line-to-ground - Connectivity issues - Bypass diodes	- DC Current - DC Voltage
[19]	- Twin model - Decision Tree - Random Forest - k-Nearest Neighbours - Artificial Neural Network	- DC cable degradation - DC cable open-circuit - Switch degradation - Switch open-circuit	- Irradiance - Ambient temperature - DC Currents - DC Voltages - Power
[20]	- Time series analysis	- Random shading	- DC Voltage

	- Support vector machine	- Fixed shading - Aging degradation	- DC Current - Radiation - Temperature
[21]	- Unsupervised vertical federated transfer learning	- Bearing - Gearing	- Vibration signals
[22]	- Transfer learning	- Power losses	- AC signal
[23]	- Convolutional neural networks - Long short-term memory - Bi-directional long short-term memory	- String-to-string - String-to-ground - Open-circuit	- Irradiance - Temperature - String voltage - System current - System voltage
[24]	- Convolutional Neural Networks - LSTM	- String faults	- DC voltage - AC voltage
[25]	- 1D convolutional neural network - IoT	- Short-circuit - Open-circuit - Partial shading - Inverter bypass diode	- Voltage sensors
[26]	- Densely connected convolutional network	- Line-to-line - Open-circuit - Degradation - Partial shading	- IV curve
[27]	- One-Shot Aggregation network - Support Vector Data Description	- Short-circuit - Open-circuit - Shade - Degradation - Dust - Combination faults	- IV curve
[28]	- Convolutional neural network	- Partial shading - Open-circuit - Short-circuit - Degradation - Combination faults	- IV curve
[29]	- Ensemble learning algorithm - Mechanistic Performance Model - Bayesian Neural Networks - eXtreme Gradient Boosting	- Shading - Cell cracks - Inverter open-circuit - Clipping	- In-plane irradiance - Ambient temperature - Wind speed - Wind direction - Module temperature - DC current - DC voltage - DC power - AC power
[30]	- Deep learning algorithm	- Open-circuit - Short-circuit - Shading - Overlapping	- IV curve

[31]	- Heuristic optimization - Convolutional Neural Networks - Bidirectional Gated Recurrent Units	- Open-circuits - Short-circuits - Partial shading	- DC current - DC voltage - Power - Solar irradiation - Module temperature
[32]	- Coyote Optimization Algorithm - Auto-Encoder - Artificial Neural Network	- Open-circuit - Short-circuit	- DC current - DC voltage - Power - Temperature - Irradiance

Table 1. Lits of related works based on Machine Learning techniques.

[7] presents a novel embedded system designed for remote monitoring and fault diagnosis of PV systems. The system integrates machine learning algorithms into a cost-effective edge device for real-time deployment. This study mainly focuses on open circuited diode, dust, shading, and shunted diode errors. Experimental results highlight the system's effectiveness, showcasing its high accuracy in diagnosing and monitoring the PV array's performance.

The authors in [8] propose a fault diagnosis model based on graph neural networks that monitors multiple PV systems by analysing their current and voltage production over the last 24 hours. This approach eliminates the need for dedicated sensors, as the hourly measurements of current and voltage are obtained directly from the PV systems' inverters. Notably, the model achieves high accuracy even without weather data, and further improves when satellite weather estimates are incorporated. Furthermore, the results of this study indicate that the proposed method can generalize well to PV systems it is not specifically trained on and maintain high diagnosis accuracy even when multiple systems are simultaneously impacted by faults.

In [9], the authors propose a novel metaheuristic approach using the Binary Firefly Algorithm to optimize the reconfiguration of partially shaded PV arrays. Its integration with machine learning for array reconfiguration and fault detection shows significant improvements in mitigating shading effects and detecting faults.

[10] introduces a data-driven methodology to assess fault mechanisms and reliability degradation in outdoor PV string inverters. This method effectively identifies key degradation mechanisms, such as humidity cycling and temperature fluctuations, as primary contributors to inverter faults. These findings emphasize the importance of implementing time-bound preventive measures to improve the long-term reliability of PV inverters, particularly in diverse outdoor environments.

The study in [11] focuses on analysing common fault types in PV modules and employs machine learning-based fault diagnosis methods to enhance the accuracy and efficiency of fault detection in PV systems. This method closely aligns estimated power output with actual power output, highlighting the significant impact of dust on the efficiency of PV systems. Additionally, by integrating voltmeters and support vector machines into the PV array modules, it can quickly measure and locate short-circuit and open-circuit faults in bypass diodes.

[12] focuses on enhancing fault detection in PV systems by leveraging contextual information through machine learning-based models. This study trains Linear Regression, Multilayer Perceptron, Random Forest Regressor, Support Vector Regressor, and Long Short-Term Memory to compare the context-handling strategies to detect 13

different faults. According to the reported results, out of 13 faults, the system successfully detects 6 faults early, while 7 are detected late.

In [13] the authors propose an intelligent, automatic fault diagnosis method that requires less data for training by leveraging feature extraction and selection algorithms, along with an ensemble learning algorithm to classify open-circuit and line-to-line faults in PV systems. The proposed model first extracts key features from the operating current and voltage of PV arrays that are received from sensors. In the classification stage, this study proposes an ensemble learning algorithm that combines three individual classifiers: Logistic Regression, Support Vector Machine, and k-Nearest Neighbours, using a weighted voting approach. Also, this study employs genetic algorithms to optimize the weights assigned to each machine learning method, enhancing fault detection accuracy.

Supervised Machine Learning algorithms are increasingly being developed for PV fault diagnosis. However, these algorithms typically require the extraction of relevant features to eliminate irrelevant or redundant data, reducing the computational load. Recognizing that different dimensionality reduction techniques can significantly affect FDD performance, [14] introduces a novel Data Dimensionality Reduction Strategy. The strategy is based on Information Gain score and integrates Principal Component Analysis to identify the optimal subset of features, minimizing dimensionality during the training and testing phases of various supervised machine learning algorithms. To evaluate the quality of the proposed method, three machine learning algorithms are trained in this study: Random Forest Classification, Regression Trees, and K-Nearest Neighbour. In this study, 15 datasets are considered where 8 of them represent normal operation and 7 datasets including faults. Additionally, this study considers 7 different faults: inverter fault, current sensor fault, grid anomaly, partial shading, open circuit, MPPT/IPPT controller fault, and Boost converter controller fault. Experimental results in this study confirm that the proposed reduction method increases accuracy. Also, the results show that K-Nearest Neighbour has the best performance in terms of accuracy among the others.

[15] proposes an open-circuit fault diagnosis scheme for the power switches in the output inverters of a cascaded H-bridge multilevel converter, designed for use in a large-scale PV system with a 1 MW capacity and a voltage of 13.2 kV. The proposed technique utilizes a two-stage classifier, combining (1) a signal processing algorithm and (2) a machine learning approach, with an artificial neural network as the primary classification model. The fault diagnosis scheme is developed on a per-phase basis for the converter, using three neural network classifiers to handle the three-phase configuration. This study uses irradiance, temperature, parametric variation, and irregular distribution of input irradiances information to train the neural networks. Experimental results show that the proposed method is capable of diagnosing open-circuit faults under varying irradiance conditions and parametric changes, without increasing the system’s complexity even if the converter topology is modified. Importantly, the algorithm does not require additional sensors, as it utilizes the existing individual DC bus voltages and output currents necessary for the control system.

In general, supervised machine learning strategies offer a promising approach to diagnosing PV system faults. However, obtaining sufficient labelled data for training these models is a significant challenge. To address this, [16] introduces a novel strategy that combines an ensemble learning framework with a semi-supervised learning approach based on self-training. This study introduces a novel strategy capable of distinguishing various PV systems faults, including arc faults, line-to-line faults, open-circuit faults, maximum power point tracking faults, and partial shading. The proposed method leverages a self-training approach within an ensemble learning framework, which incorporates decision trees, k-nearest neighbours, and support vector machines as base learners. These models are automatically trained to label previously unlabelled data. A majority voting criterion is then applied to finalize predictions, with the pseudo-labels assigned to the unlabelled data continuously updated to improve the model's performance with new data. In this study array voltage, array current, and solar irradiance are the main data that are received from different

sensors. To measure the quality of the proposed ensemble method, it is compared with each machine learning method. The results of this study show that the proposed ensemble method outperforms the others.

In [17], advanced artificial intelligence techniques are employed to optimize operation and maintenance tasks across 150 PV plants in Taiwan, with a combined capacity of approximately 54 MW. In this study, the designed machine learning algorithm continuously monitors and analyses the performance of each inverter under maximum power point tracking, with data collected every five minutes. The designed machine learning receives solar irradiation, rated capacity, output power, voltage, and current waveforms of each inverter under the Maximum Power Point Tracking from the sensors. This study uses a supervised decision tree structure as the main machine learning method. In general, more powerful machine learning algorithms (e.g. convolutional neural network) are used to diagnose the faults. However, this study labels the data with assistance of operation and maintenance engineers. Therefore, a faster machine learning algorithm can find the faults faster. Moreover, this study focuses on four main faults including shading, inverter thermal degradation, fuse burnt, and site outage. This study conducts field tests over two years at 74 PV power stations, involving 4,792 inverters. The results in this study highlight the improved reliability and performance of the proposed system.

[18] introduces a novel approach utilizing the Salp Swarm Algorithm as a feature selection method to enhance the accuracy of fault classification in supervised machine learning classifiers. The Salp Swarm Algorithm is designed to extract only the most critical features from raw data, eliminating unnecessary and redundant information, which improves the overall classification performance of the classifier models. The raw data of this study is composed of current and voltage of the PV panels and the grid. The selected features are then used to train various supervised machine learning techniques to distinguish between different operating modes and fault types. This study considers K-nearest neighbours, Discriminant analysis, decision tree, and support vector machine as the supervised machine learning techniques. The proposed system is tested with data containing both healthy operation conditions and 20 different fault types, including line-to-line faults, line-to-ground faults, connectivity issues, and faults related to the operation of bypass diodes.

[19] introduces a novel fault detection approach for inverters using machine learning algorithms trained on a hybrid dataset. This dataset combines real operational data from the PV systems during fault-free conditions with synthetic faulty data generated through a digital twin model. To build proper faulty data, this study uses meteorological data such as irradiance and ambient temperature and SCADA data such as currents, voltages, and power to make the twin model. Moreover, this study employs Decision Tree, Random Forest, K-Nearest Neighbours, and Artificial Neural Network as the machine learning techniques to classify the faults. In addition, this study focuses on DC cable degradation, DC cable open-circuit, switch degradation, and switch open-circuit. The results of the implementation show that the use of twin models improves the quality of classification of the supervised machine learning methods.

[20] presents a diagnosis method utilizing time series analysis and support vector machines to enhance the timeliness, accuracy, and feasibility of fault diagnosis in PV systems. The method captures real-time data, including voltage, current, radiation, and temperature from different sensors, to calculate the nominal output power of the PV array. These power values are then normalized at various times throughout the day to create a comprehensive time series dataset. By using this time series data as input for a “one-to-one” multiclass classifier, the method effectively identifies and classifies common operational faults, such as random shading, fixed shading, and aging degradation of PV arrays. The algorithmic model is trained and validated against various fault conditions using datasets generated from a PV array simulation device. Experimental results demonstrate that the model exhibits good reliability and accuracy.

The performance of machine learning models depends on two key conditions: (1) the availability of a large amount of well-labelled training data, and (2) the assumption that both the training and test datasets share the same distribution. However, these conditions are often not met in real-world PV systems, where obtaining fault-labelled data can be challenging due to safety concerns or high costs. To address this data scarcity, transfer learning models, which are also called pre-trained learning, offer an effective solution to address data deficiency in data-driven fault detection and identification systems. The core concept of pre-trained models is to utilize knowledge gained from prior data classification tasks and apply it to a related new task, thereby reducing the need for extensive training in the new task.

To address this issue, [21] introduces an adversarial-based deep transfer learning model capable of detecting and classifying short-circuit faults in PV systems without relying on historical fault data. Verification tests demonstrate that this model achieves over 90% accuracy in classifying short-circuit faults in a multi-terminal PV system, with a rapid response time. Additionally, the model exhibits robustness to measurement noise and adaptability to changes in system configuration.

Traditional power loss evaluation methods often separate theoretical analysis from experimental verification, leading to discrepancies between predicted and measured outcomes. To this end, [22] introduces a transfer learning-based refinement approach for power loss evaluation. The proposed method begins by creating a comprehensive source domain dataset to train a source domain model, followed by fine-tuning in the target domain using a small set of experimental data. The proposed model demonstrated a 50% reduction in average power loss error. Furthermore, the refined model successfully identified the peak efficiency and optimal control parameters, highlighting its effectiveness in improving power loss evaluation.

