# Artificial Intelligence for Operation and Maintenance of PV Plants

## Deliverable D3.1

### Models for root-cause analysis with data analytics


| | |
|---|---|
| Lead Beneficiary | INESCTEC |
| Delivery Date | 31/03/2023 |
| Dissemination Level | Public |
| Status | Released |
| Version | 1.0 |
| Keywords | Machine Learning, Artificial Intelligence, Random Forest, Logistic Regression, Gradient Boosting, Accuracy, Fault Classification |

**Disclaimer**

This work is financed by the ERDF - European Regional Development Fund through – the Operational Programme for Competitiveness and Internationalisation COMPETE 2020 under the Portugal 2020 Partnership Agreement within project AI4PV, with reference POCI-01-0247-FEDER-111936 – and Spain's Multi-regional Operational Programme 2014-2020. International collaborative project EUR 2020058 with the seal of the AI EUREKA CLUSTER.



**Document Information**

| | |
|---|---|
| **Project Acronym** | AI4PV |
| **Work Package** | WP 1 |
| **Related Task(s)** | T3.1 |
| **Deliverable** | D3.1 |
| **Title** | Models for root-cause analysis with data analytics |
| **Author(s)** | Louelson Costa (INECTEC), Ana Silva (INESCTEC), Christian Verrecchia (EDP NEW) |

**Revision History**

| Revision | Date | Description | Reviewer |
|---|---|---|---|
| 0.1 | 17 January 2023 | Outline of report content | INESCTEC |
| 0.2 | 14 February 2023 | Full draft of full content | INESCTEC |
| 0.3 | 2 March 2023 | Partner inputs | EDP NEW |
| 0.4 | 22 March 2023 | Second version | INESCTEC |
| 1.0 | 31 March 2023 | Final version | |

# EXECUTIVE SUMMARY

This deliverable includes the main results obtained in the task **T3.1 Models for root-cause analysis with data analytics from the project AI4PV**. The work carried out in this task has focused on the literature review of the factors that influence PV performance degradation and the different types of faults and failures, later focusing on power electronics-and power transformer related issues.

As a result, the fault classification model was analysed, including data cleaning, feature engineering, and the measurement of fault detection accuracy. In addition to that, initial research on viable options for Explainable Artificial intelligence (XAI) techniques has been approached.

The designs and studies here included will set the baseline for the development of the rest of the tasks within WP3, focused on fault diagnosis, and the rest of the technical tasks considered in the project.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ABBREVIATIONS AND ACRONYMS

| Acronym | Meaning |
|---|---|
| **AC** | Alternating Current |
| **AEP** | Annual Energy Production |
| **AI** | Artificial Intelligence |
| **ANN** | Artificial Neural Network |
| **CAPEX** | Capital Expenditure |
| **CB** | Combining Boxes; Junction Boxes |
| **CNN** | Convolutional Neural Network |
| **CSI** | Current Source Inverter |
| **DC** | Direct Current |
| **DL** | Deep Learning |
| **Dq** | Synchronous Direct-Quadrature Frame |
| **DT** | Digital Twin |
| **FL** | Fuzzy Logic |
| **Iabc** | Three-Phase Current |
| **Idc** | Direct Current |
| **IGBT** | Insulated-Gate Bipolar Transistor |
| **IQR** | Interquartile Range |
| **IS** | Isolation Forest |
| **kNN** | k-Nearest Neighbour |
| **LightGBM** | Light Gradient-Boosting |
| **LOF** | Local Outlier Factor |
| **MOSFET** | Metal–Oxide–Semiconductor Field-Effect Transistor |
| **MPPT** | Maximum Power Point Tracking |
| **MV/LV** | Medium Voltage/Low Voltage |
| **O&M** | Operation and Maintenance |
| **PCC** | Point of Common Coupling |
| **PV** | Photovoltaic |
| **PVPP** | Photovoltaic Power Plant |
| **Q1** | First Quartile |
| **Q3** | Third Quartile |
| **RF** | Random Forest |
| **RL** | Resistor-Inductor |
| **SC** | Short-Circuit |
| **SCADA** | Supervisory Control and Data Acquisition |
| **SVM** | C-Support Vector |
| **THD** | Total Harmonic Distortion |
| **UAV** | Unmanned Aerial Vehicles |
| **Vabc** | Three-Phase Voltage |
| **Vdc** | DC Voltage |
| **VSI** | Voltage Source Inverter |
| **XAI** | Explainable Artificial Intelligence |

# GLOSSARY OF KEY TERMS

| | |
|---|---|
| **Artificial Intelligence** | Artificial intelligence is a wide-ranging branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence. |
| **Machine Learning** | Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. |
| **Deep Learning** | Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behaviours of the human brain—albeit far from matching its ability—allowing it to "learn" from large amounts of data. |
| **Fault** | A fault is an unpermitted deviation of at least one characteristic property (feature) of the system from the acceptable, usual standard condition. |
| **Failure** | Permanent interruption of a system's ability to perform a required function under specified operating conditions. |
| **Malfunction** | Intermittent irregularity in fulfilment of a systems desired function. |
| **Fault detection** | Determination of faults present in a system and time of detection. |
| **Fault diagnosis** | Determination of kind, size, location and time of detection of a fault by evaluating symptoms. Follows fault detection. Includes fault detection, isolation and identification. |
| **Global Explainability** | Global Explainability entails explaining the behaviour of the entire model (including the features with more importance for the model output). |
| **Local Explainability** | Local Explainability aims to explain how a machine learning model makes individual predictions. |
| **Model-agnostic** | Model-agnostic techniques are designed independently of the architecture of a model. |
| **Model-specific** | Model-specific techniques designed for a single or a specific subset of the model's architecture type. |
| **Opaque Model** | Opaque Model, or "black box", is a model which requires post-hoc techniques to extract information from artificial intelligence models to generate explanations. |
| **Transparent Model** | Transparent Model, also called "white box", is a model which enables a straightforward interpretation of its results. |

# 1. INTRODUCTION

This document, deliverable **D3.1 Models for root-cause analysis with data analytics** from the project AI4PV, includes a summary of the results obtained in the process focused on the analysis of technologies, tools, mechanisms and methodologies for the initial validation process to be carried out in the context of the project AI4PV.

## 1.1 SCOPE OF REPORT

The PVPP has multiple types of equipment working in cooperation to ensure the proper operation of the asset. From meteorological stations, through electrical equipment, to the supervisory control and data acquisition (SCADA) system, every equipment is prone to faults and failures. However, regarding the electrical-related equipment, the PVPP problems can be divided into geological instability issues, PV string fuses, overvoltage, substation over temperature and inverter issues, low DC insulation, Medium Voltage/Low Voltage (MV/LV) transformer issues, and PV module issues [1]. Even though the inverters are the equipment most prone to issues [1] [2], they are not detailed and are seen as black boxes [3].

Usually, the studies regarding the faults and failures in PV systems focus on the DC side (PV modules and DC-DC converter). Some issues in the PV modules may be identified by using image processing [1] [4], as multiple issues are related to the surface of the PV modules (delamination, cracks, etc.), whilst others may be identified using electrical measurements processing [2] [4]. This condition-based monitoring is very important to avoid larger issues that may cause a large reduction in power production or even stop it. Early detection and diagnosis increase the safety and production of the PVPP.

This report is a Deliverable of Task 3.1 from Work Package 3 of the AI4PV Project. It contains a literature review on the factors that influence PV performance degradation and the different types of faults and failures, later focusing on power electronics-related issues. The literature review considers classical and recent scientific papers, as well as white papers. The main advancements and gaps in the technology are identified and discussed on how it can be improved.

It is worth noting that the PVPP have multiple configurations, scenarios, applications, etc. This review has a generalized approach at first, but in the end, focuses on the scenario that is studied in the project AI4PV. Thus, some solutions that do not fall under the scope of the project are not presented or explored as the ones that are the objectives of the project.

## 1.2 OUTLINE OF REPORT

*This report is structures as follows,*

- ***Chapter 1: Introduction*** presents an overview of the report.
- ***Chapter 2: The Factors of Influence*** summarizes the most common variables that may affect the power production of a PVPP, as well as the detection and diagnosis techniques employed in PVPP to achieve this goal. Additionally, Fault Signatures of the power converters present a

discussion of the state of the art and its gaps, as well as the most common measurements that are used for detection and diagnosis.

- ***Chapter 3: The Dataset*** briefly explains the dataset allowing to introduce the steps for the Rewrite Script, as well as the Data Cleaning, which includes the interquartile range (IQR) analyses for measurement outliers.
- ***Chapter 4: ML Algorithms for PVPP inverter*** focuses on the Machine Learning Algorithms used in the hybrid dataset implementation for the fault type classification. The Timeseries Dataset allows to reduce the processing time and removes any redundancy that could compromise the model's performance. Additionally, the test accuracy was evaluated in Fault Detection Accuracy, and for future study in Explainable Artificial Intelligence Techniques are presented three strategies in the context of root-cause analysis.
- ***Chapter 5: ML Algorithms for PVPP power transformer*** focuses on the Machine Learning Algorithms used in the hybrid dataset implementation for the fault type classification in power transformers.
- ***Chapter 6: Conclusions*** summarizes what is presented in this report.

# 2. THE FACTORS OF INFLUENCE

The PVPP is composed of multiple subsystems. If one of these systems malfunctions, the problem will spread throughout the whole system. Depending on the problem, it can propagate forward (grid direction) or backward (PV modules direction). For instance, a problem with the PV modules that reduce the incident sunlight will reduce the power production, thus, the expected delivered power based on the availability of solar resource will not be matched if a measurement is made at the transformer. On the other hand, a switch failure (for instance, an open circuit), will result in problems on the AC side (Total Harmonic Distortion - THD, power factor, output power, etc.), and problems on the DC side as well: an unbalanced three-phase system will result in a large DC-link voltage oscillation, compromising the Maximum Power Point Tracking (MPPT), thus the extracted power from the PV modules.

The understanding of the range of a fault or failure is very important to develop the detection and diagnosis algorithms, as a problem that originates in one of the sub-systems may be observed on multiple measurement points. At the same time, it is also important to understand how a problem at a given point may affect the adjacent sub-systems (forward- or backwards-wise).

The PV modules may present failure in their mechanical assembly or in their electrical components (PV cells, diodes, and cables). Some of the mechanical/structural faults may lead to electrical problems too, as well as electrical problems may cause overheating, compromising the PV module structure. The meteorological phenomena can also damage the PV modules of a PVPP, sometimes affecting all of them uniformly, or sometimes affecting only some of them.

The mismatch faults of PV modules happen when a PV module, or a set of them, for some reason, are not operating in the same conditions as the other sets connected in series or parallel. In a series connection, it is expected that all of the PV modules work at the same current. Similarly, in a parallel connection, it is expected that all of the PV modules (or PV strings) work at the same voltage. Assuming that all the PV modules of a PV string are the same (or as similar as possible, considering fabrication errors), a variation of the incident sunlight may cause mismatch faults [5]:

- Partial shading: trees existence, overhead supply lines, nearby structures;
- Uniform irradiance distribution: non-uniform nature of irradiance in the day;
- Soiling: dirt accumulation, snow, and droppings due to birds;
- Hot spot: immense change in temperature in tropical regions.

The degradation faults are related to multiple issues that may lead to the sunlight blocking (partial or total) of specific areas of PV modules: cell coating, delamination, yellowing, browning, bubble, and interconnection [5]. They may also lead to a series resistance increase, causing mismatch as well [1] [5]. In dire cases, the heating, both from the environment or from the Joule effects on the PV modules, may lead to some degradation too.

The PV module faults can be detected by electrical and meteorological measurements (real vs. expected out power, for instance). Although, their diagnosis, are more commonly achieved by visual

inspection (by humans or by image-processing algorithms), some of the mechanical/structural faults may lead to similar alteration in the electrical signals from the PV module.

