Artificial Intelligence for Operation and Maintenance of PV Plants

Deliverable D4.3

Validation results and Cost-Benefit analysis report

Lead Beneficiary	EDP NEW
Delivery Date	30/06/2023
Dissemination Level	Public
Status	Released
Version	1.0
Keywords	Validation, Cost-Benefit Analysis, Machine Learning, PV

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.

This Deliverable reflects only the author's views and the Agency is not responsible for any use that may be made of the information contained therein. The AI4PV consortium cannot warrant that information contained in this document is free from risk and, neither the Agency nor the AI4PV consortium parties are responsible for any use that may be made of the information contained therein. This document may contain material, which is the copyright of certain AI4PV consortium parties, and may not be reproduced or copied without permission. The commercial use of any information contained in this document may require a license from the proprietor. The sole responsibility for the content of this publication lies with the authors and all AI4PV consortium parties have agreed to full publication of this document.

Document Information

Project Acronym	AI4PV
Work Package	WP 1
Related Task(s)	T4.3
Deliverable	D4.3
Title	Validation results and Cost-Benefit analysis report
Author(s)	Christian Verrecchia (EDP NEW), Louelson Costa (INESCTEC), Ana Silva (INESCTEC), Miguel Angel Delgado (ISOTROL), Jose Garcia Franquelo (ISOTROL)

Revision History

Revision	Date	Description	Reviewer
0.1	27 April 2023	Outline of report content	EDP NEW
0.2	11 May 2023	Full draft of full content	EDP NEW
0.3	18 May 2023	Partners inputs	INESCTEC, ISOTROL
0.4	15 June 2023	Second version	EDP NEW
1.0	30 June 2023	Final version	EDP NEWI

EXECUTIVE SUMMARY

This deliverable includes the main results obtained in task **T4.3 Validation of results and Cost-Benefit analysis**. The work carried out in this task aimed at validating the models developed in WP2 and WP3, integrated in Task 4.1 and tested in Task 4.2, with experimental data.

The validation took place in the demonstration site, a real operating PV park, that offered real-world operational conditions. The demonstration plan, drafted in WP4 was used as baseline for the validation of the AI4PV solutions. Key Performance Indicators were computed, according to that plan, in order to evaluate the performance of the proposed tools.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

Acronym	Meaning
AEP	Annual Energy Production
API	Application Programming Interface
CAPEX	Capital Expenditure
CBA	Cost Benefit Analysis
DC	Direct Current
DT	Digital Twin
FDA	Fault Detection Accuracy
FLA	Fault Localization Accuracy
G	Solar Irradiance
IQR	Interquartile Range
KPI	Key of Performance
LightGBM	Light Gradient-Boosting Machine
MeasID	Measurement Tag
ML	Machine Learning
MPPT	Maximum Power Point Tracking
O&M	Operation and Maintenance
PLL	Phase-Locked Loop
PR	Performance Ratio
PV	Photovoltaic
PVPP	Photovoltaic Power Plant
RadMod,T	PV Modules Temperature
RadPl	Plane Irradiance
REST API	Representational State Transfer Application Programming Interface
RSL	Reduce Soiling Losses
SCADA	Supervisory Control and Data Acquisition
VarType	Variable Type
$\mathbf{V_b}$, $\mathbf{V_{norm}}$	Base value
$\mathbf{V_{max}}$	Maximum value
$\mathbf{V_{min}}$	Minimum value
XAI	Explainable Artificial Intelligence

COMPETE 2020 PORTUGAL 2020 CDTI @CDTIoficial European Union European Regional Development Fund

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

COMPETE 2020 PORTUGAL 2020 CDTI @CDTIoficial European Union European Regional Development Fund

1. INTRODUCTION

This document, deliverable **D4.3 Validation results and Cost-Benefit analysis report**, includes the results of the validation of the AI4PV solutions developed in WP2 and WP3, integrated in Task 4.1 and tested in Task 4.2.

1.1 SCOPE OF REPORT

D4.3 - Validation of Results and Cost-Benefit Analysis report presents the outcomes of a crucial task undertaken to assess the effectiveness and reliability of the solutions developed in WP2 and WP3. Task T4.1 integrated these models, which were subsequently tested in task T4.2 using experimental data. The primary objective of this validation process was to determine the degree to which the developed models accurately represented real-world scenarios and to evaluate the overall efficacy of the proposed solutions.

In order to establish the reliability of the developed models, a comprehensive comparison was made between the numerical results obtained from simulations and the corresponding experimental data. By undertaking this rigorous analysis, we gained valuable insights into the performance of the models and identified areas that required fine-tuning to ensure their robustness.

This report goes beyond model validation, as it also explores the practical implementation of the developed solutions at demonstration sites. These sites provided a unique opportunity to assess the real-world impact of each solution, offering a glimpse into their effectiveness under varying conditions. Furthermore, the report delves into a comprehensive cost-benefit analysis, examining the financial implications associated with the application of these solutions.

Throughout this report, we present the validation results, which encompass an assessment of the solutions' effectiveness and reliability. Various key performance indicators (KPIs) defined in the demonstration plan are utilized to gauge the overall performance of the models. The subsequent cost-benefit analysis further validates the viability of the developed solutions, shedding light on their economic feasibility and long-term sustainability.

1.2 OUTLINE OF REPORT

This report is structures as follows:

- ***Chapter 1*** introduces the scope of the report.
- ***Chapter 2*** provides an overview of the improvements made on the inverter fault detection and classification tool, compared to the one proposed in [1].
- ***Chapter 3*** presents the results of the validation. KPIs and metrics defined in the demonstration plan are computed and compared to their targets.
- ***Chapter 4*** presents the Cost-Benefit Analysis performed to compare the AI4PV cleaning module against traditional practises.
- ***Chapter 5*** summarizes what is presented in this report.

2. IMPROVEMENT ON THE PV INVERTER FAULT CLASSIFICATION TOOLS

In this chapter, the improvements on the PV inverter fault classification tools, introduced in [1], are described.

The PV inverter fault classification is composed of multiple stages, starting with the raw data until the diagnosis is achieved. Such framework is described in this chapter.

2.1 OVERALL PIPELINE

The PV inverter pipeline, i.e., the subsystem within the AI4PV solutions responsible for the fault and failures detection and diagnosis of the PV inverter is presented in Figure 2-1.