Deep learning methods, which are multilayer neural networks, have shown remarkable performance in classification tasks by learning features implicitly from training data, thus bypassing the need for explicit feature extraction. To this end, these methods are used in many PV fault diagnosis studies.

In [23], the authors introduce an approach for detecting, classifying, and locating string-to-string, string-to-ground, and open-circuit faults using multi-output deep learning algorithms. Specifically, this study utilizes convolutional neural networks, long short-term memory, and bi-directional long short-term memory. This study reports that fault classification and localization are achieved with accuracies of 99.94% and 99.54%, respectively.

[24] proposes three distinct deep learning models, which are Convolutional Neural Networks, Long Short-Term Memory networks, and a Hybrid model based on the last two models that is named CNN-LSTM model. These models focus on the line fault identification, fault classification, and fault location estimation. this study evaluates the proposed models using training and testing data from transmission line fault simulations on the IEEE 6-bus and IEEE 9-bus systems. The evaluation includes various fault classes, locations, and ground fault resistances.

[25] introduces an advanced fault detection and real-time monitoring technique for grid-connected PV systems by integrating Internet of Things technology with a one-dimensional convolutional neural network deep learning approach. The approach involves developing a temperature-dependent PV model using series resistance and ideality factor, collecting real-time data from a 15kWp grid-connected PY system using optimally placed sensors to minimize sensor count while preserving data accuracy. The collected data from the sensors trains the 1D-covolutional neural network model to classify different fault types, and the trained model is deployed on an IoT platform for real-time monitoring and fault detection, which displays system status and generates alerts via a dashboard. Furthermore, the proposed deep learning method is mainly trained on short-circuit, open-circuit, partial shading, and inverter bypass diode faults.

In [26], the authors propose a data-driven, two-stage method for fault detection and diagnosis in PV systems. In the first stage, faults are detected based on predefined criteria that analyse variations in the maximum power point values. The second stage involves diagnosing the specific fault type using I–V characteristic curve data, processed through a densely connected convolutional network (DenseNet) model. The DenseNet is extensively trained with a large dataset of I–V curves to enable precise and efficient fault diagnosis. The approach is validated through simulations and hardware tests using a 5 × 3 PV array, which initially operates under normal conditions but later experiences line-to-line faults, open-circuit faults, degradation faults, and partial shading faults. Comparative analyses with state-of-the-art PV FDD models demonstrate that the proposed DenseNet-based model accurately detects and diagnoses various PV system faults.

In addition to known faults in PV systems, in real operation, it is of interest that anomalies are detected even if they are not classifiable. Therefore, making accurate diagnosis of both single and compound known faults and identifying unknown faults is crucial for efficient operation and maintenance of PV systems. To address this, a 1D One-Shot Aggregation (VoVNet) and Support Vector Data Description (SVDD) based fault diagnosis model for PV arrays is proposed in [27]. This two-stage model consists of a 1D VoVNet network that automatically extracts fault features from raw I-V curve data, followed by a multi-classification Support Vector Data Description that combines these features with environmental parameters. The Support Vector Data Description constructs hyperspheres for each known fault type, and any fault type not classified into these hyperspheres is considered an unknown fault, enabling unknown faults diagnosis. To test the quality of the proposed method, 18 different faults are considered where 7 of them are single faults and the rest are compounds faults resulting from combining the different singles. Experimental results in this study demonstrate that the proposed model accurately classifies known faults across three test tasks while successfully identifying unknown fault types.

In real-world conditions, a combination of faults occur for PV systems simultaneously. To design machine learning based method with ability of diagnosing compound faults, [28] proposes a novel global–local dual-stream collaborative framework for multiclass PV compound fault diagnosis. This method introduces a local mining algorithm specifically designed to efficiently extract detailed characteristics from current–voltage data. A shared convolutional neural network is then employed to capture fault features from both global and local data streams, enhancing modelling efficiency. Additionally, two model fusion mechanisms are proposed to collaboratively integrate global and local fault features, depending on different data conditions. This study considers 4 basic faults: partial shading, open-circuit, short-circuit, and degradation. Then the proposed method is tested on these faults and different combinations of them. Reported results in this study validate the effectiveness of this approach, demonstrating improved fault diagnosis performance in complex PV systems.

A major challenge in the PV diagnosis field is the lack of accurate, scalable, and location-independent data-driven diagnosis algorithms for PV systems. [29] addresses this challenge by proposing a unified PV system health-state architecture aimed at predicting common array faults. The architecture integrates data quality routines, digital twin models, and AI-driven fault diagnosis algorithms to enhance predictive accuracy. In this study Mechanistic Performance Model, Bayesian Neural Networks, and eXtreme Gradient Boosting algorithms are merged to make an ensemble fault diagnosis system. This study employs two categories of data: meteorological and electrical. The meteorological data includes in-plane irradiance, ambient temperature, wind speed, wind direction. Also, electrical data includes module temperature, array DC current, voltage, DC power, and AC power. This study uses different sensors to gather the mentioned data. Also, to prove the scalability and location-independence of the proposed structure, the authors validate it by using historical data in both hot and cold climates. Additionally, the proposed structure focuses on shading, cell cracks, inverter open-circuit, and clipping faults. According to the reported results, the proposed AI-driven diagnosis algorithm achieves detection accuracies exceeding 90% for faults with magnitudes

greater than 8%. Furthermore, the classifier shows robust performance in diagnosing common PV faults, with classification accuracies surpassing 95%.

[30] introduces a deep learning-based Transformer model designed for accurate fault prediction in PV systems. The Transformer model leverages an attention mechanism, treating data points like language units or "words" to learn dependencies between them for predicting future data points. This study focuses on forecasting and classifying open circuit, short circuit, shading, and overlapping faults in PV systems. It also classifies faults based on severity, helping identify the level of maintenance required. Reported results in this study show that the proposed method outperforms traditional machine learning-based regression and classification techniques.

[31] presents an innovative application of deep learning for fault detection and diagnosis in PV systems through a structured three-step approach. First, a robust PV system model is developed and optimized using a heuristic optimization technique to ensure accuracy and adaptability. Next, a comprehensive dataset is created, integrating PV system data along with monitored parameters such as module temperature and solar irradiance under both healthy and faulty conditions. The final step involves fault classification, leveraging features extracted using a hybrid neural network combining Convolutional Neural Networks and Bidirectional Gated Recurrent Units. This combination of parallel convolutional processing and sequential recurrent processing allows the neural network to exploit both the spatial and temporal aspects of the data, improving the precision of fault detection and diagnosis. Specifically, this study receives current, voltage, power at the maximum point, solar irradiation, and module temperature from the sensors as the main data for training and testing the proposed model. Experimental results show that the proposed method effectively identifies and classifies various fault types, including open circuits, short circuits, and partial shading.

[32] introduces a deep learning-based method for fault detection, diagnosis, and classification in a 9.54 kW PV system in Algiers. The approach involves four key steps: first, the Coyote Optimization Algorithm is used to estimate the electrical parameters of the PV system, followed by PSIM-based simulations to model the system's operation. Next, a comprehensive dataset is constructed, including data such as current, voltage, power, temperature, and irradiance under both normal and faulty conditions. Using an Auto-Encoder, new features are extracted from the dataset, which are then utilized in an Artificial Neural Network to classify and detect faults like short circuits and open circuits. To make a deep learning strategy, this study combines five Auto-Encoders and Artificial Neural Networks.

In addition to machine learning-based methods, various alternative approaches, such as statistical, decision trees or fuzzy-based techniques, have been employed in different studies to detect and diagnose PV faults. Furthermore, several studies have integrated these methods with machine learning algorithms to enhance fault detection accuracy and improve overall performance. This combination of techniques has very promising results, where its operation can be more easily interpreted by the operation of the plant itself. Table 2 highlights some of the latest research where statistical methods, fuzzy logic, or a combination of these techniques with machine learning have been utilized to optimize the quality and reliability of PV fault diagnosis.

Ref.	Proposed techniques	Type of faults	Sensor data
[33]	- Multi-segment spectral similarity - Adaptive threshold model - Decision tree	- String faults	- DC current
[34]	- Fast Fourier Transform - Discrete Wavelet Transform	- String faults	- DC current

	- Decision tree		
[35]	- Machine learning models - 3-sigma rule - cumulative sum control charts	- Short circuit - Open string - Snow	- AC signal
[36]	- Neural network - Radial basis function	*- Shading* *- Line-to-line faults* *- Open string* - Hybrid faults	*- IV curve* *- DC voltage* *- DC current*
[37]	- Gaussian Process Regression - Machine learning	- Multiple operation faults	- Irradiance - Temperature - DC current - AC current
[38]	- Deep Belief Network -	- Short circuit	- DC current
[39]	- Fuzzy logic - Decision tree - Neural network	- Inverter fault - Sensor fault - Grid anomaly - PV array mismatch - MPPT controller	- DC current - DC voltage
[40]	- Wavelet transforms - Statistical alienation - Fuzzy logic	- Short circuit	- DC current - DC voltage
[41]	- Neural network - Fuzzy logic	- Shading - Interconnection faults - Open circuit - Short circuit - MPPT fault Charging fault	*- IV curve* *- DC current* *- DC voltage*
[42]	- Neural network - Fuzzy logic	*- Open circuit faults*	- AC signal
[43]	- LSTM - Fuzzy logic - Decision tree	- Production faults	- DC current - DC voltage

Table 2. Lits of related works based on Machine Learning and combination of other techniques

[33] presents a practical adaptive method for detecting series DC arc faults in PV systems, designed to be more adaptable to the complex and noisy environments of PV systems. The approach in this study is based on analysing the noise spectrum of the current before and after a DC arc fault, using the adjacent multi-segment spectral similarity characteristic to detect changes. A principal component analysis of adjacent multi-segment spectral similarity, combined with a φ-statistic, is employed to create an adaptive threshold model.

[34] focuses on diagnosing faults that may occur in one or more strings. To assess the performance of the microgrid system, this study captures output currents from the inverter's output terminals. This study employs Fast Fourier Transform (FFT) to analyse DC components and total harmonic distortions (THD). Additionally, discrete wavelet

transformation (DWT) is utilized to examine both approximate and detail coefficients of the inverter output currents, with calculating statistical parameters such as skewness and kurtosis. This study does different tests under normal operating conditions as well as various fault conditions within the strings. According to the findings of this research, there are notable relationships between the percentage of string faults and the DC components, THD, kurtosis, and skewness at specific DWT levels. Therefore, the faults in PV strings, particularly open and short circuits are detected based on these analyses.

[35] proposes a fault detection scheme specifically tailored for distributed PV systems, which solely relies on monitored AC output data and remote weather sources, eliminating the need for additional sensors or detailed information on the PV configuration and local shading conditions. This study combines machine learning and statistical methods to reach the goal. The method follows a two-step approach. First, historical monitoring data is used in a bootstrap fashion to develop machine learning models that establish baselines for normal PV output, accounting for estimation uncertainties. Second, a dynamic PV power benchmark is constructed based on the 3-sigma rule, and cumulative sum (CUSUM) control charts are simultaneously employed to detect system malfunctions. The performance of the proposed method is checked on short circuit, string open circuit, and the temporary cover of thick snow in different climate conditions.

[36] investigates different types of faults in PV systems, including partial shading fault (PSF), line-to-line faults (LLF), both Intra String (IS) and Cross-String (CS), open circuit fault (OCF), and hybrid faults, which are combinations of these types. Furthermore, the characteristic curves of PV systems under various fault conditions are analysed, and the mathematical equations for the extreme points of each curve are presented. These equations help illustrate the impact of different faults on the power–voltage (P–V) curve, allowing for a comparison with the healthy operating mode of the system. Using these findings, the paper proposes a fault detection algorithm that leverages defined fault indexes for fault identification, diagnosis, and localization. The proposed method is a hybrid approach based on artificial neural network and Radial basis function.

[37] introduces a machine learning-based approach that is equipped with a statistical approach. In this study, a non-linear squared exponential Gaussian Process Regression algorithm is added to the structure for fault detection, classification, and localization in PV systems. The proposed method employs solar cell parameters along with voltage and current to classify faults based on their severity, ensuring more precise fault identification. Moreover, this study detects and classifies Line-to-line faults, arc faults, line-to-ground faults, partial shading faults, and open-circuit faults. Experimental results validate that the proposed method reaches a good level of accuracy while the training time is reduced because of the statistical section.