Despite the possibility of numerous configurations, in a PVPP the PV modules are usually connected in series (to increase their output voltage) and, subsequently, connected in parallel at the combiner boxes (increasing the current). The cables, connections and combiner boxes may be a source of problems. For instance, in [1], combiner box fuse tripping may occur during periods of high irradiance with low temperature, resulting in unexpected overcurrent. Even though they are far behind the PV modules and power electronics in terms of production loss in case of a problem [1].

The problem associated with the DC connection and combiner boxes can be summarized as [5]:

- **Ground faults:** an unnatural ground path with no impedance;
- **Arc faults:** conductors having discontinuity caused by solar disjoint, damage of a cell, connector's corrosion or insulation breakdown;
- **Line to line faults:** short circuit among the two joints with unlike potentials;
- **Bypass diode faults:** short circuit due to wrong connections.
- **Bridging fault**: a loose connection between the different joints having different potentials.
- **Open circuit fault**: connection breaks down between the solar panels.

The consequence of those faults and failures is, at the least, a disconnection of the affected string or combiner boxes. Those short circuits may even lead to fire and result in cable and/or circuit disconnection. Those cable and combiners boxes are reliable and hardly prone to failure, but in case of a problem, as they connect the multiple parts of the PVPP, the production associated with them will be stopped. This can go from a single PV module, through entire strings or even the whole PV panels that are feeding the inverter.

Some papers are focused on the problem investigation of the PV modules, reducing the power electronics and reactive components to a single box labelled as inverter [2] [3]. On the other hand, the inverter is composed of multiple interconnected systems, and a single problem in those components may lead to a cascade effect that will result in poor performance on the entire PVPP system.

Being, literally, the central asset in a PVPP, the inverter is one of the assets with the highest problem rating and the equipment that, in case of fault or failure, will lead to the highest power loss in PV systems [1] [4]. In that sense, it is interesting to develop fault and failure detection and diagnosis tools that better evaluate the data related to the power electronics of such systems, i.e., currents, voltages, and temperatures.

General-application inverters (grid-connected, motor drive, etc.) and PV module-related faults and failures are already investigated [6] [7]. However, the lack of investigation of AI, and ML applied to a fault and failure analysis in the inverter of PV systems shows that there may be a large field of exploration, mainly to better understand how to detect and associate the consequences of a DC-DC converter input problem (MPPT, PV module disconnection, etc.) with the DC-AC converter output (THD, Park's analysis, etc. [8]). Mostly, the effort of the research towards fault and failure detection

in DC-AC converters is on the open or short circuit analysis, as those conditions can lead to power loss increase, THD increase, and current, voltage or thermal stress over the components (semiconductors and reactive), etc. [9] [10].

Also, some authors present the faults and failures associated with sensors, drivers, bond wire and substrate level, etc. [4] [11] . Such a level of detail would not be applicable to this project. The access level of this project is limited to input and output measurements; thus, it does not have access to the printed circuit board (PCB) nor the programming of the microcontrollers (firmware).

In their assembly, the power electronics converters have capacitors and inductors. These elements are also prone to problems, mostly due to degradation caused by thermal stresses [4] [11] . The result of this thermal degradation may result in the increase of the THD (worsening) at the output of the inverter (AC-side), which is an indication of poor power quality that can cause a false trip signal [12]. It is worth noting that a problem in the reactive components of the DC-side of the inverter will lead to a current and voltage ripple increase, which may affect the control and MPPT algorithms, THD, electrical stress, etc. This scenario is an example of the cascade effect that may happen. The problem associated with the power electronics and reactive components can be summarized as:

- Semiconductors: open or short circuit (considering both DC-DC and DC-AC converters), which may be caused by the bond wire thermal stress;
- Sensors: open circuit or wear-out (tuning parameter drift);
- Drivers: open or short circuit due to electrical or thermal stress, leading to switch open or short circuit, or even turning the transistor into a load;
- Capacitors and inductors: thermal stress that may lead to dielectric cracks, leakages, overheating, etc.

It can be noticed that power electronics have a lot to be explored regarding the faults and failures in PV systems. Even though the inverter presents the highest problem rating in the PV plant, it is not proportionally addressed. The PV modules are also a large source of problems, but they have a lot of investigation regarding their issues, while the same ought to be done with power electronics.

## 2.1 DETECTION AND DIAGNOSIS TECHNIQUES

There are multiple faults and failure detection and diagnosis techniques for each part of a PVPP. Besides the specific object of study within a PVPP, multiple strategies can be used to do such a task. Regardless of the technique or strategy, they have some clear steps to be followed. In [13] [14] [15], the fault and failure detection and diagnosis processes can be summarized as:

- First, detection: by the evaluation of the input data from the PVPP, it is possible to detect if there is a problem happening. At this stage, it is not possible to characterize the problem (where, when, how);
- Second, identification or diagnosis: by applying AI algorithms, it is possible to identify the nature of the problem. This is a hard task in complex systems, such as PVPP;

- Third, localization and isolation: the developed systems should localize and isolate the fault, which is a challenging task that requires information and expert knowledge.

The main three techniques are named signal-based, model-based, or AI-based. For this project, the two latter are being investigated.

The model-based solutions can be online or offline, where the online solutions are often referred to as digital twins (DT). In a very simplified way, this technique compares the measured output (real) vs. the ideal/simulated output (virtual) and flags a fault or failure if there is a deviation between those two signals [4]. Different approaches can be employed for residual generation including parity equations, parameter estimations, state observers, etc. [16] [17].

The downside of the model-based solutions is that they require a high expert level in power electronics for development, and, in some cases, the measurement of key signals within the inverter (firmware and power board level). This would not be a problem for products that are designed taking this into consideration, but this is not the current industry standard [12]. This type of measurement granularity is achieved only in laboratory prototypes [4] [14]. Nevertheless, they present some promising results and a high level of problem isolation, i.e., they can precisely point out where the problem occurred.

To develop a model-based solution for industry scenarios, it is necessary to create a model as close as possible to the real asset, considering current and voltage ratings, temperature levels, and using tailored solutions for such simulations, etc. However, once achieved, the results are of high accuracy. Once the model is developed, the offline or online application can be employed. However, the online application (a.k.a. digital twins) needs the backbone to keep the data owing: real-time (or as close as possible) measurement being fed to the model, and insightful information being sent back to the real asset or to its Operation and Maintenance (O&M) station.

Also, it is worth noting that the model-based has a limited capability for replication, as each PVPP can be unique, even though it has similar problems. Thus, every new asset to be improved with the model-based detection and diagnosis must go through a careful study of its behaviour and particularities, as a poor model can generate multiple false alarms, resulting in a waste of time and money.

Inside AI solutions, two major data types can be listed: visual and thermal, or electrical. It is worth noting that for the AI techniques, a large database is required for training and testing the AI algorithms [15]. Even though AI has multiple applications in both fields of PV systems and power electronics systems [6], the application of AI to a fault and failure detection and diagnosis has relatively fewer papers [13]. The three main approaches using AI and their methodologies are:

- Electrical measurements (currents, voltages, and their related parameters, i.e., power, frequency, etc.) from PV modules, DC-DC converters, or DC-AC converters: artificial neural network (ANN), fuzzy logic (FL) and random forest (RF) are employed;
- Image analysis, mainly infrared shots taken by unmanned aerial vehicles (UAV): deep learning (DL) and convolutional neural network (CNN) are employed;

- Clustering-based using unlabelled data: k-nearest neighbour (kNN), one-class support vector machine (1-SVM), isolation forest (IS), and local outlier factor (LOF) are employed.

In [5], the concern about the lack of real datasets available for research purposes has risen. Besides that, the datasets containing real data are unlabelled and present a few faults and failures. This unbalanced data may lead the research to look for out-of-normality analysis. To overcome the unbalanced and unlabelled problem, some papers use simulated data. A combination of the real and the simulated data may be the best outcome for training the algorithms.

Regarding power electronics, some papers that tackle fault and failure detection and diagnosis solutions from a generalized point of view can be applied to PV systems as well [11]. For instance, a methodology that evaluates the data from an inverter feeding an RL load can be applied to PV systems, however, the dynamics of replacing an ideal DC-link with a PV module is different. Besides the electrical measurements and their related parameters (power, frequency, etc.), the temperature is also a key factor. Overheating may be a signal of both fault and failure. Meteorological readings are also used by a variety of solutions, as the irradiance and ambient, or module, temperature are important factors that will influence the output power of a PVPP. Also, in [11], the AC filter is highlighted as an object of study by the AI algorithms. This corroborates the questions risen in [12] for the investigation of THD as an indicator of faults or failures.

Outside of the PV systems, some innovative solutions are exploring novel methodologies for fault and failure detection and diagnosis [10] [17]. Whilst some analyse the time-series data or use wavelet transformation, other concerns are to convert the time-series data to an image [18], and then apply image processing consolidated techniques. Even though those solutions have a generalized approach and simplified scenario, i.e., the DC-side of the converters is idealized, they can be adapted and explored in PV systems scenarios.

It can be noted that there are multiple AI algorithms that can be applied to evaluate multiple measurements from different points of a PVPP. Most of these solutions are concerned with the PV modules, whilst the investigation of the power electronics and reactive components is not as explored. Further, to develop a root cause analysis solution, some input vs. output evaluation algorithms is not enough: multiple algorithms that evaluate data from different points of the PVPP are required to do a proper diagnosis of the fault or failure. Also, some researchers are trying novel ways to convert a time-series current data to an image and applying image processing techniques for fault and failure pattern recognition. This shows that even though there are some consolidated solutions, even already employed by the industry, there is a lot to explore for innovative solutions to pre-process the data feeding the AI algorithms.

## 2.2 FAULT SIGNATURES

The fault signatures are defined as the device parameter that provides the indication of a fault event, as in [19]. Considering the inverter, the level of monitoring those variables can go from the most common input and output currents and voltages alongside temperature (typical measurements of a SCADA system), to the measurement of the gate signal of the switch or the sensor signal before

calibration. Usually, the input and output variables and temperature are selected for fault and failure detection and diagnosis techniques. Besides that, the current and voltage metadata may be stored as some key features, such as RMS values, frequency, power factor, THD, etc. The sinusoidal waveform will not always be available for processing. A similar assumption can be made for the DC-DC converter as well. It is worth noting that some PVPPs have central inverters, where the strings are directly connected to the DC-link of the DC-AC converter, thus, they do not have a DC-DC converter stage.

On the other hand, the reactive components (capacitors mainly, and inductors), are also prone to failure and are part of the power converters. They are elusive for fault detection, as some solutions focus on non-invasive techniques trying to estimate the state of the capacitor [20]. Even though, they are not as explored as the solutions for semiconductors-related faults and failures.

The investigation around a fault and failure and how they are reflected in the multiple variables of the PV systems, tries to identify the fault signatures, how to process them and use them for detection and diagnosis. Even still, when compared to the PV-module-related solutions, the power converter is not studied as much. Mostly, they are concerned about a switch fault or failure based on measurements of the output currents and voltages, with some considering the collector-emitter voltage for IGBTs or drain-source voltage for MOSFETs.

It can be noticed that both model-based and AI-based techniques have current, voltages and temperatures as the fault signatures. However, it can be pointed out that some research can be done regarding the input currents and voltages of the inverter, the THD of the output currents, or the Lissajous figure of the output currents and voltages, etc. Most of the fault signatures already pointed out in the literature are reliable and unlikely to be changed, however, the information being extracted from those measurements can be further explored.

The Idc (DC, PV or input current) and Vdc (DC, PV or input voltage) are measurements that are directly related to the MPPT algorithm. Depending on the converter topology, they will have a particular characteristic: in VSIs, the current is switched, and the voltage is continuous; in CSIs, the current is continuous, and the voltage is switched. Even though, the continuous variable has a ripple at the switching frequency. They may also present twice the grid frequency in single-phase systems. In three-phase balanced systems, the low-frequency ripple disappears.

However, the consequences of a switch open circuit in the DC-link can mislead the algorithms when searching for issues in the inverter. A high-frequency or low-frequency ripple variation will, consequently, result in poor MPPT performance, which will lead to reduced power production or reduced efficiency. Also, an increased DC-link ripple will reflect a worse THD measured at the output. This example shows the close relation between input and output variables that can be explored and, to the author's knowledge, has not been extensively investigated.