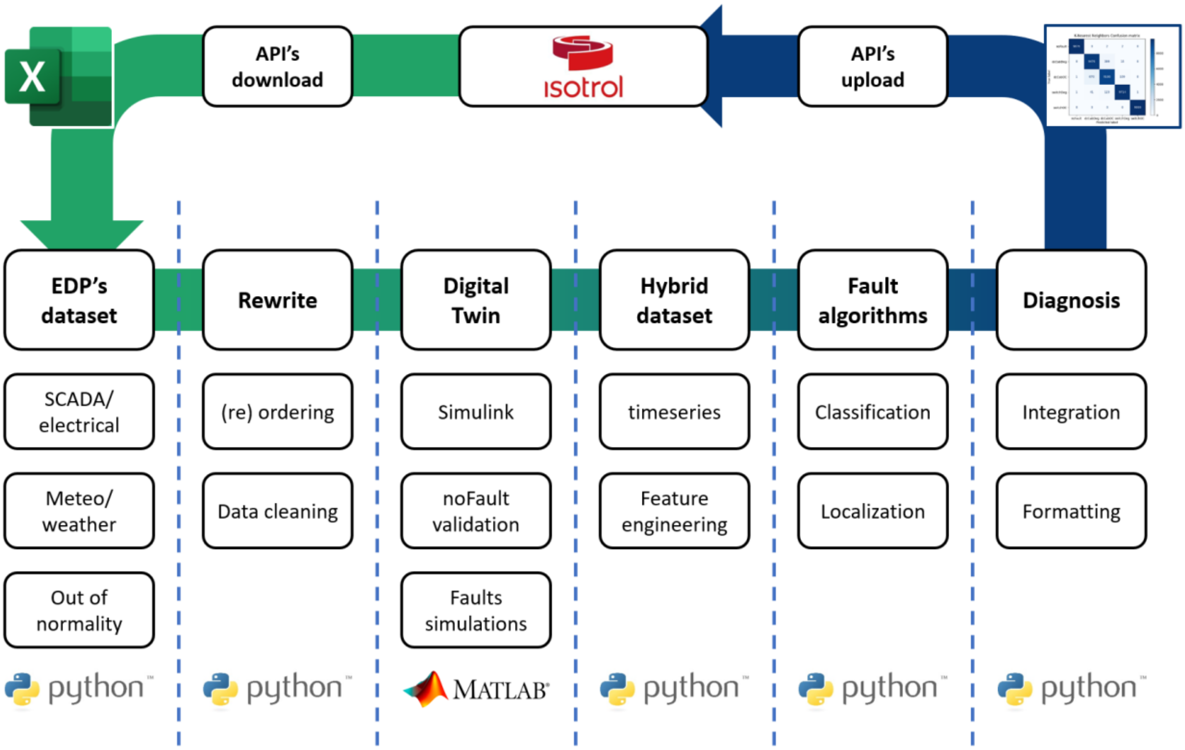


FIGURE 2-1: PV INVERTER DIGITAL TWIN DATA PIPELINE

The pipeline starts with the download of the dataset (which was first done for the whole data previously available from the PV power plant, later downloading data on a daily basis). Having the EDP's dataset available (including weather and SCADA data), a rewriting of the data is carried out, doing a second verification regarding outliers, formatting, etc. Then the digital twin is fed the time series for fault validation (which was done during the previous work packages) and random fault scenario generation (the current state at the time of the validation), which results in the hybrid dataset (composed of real and synthetic data).

Besides the digital twin, the fault algorithms, i.e., classification, localization, and out-of-normality, do the analysis of the given dataset (day, month, year, etc.) and generate multiple outputs regarding the state of the inverter and some equipment connected to it (for instance, the junction boxes). Those various reports from the different algorithms are then summarized and formatted during the diagnosis, providing insightful information for the PV power plant operators. Lastly, the report is uploaded through the API once again (which is the same one used during the download).

Most of the pipeline is developed using Python libraries, i.e., open-source solutions. However, due to the level of detail required for the implementation of the PV inverter itself, Simulink/MATLAB® is still applied, thus being the only licensed software needed for the pipeline's proper operation. MS Office Excel is depicted in the pipeline. Still, it was primarily used for visualization and storage, as all of the manipulations of the data were done using Python or MATLAB scripts.

The following sections tackle the technicality of each stage of the pipeline, providing pseudo-codes, a draft, an explanation of the implementation of the solutions, etc. It is worth noting that even though this solution is tailored to the AI4PV validation site, the pipeline and its methodology can be applied to any PVPP, considering that another PVPP might have another configuration, power level, PV modules, inverter technology, etc.

2.1.1 DEFINITIONS

There are some key terms that are being applied to the digital twin of the PV inverter. Besides the term digital twin, some other definitions are essential for adequately developing tools such as the recommender system. The digital twin was first introduced by Michael Grieves [2], stating that it is a "*Virtual representation of real-world entities and processes, synchronized at a specified frequency and fidelity*". Thus, besides having a real asset and a digital asset, there must be a data flow between them at a specified frequency. In the case of the AI4PV project, the real asset is the PV power plant and its multiple subsystems (PV modules, PV inverter, transformer, etc.). In that sense, beyond a regular simulation, the level of detail and the information trade between the real twin and the digital twin as critical features that will define the tool.

Also, the definition was already tackled in [3] regarding fault and failure, detection and diagnosis, etc. But it is worth remembering that, in general, a fault is a problem that reduces a system's performance but doesn't make it stop. On the other hand, a failure is such a dire problem that it will cause a system or a subsystem to have its working completely halt. For example, a degradation can be seen as a fault, whilst an open circuit issue can be seen as a failure.

Lastly, the diagnosis is the outcome of a series of steps to diagnose a problem (fault or failure) properly. The first step is detection, where it will be pointed out if there is a fault, thus a species of binary classification. Nevertheless, the AI4PV project performs further, including fault classification and localization, thus providing an insightful report of the condition of the assets of the PV power plant, such as the PV inverter.

Such performance shows that the AI4PV tools are not only on par with the start of the art but also pushing the boundaries of this technology by providing diagnosis to multiple assets of the PV power plant, mainly the ones related to electrical power engineering, i.e., PV modules, power electronics, power transformers, etc.

2.1.2 DATA FLOW

The data flow of the PV inverter, or the pipeline for the fault and failure detection and diagnosis of the PV inverter, follows a logical approach that combines data analytics methodologies, power electronics simulation and machine learning models.

Having prior knowledge of the PV power plant assets and the dataset formatting, it is possible to develop a pattern that will permeate the whole pipeline. Of course, as the digital twin solutions are tailored to a specific asset, creating a pipeline for another PV power plant is possible, and it can use the same pipeline. However, the details of each building block of the pipeline can be different. For instance, the digital twin of the PV inverter takes into consideration the technical specifications of the power electronics converter, or the machine learning models take into consideration the configuration of the PV power plant to provide a proper fault classification, etc.

The data flow starts with the download of the real data (daily, monthly, yearly, etc.) through an API developed by Isotrol, which retrieves the data from EDP’s SCADA system. After the download, which can include weather, inverter, string box, transformer data, etc., a rewrite building block takes place. This block performs some simple data cleaning, which at the validation stage is redundant as the data being provided is already clean. Still, it was necessary during the first stages of developing the digital twin of the PV inverter.

The digital twin uses the time-series real data to perform multiple functions, such as model validation using fault-free data, exploring future configuration scenarios, or generating faulty data, i.e., simulated data of multiple conditions that the PV inverter is susceptible to (such as switches faults, DC cable disconnection, short-circuits, etc.). The outcome is a hybrid dataset, whereas the real data can be combined with the synthetic data to generate the hybrid dataset. Also, under the hybrid dataset, some pre-processing is done by adding some weather/climate-related features that will improve the performance of the classification and localization algorithms.

The developed algorithms perform the classification (consequently, detection) through a REST API, followed by localization. In that way, besides pointing out if there is a problem, it can indicate the type of the fault and in which equipment it occurred (string box, inverter, etc.). Lastly, all of the pipeline outcomes depicted in Figure 2-1 are summarized in the diagnosis building block, performing the integration of the whole system and formatting the output to a time-series pattern achieved in agreement with the project's other partners. After that, the report is uploaded back to the PV power plant operators.

2.2 THE AVAILABLE DATA AND INFORMATION FROM THE PV INVERTER

To build a digital twin of a given asset, it is necessary to have information about its technical specifications (i.e., electrical ratings), the configuration (i.e., how the PV modules, junction boxes, PV inverter and transformer are connected), and input data (i.e., electrical and weather data).

COMPETE 2020 PORTUGAL 2020 CDTI @CDTIoficial European Union European Regional Development Fund

It is worth noting that the more detailed information and data with high granularity, the more precise the digital twin will be. However, it is necessary to make a trade-off between the available data and how detailed the digital twin needs to be. If a certain level of detail is enough, a discretization of the available data can occur, or using generalized/classical models for some subsystems might be a better approach.

2.2.1 PV MODULES, PV INVERTER, AND CONFIGURATION

The PV modules datasheet provides the information needed to use an existing PV module model, such as the ones supplied by Simulink/MATLAB® or to build your mathematical model using a single-diode or two-diode model [4]. This model must be carefully developed, as it is responsible for the power source of the PV inverter digital twin. Thus, properly validating the model by comparing it with real data is advised. If one is using real irradiance and temperature data as input, fine-tuning might be required if the calibration state of the sensor is still being determined. Of course, it is not expected that the error between the real and simulated data to be zero, but a KPI must be indicated before moving on to the validation of the PV modules (or the validation of any subsystem, actually) [5].