[38] proposes an improved deep belief network method based on statistical feature extraction to classify short-circuit fault in Modular Multilevel Converter-based High Voltage Direct Current (MMC-HVDC) grids. The method involves extracting three key features, standard deviation, information entropy, and kurtosis from the fault current training samples of a single terminal within the MMC-HVDC grid. These features are then fused as inputs to the improved Deep Belief Network (DBN). Simulation results demonstrated that the proposed method outperforms these traditional techniques in terms of overall classification accuracy, kappa coefficient, Jaccard distance, and detection speed.

[39] introduces an intelligent algorithm-based fault detection scheme aimed at enhancing the reliability and sustainability of PV systems. In this study an adaptive neuro-fuzzy inference system is developed to distinguish between normal and faulty operations in grid-connected PV systems. A large dataset, gathered from real-time laboratory experiments using TBD125x125-36-P PV modules, including current and voltage characteristics, is extracted, pre-processed, and utilized for training the proposed algorithm. This research compares the performance of the proposed fuzzy-based fault detection scheme with several other popular machine learning algorithms, including

k-nearest neighbours, decision trees, Naïve Bayes, ensemble methods, linear discriminant analysis, support vector machines, and neural networks. To draw the comparison, this study uses inverter fault, feedback sensor fault, grid anomaly, PV array mismatch, MPPT controller fault, and boost converter controller fault. According to the reported results, the proposed method outperforms the others in terms of accuracy.

[40] proposes a self-adaptive fault identification scheme that considers the high penetration of PV systems, and the uncertainties introduced by varying temperature and irradiation conditions. The scheme uses real-time voltage and current measurements collected locally, applying wavelet transform, statistical alienation, and fuzzy logic for fault detection. The proposed method in this study particularly designed for detecting short-circuit faults. Various fault parameters, including fault types, locations, and resistances, as well as uncertainties from PV systems and different loading conditions, are thoroughly tested. Simulation results demonstrate the scheme's effectiveness. Additionally, the authors tested the proposed method under noisy data. The results show that the method performs well under white Gaussian and impulsive noise.

[41] proposes a fault detection and localization technique specifically designed for PV systems. The technique employs an adaptive neuro-fuzzy inference system, combining the strengths of artificial neural networks and fuzzy logic to detect and classify faults accurately. Nine distinct types of faults are investigated in this study, covering issues related to the PV array, DC-DC converter, and battery components. Shading, Temperature increase, Series resistances, Shunt resistances, Interconnection faults, open-circuit, short-circuit, MP controller fault, and Charging fault are the considered faults in this study. Moreover, to determine the faults of a PV systems, the proposed method extracts properties of current-voltage and power-voltage curves. The extracted properties in this study are maximum power, short-circuit voltage, and open-circuit voltage. In addition to the mentioned properties, this study uses converter voltage and battery voltage as the inputs of the designed system, too. Simulation results using MATLAB-Simulink demonstrate the quality of the proposed method.

[42] addresses the detection of single- and double-switching open-circuit (O-C) faults in the inverter of a photovoltaic solar pumping system. The system comprises a photovoltaic module, a DC/DC step-up converter controlled by a perturbation and observation (P&O) maximum power point tracking (MPPT) technique, a three-phase DC/AC inverter regulated by sinusoidal pulse width modulation (SPWM), a three-phase induction motor, and a water pump. To detect O-C faults in the inverter, the study employs artificial intelligence techniques, specifically neural networks and neuro-fuzzy networks, as observers of the inverter. These techniques detect faults by analysing features extracted from the inverter's output currents. The integration of a fuzzy system with an artificial neural network (ANN) is driven by the advantage of leveraging the ANN's learning and training capabilities to enhance the performance of the fuzzy model. The reported results in this study indicate that the proposed structure has good quality in terms of accuracy in fault diagnosing.

[43] presents a deep learning model combining Long Short-Term Memory and Adaptive Neuro-Fuzzy Inference System to detect and classify faults in a smart distribution grid supported by communication systems using smart meter data. The proposed model first utilizes LSTM to train data samples extracted from voltage and current, learning the temporal dependencies of the data. Next, ANFIS is employed to detect and classify faults from the trained data, leveraging its adaptive learning capabilities and fuzzy inference system. The model effectively identifies single-phase, two-phase, and three-phase faults with high precision while requiring a relatively small dataset for fault classification. To validate its effectiveness, the authors apply the intelligent model to the IEEE 13-node network, with several performance metrics such as accuracy, precision-recall, F1-score, Receiver Operating Characteristic (ROC) curve, and computational complexity used for evaluation. The results demonstrate that the proposed model performs well.

There is a wide range of techniques applied to the diagnosis of faults and anomalies in PV plants. Both approaches where ML is applied directly and those using mixed techniques are very promising. Both lines will be worked on in this project, being the specific techniques that will be used explained in Section 5. On the other hand, it is important to note that although there is a large amount of work in this line, there is no common characterization of the input data and the diagnoses, making it difficult to intercompany the results and the possible standardization of these techniques. Section 4 will define these elements for this project.

4. Data structure proposal

A Fault and Anomaly Diagnosis (FAD) System has a structure and information flow similar to any continuous data processing system. This can be divided into three basic elements: i) input data, ii) data processing and iii) output data. This system must be continuously receiving information, adapting it, processing it and, if it has sufficient valid information, delivering output data or results. This information flow can even have a feedback stage in more advanced processing technology, which can perform a certain evaluation of its own results and adapt to the environment in which it is located. In this section, the structure and information flow of the FAD system of the PVOP project is defined.

Figure 3**Error! Reference source not found.** shows a general diagram of the information flow of the system in a plant. There is a first stage of input data, where the following elements can be differentiated:

- ***Raw Data****: These data come from any sensor system or data collection system in a plant. This data can come from different sources, and even be of a different nature. It is often noisy, incomplete or even contains errors.*
- ***Data Adequacy****: This is a data processing stage that is necessary for any system that works in this type of environment. It must adapt the raw data by performing cleaning and validation operations, delivering processed data.*
- ***Processed Data****: The data adaptation stage will deliver processed data that must ensure consistent information in order to be processed by FAD algorithms.*

Figure 3. General information flow of algorithm in operation.

Once consistent input data are available, the main processing of the information will take place. This processing is carried out by algorithms of different nature and objectives, being this project focus on those based on artificial intelligence. In Section 5, the objectives of the algorithms are defined through a classification of the problems to be diagnosed based on current regulations and literature, without being a barrier to other possible classifications that may appear during the execution of the project.

The output data will be the diagnoses and evaluation of the faults and anomalies detected by the algorithms. These also respond to a specific classification described in Section 4.2.

4.1. The input data

PV plants have different sources of data: measurements from devices, specialized sensors, analysers, alarms, etc. This information varies in nature, from high-precision measurements of physical quantities to device status information. These measurements may contain noise or errors, from a calibration to the temporary disconnection of a device. The nature of the information and its quality are key to the operation of a FAD system, as seen in the State of the Art (Section 3).

Figure 4 shows a more detailed scheme of the data flow in the FAD system. The first element within the input data is the raw data. These are those directly obtained from the data acquisition systems, which will be divide into 3 types for the present project:

- ***Time series***: *It is a sequence of numerical data collected and recorded at regular time intervals. Each point in the series represents a time-dependent observation, which implies that the order of the data is crucial for the understanding and analysis of the underlying behaviour. They typically represent measurements of physical quantities.*
- ***Structural data***: *This is static information that defines the PV plant itself, such as elements hierarchy, nominal power or operating ranges.*
- ***Events***: *This encompasses asynchronous information that may or may not occur. It can be automatically generated, such as an alarm or a change in the status of a device, or even manually generated, such as the notification of a maintenance action by an operator.*

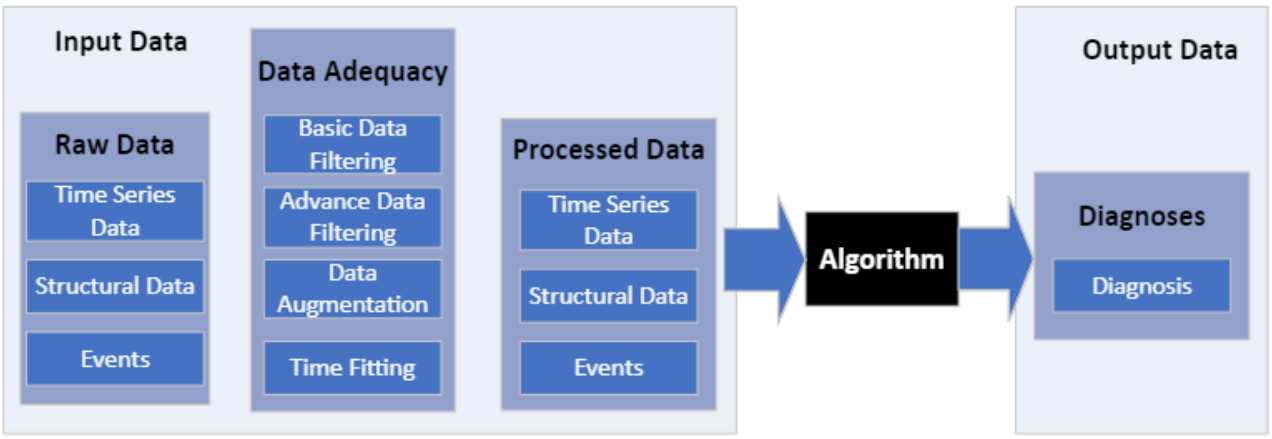

Figure 4. Input and output data structure scheme.

Data adequacy is critical to the success of a FAD system, mainly based on IA. If the data does not meet the necessary criteria, the results of models will be unreliable, and its ability to generalize to new cases will be limited. In addition, proper data adequacy helps reduce bias, improves the accuracy of the model, and ensures that the conclusions drawn are useful and applicable to the real problem. Here are the main types of processing that can take place at this stage of data adequacy:

- ***Basic data filtering***: *These respond to simple filtering, usually applied to individual variables. These would be: range filtering, frozen value, abrupt changes or stability. There are standards that define them for the main variables, such as the IEC61724 standards.*

- ***Advanced data filtering:*** *These filters perform multivariate or multi-device processing, applying heuristic rules that look for patterns that are not possible at the physical level or that show anomalous behaviour. For example: irradiance during the night or irradiance data from a pyranometer placed on a stopped tracker.*
- ***Data augmentation:*** *It is defined as the process of modifying, transforming or generating new data instances from an original data set in order to improve generalization and robustness. These processes often already include AI algorithms or advanced statistical processes.*
- ***Temporal adjustment****: Data sources can have temporal consistency problems, usually of 2 types: synchronization, data do not match in time, or time base, data are on different time base.*

It should be noted that it is not the objective of this work package to develop data adequacy process algorithms. However, it is necessary to know which ones are applied to the data, as modification of the original data may modify the behaviour of the FAD algorithm itself.

After the adequacy of the raw data, a set of pre-processed data will be available, being of the same nature as the raw data.

4.2. The output data

The classification of the diagnoses should answer the questions asked in the introduction, dividing them into 3 main groups:

- ***Data anomalies*** *(**Can I trust what I see?**) Even after a data adequacy phase, faults or anomalies can be detected in the data collection system of the plant. These refer to anomalous behaviour that has no physically coherent pattern and can only be explained by the source data. An example of this is poorly positioned or incorrectly connected sensors.*
- ***Production faults (Something is not working well?)*** *They refer directly to problems in production, whether due to malfunction or even breakage of some element of the plant. In photovoltaic plants it must be considered that detailed information is not always available for each element, some are not even monitored directly. These can range from a voltage drop on the input line of the plant to a short-circuited diode.*
- ***Predictive analysis*** *(**Is there going to be a problem?**) Predictive analysis and algorithms that detect this type of anomalies are increasingly used in the industry, and their need is also growing in the PV industry. These detect anomalous behaviours that may not be causing energy loss but prevent greater loss or element breakage. A clear example is algorithms that analyse the thermal behaviour of inverters to see if a problem is starting to occur.*

Starting from these 3 large groups, many types of faults can be defined. There are different examples in the literature, usually associated with specific problems of well-known devices. There is also a standard that defines certain types of faults, although with the objective of defining the availability of the plant, it is IEC 63019 standard. The project seeks to make its developments comparable with the state of the art, being also a reference for classification and development of fault diagnosis algorithms. To achieve this objective, it is important to define a classification that allows developers and researchers to compare these results, this is the objective of the Section 6. In addition, should be simple and intuitive so that it can easily reach an operational phase. Table 3 shows the rating proposed by PVOP.