It is worth noting that the ripple analysis would require a real-time measurement of the input variables, which is not always the case as most SCADA presents RMS values measured in every minute, or even higher, intervals. On the other hand, the model-based approach for a closer look at the DC-link measurements may bring promising results.

The iabc (AC, grid or output current) and vabc (AC, grid or output current) are measurements that are directly related to the modulation and control of the DC-AC converter. Also, depending on the converter topology they will have characteristics, but, mostly, the measurements are regarding the filtered waveform, i.e., sinusoidal currents and voltages. Here, it is possible to evaluate how the DC-link ripple will affect the waveform. Also, the THD is an interesting reading that can be used for detection and diagnosis: the THD may indicate poor DC-link performance, switch problems or even capacitor or inductor degradation. Of course, information from other measurements can be used to do disambiguation.

Like the input measurements, the waveform analysis would require a real-time measurement of the input variables. In the same way, the model-based approach for a closer look at the AC-link measurements may give promising results. It is worth noting that the THD is a valuable measurement that can be stored in a SCADA, for instance.

Regarding the IGBTs and MOSFETs, a switch failure will result in a non-balance output (in the case of three-phase systems), resulting in one of the phases having a lower RMS value. This would be also noticeable by a power measurement, as one of the phases would not be supplying as much power as the others. Like the DC-link issues, real-time or model-based and AI-based solutions combinations can improve the performance of a detection and diagnosis system.

Parallel to those current and voltage measurements, the temperature measurement of some equipment, such as combiner boxes and inverters, is a good indicator of the assets condition The temperature of capacitors, inductors and switches can be an additional, or first, indicator if there is an issue with the converter, thus, any solution should include the temperature if it is available.

It can be noticed that the combination of model-based and AI-based solutions can provide the most resilient solution, both by working in parallel or at least using the model-based results to feed the AI-based algorithms. Nevertheless, there are some gaps in the state of the art that can be explored, mainly for fault and failure diagnosis, as the detection (first step) is already successfully employed in the industry. The new challenge is to combine tested and validated solutions in a way to reduce the time to detection and to do an accurate diagnosis of the detected problem, using the multiple available data to identify the issue.

It is worth noting that irradiance and ambient temperature are the input variables for any PV system model, thus these meteorological variables are always considered for both DC-side and AC-side solutions development. The most common AI algorithms are also listed, showing that multiple experiments are being researched and that a deeper investigation regarding where they are being applied in PVPP is needed [14].

The considered measured signals and variables for the PV strings are:

- PV currents: output current of the PV string;
- PV power: output power of the PV string;
- PV voltages: output voltage of the PV string;
- Strings temperatures: temperature of the PV modules.

The considered measured signals and variables for the combiner boxes are:

- CBs currents: output current of the combiner boxes. Similar to the PV currents;
- CBs power: output power of the combiner boxes. Similar to the PV power;
- CBs temperatures: internal temperature of the combiner boxes;
- CBs voltages: output voltage of the combiner boxes. Similar to the PV voltages;

The considered measured signals and variables for the power converters are:

- AC currents: output currents of the inverter, or grid-side currents;
- AC frequency: frequency of the output currents and voltages;
- AC voltages: output voltages of the inverter, or grid-side currents;
- AC powers: output power of the inverter, or grid-side power;
- DC currents: input currents of the inverter. Similar to the CBs currents;
- DC power: input power of the inverter. Similar to the CBs power;
- DC voltages: input voltages of the inverter. Similar to the CBs voltages;
- Power converter temperature: internal temperature of the inverter housing;
- THD: total harmonic distortion of the AC currents. Calculated from the sinusoidal current signals;

The considered measured signals and variables for the AC filters are:

- AC currents: output currents of the inverter, or grid-side currents;
- AC frequency: frequency of the output currents and voltages;
- AC power: output power of the inverter, or grid-side power;
- AC voltages: output voltages of the inverter, or grid-side currents;
- Power factor: power factor of the AC power. Calculated from the active and reactive AC power;
- THD: total harmonic distortion of the AC currents. Calculated from the sinusoidal current signals

# 3. THE DATASET

The original dataset consists of two main dataset types: meteorological, which defines weather conditions such as ambient temperature and daily irradiation, and electrical data provided by the SCADA control system (e.g., inverter frequency). For its better understatement and study, a characterization for each variable was created (Table 3-1). Furthermore, the description of each variable type for both dataset types is in Table 3-2 and Table 3-3.

**TABLE 3-1: VARIABLES CHARACTERIZATION**

| Parameter | Description |
|---|---|
| **VarName** | Name of the variable |
| **Datatype** | Type of dataset ("meteo": weather data; "inv": electrical data) |
| **VarType** | Variable Type (See Table 3-2 and Table 3-3) |
| **TransfID** | Transformer Number |
| **InvID** | Inverter Number |
| **JbID** | Junction Box Number |
| **StrgID** | String Number |
| **SensID** | Sensor Number |
| **MeasID** | Measurement Tag |
| **MeasDescp** | If MeasID is not None, Measurement Description (MeasDescp). Else: None |
| **State** | If State == 1 and MeasID != None, the variable is used to find IQR's value |
| **$V_{norm}$** | Base value in SI. For mode and control variables, such as Reactive Power Control Mode, the normalisation operation isn't applied. In consequence, [$V_{norm}$, $V_{min}$, $V_{max}$] = [0, 0, 0]. Additionally, if normalization isn't considered as input, Vnorm, Vmin and Vmax are set with default values. |
| **$V_{min}$** | Minimium value, in per unit |
| **$V_{max}$** | Maximum value, in per unit |
| **MachineID** | Specific equipment associated to the variable:<br>• TRANSF_A: Transformer A<br>• INV_A.B: Inverter B of TRANSF A<br>• JB_A.B.C: Junction Box C of INV A.B<br>• STRG_A.B.C.D: String D of JB A.B.C<br>• [VarType]_A.B.C.D.E: Sensor E of type [VarType] |

**TABLE 3-2: WEATHER'S VARIABLES**

| Type | Datatype | VarType | Units |
|---|---|---|---|
| **Ambient Temperature** | *Meteo* | TempAmb | ºC |
| **Average Direct Normal Irradiance** | | RadDirAv | W/m² |

| | | | |
|---|---|---|---|
| **Average Plane Irradiance** | | RadPlAv | W/m² |
| **Average Module Temperature** | | TempModAv | ºC |
| **Daily Irradiation** | | Irrad | kWh/m² |
| **Direct Normal Irradiance** | | RadDir | W/m² |
| **Horizontal Irradiance** | | RadH | W/m² |
| **Module Irradiance** | | RadMod | W/m² |
| **Module Temperature** | | TempMod | ºC |
| **Plane Irradiance** | | RadPl | W/m² |
| **Wind Direction** | | WindDir | º |
| **Wind Speed** | | WindS | m/s |

**TABLE 3-3: SCADA'S VARIABLES**

| Type | Datatype | VarType | Units |
|---|---|---|---|
| **Availability** | *Inv* | AVL | - |
| **Active Power** | | Pac | kW |
| **Active Power Control Mode** | | PMAXmod | - |
| **Active Power Limit Set-Point** | | PMAXsp | kW |
| **Alternating Current** | | Iac | A |
| **Apparent Power** | | Sac | kVA |
| **Daily Energy Produced** | | ENGDay | kWh |
| **Efficiency** | | EF | % |
| **Frequency** | | Fac | Hz |
| **Internal Temperature** | | TempInt | ºC |
| **Inverter DC** | | IdcI | A |
| **Inverter DC Power** | | PdcI | kW |
| **Junction DC** | | IdcJB | A |
| **Junction DC Power** | | PdcJB | kW |
| **Power Factor** | | Fpac | - |
| **Type** | **Datatype** | **VarType** | **Units** |
| **Power Factor Set-Point** | *Inv* | FPsp | % |
| **Priority of Current Injection** | | QCTRref | - |
| **Quadrant Set-Point** | | QQUADsp | - |
| **Reactive Power** | | Qac | kvar |

| | | | |
|---|---|---|---|
| **Reactive Power Control Mode** | | QCTRmod | - |
| **String DC** | | IdcS | A |
| **String DC Power** | | PdcS | kW |
| **Total Energy Produced** | | ENGTot | kWh |
| **Voltage AC** | | Vac | V |
| **Voltage DC** | | Vdc | V |

## 3.1 REWRITE SCRIPT

The main purpose of rewriting the datasets was to reduce the original size of files and consequently speed up the processing time. As will be mentioned in Subsections 3.1.1 - 3.1.3, it consists in:

- Define a uniform granularity.
- Fill in the possible missing data.
- Merge the datasets (weather and SCADA)

### 3.1.1 UNIFORM GRANULARITY

The initial dataset presents a non-uniform granularity. In the first attempt, a time frame of 15 minutes per measurement was considered for the data processing of each signal, respecting the minimum requirement of 5 minutes.

### 3.1.2 FILLING MISSING DATA

In the case of a lack of data, the pad method [20] is used to fill in the missing values. This methodology consists of filling the gaps with the previous valid value. In the case, it occurs in the first index, '0' was assumed as the default value.

### 3.1.3 MERGE OF SCADA AND WEATHER DATASETS

A multiclass classification is described by multiple possible outputs (classes) for each target, although each sample can only be labelled as one class [21]. This type of problem characterizes the fault types classification under study, which requires a new dataset composed of the following fields:

- 'date': Column of the type 'string'. It defines the reading date of the signals. E.g., 2020-03-05 05:00:00+00;
- [Weather_Features]: Set of columns ordered alphabetically, and related to meteorological conditions;
- [SCADA_Features]: Set of columns associated with SCADA signals, and ordered alphabetically;
- 'faultType': Column which identifies the fault's type.

## 3.2 DATA CLEANING

Data cleansing, or data cleaning, is a critical step in machine learning (ML) that involves identifying and removing any missing, duplicate, or irrelevant data. The advantages of cleaning data include improving model performance, reducing bias, and saving processing time and resources.

### 3.2.1 PER-UNIT SYSTEM

The per-unit system, or p.u. system, consists of electrical quantities normalisation (e.g., voltage, current, power, etc.) based on predetermined values. For a given quantity (V), the per-unit value (Vnorm) is the value related to a base quantity (Vb) by the expression Vnorm = V/Vb [22]. In the current section, as is presented in Table 3-4, the features normalisation was achieved based on the system per unit.

**TABLE 3-4: PARAMETERS OF PER-UNIT NORMALISATION**

| VarType | Vnorm | Vmin (pu) | Vmax (pu) |
|---|---|---|---|
| AVL, PMAXmod, QCTRmod, QCTRref, QQUADsp, PMAXsp, WindDir | - | - | - |
| EF, FPsp | 100% | 0.0 | 1.0 |
| ENGDay | 4000 kWh | 0.0 | 1.5 |
| ENGTot | 1116 kWh | 0.0 | 100.0 |
| Fac | 50 Hz | 0.95 | 1.03 |
| FPac | 1 | -1.0 | 1.0 |
| Iac | 1310 A | 0.0 | 1.5 |
| IdcI, IdcJB, IdcS | 1300 A | 0.0 | 1.5 |
| Irrad | 7 kWh/m² | 0.0 | 1.5 |
| Pac, PMAXsp | 630 Kw | 0.0 | 1.5 |
| PdcI, PdcJB, PdcS | 725 kW | 0.0 | 1.5 |
| Qac | 630 kvar | 0.0 | 1.5 |
| RadDir, RadDirAv | 1000 W/m² | 0.0 | 1.5 |
| RadMod, RadH, RadPl, RadPlAv | 1000 W/m² | 0.0 | 1.5 |
| Sac | 630 kVA | 0.0 | 1.5 |
| TempInt, TempMod, TempModAv | 25 °C | -0.1 | 3.0 |
| Vdc | 1000 V | 0.0 | 1.5 |
| Vac | 315 V | 0.0 | 1.5 |
| WindS | 10 m/s | 0.0 | 3.0 |

### 3.2.2 IQRED VALUES

The interquartile range is used to measure the dispersion of a distribution. This parameter defines the difference between the 75th and 25th percentiles of the data (Q3 and Q1), allowing the detection of outliers outside the range between the minimum and maximum limits [23]. Additionally, this method has been used to obtain the average value of each type of measurement (Figure 3-1- Figure 3-3).