The PV inverter datasheet provides valuable information to build the digital twin like the PV modules. Minimum and maximum current and voltage ratings, nominal values, power ratings, etc., will define the operating conditions of the PV inverter. Suppose details about the reactive components (capacitors, inductors) and control (MPPT, PLLs, current control, etc.) are available. In that case, they can be added or implemented on the simulation platform, but this is only sometimes the case. Having at least the datasheet information should be enough, though, since in the case of lacking details about the hardware and firmware of the PV inverter, a generalized/classical approach can be taken: applying classical current control and MPPT, using solid references for the design of capacitors and inductors, etc. After all, a validation of the model is still necessary. In that case, usually, the current will have a smaller error when compared to the real fault-free data, whilst voltages (mainly the DC-link voltage and output of the MPPT control) will present a larger error.

Lastly, the configuration of the PV power plant as a whole, or at least of the electrical equipment directly connected to the PV inverter, must be provided. The reason is that there are multiple configurations of PV systems that can take place (central inverter, microinverters, etc.), and such configuration will impact the current, voltage and power levels in the multiple nodes and connections of the PV power plant. Besides that, the configuration has a crucial role when implementing the fault and failure detection and diagnosis algorithms, as it is necessary to understand the behaviour of the real asset to develop proper machine learning solutions.

2.2.2 WEATHER AND SCADA DATA

Besides the technical specifications of the PV modules, PV inverter and PV power plant configuration, it is essential to have access to SCADA and weather data of the PV power plant. This is necessary to do a proper simulation using real data as input, such as irradiance, temperatures, voltages, etc., and to validate the model by comparing the output generated by the simulation with the real data. If a specific KPI error is achieved, the model is considered reasonable.

The weather data provides information about irradiance, temperature, humidity, etc. Still, the most important ones are the irradiance and ambient temperature, as they are the input for any PV cell/module model. If there are multiple irradiance and temperature sensors spread around the PV power plant, they can be used as input to the PV modules closest to them. Otherwise, a "general" irradiance and temperature profile can be applied to the model without compromising the results.

The SCADA data provides information about electrical measurements, such as currents, voltages and powers. Still, it can also provide information about the power factor, frequency, efficiency, and PV inverter operating state (power level reference, reactive control input, etc.). All of this data can be used to validate the model. Thus, it is interesting to have which of these data are being adequately considered in the simulation to compare later with the real data.

Nevertheless, it can be noticed that besides having the technical information about the PV power plant assets, the data have a fundamental role during the simulation and validation, as it is being used as input and output features of the digital twin. Access to this type of data is necessary to validate the digital twin. Thus, even the fault-free operating condition would not be reliable.

2.2.3 DATASET CHARACTERIZATION

The use of different types of variables (VarType) for each measurement group (MeasID) is the cause for their characterization through the creation of two dictionaries (Table 2-1 e Table 2-2), as it will contribute to a better understatement and study of the dataset available.

TABLE 2-1: MEASURES CHARACTERIZATION

Parameter	Description
MeasID	Measurement Tag
MeasDescp	Measurement Description
VarType	Dictionary of VarTypes
V_{min}	Minimum value, in per unit
V_{max}	Maximum value, in per unit
$IQR_{, threshold}$	Interquartile Range, in per unit
V_{norm}	Base value in SI

TABLE 2-2: VARIABLES CHARACTERIZATION

Parameter	Description
VarName	Name of the variable
Datatype	Type of dataset ("meteo": weather data; "inv": electrical data)
VarType	Variable Type [1]
TransfID	Transformer Number
InvID	Inverter Number

JbID	Junction Box Number
StrgID	String Number
SensID	Sensor Number
Units	Variable Units in SI
VarDescp	Variable Description
MeasID	List of Measurement Tags
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: • TRANSF_A: Transformer A • INV_A.B: Inverter B of TRANSF A • JB_A.B.C: Junction Box C of INV A.B • STRG_A.B.C.D: String D of JB A.B.C • [VarType]_A.B.C.D.E: Sensor E of type [VarType]

2.3 PER UNIT SYSTEM

As it was reported on [1], the per-unit system (or pu 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 (V_{norm}) is the value related to a base quantity (V_b) by the expression $V_{norm} = V/V_b$ [6]. In the current section, a brief features normalization update is shown in Table 2-3.

TABLE 2-3: PARAMETERS OF PER-UNIT NORMALISATION

VarType	Vnorm	Vmin (pu)	Vmax (pu)
AVL, PMAXmod, QCTRmod, QCTRref, QQUADsp, PMAXsp, WindDir	0	0	0
EF, FPsp	100%	0.0	1.0
ENGDay	4000 kWh	0.0	1.5
ENGTot	1116 kWh	0.0	100.0
Fac	50 Hz	0.99	1.01
Fpac	1	-1.0	1.0
Iac	1310 A	0.0	1.1
IdcI, IdcJB, IdcS	1300 A	0.0	1.1

Irrad	7 kWh/m^2	0.0	1.5
Pac, PMAXsp	630 Kw	0.0	1.1
PdcI, PdcJB, PdcS	725 kW	0.0	1.1
Qac	630 kvar	0.0	1.1
RadDir, RadDirAv	1000 W/m^2	0.0	1.1
RadMod, RadH, RadPl, RadPlAv	1000 W/m^2	0.0	1.1
Sac	630 kVA	0.0	1.1
TempMod, TempModAv	25 ºC	-0.1	3.0
TempInt	25ºC	-0.1	2.0
Vdc	1000 V	0.48	1.0
Vac	315 V	0.9	1.1
WindS	10 m/s	0.0	3.0

2.4 DIGITAL TWIN

The PV inverter digital twin is revisited, since the model presented in [5] has been updated during the validation phase of the project. The main modifications are: [5]

- the addition of more junction boxes, now precisely matching the number present in the real asset, thus allowing for a proper analysis of the faults on the DC side of the PV inverter;
- the addition of the transformer to the simulation, which now replaces an ideal grid. This results in more AC-link voltage fluctuations, which is an expected behaviour once it is within the proper limitation values;
- the addition of three more faulty conditions, following the previous implementation based on some of the most common faults and failures of the PV power plant;
- and the randomly generated conditions, whereas now it is possible to apply predetermined faults or fully randomly generated scenarios.

2.4.1 SIMULINK MODEL

The first versions of the PV power plant centred around the PV inverter used a simplified version of the original configuration. Even though the power and voltages rating were the same, the DC current level distribution across the junction boxes was not. In the latest version, it is possible to measure each junction box's current, allowing fault classification and localization of those assets as well.

Besides that, adding the transformer to the PV inverter simulations adds some key interactions that can improve the analysis of the AC side of the inverter. The outcome of the digital twin of the transformer [5] is imported to the PV inverter simulation, approaching the virtual representation of the real PV power plant. This integration, for instance, allows for evaluation in the future of the interactions between the two inverters through the transformer under a severe fault on their AC side.

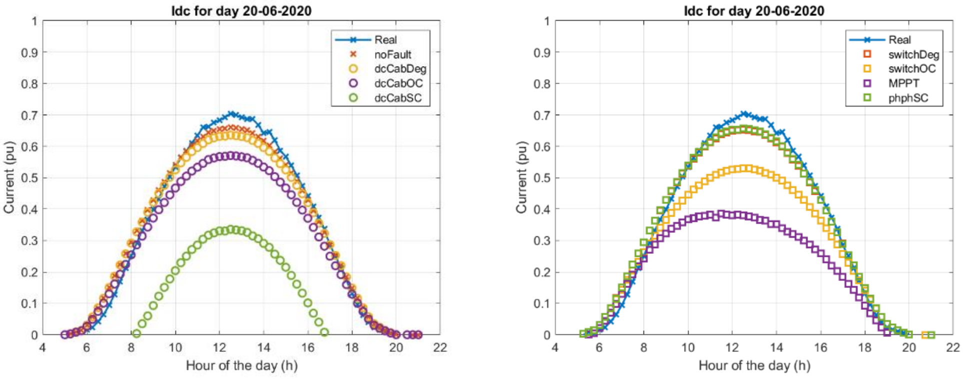
2.4.2 FAULTS AND FAILURES IMPLEMENTATION

Whilst the first versions of the PV inverter digital twin had one fault-free condition and four more faulty conditions, the latest version added three more faults. The complete list of the implemented faults can be found in Table 2-4.