Diagnosis	Description
Data[1]	
No data	The presence or absence of enough good quality data. It is usually linked to the concept of data availability.
Sensor malfunctioning	Malfunctioning of a sensor because its physical measurement does not make sense with the rest of the behaviour of the plant elements.
Sensor crossover	The plant information capture system may not correctly identify the physical elements, as their variables are crossed.
Production	
Power grid outage	Plant outage because of an external cause.
Grid constriction	Limited plant operation due to a control signal.
POI limit	Limited plant operation due to reaching maximum power output.
Inverter stop	Inverter stopped during plant operation.
Late start	Inverter does not start when it should.
Temperature derating	Inverter power limitation due to overheating.
MPPT deviation	Deviation from the maximum power point of the inverter.
Inverter limit	Reduction of production due to technically uncontrolled causes.
Open string box	String box is not producing.
Open string	String is not producing.
Damaged string	String producing below expectations.
Vegetation	Losses due to shading of vegetation.
Snow	Losses due to shading of snow.

[1] Note that the data adequacy phase implements filtering that will also report faults in the data system. The FAD system will give the ones indicated here, which are of a high level of complexity and are usually detected during the data analysis process.

Backtracking	Losses due to shading of backtracking.
Tracker stop	Losses due to tracker stop.
Tracker deviation	Tracker deviated from target position.
Tracker target error	Incorrect tracker target position, deviating from the optimal one.
Flag position	Tracker in flag position.
Predictive	
Shadows	Degradation or local damage due to shadows can be divided into subgroups such as backtracking, vegetation or snow.
Degradation	Abnormal degradation of element, modules or inverter.
Degraded battery	Degraded tracker battery with risk of stopping.
Electrical instability	Electrical signal anomaly, most of them defined by IEC 61000-4-30.
Anomalous temperature	Detailed analysis of the thermal behaviour of the inverter that can result in multiple types of faults or incidents.
Temperature imbalance	Unbalanced temperatures between inverter elements that can create risky situation.

Table 3. Classification of fault types

The output data shall be a set of diagnoses within the above classification. The diagnoses also have some additional information available to them that will help to make use of it in the operation of the plant. Table 4 shows the data structure of an output diagnosis.

Field	Description
Diagnosis	One of the previous classifications
Time range	information about the times when the diagnosis has been detected

Element	Element of the planta affected
Component (Optional)	Component of the affected element. For example, in a thermal analysis, which thermal sensor has presented the anomaly.
Energy losses (Optional)	Depending on the type of diagnosis, the losses associated with this can be estimated.
Severity (Optional)	Depending on the type of diagnosis, the severity or weight associated can be calculated.

Table 4. Data structure of an output diagnosis

5. FAD algorithms and development methodology proposal

The main core of the FAD system are the algorithms. These will be able to use different techniques to process the information and deliver the desired diagnoses. The detection capability and the accuracy of their analysis will be the main figures of merit of these algorithms, which will have to reach high accuracy values if they are going to be used in practical applications.

There are many techniques that can be used to implement algorithms with these objectives and their evolution will be a continuous work that transcends this project. In this respect, this project seeks to lay a good foundation that allows to create an objective environment for the development and evaluation of these kind of algorithms. Thanks to this, standardization of their use and a better intercomparison between them can be achieved, improving their usability and integration in the market.

Before defining the algorithms to be developed, it is necessary to identify the problem. The diagnosis of faults and anomalies in PV plants has some characteristics that are well known in the world of artificial intelligence:

- *From a set of source data an output must be delivered within a discrete set of possibilities.*
- *The input is a set of data with characteristics or attributes, usually being time series assigned to a particular physical quantity.*
- *The output are discrete classes or categories associated with some characteristics of the input. The output may be combinations of different classes or categories, although the input-output relationship is unique and exclusive.*
- *The objective is to predict the correct class for new entries based on learning from a set of labelled data.*
- *Well-known input-output assemblies are available. Not in all cases they are sufficiently large that they can be generic with initial training.*

Therefore, the faced problem is a classification one, where multiple solutions exist for a set of input values. This is a problem that has been widely addressed in the AI and there are multiple previous works, as seen in the State of the Art (Section 3). One of the most powerful capabilities of AI is the ability to learn, known as Machine Learning (ML). This is a process that usually consists of two elements:

1) *A set of data on which this learning is to be performed.*
2) *A training process, where this data is used to tune the algorithm.*

The data set to train the algorithm is key to its learning. Its representativeness of the problem and the quality of the information will allow an algorithm to be successful in its operation. This will be even more relevant in supervised learning processes, where a relationship between input and output data is established. For example, when a set of operating variables of a piece of equipment is identified, a specific fault is identified. If there is a large set of cases, the algorithm will be able to "learn" the pattern of input data, relating it to that fault, so that when it finds that pattern in a normal operation, it can warn that the fault is occurring. At the same time, the experience in the sector is key for the selection and adaptation of this data sets, to ensure that the algorithm will learn with a number of representative cases. This project has covered this aspect with a comprehensive set of validated diagnoses, as explained in Section 6.

Not all learning has to be supervised, there are also work with algorithms with unsupervised learning in search of patterns or hidden structures in unlabelled data. An example would be the search for hidden patterns or structures in a large data set. In this case, the algorithm does not provide a diagnosis, but rather it will return the behaviours that have statistical differences. The use of this type of algorithm is very useful in a data exploration stage, where an expert can subsequently study and classify these patterns found. The results could then feed a heuristic algorithm based on thresholds, creating a combination of techniques of a different nature. These combinations are common when algorithms are developed with a strong focus on a specific use, as is the case of this project. Although the algorithm has an AI base, it is usually combined with heuristic rules, where experience in the sector is a key factor. This can be useful for a prior classification, definition of general design, or for some interpretation or evaluation of the output. For example, if we are looking for thermal anomalies in inverters, we can pre-sort the data into inverters that are performing nighttime reactive compensation for those that are not.

The classification algorithms to be developed in this project are divided into 2 main groups, with a different structural approach:

1) *Heuristic operation algorithms: A main structure based on heuristic rules or deterministic parametric systems for an operation phase. The design or adjustment phase will be carried out with AI algorithms. These will be the first to be addressed in the project, as there is previous work by the team of participants, and they allow for a simpler view of the selection process.*
2) *End-to-end learning algorithms: the operation is a block with its integrated learning system, there being no two distinct parts. They use machine learning techniques that do not require as much knowledge about the application.*

5.1. Heuristic operation algorithms

These algorithms aim to be easy to analyse in operation, and their calculations and decisions to be understandable. The operation will be performed by a static algorithm based on thresholds, parameters and fixed rules. Their design and tuning will use AI techniques from the training data sets. Figure 5 shows a conceptual scheme of this structure:

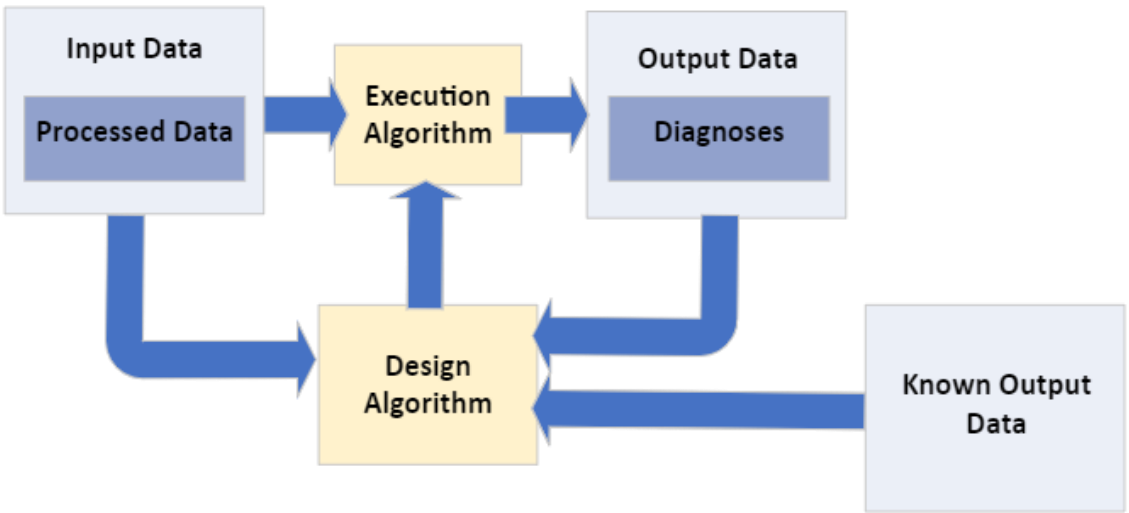

Figure 5. Heuristic operation algorithm scheme.

1) *Execution algorithm. Based on heuristics, threshold rules or advanced statistics. A parameterised structure is designed that establishes an input-output relationship. Knowledge of the problem and the technology is key to this design. This structure with its parameters defines the algorithm in execution. The parameters may be static.*
2) *Design algorithm. Based on unsupervised training and optimisation algorithms. Data exploration and structural definition will be performed with unsupervised methods. Parameter optimisations will be carried out with genetic algorithms.*

The execution algorithms to be developed in this project will be multi-class decision trees, where heuristic concepts are used to define their branches and their characteristics. Combinations of trees, both independent and forming a random forest, will give complete solutions to the FAD system. The tree design and preprocessing process will make use of clustering algorithms for data grouping, feature exploration and hyperparameter optimisation. Mainly two clustering techniques will be used: K-Means and EM-GMM (Expectation-Maximization for Gaussian Mixture Models). This combination will form the design algorithm.

Example of procedure with analysis of daily temperature

A simple example of algorithm development for temperature anomaly analysis is shown. Suppose that a set of temperature data is available for the three phases of the IGBT bridge of the inverters of a plant. The state of this IGBT bridge can be characterised by the average temperature and the deviation of the three phases. Figure 6 shows a point cloud of the average of the mean and deviation of temperatures of all inverters of a large power plant for one day.

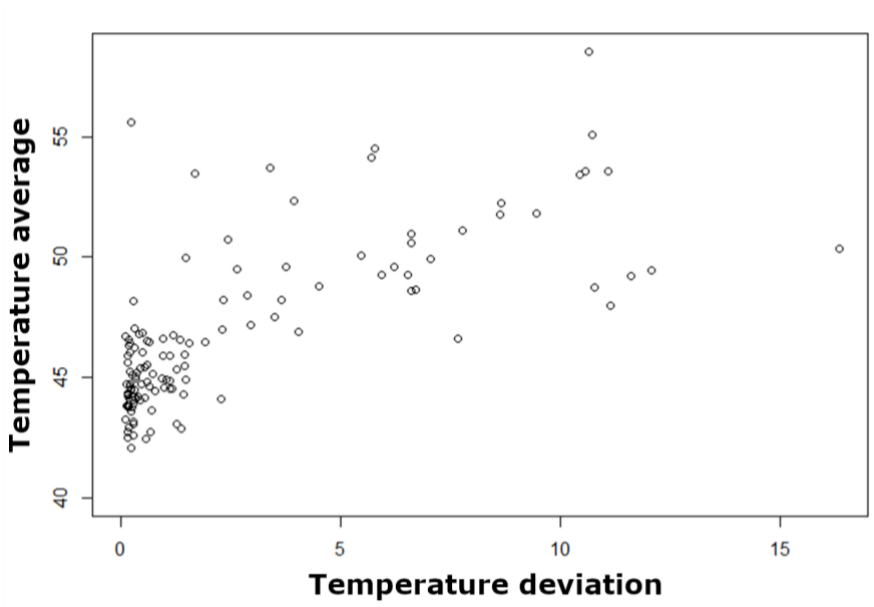

Figure 6. Average and deviation of temperature classification example.

As we can see, some inverters have very high average temperatures and others have very high deviations. Both cases are indicative of a thermal problem in the equipment. If we apply a k-means clustering algorithm to this set and

tell it to generate 5 different groups, we will have something similar to what is shown in the Figure 7. This algorithm groups the elements based on centroids and their distance from them.

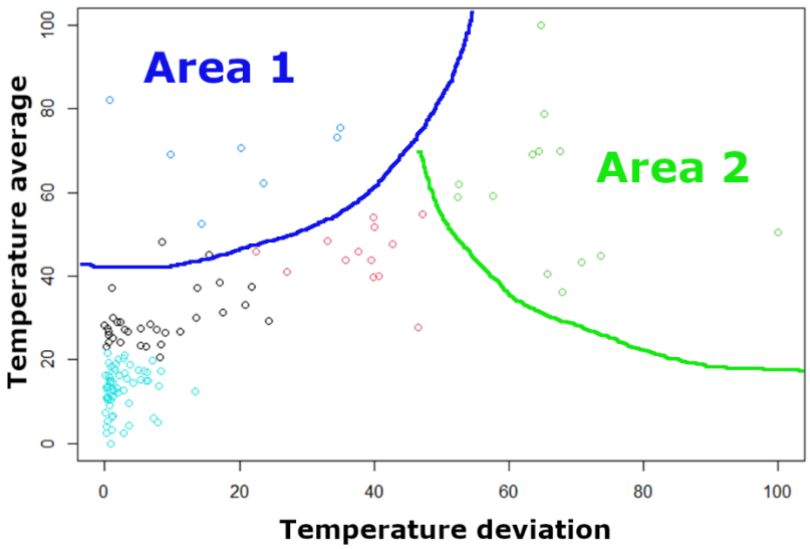

Figure 7. Clustering of data of temperature example.