It is also important to emphasize that the average measurements of temperature and irradiance weren't considered in the mean value estimation because of measurement errors. Moreover, in the case of module temperatures (TEMPM), it was necessary to disregard the ambient temperature due to its reduced daily deviation.

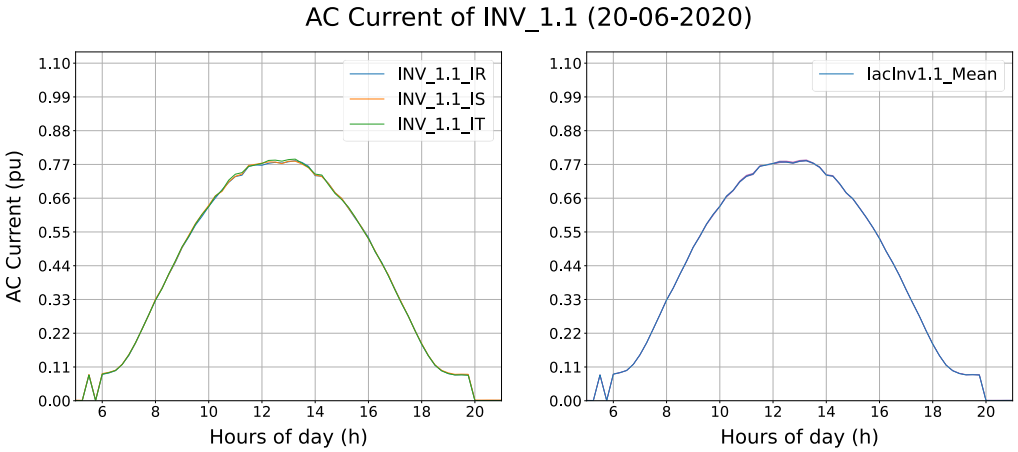



**FIGURE 3-1: AC CURRENT OF INV_1.1 FOR 20-06-2020 ($I_{AC,N}$ = 1310 A).**

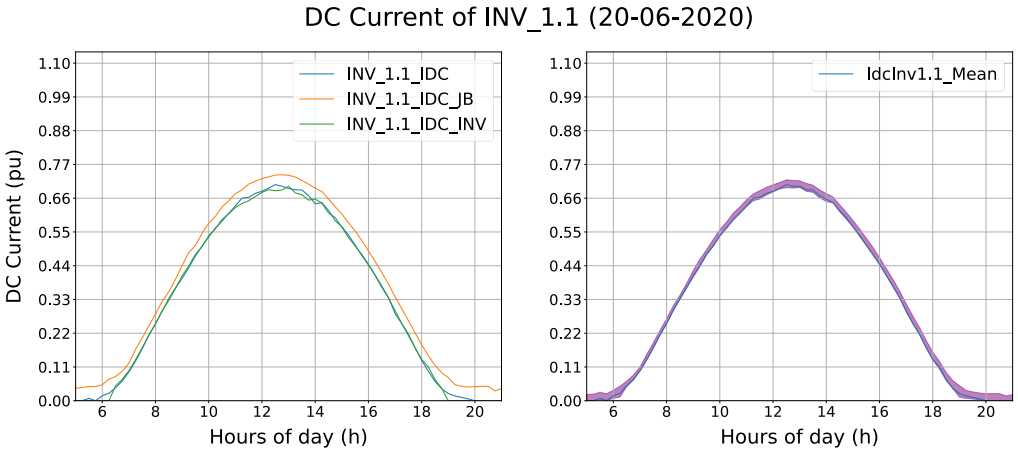



**FIGURE 3-2: DC CURRENT OF INV_1.1 FOR 20-06-2020 ($I_{DC,N}$ = 1300 A).**

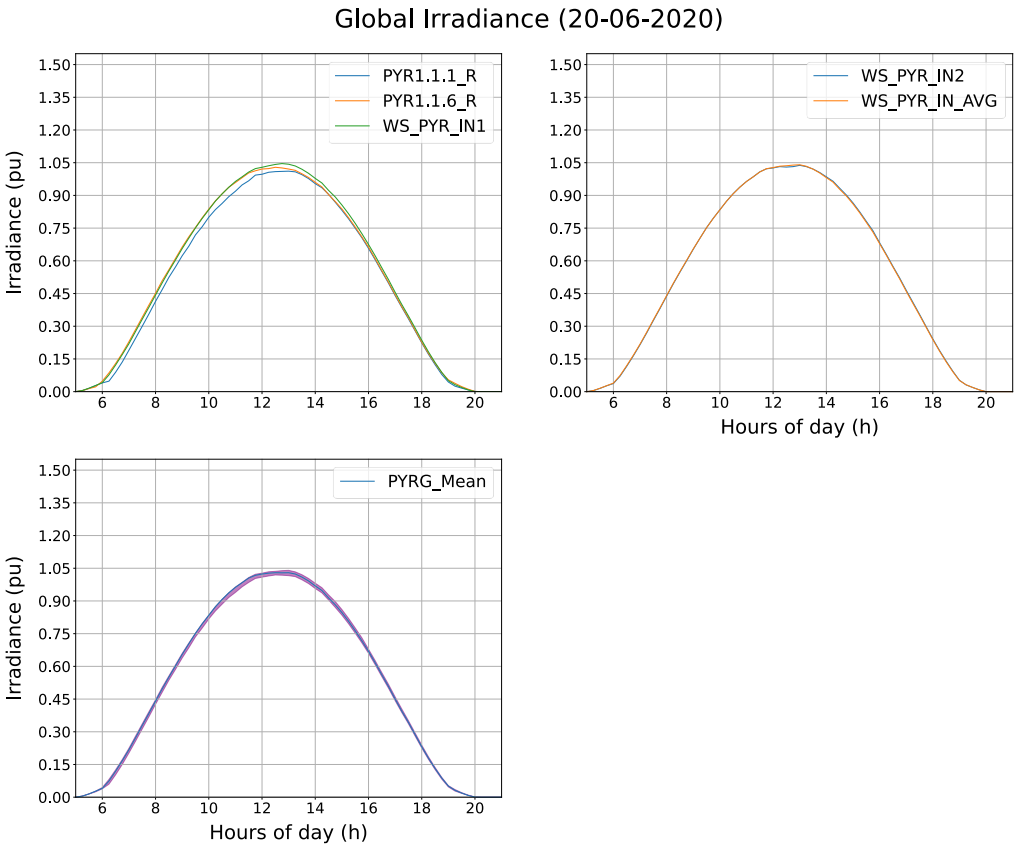



**FIGURE 3-3: GLOBAL IRRADIANCE OF INV_1.1 FOR 20-06-2020 ($G_N$ = 1000 W/M²).**

## 3.2.3 OUTLIERS

Initially, the Simulink/MATLAB® tool was used to rewrite the dataset considering all the data between 5 AM and 9 PM. Although, the datasets merge was later implemented by recurring to Python. In this case, were only considered dates containing the 5 am. The possibility of missing days in both cases set the registration of some errors for each month:

- Invalid File: e.g.,' Invalid File inv_Evora_2018-11-07.csv (2023-01-09 09:30:49)';
- Invalid Timestamp: e.g.,' Invalid Timestamp 2018-11-09 05:00:00+00 (2023-01-09 09:30:50)';
- Data Missing (Only Header): e.g., 'Invalid Timestamp 2018-11-09 05:00:00+00 (2023-01-09 09:30:50)'.

The more restrictive the data processing, the more likely outliers will occur. As shown Table 3-5, the consideration of days with both types of datasets will lead to a total of 587 days with data failures, with the year 2022 being the most critical.

**TABLE 3-5: NUMBER OF MISSING DAYS**

| Month/Year | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **January** | 31 | - | - | - | - |
| **February** | 21 | - | - | - | - |
| **March** | 28 | 31 | 4 | 4 | - |
| **April** | - | - | 1 | - | 30 |
| **May** | 31 | - | - | - | 31 |
| **June** | 30 | - | - | - | 30 |
| **July** | 31 | 2 | - | - | 31 |
| **August** | 31 | - | - | - | 31 |
| **September** | 30 | - | - | - | 30 |
| **October** | - | - | - | 7 | 31 |
| **November** | 3 | - | 1 | 25 | 30 |
| **December** | 1 | - | - | - | 31 |
| **Total** | 237 | 33 | 6 | 36 | 275 |

**TABLE 3-6: NUMBER OF DAYS AVAILABLE**

| Month/Year | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|
| **January** | - | 31 | 31 | 31 | 30 |
| **February** | 7 | 28 | 29 | 28 | 28 |
| **March** | 3 | - | 27 | 27 | 29 |
| **April** | 30 | 30 | 30 | 30 | - |
| **May** | - | 31 | 31 | 31 | - |
| **June** | - | 30 | 30 | 30 | - |
| **July** | - | 29 | 29 | 31 | - |
| **August** | - | 31 | 31 | 31 | - |
| **September** | - | 30 | 30 | 30 | - |
| **October** | 31 | 31 | 31 | 24 | - |
| **November** | 27 | 30 | 30 | 5 | - |
| **December** | 30 | 31 | 31 | 31 | - |
| **Total** | 128 | 332 | 360 | 329 | 87 |

Besides the missing days, some outliers were also detected in the mean value for each measurement, as are the examples of Table 3-7.

**TABLE 3-7: EXAMPLE OF OUTLIERS DETECTED BY IQR.**

| Type of Error | Description | Period |
|---|---|---|

| | | |
|---|---|---|
| **Error in PYR_1.1.1_R** | A significant deviation of IQR for the global irradiance of INV_1.1, due to a faulty operation of PYR_1.1.1_R | [21-05-2018;16-01-2020] |
| **Error in JB_1.1.1_AN1** | Error in the calculation of the mean value of module temperature, due to a faulty operation of JB_1.1.1_AN1. | [02-06-2020;02-11-2020] |
| **Error in JB_1.1.1_U, JB_1.1.3_U, JB_1.1.5_U, JB_1.1.7_U** | A significant deviation of IQR for the DC Voltage of INV_1.1, due to a faulty operation of JB_1.1.1, JB_1.1.3, JB_1.1.5, and JB_1.1.7. | [14-09-2020;07-10-2020] |

Finally, a full version of the rewrite dataset is expected to be implemented in Python, containing only days with all timestamps between 5 AM and 9 PM.

# 4. ML ALGORITHMS FOR PVPP INVERTER

The fault classification requires the implementation of a hybrid dataset. This dataset is composed of real and synthetic data for fault-free and faulty conditions (without noise), respectively. Thus, after validating the fault-free data [24], a digital twin (DT) was implemented in Simulink/MATLAB® to generate a faulty dataset (due to the lack of real faulty data) [25].

## 4.1 TIMESERIES DATASET

To reduce the processing time and remove any redundancy that could compromise the model's performance, it was necessary to reduce the total number of features from the hybrid dataset. Although, some features were added to overcome the model's low performance under low irradiance and fast weather transitions.

### 4.1.1 FEATURE ENGINEERING

To improve the accuracy of the ML algorithms, some additional features were added to the algorithms, to add some contextualization of the data. Most of the new features are weather-related, as the weather will dictate the operating conditions of the PV inverter.