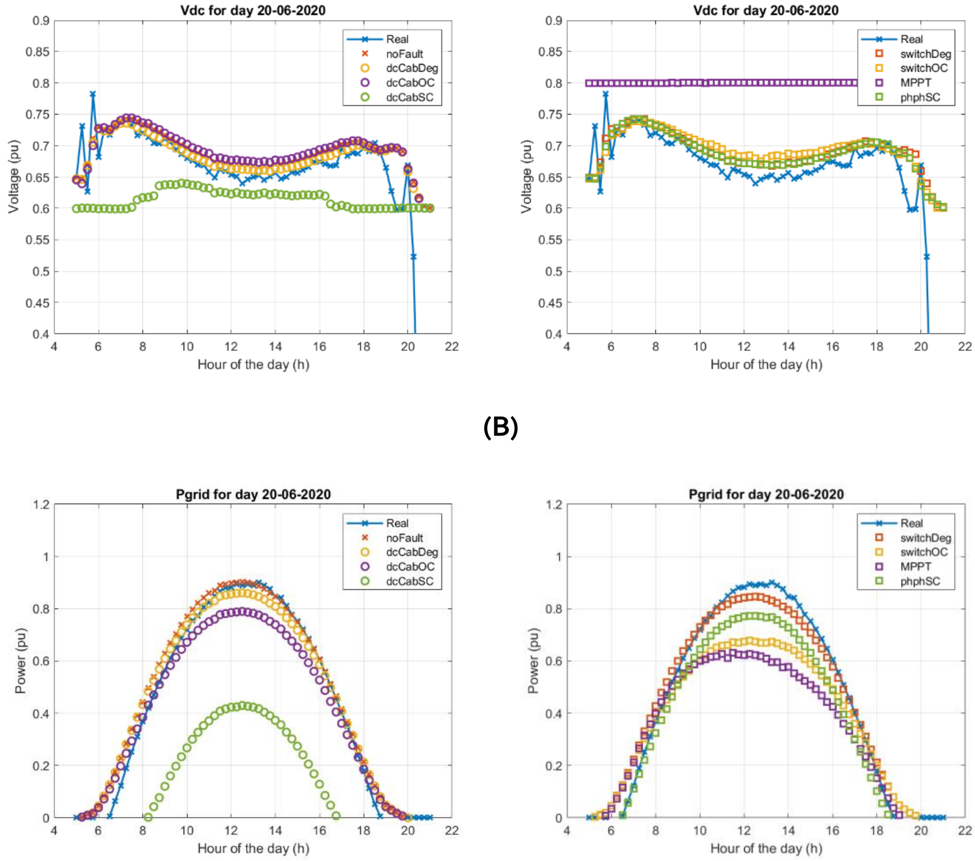
TABLE 2-4: PV INVERTER IMPLEMENTED FAULTS AND FAILURES

Fault number	Fault acronym	Description
00	noFault	Regular operation, used for model validation
01	dcCabDeg	A series resistance in the cables of a given junction box
02	dcCabOC	An sudden open circuit in the cables of a given junction box
03	switchDeg	An increased on resistance in a given switch of the inverter
04	switchOC	An open circuit in a given switch of the inverter
05	dcCabSC	A short circuit between positive and negative poles of a given junction box
06	phphSC	A short circuit between phases on the point of common coupling of the PV inverter
07	MPPT	A saturation of the output of the MPPT algorithm

All of those faults and failures will have a direct impact on the operation of the inverter; thus, the electrical measurements will present a pattern that allows the ML algorithms to classify those problems, followed by the localization algorithms to the point where the fault happened. For exemplification, all of those conditions are simulated and displayed in Figure 2-2 for DC-link current, DC-link voltage, and AC power.



(A)



(B)

(C)

FIGURE 2-2: SIMULATION RESULTS FOR THE PV INVERTER: (A) DC-LINK CURRENT; (B) DC-LINK VOLTAGE; AND (C) AC ACTIVE POWER

As shown in Figure 2-2, some conditions are easily distinguishable from the others in the DC-link current, the DC-link voltage, etc. However, others are not easily detectable, even using those three measurements. Thus, multiple measurements of the PV inverter are considered for fault classification at the end of the day. Measurements such as frequency, power factor, AC currents and voltages, etc., are all turned into features to be fed to the ML models so they can do the pattern recognition, thus classifying the faults.

Besides that, trying to achieve a behaviour more like a real PV inverter, some functions for randomly generated scenarios were added. It is worth remembering that a digital twin is a powerful tool for benchmarking and exploring different scenarios of the asset. In that sense, other than the deterministic fault and failure events, it is possible to generate the scenarios randomly. The randomness can be related to the fault of the day, fault starting time, etc., and they are listed in Table 2-5.

COMPETE 2020 PORTUGAL 2020 CDTI @CDTIoficial European Union European Regional Development Fund

TABLE 2-5: PV INVERTER RANDOM SCENARIOS GENERATION FUNCTIONS

Random scenario function	Description
Inverter under fault	For a set of inverter parallel-connected to a power transformer, it is possible that the fault will occur in one of them
Fault of the day	From the faults pool, it is possible that one of them will happen (or no fault will happen at all)
Starting time of the fault	A fault can start anywhere between 05 AM and 09 pM
Fault location	The location of the fault, i.e., junction box, AC phases, switch of the inverter, etc., is randomly selected
Degradation evolution	A degradation usually spams accros multiple days, weeks or months, evolving from a harmless output power reduction until a failure like a subsystem disconnection

In that sense, the digital twin can achieve randomly generated scenarios. Of course, if it is needed, it is still possible to specify which scenarios are wanted to be simulated. However, from a training dataset point of view, having those randomly generated scenarios can place the machine learning models under unexpected conditions, thus stressing and validating its capability of predicting/classifying faults and failures.

For instance, in Figure 2-3, it can be seen that around 07 AM, an open circuit fault happens in the DC cable of a given junction box. As for this example, the fault only happens on one of the inverters; the AC power of the second inverter (plot on the right) is unaffected, whilst the AC power of the first inverter (plot on the left) is significantly reduced. Also, it is worth noting that the gap between the fault-free and the faulty condition is more discernible during periods of higher irradiance levels; in Figure 2-3, this happens around noon. Similarly, an MPPT fault is depicted in Figure 2-4, whereas the output power reduction can be clearly seen on the left plot.