The 5 groups could be classified as follows: i) light blue, black and red normal behaviour, ii) dark blue thermal anomaly, iii) green temperature unbalance. The groups generated by the clustering algorithm allow the definition of areas: i) area 1 thermal anomaly and ii) area 2 temperature imbalance. These areas can be used to define a decision tree where:

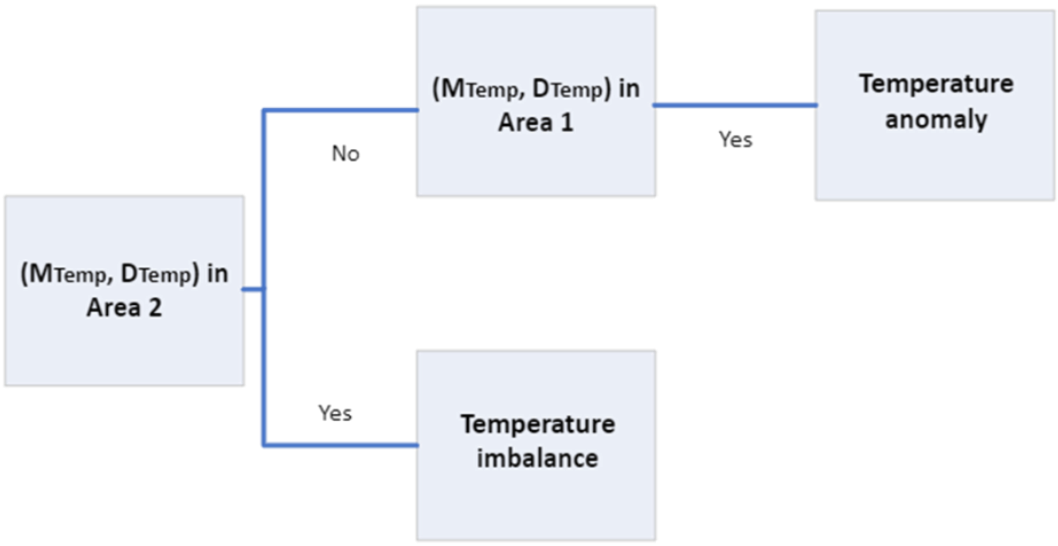

Figure 8. Decision tree of the temperature example.

This decision tree would be the execution algorithm operating in the plant. This would be a thermal anomaly algorithm, capable of differentiating between elevated temperatures and temperature imbalances in the IGBT bridge. Once the algorithm is defined, it would be applied to known diagnosis cases from the available project dataset (see Section 6) and assess their similarity to reality according to the equations in Section 6.3. Variations in the procedure or the defined characteristics would result in different versions of the algorithm, which can also be evaluated. All of these can be ranked, as defined in Section 6.4, which will allow the most suitable algorithms to be implemented for the solution of each problem.

5.2. End-to-end learning algorithms

The execution algorithm shall be a machine learning based algorithm (ML algorithm). The structure or configuration of the algorithm itself will have to be defined, together with a supervised learning procedure. As shown in Figure 9, this may have a training or pre-training phase with a post-reward adjustment procedure.

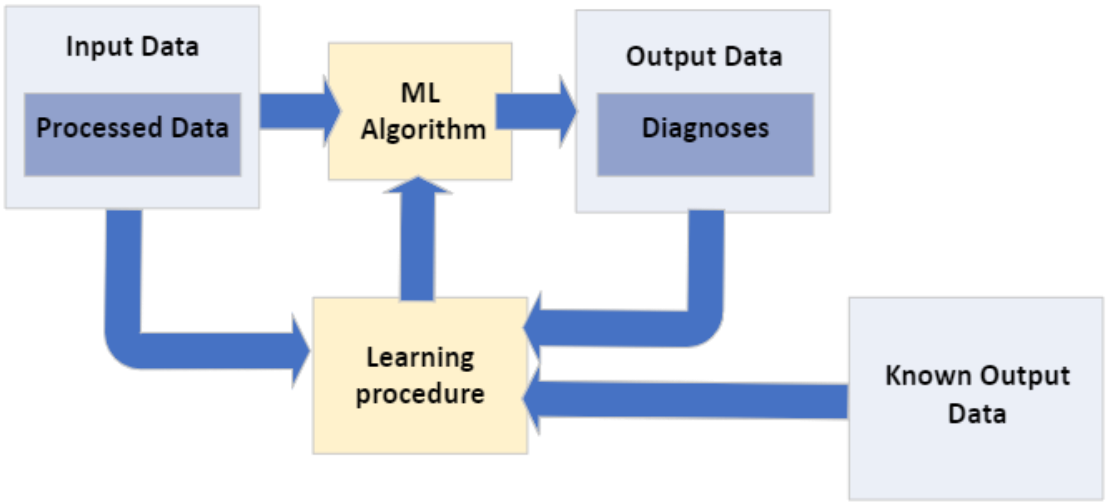

Figure 9. End-to-end learning algorithm scheme

ML algorithms will be developed with the following 2 technologies:

- *Support Vector Machines (SVM): The use of SVM will be key in the initial fault classification phase. This supervised learning algorithm is particularly effective in scenarios where the data has a significant number of dimensions and accurate results are required. In this phase of the development, SVM will be used for:*
  - *Classifying faults from historical sensor and signal data, identifying anomalous patterns.*
  - *Maximising the margin between fault classes and normality, ensuring accurate detection in noisy scenarios or with non-linear data.*
- *Artificial Neural Networks: especially Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN).*
  - *RNNs will be used mainly for processing temporal data, such as time series from sensors in mechanical, electrical or electronic systems.*

- *CNN are designed to process data that have a spatial or hierarchical structure, such as images or signals that have spatial relationships between features. Spatial maps of the hierarchical structure of the plant shall be made to be able to analyse anomalies between elements.*

In the development of SVM algorithms, features of the model shall be measured from the plant's sensors. The data labelling process and the training of the algorithm itself will be done with the available dataset. Compared to the previous example, the clustering process would allow to see the relevance of the features used by the SVM and the SVM training itself would define the areas that would then identify the diagnoses from the input data. Once the algorithm is ready, its diagnoses will be evaluated against reality, in the same way as the heuristic previous case.

The procedure is similar for the case of RNN use, but with a time series-oriented approach, where the diagnoses will be estimated for the whole series. In this case we will use LSTM networks, which have been seen in the literature as the most promising. We will explore how many hidden layers and how many neurons per layer to use, considering including dropout layers to prevent overfitting. Different activation functions will be tested, such as ReLU or tanh.

In the case of CNNs, we will make use of the data at the spatial level, which will allow us to have a picture of all elements of the plant with respect to different variables. Figure 10 shows an example of the temperature deviation distribution in the same plant as the heuristic case example. In this case, an RNN could interpret the time series of each investor, a row of the graph. The CNN could interpret the image as a whole and find patterns of anomalous behaviour.

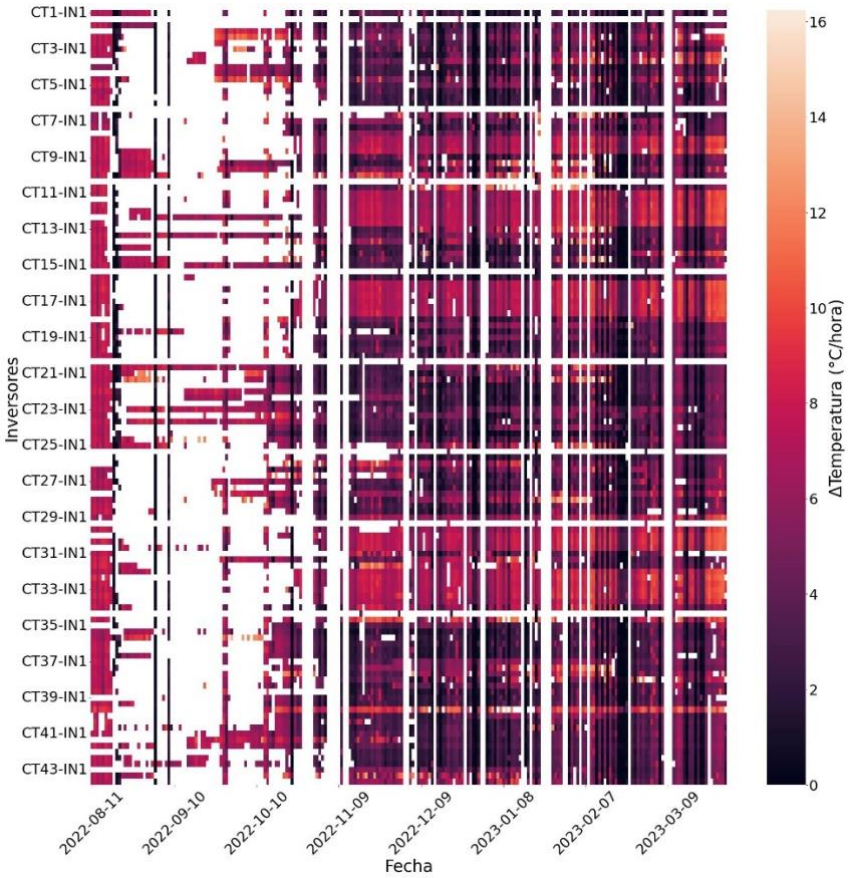

Figure 10. Example of temperature deviation distribution in the same plant as the heuristic case example.

However, time series classification remains a challenging issue due to the large data volumes and continuous updates of time series data. To enhance the performance of traditional feature-based methods, a convolutional neural network (CNN) based approach is proposed for time series classification. CNNs are particularly effective in fault detection due to their ability to automatically extract features and patterns from complex data.

CNNs are designed to recognize spatial hierarchies in data, making them well suited to analysing structured information such as images. To provide image data from time series to CNNs, the time series data collected from the PV real components associated to the digital twins can be converted into images using Markov Transition Field transform (MTF). The MTF transform is a method in time series analysis that converts temporal data into a 2D matrix. This process begins by discretizing the time series into a finite set of states. Next, the probabilities of transitioning between these states are calculated, forming a Markov transition matrix. The MTF is generated by mapping these transition probabilities onto the original time series, creating a matrix that encodes state transitions over time. This matrix captures the temporal patterns and structure of the time series, in this way enabling the use of image-based machine learning techniques, like convolutional neural networks (CNNs), for tasks such as classification or pattern recognition. Therefore, the MTF allows for effective application of advanced image processing methods to time series data.

5.3. Developing plan

This section will specify the development of AI algorithms for the FAD system in tasks that will allow to control the workflow of the project. Algorithms will be developed that are able to partially or fully diagnose the problems of a solar power plant, following the diagnoses approach described in Section 4. For this purpose, the diagnoses are grouped according to Table 5.

Name	Description	Diagnoses groups
Production		
Production-generator	Generator losses, both due to electrical and mechanical problems	Open string box Open string Damaged string Backtracking Tracker stop Tracker deviation Tracker target error Flag position
Production-AC	Losses from AC/DC conversion to plant output	Power grid outage Grid constriction

		POI limit Inverter stop Late start Temperature derating MPPT deviation Inverter limit
Production-complete	Combination of the above algorithms	Open string box Open string Damaged string Backtracking Tracker stop Tracker deviation Tracker target error Flag position Power grid outage Grid constriction POI limit Inverter stop Late start Temperature derating MPPT deviation Inverter limit
Production-shadows	Production losses due to shadows	Vegetation Snow Backtracking
Predictive		
Predictive-shadows	Preventive detection of shading problems due to vegetation snow or backtracking	Shadows (predictive)

Predictive-degradation	Unexpected degradation of plant elements	Degradation (predictive)
Predictive-TR-battery	Degradation or abnormal behaviour of the batteries of the tracking system	Degraded battery
Predictive-ACQ	Problems with AC signal quality	Electrical instability
Predictive-temperature	Anomalous thermal behaviour	Anomalous temperature
Predictive-aggregation	Combination of previous predictive algorithms	Shadows (predictive) Degradation (predictive) Degraded battery Electrical instability Anomalous temperature Temperature imbalance
Complete		
Total-aggregation	This will be a complete integration of the above algorithms	All diagnoses

Table 5. List of algorithms to be developed

All algorithms listed in Table 5**Error! Reference source not found.** will also output diagnoses of data faults, which are part of the possible output cases by the very nature of data problems.