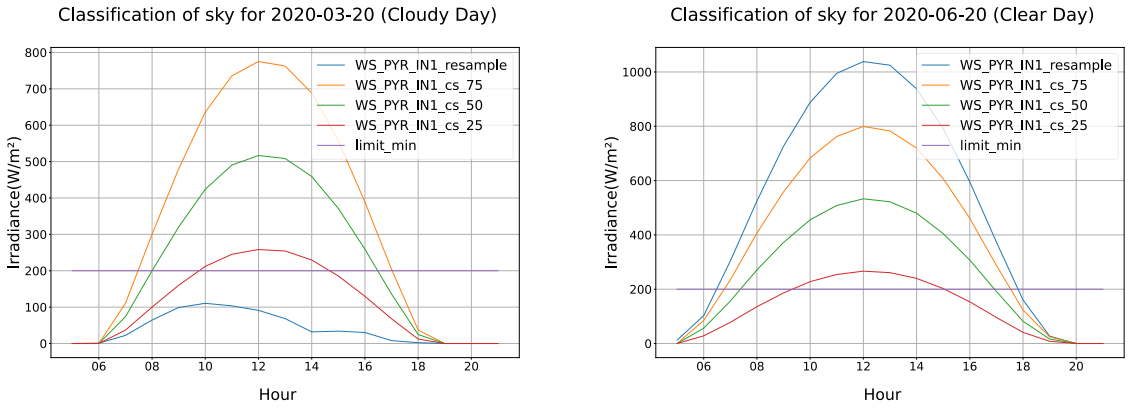
#### 4.1.1.1 SKY'S TYPE

Shadows in photovoltaic systems due to moving clouds are one of the major causes of losses. Fast weather transitions and low irradiance may disable the photovoltaic system operation at maximum power point (MPP) [26]. Thus, it is imperative to include the sky's type as a feature. Table 4-1 presents the sky's type classification under consideration for the AI4PV system for each hour. It includes the okta cloud cover scale [27].

**TABLE 4-1: SKY'S TYPE CLASSIFICATION FOR EACH HOUR**

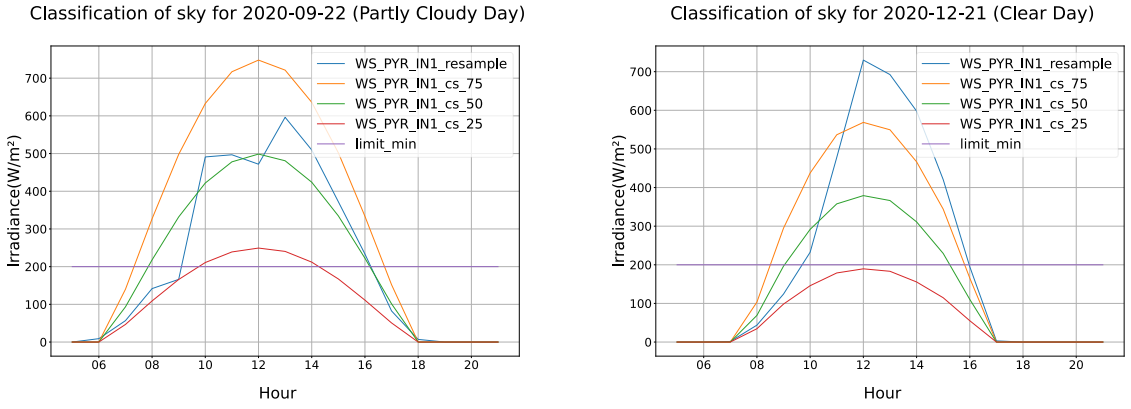
| Sky's Type | Okta | Definition | Condition |
|---|---|---|---|
| Clear | 0-2 | Sky clear- Few Clouds | IF (value ≥ 75% Clear Sky) |
| Partially Clear | 3-4 | Scattered | IF (value ≥ 50% Clear Sky)<br>AND (value < 75% Clear Sky) |
| Partilally Cloudy | 5-6 | Broken | IF (value ≥ 25% Clear Sky)<br>AND (value< 50% Clear Sky) |
| Cloudy | 7-8 | Broken-Overcast | IF (value < 25% Clear Sky)<br>OR (value ≤ minimum limit) |

The clear's sky hourly estimation used on AI4PV project doesn't depend on altitude, longitude, and latitude. For the daily classification (Clear, Partially Cloudy and Cloudy) was necessary to calculate the weighted average value, only for sunset – 3 hours ≥ timestamp ≥ sunrise + 3 hours, allowing a better distinction between the different sky's types. In Figure 4-1 and Figure 4-2, are represented some examples of the daily classifications for the year 2020.




**(A)** **(B)**

**FIGURE 4-1: DAILY CLASSIFICATION FOR 2020-03-20 (A) AND 2020-06-20 (B).**




**(A)** **(B)**

**FIGURE 4-2: DAILY CLASSIFICATION FOR 2020-09-22 (A) AND 2020-12-21 (B).**

In the future, the classification of the sky's type will be implemented using PVLIB/Python library, allowing the comparison with the current clear's sky estimation.

#### 4.1.1.2 SEASON

In the northern hemisphere (latitude > 0º) is expected that during the summer season, when sunny days are more frequent, photovoltaic production will be maximized due to the high irradiance. The opposite occurs during the winter when the output production may be almost inexistent. Therefore, the classification of seasons was considered as a feature (Table 4-2) [28].

**TABLE 4-2: SEASON CLASSIFICATION FOR EACH DAY**

| Season | Start Day Condition |
|---|---|
| **Spring** | 'spring_daystart' = 20 |
| **Summer** | IF (Year is Leap): 'summer_daystart' = 20.<br>ELSE: 'summer_daystart' = 21 |

| | |
|---|---|
| **Autumn** | IF (Year is Leap) OR (Year-1 is Leap): 'autumn_daystart' = 22.<br>ELSE: 'autumn_daystart' = 23 |
| **Winter** | IF (Year is Leap) OR (Year-1 is Leap) OR (Year-2 is Leap): 'winter_daystart' = 21.<br>ELSE: 'winter_daystart' = 22 |

#### 4.1.1.3 WEATHER AND SCADA VARIABLES

For the current deliverable, the digital twin is composed of one inverter (INV_1.1) and two junction boxes (JB_1.1.1 and JB_1.1.2). Additionally, the reduction of the total variable number from 1224 to 21 was made by the following steps:

- Considering only two weather features: Ambient Temperature and Plane Irradiation (Sensor 1);
- Including only SCADA variables from the inverter side. Exceptions: All set-point and control features; Daily Energy Produced; and Total Energy Produced.

## 4.2 FEATURES SCALING AND ENCODING

Two main steps for data pre-processing are encoding categorical features and scaling numerical features. Many methods, such as the K-Nearest Neighbours (K-NN), can't process categorical features. Therefore, encoding is necessary to transform these types of features into numerical data. One of the most common methods is the One Hot Encoder, which maps each category in a variable with binary values (0 or 1) [29].

Additionally, scaling numerical features is required to normalize their range and improve the model performance. Rescaling, also known as Min-Max Normalization, is defined by rescaling the range of features that are not standard normally distributed to a given range (default range is [0;1]) [30].

## 4.3 HYPERPARAMETERS TUNNING

Grid Search and Random Search are two of the most common methods for hyper-parameter tuning. Their purpose is to determine the estimator with the most accurate predictions. For the classifiers C-Support Vector (SVM) and Logistic-Regression, was implemented the function GridSearchCV of the Scikit-Learn/Python library [31], which includes the cross-validation of the training dataset.

For the other classifiers, for instance, Light Gradient-Boosting (LightGBM), the Random Search was applied through the function RandomizedSearchCV [32]. Compared to Grid Search, Random Search allows a faster optimization by selecting random combinations, although it doesn't guarantee the optimal hyper-parameters.

## 4.4 FAULT DETECTION ACCURACY

Multi-class classification can be executed through distinct algorithms such as:

- Logistic Regression: an algorithm used to predict the probability of a target variable. In multi-class classification, it requires splitting the model into N classes (one-vs-rest). The final output will be the class with a higher probability [33] [34].
- Random Forest: an ensemble of decision trees which improves the predictive accuracy and controls the overfitting [35].
- LightGBM: a free and open-source distributed gradient-boosting framework based on decision trees to increase the efficiency of the model and reduce memory usage [36].

The logistic regression model, illustrated in Figure 4-3, was initially implemented for fault classification. The model achieved a Fault Detection Accuracy (FDA) of 80.8%, above the minimum value expected of 80% for this key of performance and enabled a perfect prediction of the switch open-circuit condition for the dataset spanning from 2021 to 2022. Subsequently, as presented in Figure 4-4, the FDA improved to 95.2% through the random forest ensemble method, despite the difficulty of distinguishing the fault-free condition from DC cable degradation.

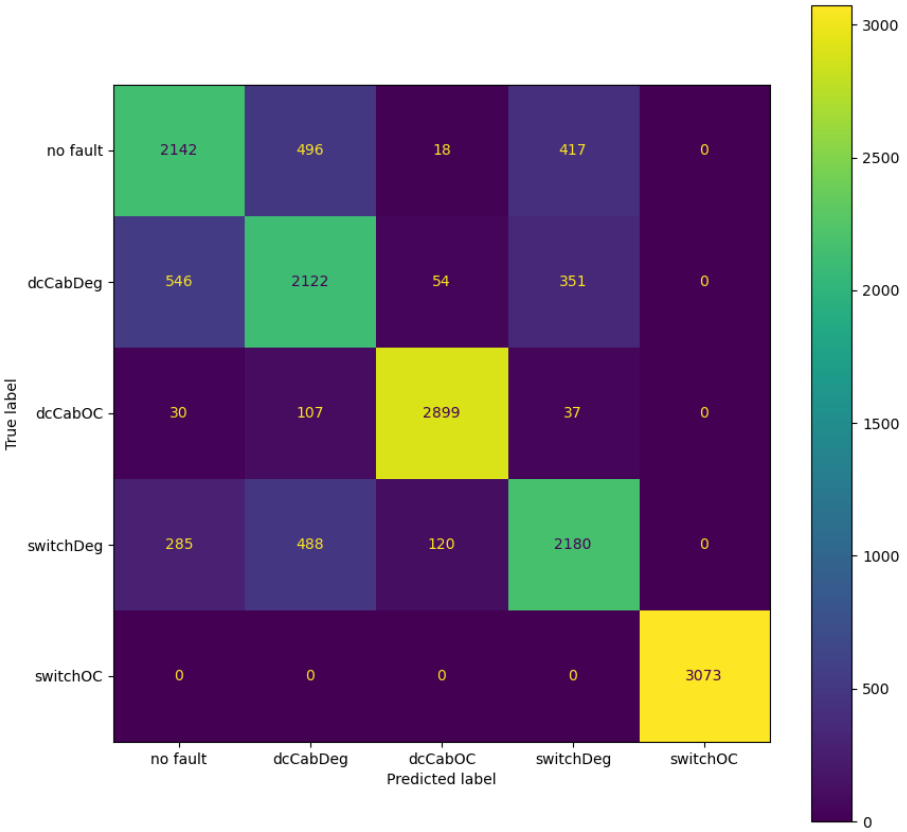



**FIGURE 4-3: CONFUSION MATRIX FOR FAULT CLASSIFICATION USING THE LOGISTIC REGRESSION CLASSIFIER WITH THE DATASET FROM 2021 TO 2022**