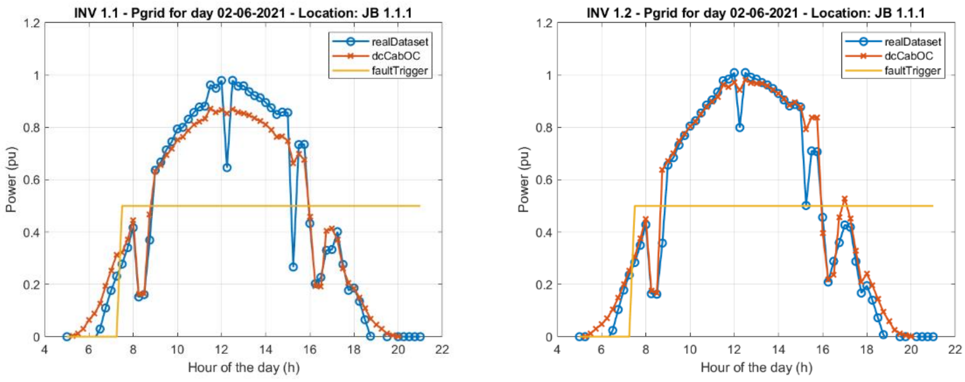


FIGURE 2-3: AN OPEN CIRCUIT FAULTS HAPPENING ON THE FIRST INVERTER AROUND 07 AM OF A GIVEN DAY

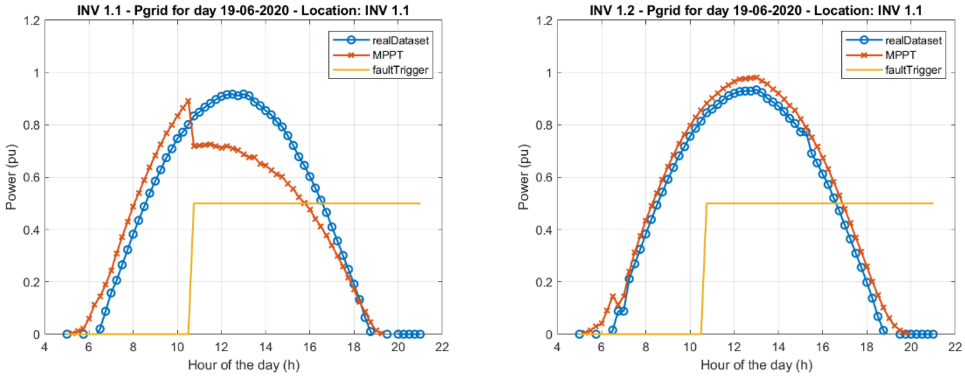


FIGURE 2-4: A MPPT CIRCUIT FAULTS HAPPENING ON THE FIRST INVERTER AROUND 10 AM OF A GIVEN DAY

Those results show that the digital twin platform is very flexible, allowing the exploration of multiple future or faulty scenarios. Such a platform can be expanded to accommodate weather-related faults, threshold-triggered faults, etc. Thus, if a correlation between a specific event and a fault can find out, it can be replicated in the digital twin (such as the fuse tripping during spring days [7]).

2.5 HYBRID DATASET

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, a digital twin (DT) was implemented in Simulink/MATLAB® to generate a faulty dataset (due to the lack of real faulty data), including random daily parameters (fault's start time, fault type, and fault's location).

2.5.1 TIMESERIES DATASET

To reduce the processing time and remove any redundancy that could compromise the model's performance, reducing the total number of features from the hybrid dataset was necessary. More details about those first implementations can be found in [1].

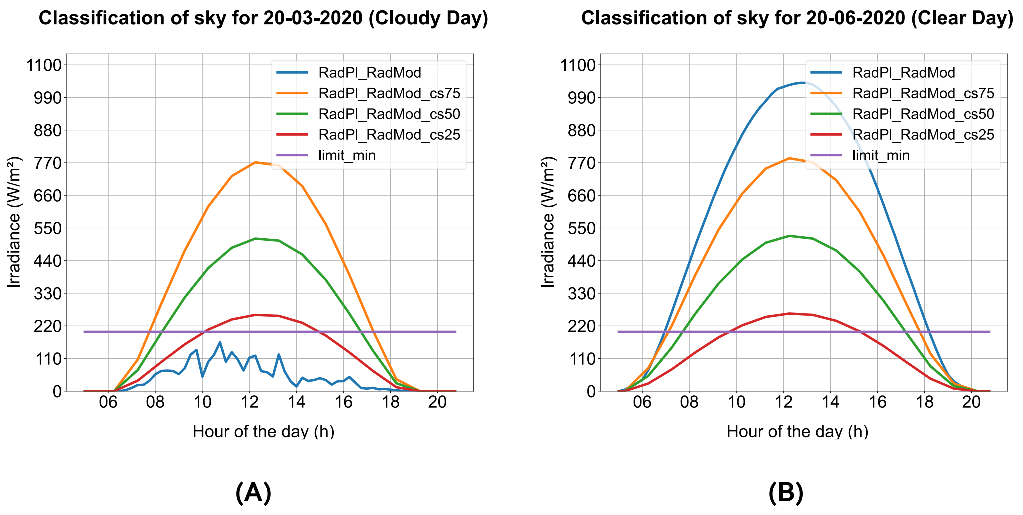
2.5.1.1 FEATURE ENGINEERING

As it was implemented in [1], to improve the accuracy of the ML algorithms, some additional features were included to add some contextualization to the data. Most of the new features are weather-related, as the weather will dictate the operating conditions of the PV inverter.

2.5.1.2 SKY'S TYPE

The clear's sky classification for each timestamp was based on the hourly estimation implemented in [1], with time-wise interpolation [8] [8] and by filling missing values with '0' [9]. Two variable types were considered for the classification of two inverters: PV Modules Irradiance (RadMod) and Plane Irradiance (RadPl). The pyranometer PYR1.1.1_R was excluded due to its measurement errors. In

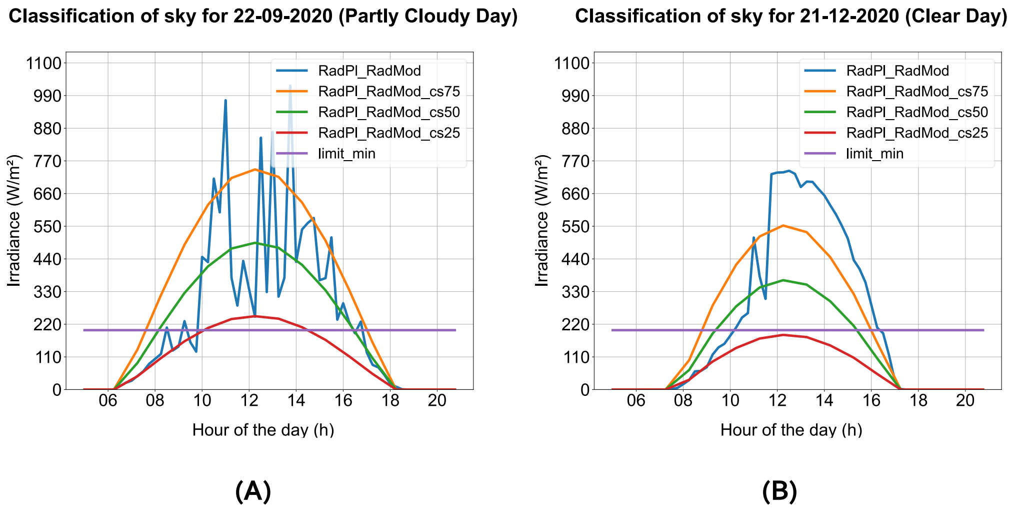
Figure 2-5 and Figure 2-6, are represented some examples of the sky's type classifications for the year 2020.



(A)

(B)

FIGURE 2-5: CLASSIFICATION OF SKY FOR: (A) 2020-03-20; AND (B) 2020-06-20



(A)

(B)

FIGURE 2-6: CLASSIFICATION OF SKY FOR: (A) 2020-09-22; AND (B) 2020-12-21

2.5.1.3 WEATHER AND SCADA VARIABLES

For the current deliverable, the digital twin comprises two inverters connected by the same transformer (INV_1.1 and INV_1.2). Which inverter is composed of eight junction boxes (JB_1.X.1 - JB_1.X.8). Additionally, the reduction of the total variable number from 1224 to 85 was made by the following steps:

- Considering only the Ambient Temperature, the Plane Irradiance (Sensor 1 and Sensor 2), and PV Modules Irradiances.
- Including electrical variables from the inverter side. Exceptions: All set-point and control features; Availability; Internal Temperature; Daily Energy Produced; Total Energy Produced; Apparent Power and Power Factor.

- Add DC Voltage and DC Current variables from the junction boxes.

On Table 2-6 is described the groups of features used.

TABLE 2-6: GROUPS OF FEATURES

Group of Features	Description
'weather'	All features related to the weather station (WS_[X]), and to the pyranometers (containing 'PYR') Note1: Average Module Temperature included. Note2: Exclusion of the variables associated with the Plane Irradiance used on the clear sky's estimation.
'junctionBox'	All features related to the junction boxes connected to the classified inverter (JB_[X])
'inverter'	All features related to the classified inverter (INV_[X])
'skytype'	Sky's type features
'calendar'	All features associated to date, daytime, sunrise, and sunset. Note: Exclusion of 'date', 'year', 'day', 'minute'
'statistical'	Mean and standard deviation of measurements related to pyranometers and junction boxes.