The proposal is to develop an algorithm for each of the techniques defined in previous sections (Section 5.1 and Section 5.2) and for each of the diagnoses groups of the previous table. The specific technologies or combinations of technologies are summarised as follows:

a) *Decision trees design with K-Means clustering algorithms.*
b) *Decision trees with fuzzy logic rules and design with EM-GMM clustering.*
c) *SVM, where the features will start with the same features as the tree cases and then expand to more sensors.*
d) *RNN, where time series of the previously algorithms features will be used.*
e) *CNN, where spatial format with the hierarchical structure of the plant of the previously algorithms features and Markov Transition Field will be used.*

This makes a combination of 55 possible different algorithms. Certain combinations shall be discarded during development where there is evidence that they do not perform well. These in turn may have variants, using different

training techniques or parameterisations. In any case, the process of execution and evaluation of algorithms will be automated, with the aim of being able to develop as many as possible and to make a wide ranking that allows to see a great variety of possibilities and to have reliable solutions.

Initially we will work on the algorithms of points a) and b). This will allow us to explore the data and its characteristics in a more supervised and intuitive way. With the first results of these algorithms, we will be able to start the development of those indicated in points c), d) and e). As a guideline, this would be the chronogram for the next 12 months:

Task	1	2	3	4	5	6	7	8	9	10	11	12
Initial data exploration and characterization												
Dev. algorithms a)												
Validation and tunning algorithm a)												
Dev. algorithms b)												
Validation and tunning algorithm b)												
Dev. algorithms c)												
Validation and tunning algorithm c)												
Dev. algorithms d)												
Validation and tunning algorithm d)												
Dev. algorithms e)												
Validation and tunning algorithm a)												

Table 6. Algorithm development timetable. Green data related work and blue pure algorithm development.

The second phase of the project involves the implementation of these algorithms in real environments. This will be done in plants available in the project for analysis. The aim is to be able to validate the algorithms in real environments. To achieve this objective, it is necessary to be able to have some feedback on their performance. This will be done both by internal project analysts and by plant personnel. The work should be organised through a working protocol, which considers the validation process of the first phase and the evaluation needs of the second phase. In addition, the algorithms shall be adapted to the possible peculiarities of the plants and the real-time working environment. As a guideline, this would be the chronogram for the second phase:

Task	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Review protocol																		
Algorithm adaptation																		
Algorithm integration																		
Feed-back																		
Results																		

Table 7. Algorithm integration timetable, second phase of PVOP

In addition to the algorithm adaptation stage, an integration stage is considered, where the algorithm is applied and adjusted to the specific plant and the review protocol is established. Once the algorithm is adapted, integrated and the review protocol is operational, the feedback stage will begin, where the operation of the algorithm will be monitored and evaluated. Possible adjustments are also envisaged. Finally, there will be a results stage where all the operational information will be collected, and the results of the project will be elaborated.

6. Analysis data, classification and KPI of algorithms

The development of an algorithm can be divided into the following stages:

1. *Data and statistical exploration to define the inputs for each type of algorithm.*
2. *Adaptation of the input data to the specific needs of each technology.*
3. *Design of the algorithm defining all its possible variants.*
4. *Implementation of the algorithm and its variants.*
5. *Training or adjustment of the algorithm with training data sets.*
6. *Evaluation of the algorithm with validation and testing data sets.*

Stages 3 and 4 are the design and implementation stages, which are defined in Section 5. The other stages are not the core of development but are just as important. The data for training and validation, as well as the actual exploration of possible solutions for design, depends on a large amount of quality data. Furthermore, an objective evaluation and classification of the algorithms at their output is key to being able to compare them and standardize their use.

One of the PVOP's consortium members commercial solutions is PVET®, a cloud-based digital service for the continuous analysis of PV plants' operation. This service is based on a Big Data system that combines statistical and parametric algorithms and which allows optimizing the long-term production and O&M costs of the PV plant. This service is currently analysing data in real time in about 6 GW of plants around the world with between 1 and 10 years of operational data. It includes PV plants with monofacial and bifacial modules; installations in climates as different as dry/desertic (Atacama, Kalahari, Sonora, Sahara), mediterranean (Spain, Italy, Portugal) or tropical (Brazil, Dominican Republic); static and 1-axis horizontal (multirow, bi-row and mono-row) and azimuthal tracking systems; central and string inverters...

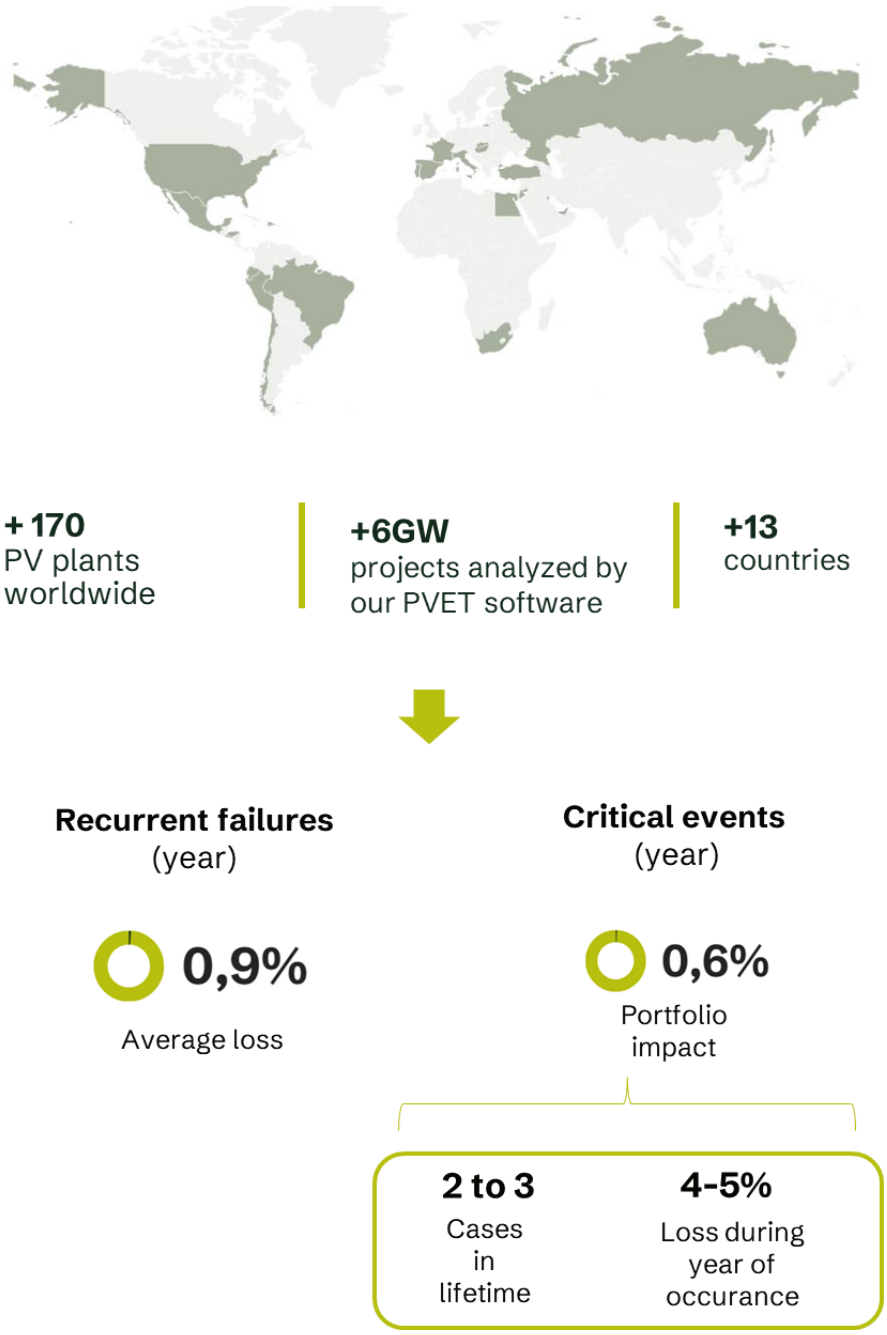

Figure 11. Information available from the PVET system.

From an evaluation point of view, this section defines the main KPIs and a proposal for a classification and ranking methodology for the developed algorithms. Furthermore, it is expected that these evaluation and comparison methods will serve as a reference for the sector and that they will serve, both during this project and in the future, as a source of reference and an accelerator for these technologies.

6.1. Analysis data sets

The plants analysed by PVET have been analysed by its anomaly and fault diagnosis system. Mainly heuristic systems that have a good detection capability but are poor in diagnosis. Still, they are a good basis of comparison for the project. In addition to the existing automatic detection in PVET, there are many anomaly and fault diagnoses that have been validated by an analyst or even verified in the field. Therefore, the pre-diagnosis available to the project have 3 types of verification levels:

1. *Automatic validation: Current algorithms have generated the diagnoses and have been automatically filtered according to their energy loss or duration.*
2. *Manual data validation: A solar analyst has reviewed, validated or corrected the diagnosis.*
3. *Field validation: A solar expert together with field staff at the plants have validated the diagnosis and corroborated that it has happened in the field.*

These diagnoses are related to the list in Section 4.2, being the outputs of the algorithms and composing a comprehensive set of input-output for the development, training, validation and test of the project's algorithms. For reference, this is a summary of the available diagnoses and their validation status at the start of the project:

Diagnosis	Automatic validation	Manual data validation	Field validation
Data			
No data	1622	162	81
Sensor malfunctioning	5405	541	270
Sensor crossover	1081	108	54
Production			
Power grid outage	541	162	11
Grid constriction	757	227	15
POI limit	15784	4735	316
Inverter stop	14270	4281	143

Late start	16865	5059	169
Temperature derating	4865	1459	49
MPPT deviation	12432	3730	124
Inverter limit	1622	486	16
Open string box	263351	26335	790
Open string	728497	72850	2185
Damaged string	72324	7232	217
Vegetation	541	216	11
Snow	216	86	4
Backtracking	114195	5710	114
Tracker stop	501081	15032	501
Tracker deviation	272865	8186	273
Tracker target error	3676	368	37
Flag position	889503	26685	890
Predictive			
Shadows	389	39	0
Degradation	541	54	5
Degraded battery	1622	324	16
Electrical instability	865	173	9
Anomalous temperature	27568	4135	551

Table 8. Reference of available diagnoses examples.

These diagnoses provide complete information on the plant and its operating elements: voltages, currents, power, limitations, irradiation, etc. Below there is an example, where **Error! Reference source not found.**2 shows the yield

of some string boxes and the power limitation of the plant for one day. For this day, the diagnoses in Table 9 were verified in field.