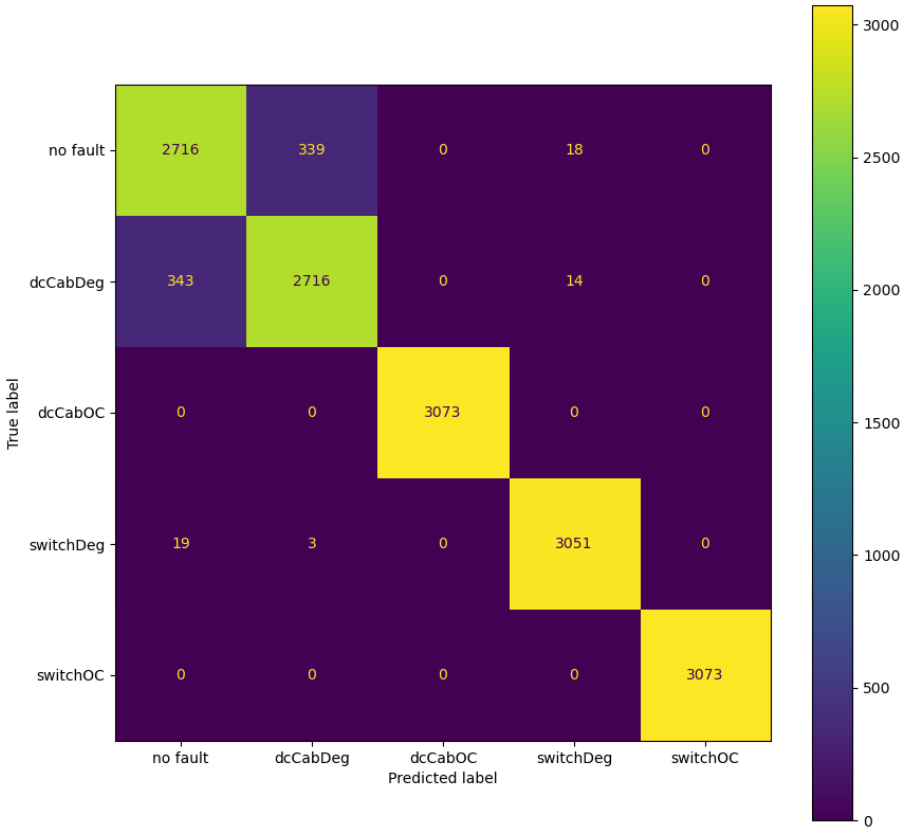
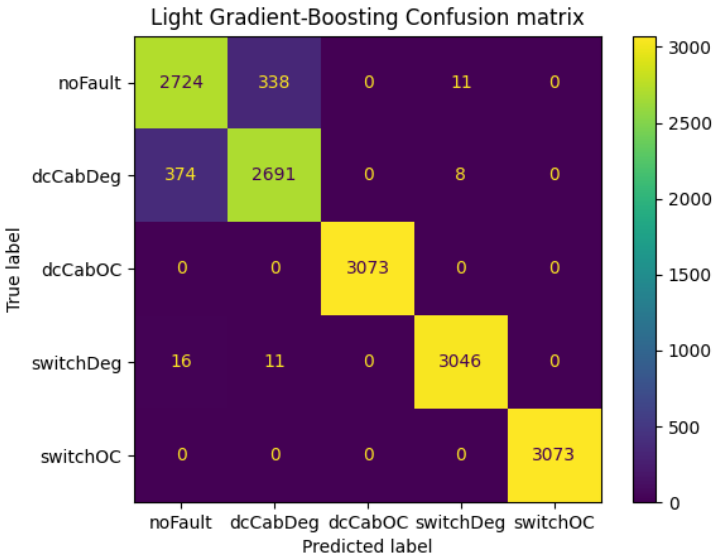
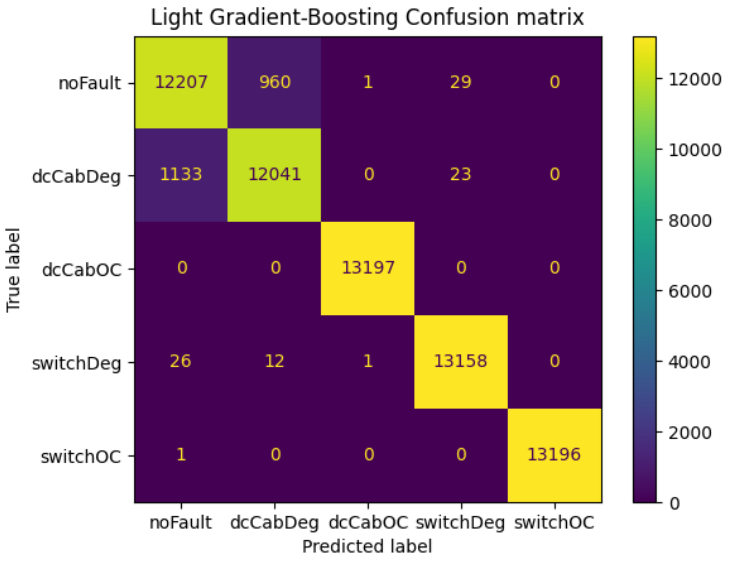



**FIGURE 4-4: CONFUSION MATRIX FOR FAULT CLASSIFICATION USING THE RANDOM FOREST CLASSIFIER WITH THE DATASET FROM 2021 TO 2022.**

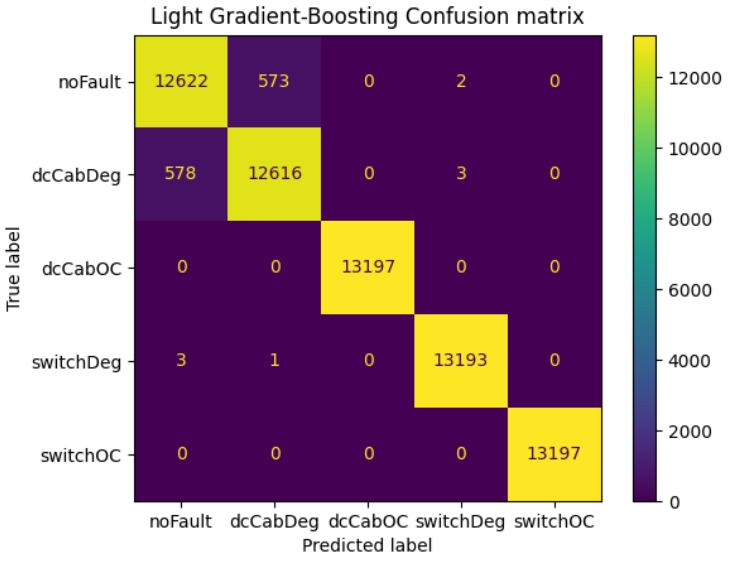
Additionally, the LightGBM algorithm was implemented for fault classification due to its low memory usage, resulting in an FDA of 95.1% (Figure 4-5 (a)).




**(A)**




**(B)**




**(C)**

**FIGURE 4-5: CONFUSION MATRIX FOR FAULT CLASSIFICATION USING THE LIGHTGBM CLASSIFIER WITH DIFFERENT DATASETS: 2021-2022 (A), MARCH 2018-2022 (B), AND MARCH 2018-2022 INCLUDING THE GROUPS OF FEATURES 'LAGS' AND 'STATS' (C) .**

The LightGBM allowed the testing of the dataset from 2018-03 to 2022, with and without including the feature groups 'lags' and 'stats', resulting in an improved FDA of 96.7% and 98.2%, respectively. On Table 4-3 are described the groups of features used on Figure 4-5 (c).

**TABLE 4-3: GROUPS OF FEATURES**

| Group of Features | Description |
|---|---|
| 'irradiance' | Daily Irradiation and Horizontal Irradiance |
| 'weather' | All features related to the weather station (WS_[X])<br>Note1: Horizontal Irradiance excluded.<br>Note2: Average Module Temperature included.<br>Note3: Exclusion of the variables associated with the Plane Irradiance used on the clear sky's estimation. |
| 'inv' | All features related to the inverter (INV_[X]) |
| 'skytype' | Sky's type features |
| 'time' | All features associated to date, sunrise, and sunset. |
| 'stats' | Mean and standard deviation of measurements for all pyranometers in the park<br>Note: Pyranometers are all features starting with 'PYR', but don't belong to 'irradiance' or 'weather' |
| 'lags' | All features containing 'lag' or 'windowmean':<br>• 'lag_N': value of a specific variable (N*Granularity) minutes ago;<br>• Sliding window that calculates a mean value for the measurements within the last 1 hour (for each fault type);<br>• Additional sliding window for the Plane Irradiance used on the clear sky's estimation, for each hour (over ≈ 1h)<br>Note: Include all features that don't belong to 'skytype', 'time', 'stats' |

## 4.5 EXPLAINABLE ARTIFICIAL INTELLIGENCE TECHNIQUES

Although Interpretability and Explainability are commonly interchangeable, it is possible to define both terms separately. A model is interpretable if it is capable of being understood by humans on its own. On the other hand, a model is explainable if it is too complex for a human to comprehend and requires additional methods/techniques (Explainable Artificial Intelligence Techniques, also called XAI Techniques). XAI Techniques "enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners". In Figure 4-6 is represented a diagram of the XAI approaches.

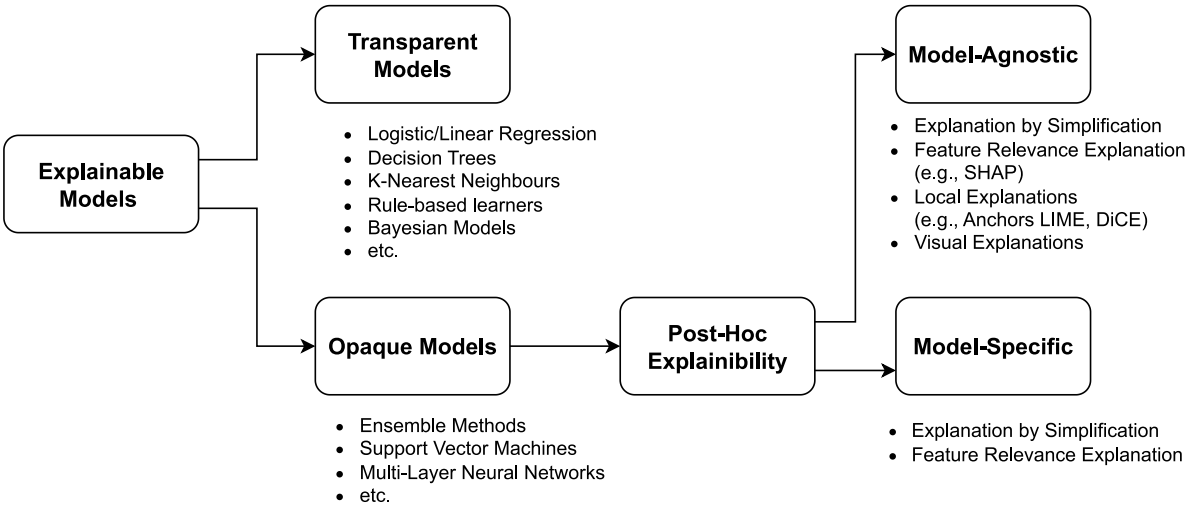



**FIGURE 4-6: EXPLAINABLE ARTIFICIAL INTELLIGENCE APPROACHES**

The authors of [37] denoted the emergence of XAI techniques in the energy sector by proposing the implementation of local, model-agnostic post hoc explanation approaches in the context of PV fault detection from a multi-layer perceptron (MLP) model, such as [38] [39]:

- Feature relevance explanations: aim to measure the importance of a model's inputs to its output . This results in an importance score ranking, where higher scores mean that the corresponding variable was more relevant for the model . E.g., SHapley Additive exPlanation (SHAP);
- Local explanations: approximate the model in a narrow area around a specific instance of interest . The resulting explanations do not necessarily generalize to a global scale but approximate the model around the instance the user wants to explain . E.g., Anchors and Diverse Counterfactual Explanations (DiCE).

### 4.5.1 ANCHORS

Anchors is a model-agnostic explanation based on if-then rules, also known as a rule-based learner. This method explains individual predictions by finding a set of rules that "anchors" the output sufficiently and independently of other features change [40].

### 4.5.2 DICE

DiCE is a counterfactual explanation and produces feature-perturbed versions of the original observations, which result in a change of prediction [41]. The main results are the ones which cause a relevant change in the model's output, like a flip in a predicted class. Nevertheless, DiCE can produce highly varying explanations due to its diverse nature, which might be contradictive but could also be useful for model debugging [37].

### 4.5.3 SHAP

In the case study of [37], and by evaluating the different approaches according to their stability and the generated explanations consistency, SHAP recorded the best performance, followed by Anchors and DiCE.

SHAP is a game-theoretic approach to XAI techniques, which computes the contribution of each feature on a decision by evaluating their additive measure of importance, also called Shapley value. Although, and due to its nature, for consistence and transparent results in local and global interpretations, the predictive model may have only independent features [38]. In addition, this technique is available in Python [42] and R languages.