2.6 HYPERPARAMETERS TUNING

The Bayesian Optimization is a sequential design strategy for the global optimization of black-box functions [10] [11], based on the Bayes Theorem [12], and shown in Equation 2-1.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ **EQUATION 2-1**

Where:

- P(A|B): The probability of event A occurring given that B is true. It is also called the posterior probability.
- P(B|A): The probability of event B occurring given that A is true. It is also called the likelihood probability.
- P(A) and P(B): The probabilities of observing A and B, also known as the prior probability and marginal probability.

In contrast to the most common methods for hyper-parameter tuning, Grid Search and Random Search, it uses the results from the previous iteration to decide the next hyper-parameter value candidates. Additionally, Bayesian Optimization was implemented with the function BayesSearchCV of the Scikit-Learn/Python library [10], which includes the cross-validation of the training dataset. Initially, a wider parameter range was used for optimization on 150 iterations. With the borders of search reduced, 50 iterations were made to determine the estimator with the most accurate predictions.

2.7 ALGORITHMS FOR DIAGNOSIS

Fault Diagnosis is defined by the determination of kind -or, fault classification-, size, location -also known as fault localization- and time of detection of a fault by evaluating symptoms [13]. A prior out of normality analysis will be implemented to identify any possible incorrect measurements, which could compromise the fault diagnosis performance, or some variables redundancy.

2.7.1 OUT OF NORMALITY ANALYSIS

The interquartile range (IQR) has been used to obtain the mean value of each group of measurements (MeasID), and for out of normality detection of PV Modules Temperature, Global Irradiance, and Direct Irradiance. As it was noted in [1], the average measurements of temperature and irradiance weren't considered in the mean value estimation because of measurement errors. Moreover, in the case of PV module temperature, it was necessary to disregard the ambient temperature due to its reduced daily deviation.

Initially, the out of normality detection was implemented as an IQR function for each timestamp. If the IQR's value is above or equal to a threshold, the outlier detection is defined as 1– otherwise, 0. Subsequently, three criteria were evaluated to identify the possible measurements for each MeasID:

- Criterion 1: $\mathrm{Min}(|Y_i\text{-}\mathrm{MeasID}_{\mathrm{Min}}|, |Y_i\text{-}\mathrm{MeasID}_{\mathrm{Max}}|)$.
- Criterion 2: $\mathrm{Max}\left(\left|\frac{Y_i\text{-}Y_{i\text{-}1}}{\Delta t}\right|\right)$, Δt=15min.
- Criterion 3: $\mathrm{Max}(|Y_i\text{-}\mathrm{MeasID}_{\mathrm{Mean}}|)$.

The out of normality identification is determined by the maximum number of criteria equal to 'True'. Any variable identified will be excluded from the calculation of the average value. Furthermore, the final mean value will be constrained between the minimum value V_{min} and the maximum value V_{max} [14]. Table 2-7 describes the parameters for each MeasID.

TABLE 2-7: OUT OF NORMALITY PARAMETERS

MeasID	Definition	VarType	$[V_{norm}, V_{min}, V_{max}]$	Threshold	Group By
pyrGlobalPV	Global Irradiance	RadPl, RadMod	[1000; 0; 1100] W/m²	0.06 p.u.	Inverter
PyrDirPV	Direct Irradiance	RadDir	[1000; 0; 1100] W/m²	0.16 p.u.	None
tempModsPV	PV Modules Temperature	TempMod	[25.0; -0.1; 3.0] ºC	0.20 p.u.	Transformer

In Figure 2-7-Figure 2-9 are presented some examples of the mean value estimation.

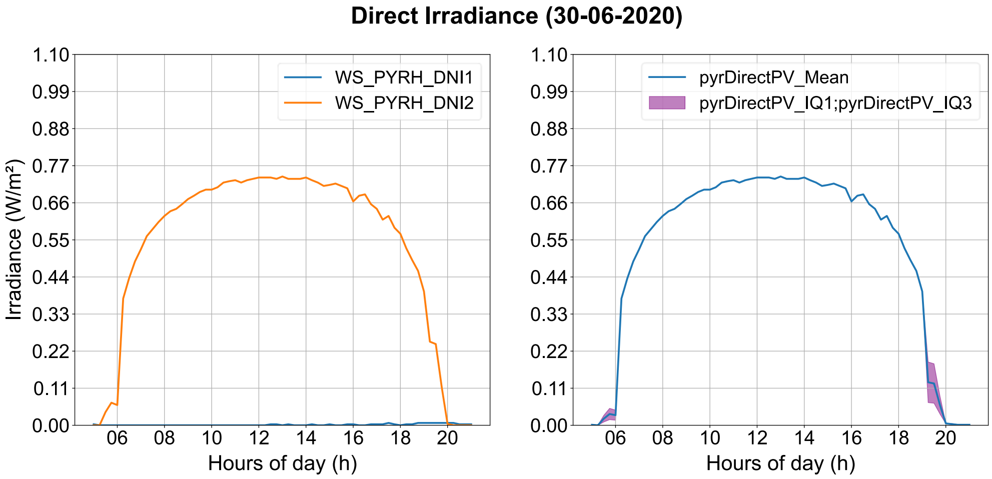


FIGURE 2-7: DIRECT IRRADIANCE FOR 30-06-20202 (V_{NORM} = 1000 W/M²)

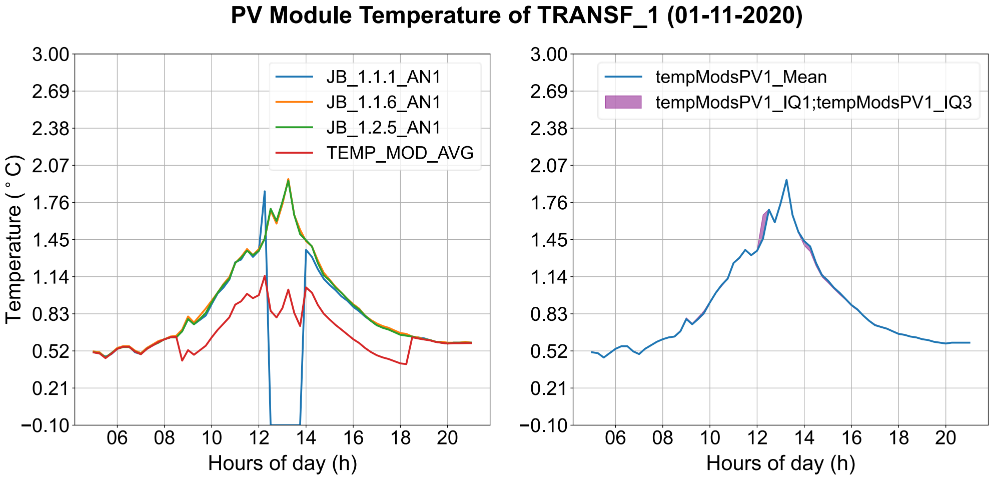


FIGURE 2-8: PV MODULES TEMPERATURE OF TRANSF_1 FOR 01-11-2020 (V_{NORM} = 25ºC)