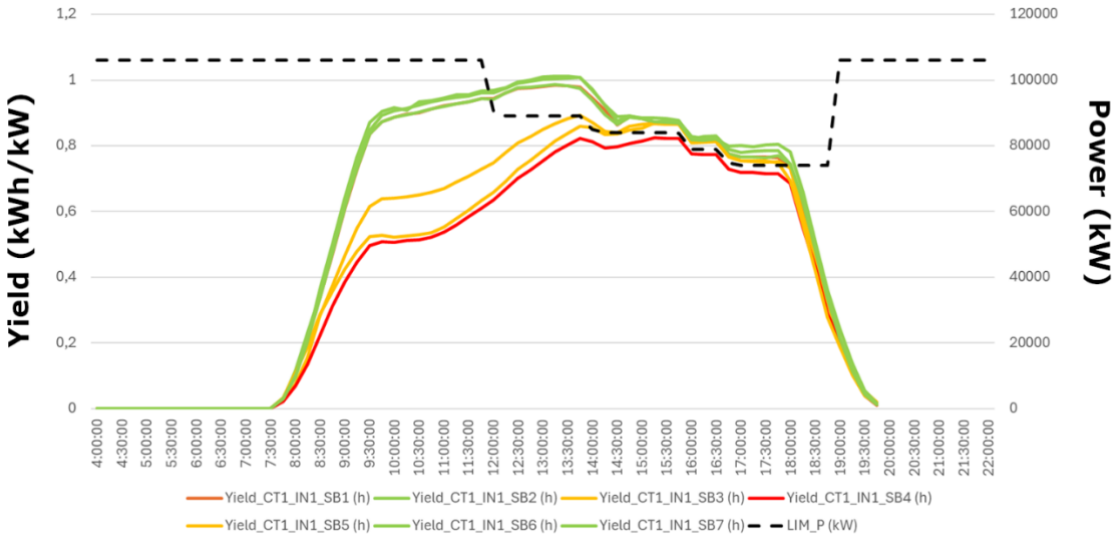

Figure 12. Yield and Plant limitation for an example day related to diagnoses of Table 9

Diagnose	Time range		Element	E. Losses
	Ini time	End time		
Grid constriction	12:00	19:00	Plant	196 MWh
Open string	7:30	19:30	CT1-IN1-SB4	6.2 kWh
Tracker deviation	7:30	13:30	CT1-IN1-SB3	24 kWh
Tracker deviation	7:30	13:30	CT1-IN1-SB4	28 kWh
Tracker deviation	7:30	13:30	CT1-IN1-SB5	28.5 kWh

Table 9. Diagnosis example for the day of Figure 12

During the project, this amount of information will increase due to the actual operation of the plants in the system. Many of the faults validated automatically can be validated manually, even for past faults. Field validation cannot be done in the past, because it is usually linked to the O&M team's own process of solving the problem and its

interrelation requires that an analyst of our team coordinates with O&M. Even so, they will continue to increase due to the collaboration of the project staff with the operation of the plants. This means that a large set of input-output data is available for the algorithm development process. These data will be used in different ways during the project:

- *Solution exploration process of unsupervised algorithms (stage 1 and 2): This is one of the first uses to which the data is being put. Analysts working on the project use unsupervised algorithms, such as clustering, to explore features or patterns that can be used to design algorithms.*
- *Training of supervised algorithms (stage 5): Algorithms that require a supervised training process will use these diagnoses for their training phase. They shall be trained with field validated diagnoses and validated with the rest.*
- *Optimization of parameters or structure of heuristic algorithms (stage 5): Similar to a training process, diagnoses shall be used as part of the evaluation function of optimisation algorithms, such as genetic algorithms.*
- *Evaluation of algorithms (stage 6): The whole process of evaluation and ranking described in Section 6.3 will be done with these diagnoses.*

Plant operation data can be sensitive and, in some cases, may be compromised by intellectual property. The project will create a database where 3 types of access will be differentiated:

- *Public: open access to the data so that any developer can use it.*
- *Project: access to project members or collaborators.*
- *Member: the data will be accessible by the project member who has permission, the rest will not be able to use it.*

6.2. Algorithm classification

There are a large number of AI algorithms applied to solar photovoltaics, as seen in the State of the Art. Typically, the development is problem-oriented and is not framed in any classification, or when they are attempted to be classified, they focus merely on the AI technology that has been used. In this project, a classification is defined based on the concepts already described, which helps to compare the algorithms with each other and easily know the problems covered in the plants. Table 10 defines the characteristics with which algorithms are classified.

Feature	Description
IA technology	Set of technologies that it uses, both for its implementation and for its optimization.
Online learning	If the algorithm has adaptation or training in the operation phase.
Type of input data	Types and sources of input data that you use for your operation.
Pre-processed input data	Pre-processing of the data required
Covered diagnoses	Diagnoses of the proposed classification that the algorithm.

Table 10. Features of the algorithm classification

Every algorithm developed related to the project will be tagged with these characteristics. This will allow for easy classification and comparison.

6.3. Algorithm assessment

The evaluation of algorithms is based on comparing the results of their outputs with previously validated diagnoses. Input-output data sets presented in Section 6.1 are the basis of this evaluation process. These have quantified diagnoses for different cases, with geographical and operational differences. In addition, the evaluation must be done in a closed evaluation period. Once an algorithm is applied to the known data set, it can be evaluated by comparison with the known output. This allows us to evaluate both algorithms developed in this project and others existing in the state of the art. This section defines the main KPIs for the algorithm assessment.

6.3.1. Occurrence of fault

This index measures the percentage of matching diagnoses in a certain time period. It should be noted that a FAD algorithm may not give a fault that exists, but it may also give a fault when nothing happens. These cases are summarized in the following concepts:

- ***True Positive (TP):*** *A fault has been detected and occurs in reality, is a success case.*
- ***False Positive (FP):*** *A fault has been detected but does not occur in reality, is a false alarm by the FAD algorithm.*
- ***True Negative (TN):*** *A fault has not been detected and does not occur in reality, is the neutral case that occurs in a correct operation of the plant.*
- ***False Negative (FN):*** *A fault has not been detected but occurs in reality, the algorithm has not been able to detect a problem.*

False cases (FP + FN) score the operation of the algorithm negatively. Following this logic, the following occurrence KPI is defined:

$$KPI_{ocurrence} = 1 - \sum_{Diag} k_{Diag}^{ocu} \times FC_{Diag}$$

where $Diag$ is the list of available type of diagnoses in the evaluation period, k_{Diag}^{ocu} is the weight of this diagnoses in the KPI and FC_{Diag} is the number of false cases (FP and FN). The variation of weights allows the errors in some diagnoses to count more in the evaluation, for example, it will not be the same to fail in the diagnosis of an inverter shutdown than of an anomalous cabin temperature. Even for some cases, the weight of a diagnosis could be set to zero, not affecting the evaluation of the algorithm, for example, we might not want to evaluate the effect of data faults. Annex I show a proposal of weights for a specific ranking proposed from our own previous experience.

6.3.2. Time correlation

Even if a fault or anomaly is diagnosed, it may not be detected for the same times. This would be an error on the part of the algorithm that will be quantified by the correlation KPI:

$$KPI_{correlation} = \frac{\sum_{TP} k_{Diag}^{cor} \times \rho_{TP}}{\sum_{TP} k_{Diag}^{cor}}$$

$$\rho_{TP} = \frac{\sum_{t}^{T} f(TP_t)}{T}$$

where *TP* is the list of true positive cases, k_{Diag}^{cor} is the weight of this diagnoses in the KPI, $\rho_{TP}$is the correlation of the true positive case, T is the number of samples of the assessment period and $f(TP_t)$ the coincidence function, which takes value 1 if in the sample the diagnosis of the algorithm matches that of the data set. Annex I show a proposal of weights for a specific ranking.

6.3.3. Energy losses (optional)

Diagnosis that has a lost energy calculation can also be evaluated by this factor, these are the production diagnoses. The following equation defines the KPI for energy losses:

$$KPI_{Elosses} = \frac{\sum_{TP} k_{Diag}^{losses} \times g(E_{TP}^{real}, E_{TP}^{est})}{\sum_{TP} k_{Diag}^{losses}}$$

$$g(E_{TP}^{real}, E_{TP}^{est}) = abs\left(\frac{E_{TP}^{real} - E_{TP}^{est}}{E_{TP}^{real}}\right)$$

where k_{Diag}^{losses} is the weight of this diagnoses in the KPI, $E_{TP}^{real}$is the real energy losses by the fault and $E_{TP}^{est}$is the estimated energy losses by the fault. Annex I show a proposal of weights for a specific ranking.

6.4. Algorithm comparison and ranking

Once the data sets for development and evaluation, a classification of algorithms and an evaluation method have been established, the algorithms can be compared with each other and evaluation criteria can be established based on this intercomparison. This ranking will allow us to choose which algorithms we pass from the first phase of the project to the second phase, being able to reduce the algorithms applied in the real operation in plants. In addition, it will be the main reference in the results of this WP, where the rankings will show the work done and its evaluation as a whole. Four initial classifications are established for the project. These are based on an explicit criterion of quality of the results with respect to the output diagnoses:

Name	Description	Weight table	Total KPI equation

Data ranking	Evaluate the best algorithms for data faults.	Data Ranking Table	$KPI_{Total} = 0.5 \times KPI_{ocurrence} + 0.5 \times KPI_{correlation}$
Production ranking	Evaluate the best algorithms for production faults.	Production Ranking Table	$KPI_{Total} = 0.33 \times KPI_{ocurrence} + 0.33 \times KPI_{correlation} + 0.33 \times KPI_{losses}$
Predictive ranking	Evaluate the best algorithms for predictive anomalies.	Predictive Ranking Table	$KPI_{Total} = 0.5 \times KPI_{ocurrence} + 0.5 \times KPI_{correlation}$
Total ranking	The best combined algorithms.	Total Ranking Table	$KPI_{Total} = 0.33 \times KPI_{ocurrence} + 0.33 \times KPI_{correlation} + 0.33 \times KPI_{losses}$

Table 11. Definition of the main 4 algorithm ranking

The rankings will allow to have a complete view of all the algorithms developed and tested with the project data sets. Table 12 shows an example of the four rankings.

Algorithm	KPI_{Total} **(%)**	$KPI_{ocurrence}$ **(%)**	$KPI_{correlation}$ (%)	KPI_{losses} (%)
Data ranking				
Algorithm 2	**97**	95	99	-
Algorithm 1	**96**	98	94	-
Algorithm 3	**70**	75	65	-
Production ranking				
Algorithm 3	**86**	95	98	68
Algorithm 1	**85**	85	99	75
Algorithm 2	**81**	87	78	80
Predictive ranking				

Algorithm 2	**99**	99	99	-
Algorithm 1	**87**	78	95	-
Algorithm 3	**86**	90	82	-
Total ranking				
Algorithm 2	**88**	94	92	80
Algorithm 1	**85**	87	96	75
Algorithm 3	**78**	87	82	68

Table 12. Example of the main rankings

7. Conclusions

This document describes the basis for the development of Fault and Anomaly Diagnosis (FAD) algorithms for the PVOP project. These are based on AI and the study of the current state of the art and the identification of the most promising techniques. The main objective is to have a complete solution in this aspect to optimise the operation of solar plants. On the other hand, the description of the data structure, information flow, classification and evaluation are intended to be a reference in the sector. This would allow for a more standardised development, enabling intercomparison of developments and speeding up implementation.

In this line, Section 4 described the input and output data. The output data are the diagnoses that the algorithms can give, proposing a classification based on standards and literature. This will be used during the development of the project and published to share with the sector. It may be modified during the project to improve its standardisation and adoption by other agents involved. All modifications will be published so that they can be easily shared and at the end of the project there will be a definitive classification, which brings together all the experience acquired during its development.

Section 5 has defined the developments of the project's algorithms, establishing two main groups: i) combination of heuristic techniques and exploratory AI algorithms and ii) algorithms based on Machine Learning techniques. The first group will allow for algorithms that are easier to analyse by the human eye and will allow the solar experts to apply their knowledge to the project. The second group will make use of the great potential of machine learning and its capacity for data ingestion and pattern detection. All of them will form a powerful set of FAD algorithms that hope to be a complete solution to this challenge of solar PV technology.

Section 6 describes the classification of algorithms, with simple defining characteristics. KPIs for their evaluation are also defined. The combination of both definitions allows them to be ranked and compared with each other. All algorithms that are developed in the project or externally can be classified with this method and placed in a ranking that allows to know which are the best solutions to the defined diagnoses. This list will be kept up to date, being a guide to know the state of WP evolution and will be the main result of this part of the project.

8. References

[1] S. Daliento *et al.*, “Monitoring, diagnosis, and power forecasting for photovoltaic fields: A review,” *International Journal of Photoenergy*, vol. 2017. 2017, doi: 10.1155/2017/1356851.

[2] S. R. Madeti and S. N. Singh, “A comprehensive study on different types of faults and detection techniques for solar photovoltaic system,” *Solar Energy*, vol. 158. Elsevier Ltd, pp. 161–185, 2017, doi: 10.1016/j.solener.2017.08.069.

[3] J. D. Bastidas-Rodriguez, G. Petrone, C. A. Ramos-Paja, and G. Spagnuolo, “Photovoltaic modules diagnostic: An overview,” 2013, doi: 10.1109/IECON.2013.6699117.

[4] G. B. Balachandran, M. Devisridhivyadharshini, M. E. Ramachandran, R. Santhiya, “Comparative investigation of imaging techniques, pre-processing and visual fault diagnosis using artificial intelligence models for solar photovoltaic system – A comprehensive review,” Measurement, vol. 232, pp. 114683, 2024.

[5] H. Qian, B. Lee, Z. Wu, G. Wang, “Research on DC arc fault detection in PV systems based on adjacent multi-segment spectral similarity and adaptive threshold model,” Solar Energy, vol. 264, pp. 112011, 2023.

[6] T.K. Das, S. Chattopadhyay, A. Das, “String fault detection in solar photo voltaic arrays,” IETE Journal of Research, vol. 69, no.5, pp. 2670-2682, 2023.

[7] A. Mellit, M. Benghanem, S. Kalogirou, A.M. Pavan, “An embedded system for remote monitoring and fault diagnosis of photovoltaic arrays using machine learning and the internet of things,” Renewable Energy, vol. 208, pp. 399‑408, 2023.