Currently, the XAI techniques are not implemented in the AI4PV project. However, in the context of root-cause analysis, it will be expected to use the SHAP method to ensure the transparency and consistency of the model, as well to improve the project performance on fault diagnosis and localisation.

# 5. ML ALGORITHMS FOR PVPP POWER TRANSFORMER

The fault classification requires the implementation of a hybrid dataset. This dataset is composed of real and synthetic data for fault-free and faulty conditions (without noise), respectively. Thus, after validating the fault-free data [24], a digital twin (DT) was implemented in Simulink/MATLAB® to generate a faulty dataset (due to the lack of real faulty data) [25].

## 5.1 FEATURE ENGINEERING

Similarly to what was done for the inverter-related AI algorithms, the same procedure was followed for the transformer's ones.

In particular, the AI algorithms were developed considering the following features:

- Weather features: Ambient Temperature and Plane Irradiation (Sensor 1);
- Including only SCADA variables from the inverter output and transformer side. In particular, measurements such as current, voltage, power factor of the output of the inverter were considered as this represent the connection point between the inverter and the transformer. Measurements on the grid-side were included, such as current injected, voltage at the PCC, power factor.

Additionally, scaling numerical features is required to normalize their range and improve the model performance. Rescaling, also known as Min-Max Normalization, is defined by rescaling the range of features that are not standard normally distributed to a given range (default range is [0;1]) [30].

Hyper-parameter tuning was performed to identify the estimator with the most accurate predictions.

Grid Search and Random Search are two of the most common methods for hyper-parameter tuning. For all the classifiers (Logistic-Regression, Gradient Boosting and Random Forest), was implemented the function GridSearchCV of the Scikit-Learn/Python library [31], which includes the cross-validation of the training dataset.

## 5.2 FAULT DETECTION ACCURACY

Multi-class classification can be executed through distinct algorithms such as:

- Logistic Regression: an algorithm used to predict the probability of a target variable. In multi-class classification, it requires splitting the model into N classes (one-vs-rest). The final output will be the class with a higher probability [33] [34].
- Random Forest: an ensemble of decision trees which improves the predictive accuracy and controls the overfitting [35].
- LightGBM: a free and open-source distributed gradient-boosting framework based on decision trees to increase the efficiency of the model and reduce memory usage [36].

The logistic regression model, illustrated in Figure 5-1, was initially implemented for fault classification. The model achieved a Fault Detection Accuracy (FDA) of 86.5%, above the minimum value expected of 80% for this key of performance.

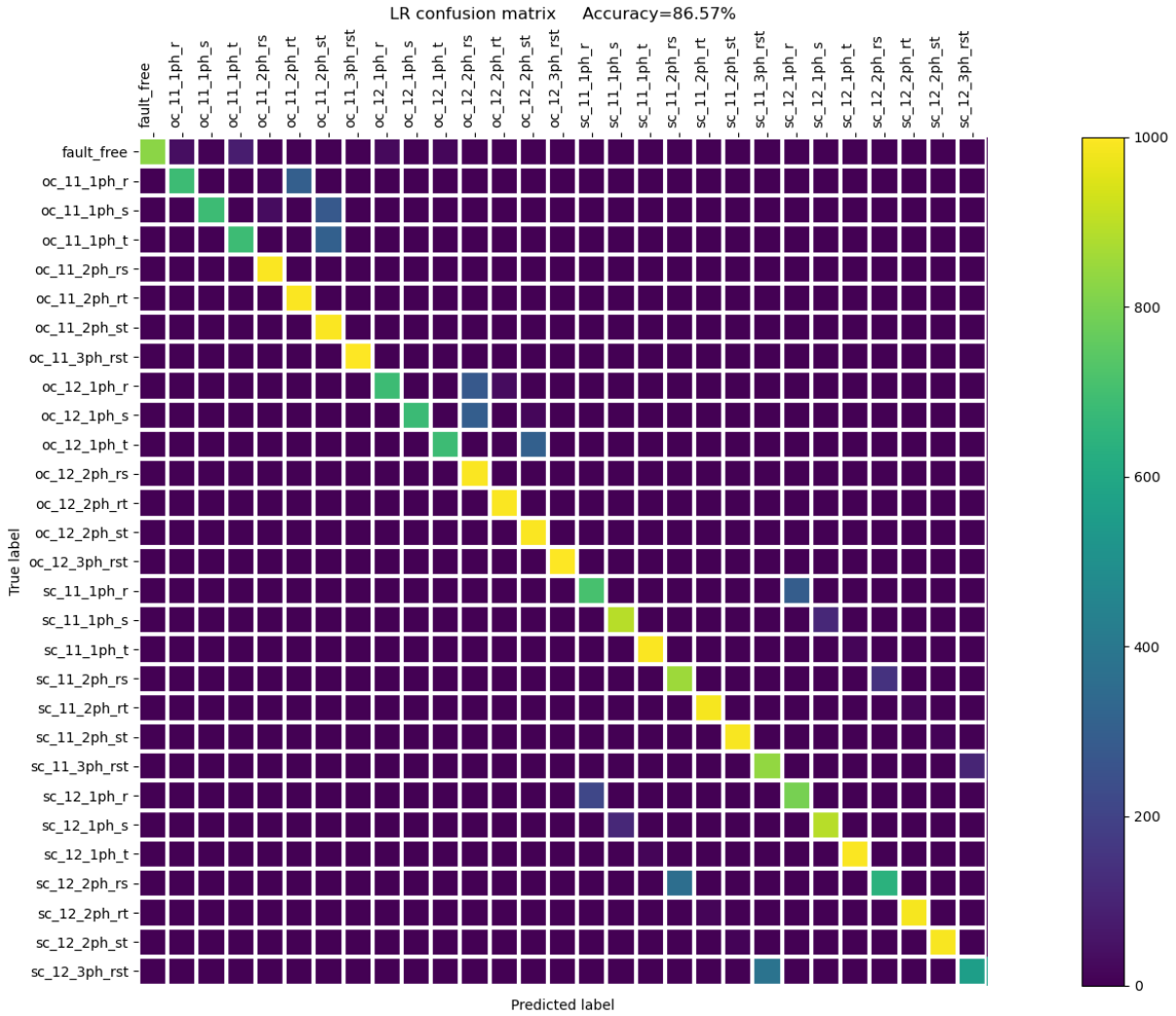



**FIGURE 5-1: CONFUSION MATRIX FOR FAULT CLASSIFICATION OF POWER TRANSFORMER'S FAULTS USING THE LOGISTIC REGRESSION CLASSIFIER**

Subsequently, as presented in Figure 5-2, the FDA improved to 91% through the random forest ensemble method, despite the difficulty of distinguishing the short circuits on the two low voltage windings of the transformer.

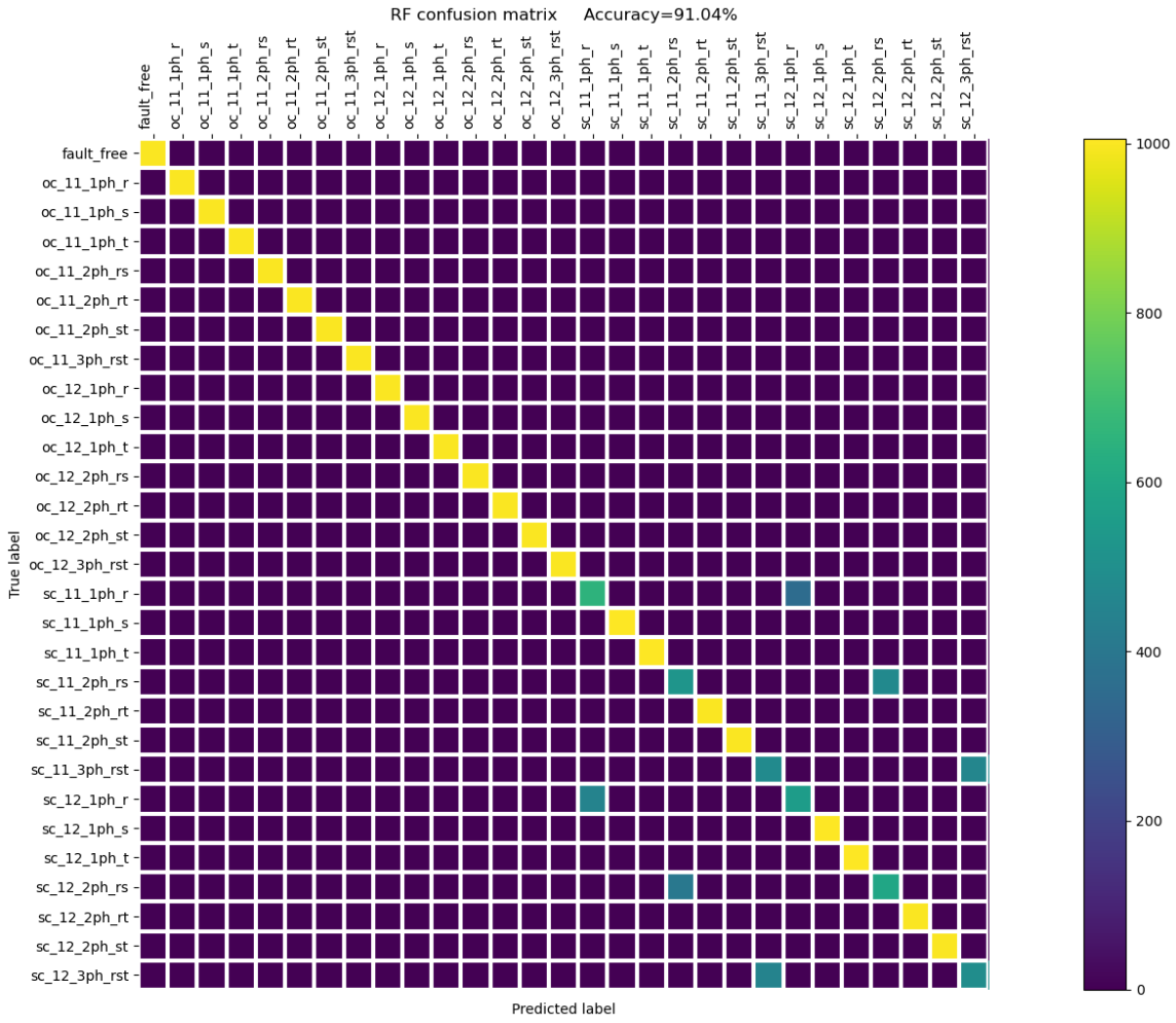



**FIGURE 5-2: CONFUSION MATRIX FOR FAULT CLASSIFICATION OF POWER TRANSFORMER'S FAULTS USING THE RANDOM FOREST CLASSIFIER**

Additionally, the LightGBM algorithm was implemented for fault classification due to its low memory usage, resulting in an FDA of 96.88% (Figure 5-3).