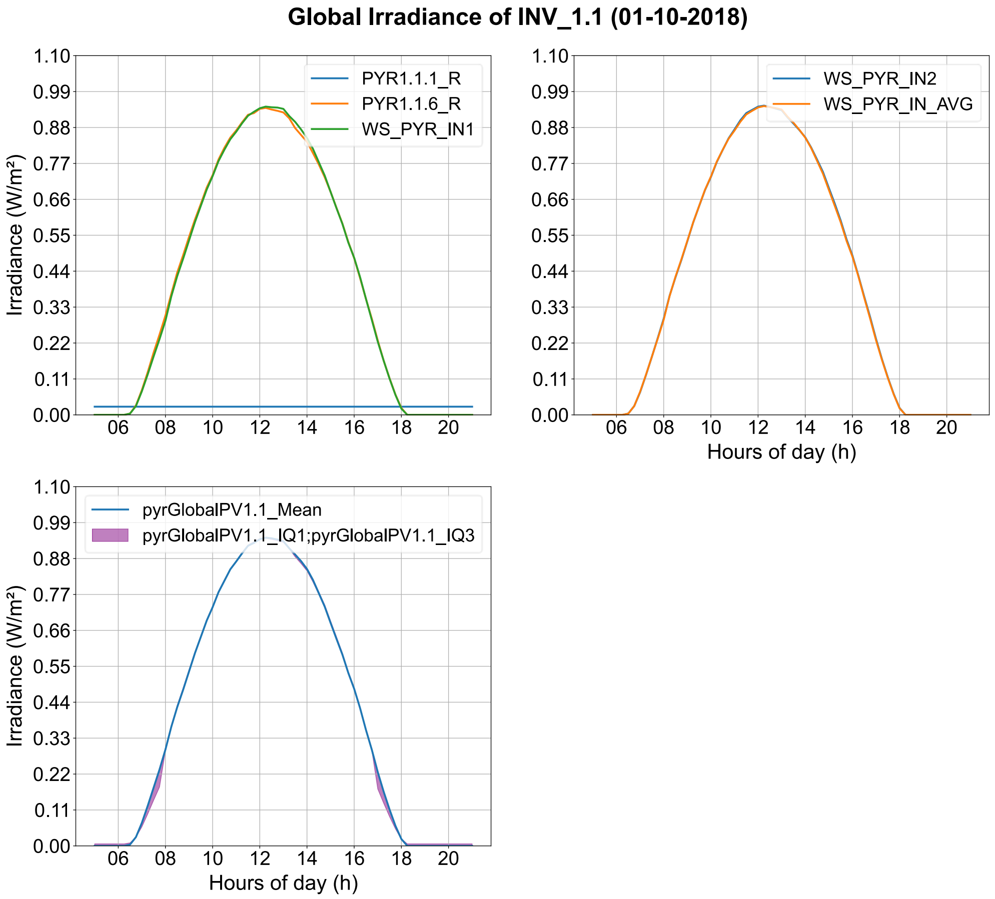


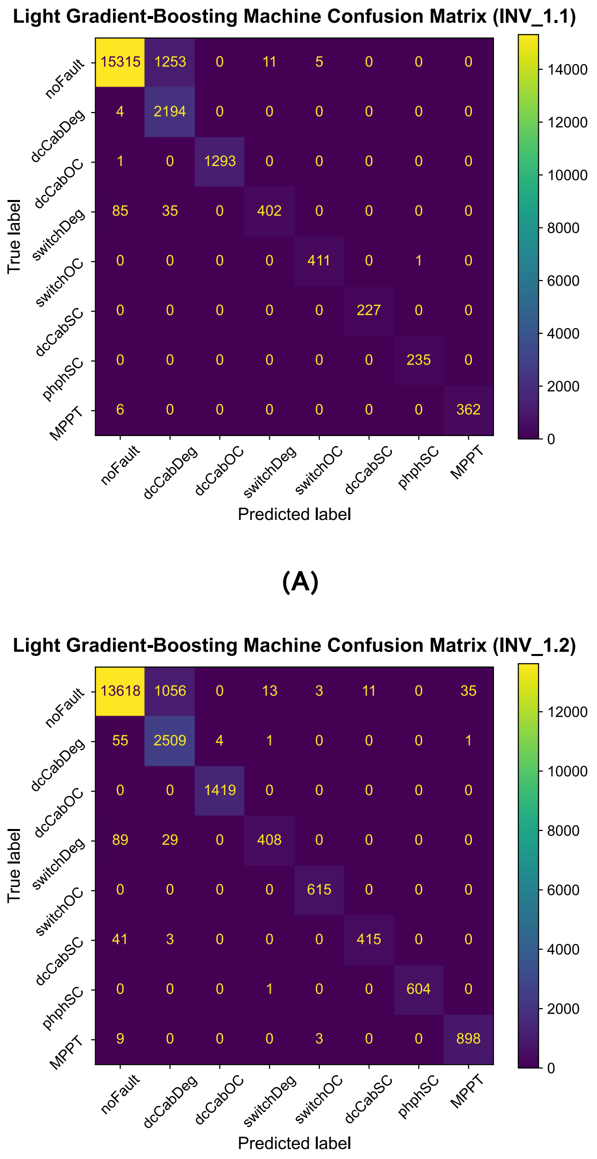
FIGURE 2-9: GLOBAL IRRADIANCE OF INV_1.1 FOR 01-10-2018 (V_{NORM} = 1000 W/M²)

2.7.2 FAULT CLASSIFICATION USING ML MODELS

The Light Gradient-Boosting Machine (LightGBM) algorithm already approached in [1], was tested with the dataset between 2022-04 and 2023-05 -with a total number of samples equal to 21840- resulting in Fault Detection Accuracy (FDA) of 93.6% for INV_1.1 and 93.8% for INV_1.2. The previous values meet its minimum requirement of 80%. During the feature selection stage, the best sets of features were defined for both inverters, namely:

- ‘weather’, ‘inverter’, ‘calendar’ and ' statistical' for INV_1.1.
- ‘inverter’, ‘calendar’ and ' statistical' for INV_1.2.

As shown in Figure 2-10, the current model presents difficulty distinguishing the fault-free condition from DC cable degradation and the switch degradation from fault-free. The previous results were observed, mainly in the day beginning (5h00 -7h30) and end (17h00-21h00).



(A)

(B)

FIGURE 2-10: CONFUSION MATRIX FOR FAULT CLASSIFICATION USING THE LIGHTGBM CLASSIFIER WITH DIFFERENT INVERTERS: (A) INV_1.1; AND INV_1.2

As last, initially, the explainability of the model was tested through the Shapley values [15]. However, it was time-consuming (about 1 hour per iteration), leading to the removal of the Explainable Artificial Intelligence (XAI) Techniques as part of the current pipeline.

2.7.3 FAULT LOCALIZATION BASED ON BENCHMARKING

The fault localization defines the physical fault's location in the function of the prior classification. Additionally, in [16], fault detection and classification are implemented by evaluating an error residual vector, which is defined as the difference between the digital twin reference estimation – with solar irradiance (G) and Module Temperature (T) as inputs - and the measured outputs. Consequently, the methodology implemented in the fault localization is determined by four main steps (from Equation 2-2 to Equation 2-5), namely:

1. Fault Detection

$$Fault\ Detection(t) = \begin{cases} 0, & Fault\ Classification\ = 'noFault' \\ 1, & others \end{cases}$$

EQUATION 2-2

2. Error Residual

$$\Delta Y_i(t_j) = Y_{i,Fault}(t_j) - Y_{i,noFault}(t_j)$$

EQUATION 2-3

Where:

- $\Delta Y_i(t_j)$: Error residual of variable i, on timestamp j.
- $Y_{i,\text{Fault}}(t_j)$: Value of variable i from the Timeseries dataset, on timestamp j.
- $Y_{i,\text{noFault}}(t_j)$: Value of variable i from Rewrite dataset, on timestamp j.

3. Cumulative Average

$$CA(t_{i,j}) = \frac{1}{N+1}\sum_{j=0}^{N} \Delta Y_i(t_j),\ \Delta Y_i(t_0) = 0$$

EQUATION 2-4

Where:

- i: Variable.
- j: Timestamp.
- N: Number of samples between the fault's start and the timestamp j, if Fault Detection $(t_j) = 1$ and $t_j \leq$ fault's end.

4. Fault Localization

$$Fault\ Localization\ (t_j)\ =\ MachineID\ where\ MAX\left[\left|CA(t_{i,j}) - \overline{CA}(t_j)\right|\right]$$

EQUATION 2-5

Where:

- CA $(t_{i,j})$: Cumulative average for each variable, on timestamp j.
- $\overline{\text{CA}}$ (t_j): Mean value of all cumulative averages for every variable considered, on timestamp j.
- MachineID: Inverters and Junction boxes. Default: 'noFault'

In Table 2-8 are described the parameters for each fault localization.