[8] J.V. Gompel, D. Spina, C. Develder, “Cost-effective fault diagnosis of nearby photovoltaic systems using graph neural networks,” Energy, vol. 266, pp. 126444, 2023.

[9] S. Saravanan, R.S. Kumar, P. Balakumar, “Binary firefly algorithm-based reconfiguration for maximum power extraction under partial shading and machine learning approach for fault detection in solar PV arrays,” Applied Soft Computing Journal, vol. 154, pp. 111318, 2024.

[10] S. Roy, S. Tufail, M. Tariq, A. Sarwat, “Photovoltaic Inverter Failure Mechanism Estimation Using Unsupervised Machine Learning and Reliability Assessment,” IEEE TRANSACTIONS ON RELIABILITY, vol. 73, no. 3, pp. 1418-1432, 2024.

[11] S.D. Lu, H.D. Liu, M.H. Wang, C.C. Wu, “A novel strategy for multitype fault diagnosis in photovoltaic systems using multiple regression analysis and support vector machines,” Energy Reports, vol. 12, pp. 2824–2844, 2024.

[12] D. Baruah, R. Roy, R. Ahmed, S. Subbiah, S. Chouhan, K. Angappan, “Contextual Approaches to Data-Driven Fault Detection in Solar Photovoltaic System,” In 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS) (pp. 1-7). IEEE.

[13] A. Eskandari, M. Aghaei, J. Milimonfared, A. Nedaei, “A weighted ensemble learning-based autonomous fault diagnosis method for photovoltaic systems using genetic algorithm,” Electrical Power and Energy Systems, vol. 144, pp. 108591, 2023.

[14] B. Chokr, N. Chatti, A. Charki, T. Lemenand, M. Hammoud, “Feature extraction-reduction and machine learning for fault diagnosis in PV panels,” Solar Energy, vol. 262, pp. 111918, 2023.

[15] Y. Leon-Ruiz, M. Gonzalez-Garcia, R. Alvarez-Salas, V. Cardenas, R.I.V. Diaz, “Fault Diagnosis in a Photovoltaic Grid-Tied CHB Multilevel Inverter based on a Hybrid Machine Learning and Signal Processing Technique,” IEEE Access, 2024.

[16] M.M. Badr, A.S. Abdel-Khalik, M.S. Hamad, R.A. Hamdy, E. Hamdan, S. Ahmed, N.A. Elmalhy, “Intelligent fault identification strategy of photovoltaic array based on ensemble self-training learning,” Solar Energy, vol. 249, pp. 122–138, 2023.

[17] M. Chang, K.H. Chen, Y.S. Chen, C.C Hsu, C.C. Chu, “Developments of AI-Assisted Fault Detection and Failure Mode Diagnosis for Operation and Maintenance of Photovoltaic Power Stations in Taiwan,” IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS, vol. 60, no. 4, pp. 5269-5281, 2024.

[18] A. Hichri, M. Hajji, M. Mansouri, H. Nounou, K. Bouzrara, “Supervised machine learning-based salp swarm algorithm for fault diagnosis of photovoltaic systems,” Journal of Engineering and Applied Science, vol. 71, no. 1, pp. 3-17, 2024.

[19] L. Costa, A. Silva, R.J. Bessa, R.E. Araujo, “PV Inverter Fault Classification using Machine Learning and Clarke Transformation,” In 2023 IEEE Belgrade PowerTech (pp. 1-6). IEEE.

[20] Y. Zhong, B. Zhang, X. Ji, J. Wu, “Fault Diagnosis of PV Array Based on Time Series and Support Vector Machine,” Energy sources, part a: recovery, utilization, and environmental effects, vol. 45, no. 2, pp.5380-5395, 2023.

[21] T. Wang, C. Zhang, Z. Hao, A. Monti, F. Ponci, “Data-driven fault detection and isolation in DC microgrids without prior fault data: A transfer learning approach,” Applied Energy, vol. 336, pp. 120708, 2023.

[22] Z. Xiao, Y. Jiang, T. Sun, Y. Wu, Y. Tang, “Refining Power Converter Loss Evaluation: A Transfer Learning Approach,” IEEE TRANSACTIONS ON POWER ELECTRONICS, vol. 39, no. 4, pp. 4313-4324, 2024.

[23] Z. Mustafa, A.S.A. Awad, M. Azzouz, A. Azab, “Fault identification for photovoltaic systems using a multi-output deep learning approach,” Expert Systems with Applications, vol. 211, pp. 118551, 2023.

[24] A.S. Al-Hanaf, M. Farsadi, H.H Balik, “Fault Detection and Classification in Ring Power System with DG Penetration Using Hybrid CNN-LSTM,” IEEE Access, vol. 12, pp. 59953 – 59975, 2024.

[25] B. Aljafari, P.R. Satpathy, S.B. Thanikanti, N. Nwulu, “Supervised classification and fault detection in grid-connected PV systems using 1D-CNN: Simulation and real-time validation,” Energy Reports, vol. 12, pp. 2156–2178, 2024.

[26] J. Ha, J.P. Ram, Y.J. Kim, J. Hong, “Data-Driven Two-Stage Fault Detection and Diagnosis Method for Photovoltaic Power Generation,” IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, vol. 73, pp. 3508611- 3508611, 2024.

[27] P. Lin, F. Guo, X. Lu, Q. Zheng, S. Cheng, Y. Lin, Z. Chen, L. Wu, Z. Qian, “A compound fault diagnosis model for photovoltaic array based on 1D VoVNet-SVDD by considering unknown faults,” Solar Energy, vol. 267, pp. 112155, 2024.

[28] Q. Liu, B. Yang, Y. Liu, K. Ma, X. Guan, “Collaborate Global and Local: An Efficient PV Compound Fault Diagnosis Scheme With Multilabel Learning and Model Fusion,” IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, vol. 72, pp. 2522816- 2522816, 2023.

[29] J. Montes-Romero, N. Heinzle, A. Livera, S. Theocharides, G. Makrides, J. Sutterlueti, S. Ransome, G.E. Georghiou, “Novel data-driven health-state architecture for photovoltaic system failure diagnosis,” Solar Energy, vol. 279, pp. 112820, 2024.

[30] I.U Khalil, A.U. Haq, N.U. Islam, “A deep learning-based transformer model for photovoltaic fault forecasting and classification,” Electric Power Systems Research, vol. 228, pp. 110063, 2024.

[31] A.F. Amiri, S. Kichou, H. Oudira, A. Chouder, S. Silvestre, “Fault Detection and Diagnosis of a Photovoltaic System Based on Deep Learning Using the Combination of a Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (Bi-GRU),” Sustainability, vol. 16, no. 3, pp.1012, 2024.

[32] A. Seghiour, H.A. Abbas, A. Chouder, A. Rabhi, “Deep learning method based on autoencoder neural network applied to faults detection and diagnosis of photovoltaic system,” Simulation Modelling Practice and Theory, vol. 123, pp. 102704, 2023.

[33] H. Qian, B. Lee, Z. Wu, G. Wang, “Research on DC arc fault detection in PV systems based on adjacent multi-segment spectral similarity and adaptive threshold model,” Solar Energy, vol. 264, pp. 112011, 2023.

[34] T.K. Das, S. Chattopadhyay, A. Das, “String fault detection in solar photo voltaic arrays,” IETE Journal of Research, vol. 69, no.5, pp. 2670-2682, 2023.

[35] B. Meng, R.C.G.M. Loonen, J.L.M. Hensen, “Leveraging dynamic power benchmarks and CUSUM charts for enhanced fault detection in distributed PV systems,” Energy Conversion and Management, vol. 314, pp. 118692, 2024.

[36] H.R. Parsa, M. Sarvi, “Online Fault Diagnosis, Classification, and Localization in Photovoltaic Systems,” IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, vol. 73, pp. 3516208-3516208, 2024.

[37] A. Javaid, I. Shafi, I.U. Khalil, S. Ahmad, M. Safran, S. Alfarhood, I. Ashraf, “Enhancing photovoltaic systems using Gaussian process regression for parameter identification and fault detection,” Energy Reports, vol. 11, pp. 4485-4499, 2024.

[38] Y. Liu, M. Mao, Y. Zheng, L. Chang, “DC Short-Circuit Fault Detection for MMC-HVDC-Grid Based on Improved DBN and DC Fault Current Statistical Features,” CPSS TRANSACTIONS ON POWER ELECTRONICS AND APPLICATIONS, vol. 8, no. 2, pp. 148 -160, 2023.

[39] M. Pa, M.N. Uddin, N. Rezaei, “An Adaptive Neuro-Fuzzy Model-Based Algorithm for Fault Detection in PV Systems,” IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS, vol. 60, no. 1, pp. 1919-1927, 2024.

[40] M.I. Zaki, R.A.E. Sehiemy, T.F. Megahed, T. Asano, S.M. Abdelkader, “A proposed fault identification-based fuzzy approach for active distribution networks with photovoltaic systems,” Measurement, vol. 223, pp. 113678, 2023.

[41] Y. Lahiouel, S. Latreche, M. Khemliche, “Adaptive neuro fuzzy inference system based method for faults detection in the photovoltaic system,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 32, no. 2, pp. 773-786, 2023.

[42] A.A. Bengharbi, S. Laribi, T. Allaoui, A. Mimouni, “Open-Circuit Fault Diagnosis for Three-Phase Inverter in Photovoltaic Solar Pumping System Using Neural Network and Neuro-Fuzzy Techniques,” Electrica, vol. 23, no.3, pp. 505-516, 2023.

[43] C.F. Mbey, V.J.F. Kakeu, A.T. Boum, F.G.Y. Souhe, “Fault detection and classification using deep learning method and neuro-fuzzy algorithm in a smart distribution grid,” The Journal of Engineering, vol. 11, pp. e12324, 2023.

Annex I: Diagnosis weight tables for rankings

Diagnoses	k_{Diag}^{ocu}	k_{Diag}^{cor}	k_{Diag}^{losses}
No data	0,04	0,04	-
Sensor malfunctioning	0,03	0,03	-
Sensor crossover	0,03	0,03	-
Power grid outage	0,07	0,07	0,092
Grid constriction	0,07	0,07	0,092
POI limit	0,07	0,07	0,092
Inverter stop	0,06	0,06	0,079
Late start	0,04	0,04	0,053
Temperature derating	0,04	0,04	0,053
MPPT deviation	0,04	0,04	0,053
Inverter limit	0,05	0,05	0,066
Open string box	0,04	0,04	0,053
Open string	0,03	0,03	0,039
Damaged string	0,02	0,02	0,026
Vegetation	0,03	0,03	0,039
Snow	0,03	0,03	0,039
Backtracking	0,03	0,03	0,039
Tracker stop	0,04	0,04	0,053
Tracker deviation	0,04	0,04	0,053
Tracker target error	0,03	0,03	0,039
Flag position	0,03	0,03	0,039
Shadows	0,02	0,02	-
Degradation	0,03	0,03	-
Degraded battery	0,03	0,03	-
Electrical instability	0,03	0,03	-
Anomalous temperature	0,03	0,03	-

Table 13. Weight of diagnoses for Total ranking.

Diagnoses	k_{Diag}^{ocu}	k_{Diag}^{cor}	k_{Diag}^{losses}
Power grid outage	0,092	0,092	0,092
Grid constriction	0,092	0,092	0,092
POI limit	0,092	0,092	0,092
Inverter stop	0,079	0,079	0,079
Late start	0,053	0,053	0,053

Temperature derating	0,053	0,053	0,053
MPPT deviation	0,053	0,053	0,053
Inverter limit	0,066	0,066	0,066
Open string box	0,053	0,053	0,053
Open string	0,039	0,039	0,039
Damaged string	0,026	0,026	0,026
Vegetation	0,039	0,039	0,039
Snow	0,039	0,039	0,039
Backtracking	0,039	0,039	0,039
Tracker stop	0,053	0,053	0,053
Tracker deviation	0,053	0,053	0,053
Tracker target error	0,039	0,039	0,039
Flag position	0,039	0,039	0,039

Table 14. Weight of diagnoses for Production ranking.

Diagnoses	k_{Diag}^{ocu}	k_{Diag}^{cor}	k_{Diag}^{losses}
No data	0,4	0,4	-
Sensor malfunctioning	0,3	0,3	-
Sensor crossover	0,3	0,3	-

Table 15. Weight of diagnoses for Data ranking.

Diagnoses	k_{Diag}^{ocu}	k_{Diag}^{cor}	k_{Diag}^{losses}
Shadows	0,143	0,143	-
Degradation	0,214	0,214	-
Degraded battery	0,214	0,214	-
Electrical instability	0,214	0,214	-
Anomalous temperature	0,214	0,214	-

Table 16. Weight of diagnoses for Predictive ranking.