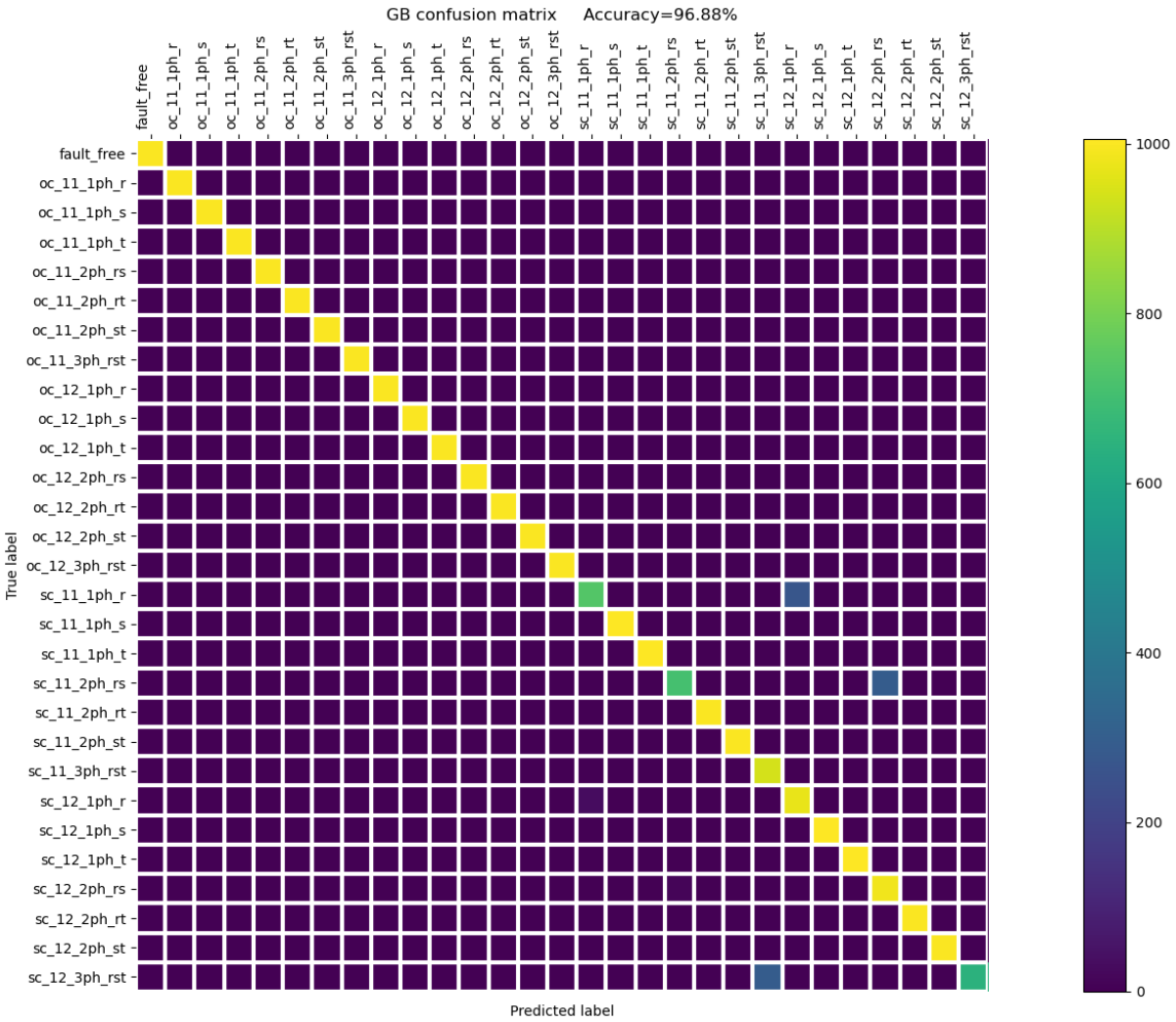



**FIGURE 5-3: CONFUSION MATRIX FOR FAULT CLASSIFICATION OF POWER TRANSFORMER'S FAULTS USING THE LIGHTGBM CLASSIFIER**

# 6. CONCLUSIONS

This deliverable presented the outcomes developed within the Task 3.1: Root cause analysis and asset replacement, which results in the factors that influence the PV performance and the fault detection and classification algorithms. Starting from the theory, the PV inverter and power transformer DT was used to generate the hybrid dataset, as real faulty data was not available. After the validation of the DT, the faults were simulated and added to the dataset, which was later processed by the ML algorithms. By pattern recognition, the ML algorithms were able to classify the faults with accuracy higher then 90%, which is a promising result. Of course, under a field testing this accuracy should be reduced due to noise, unprecedent issues, etc., nevertheless an accuracy above 80% is still achievable, as determined by the KPIs of D4.2 [43]. Such development provided two main outcomes after all: the digital twin framework, i.e., how to build a hybrid dataset, the main problems faced in such developments and their solutions, etc.; and the fault classification algorithms, which provide insightful information about the current state of the PVPP, supporting O&M staff on their decision making.

# 7. REFERENCES

[1] G. D. L. e. al., "Review of O amp;M Practices in PV Plants: Failures, Solutions, Remote Control, and Monitoring Tools," *IEEE Journal of Photovoltaics*, 2020.

[2] J. A. D. e. al., "An Effective Evaluation on Fault Detection in Solar Panels," *Energies*, 2021.

[3] S. S. R. a. W. G. M. Kais AbdulMawjood, "Detection and prediction of faults in photovoltaic arrays: a review," in *IEEE 12th International Conference*, 2018.

[4] P. J. e. al., "A Digital Twin Approach for Fault Diagnosis in Distributed Photovoltaic Systems," *IEEE Transactions on Power Electronics*, 2020.

[5] Q. N. e. al., "Fault Diagnostic Methodologies for Utility-Scale Photovoltaic Power," *Sustainability*, 2021.

[6] F. B. a. H. W. Shuai Zhao, "An Overview of Artificial Intelligence Application for Power Electronics," *IEEE Transactions on Power Electronics*, 2021.

[7] A. M. e. al., "Data Mining Applications to Fault Diagnosis in Power Electronics Systems: A Systematic Review," *IEEE Transactions on Power Electronics*, 2022.

[8] G. A. O. e. al., "Diagnosis of Electrical Distribution Network Short Circuits Based on Voltage Park's Vector," *IEEE Transactions on Power Delivery*, 2012.

[9] Y. X. a. Y. Xu, "A Transferrable Data-Driven Method for IGBT Open-Circuit Fault Diagnosis in Three-Phase Inverters," *IEEE Transactions on Power Electronics*, 2021.

[10] V. F. P. a. A. J. P. J. F. Martins, "Unsupervised Neural-Network-Based Algorithm for an On-line Diagnosis of Three-Phase Induction Motor Stator Fault," *IEEE Transactions on Industrial Electronics*, 2007.

[11] V. S. B. K. e. al., "A Review on Artifcial Intelligence Applications for Grid-Connected Solar Photovoltaic Systems," *Energies*, 2021.

[12] GreenPowerMonitor, "Position Paper: PREDICTIVE MAINTENANCEOF SOLAR PV PLANTS: THE TIME IS NOW," DNV GL GROUP TECHNOLOGY & RESEARCH, 2021.

[13] G. M. T. e. al., "A State-of-Art-Review on Machine-Learning Based Methods for PV," *Applied Sciences*, 2021.

[14] A. E. L. e. al., "A Monitoring System for Online Fault Detection and Classification in Photovoltaic Plants," *Sensors*, 2020.

[15] A. M. a. S. Kalogirou, "Artificial intelligence and internet of things to improve efficacy of diagnosis and remote sensing of solar photovoltaic systems: Challenges, recommendation and future directions," *Renewable and Sustainable Energy Reviews*, 2021.

[16] J. P. e. al., "Model-Based Fault Detection and Identification for Switching Power Converters," *IEEE Transactions on Power Electronics*, 2017.

[17] A. M. e. al., "Overview of fault detection approaches for grid connected photovoltaic inverters," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, 2022.

[18] L. M. d. O. e. al., "Siamese Neural Network Architecture for Fault Detection in a Voltage Source Inverter," in *Brazilian Power Electronics Conference*, 2021.

[19] J. K. a. B. Kroposki, "Understanding Fault Characteristics of Inverter-Based Distributed Energy Resources," National Renewable Energy Lab., 2010.

[20] Pandas, "pandas.DataFrame.fillna," 05 2023. [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.fillna.html.

[21] Scikit-learn, 05 2023. [Online]. Available: https://scikit-learn.org/stable/modules/multiclass.html.

[22] Wikipedia, "Per-unit system," 05 2023. [Online]. Available: https://en.wikipedia.org/wiki/Per-unit_system.

[23] Wikipedia, "Interquartile range," 05 2023. [Online]. Available: https://en.wikipedia.org/wiki/Interquartile_range.

[24] Miguel Angel Delgado (ISOTROL), Sergio Raigon (ISOTROL),Ricardo Morales (ISOTROL), Jose Garcia Franquelo (ISOTROL), Rubén González (ISOTROL), Christian Verrecchia (EDP NEW), Louelson Costa (INESCTEC), D2.2 - Data management and modelling tools, AI4PV project.

[25] Miguel Angel Delgado (ISOTROL), Sergio Raigón (ISOTROL), Ricardo Morales (ISOTROL), Jose Garcia Franquelo (ISOTROL), Rubén González (ISOTROL), Louelson Costa (INESC TEC), Christian Verrecchia (EDP NEW), D2.3 - Out of normality analysis report, AI4PV project.

[26] S. V. Kari Lappalainen, "Recognition and modelling of irradiance transitions caused by moving clouds," *Solar Energy*, 2015.

[27] Wikipedia, "Okta," 05 2023. [Online]. Available: https://en.wikipedia.org/wiki/Okta.

[28] Wikipedia, "Solstice," 05 2023. [Online]. Available: https://en.wikipedia.org/wiki/Solstice.

[29] Scikit-learn, "OneHotEncoder," 05 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

[30] Scikit-learn, "Compare the effect of different scalers on data with outliers," 05 2023. [Online]. Available: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#sphx-glr-auto-examples-preprocessing-plot-all-scaling-py.

[31] Scikit-learn, "Tuning the hyper-parameters of an estimator," 05 2023. [Online]. Available: https://scikit-learn.org/stable/modules/grid_search.html.

[32] Scikit-learn, "RandomizedSearchCV," 05 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html.

[33] J. Brownlee, "Multinomial Logistic Regression With Python," 05 2023. [Online]. Available: https://machinelearningmastery.com/multinomial-logistic-regression-with-python/.

[34] Scikit-learn, "LogisticRegression," 05 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression.

[35] Scikit-learn, "RandomForestClassifier," 05 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[36] LightGBM, "Welcome to LightGBM's documentation!," 05 2023. [Online]. Available: https://lightgbm.readthedocs.io/en/v3.3.5/.

[37] C. M. J. S. R. S. C. U. Christian Utama, "Explainable artificial intelligence for photovoltaic fault detection: A comparison of instruments," *Solar Energy*, 2023.

[38] P. P. S. E. A. J. R. A. N. I. &. A. P. M. Angelov, "Explainable artificial intelligence: an analytical review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2021.

[39] P. I. Belle Vaishak, "Principles and Practice of Explainable Machine Learning," *Frontiers in Big Data* , 2021.

[40] M. T. S. S. &. G. C. Ribeiro, "Anchors: High-Precision Model-Agnostic Explanations," in *AAAI Conference on Artificial Intelligence*, 2018.

[41] A. S. C. T. Ramaravind Kommiya Mothilal, "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations," in *Conference on Fairness, Accountability, and Transparency*, 2019.

[42] S. Slundberg, "SHAP," 05 2023. [Online]. Available: https://github.com/slundberg/shap.

[43] Christian Verrecchia (EDP NEW), Louelson Costa (INESC TEC), Ruben Gonzalez Bernal (ISOTROL), D4.2 - Demonstration Plan, AI4PV project.

[44] "pandas.DataFrame.fillna," 22 02 2023. [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.fillna.html.

[45] "sklearn.preprocessing.OneHotEncoder," 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

[46] "Solstício," 2023. [Online]. Available: https://pt.wikipedia.org/wiki/Solst%C3%ADcio.

[47] "Compare the effect of different scalers on data with outliers," 2023. [Online]. Available: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#sphx-glr-auto-examples-preprocessing-plot-all-scaling-py.

[48] "3.2. Tuning the hyper-parameters of an estimator," 2023. [Online]. Available: https://scikit-learn.org/stable/modules/grid_search.html.

[49] "sklearn.model_selection.RandomizedSearchCV," 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html.

[50] R. Isermann, Fault-Diagnosis Applications: Model-Based Condition Monitoring, Springer, 2011.

[51] M. L. P. R. Remus Teodorescu, Grid Converters for Photovoltaic and Wind Power Systems, John Wiley and Sons, Ltd, 2010.

[52] J. R. G. a. E. R. Marcelo Gradella Villalva, "Comprehensive Approach to Modeling and Simulation of Photovoltaic Arrays," *IEEE Transactions on Power Electronics*, 2009.

[53] S. S. R. a. W. G. M. Kais AbdulMawjood, "Detection and prediction of faults in photovoltaic arrays: A review," in *International Conference on Compatibility, Power Electronics and Power Engineering*, 2018.