TABLE 2-8: FAULT LOCALIZATION PARAMETERS

Fault Classification	VarType	Possible Localization (s)
noFault	-	'noFault'
dcCabDeg	$I_{DC, JB(Inverter)}$	Junction Boxes
dcCabOC		
dcCabSC		
switchDeg **switchOC**	$I_{DC,I}$	Inverters
phphSC	P_{AC}	Inverters
MPPT	$P_{DC,I}$	Inverters

Although the Timeseries Dataset included eighteen possible localizations (two inverters and eight junction boxes per inverter), each inverter was classified independently. Thus, it was only possible to compare different locations for the 'dcCabDeg,' 'dcCabOC,' and 'dcCabSC.' From Figure 2-11 to Figure 2-13 are presented some examples of fault localization for the previous classifications.

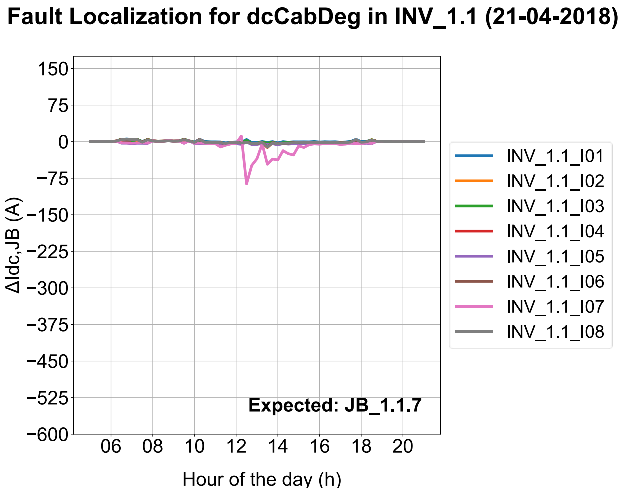


FIGURE 2-11: FAULT LOCALIZATION FOR DCCABDEG IN INV_1.1 (21-04-2018)

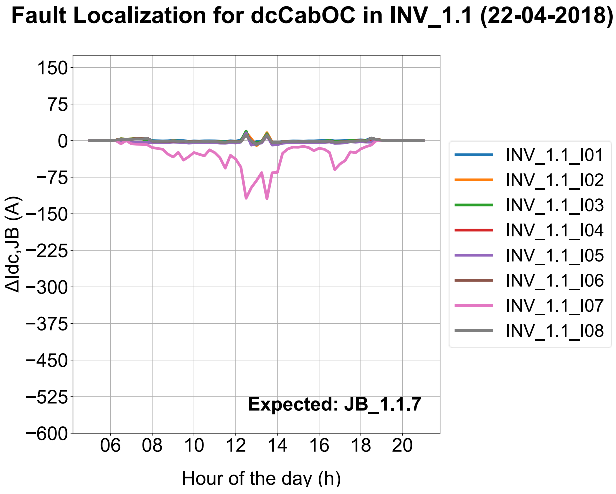


FIGURE 2-12: FAULT LOCALIZATION FOR DCCABOC IN INV_1.1 (22-04-2018)

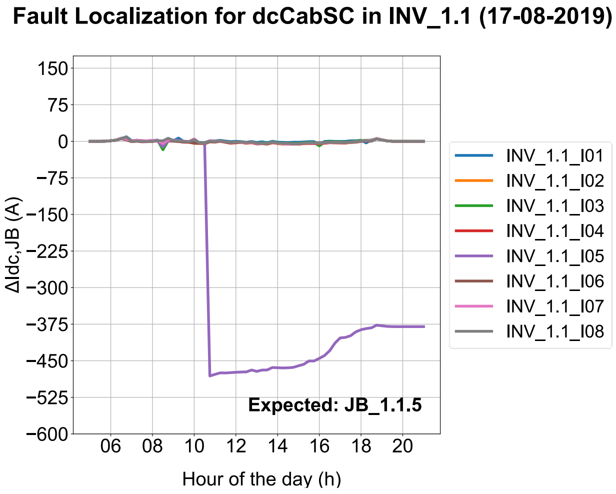


FIGURE 2-13: FAULT LOCALIZATION FOR DCCABSC IN INV_1.1 (17-08-2019)

Finally, the dataset between 2022-04 and 2023-05 presented a Fault Localization Accuracy (FLA) of 79.9% for INV_1.1 and 78.2% for INV_1.2.

2.8 DIAGNOSIS OR PIPELINE INTEGRATION

Lastly, after all of the described tools related to the PV inverter digital twin are operational, it is still necessary to summarize and format them to properly communicate with the PVP O&M staff. Thus, the last building block of the pipeline presented in Figure 2-1 is developed. To do so, it is necessary to close communication and cooperation with the PVPP owners and operators, i.e., a digital twin solution must have to account not only for the power electronics and electrical engineering aspects of the real asset but also for the "logistics" of the data flow from and to the real asset.

Depending on the needs of the other subsystems of the PVPP or the O&M staff itself, three stages of diagnosis can be achieved:

1. Fault detection: the algorithms inform whether there is a fault, thus a binary classification. Even though this can be useful for maintenance planning, it has yet to provide further details about the state of the PVPP.
2. Fault classification: the algorithms can, beyond detection, classify the fault based on a predetermined list since the ML solutions developed for this project are all based on supervised learning.
3. Fault localization: In one of the diagnosis's final stages, the algorithms can pinpoint where the fault happened. Of course, the level of detail of the localization relies on the level of detail achievable by the available dataset. Nevertheless, pointing out the equipment under fault (a specific junction box, an inverter, etc.) is usually reasonable enough.

It is also possible to record a fault's start and end, as some of them can be intermittent. The indication of the starting, end and period of a fault can help do a root cause analysis of the problem, trying to find a correlation between an event and the fault (for instance, a problem that is occurring only during rain periods or during a season of the year, etc.).

At last, the report can be assembled by combining the date and time (or time stamp as an alternative), detection, classification, and localization. Depending on the number of elements or subsystems on a PVPP, it can be interesting to aggregate the results by equipment or element. An example of a digital twin report is presented in Table 2-9.

TABLE 2-9: DIAGNOSIS REPORT TEMPLATE

Report entry	Time stamp	Date and time	Equipment	Detection	Classification	Localization
Description	An unit of tame based on a starting epoch, which varies from different systems. It can be counted in seconds, miliseconds, minutes, etc.	A combination of a given data and its multiple times during the day. It can include the timezone too. Similar to the timestamp, is dependent on the granularity of the system, which can be sceonds, minutes or even hours.	A tag that can easily indentfy the equipment under analysis	The first stage of the diagnosis, rapdily indentifying if there is a fault or not	The second stage of the diagnosis, now indenteifying if the fault is within a list of trained possible conditions	The third stage of the diagnosis, based on the classification, pinpointing the possible physical location of the fault

Example	1655812800 000 ms	21-06-2022 12:00:00	INV 1.1, INV 2.2, etc.	0 or 1	noFault, switch OC, etc., accorind to Table X.Y	JB 2.1.7, INV 1.2, etc., depending on the configuration of the PVPP

3. VALIDATION RESULTS

In this section, the validation results are reported. In order to evaluate the effectiveness of the solutions developed, the KPIs defined in D4.2 are used as acceptance criteria. Wherever the results obtained by the validation of a certain solution match the KPIs, the solution is deemed as valid.

The validation phase started on April 1st 2022 and ended in June 20th 2023. During this timeframe, the operational data of the validation site were collected and analysed by the AI4PV tools. Faults, failure and underperformance were identified by the AI4PV solutions and recommendation actions were generated.

3.1 KPI1: RMSE EMPIRICAL AND REPRODUCED I-V CURVE

This KPI represents the difference between the empirical I-V curve provided in the datasheet of the PV module and the reproduced curve through the DT modelling. Both curves, the one from the datasheet and the one produced by the DT, are reported in Figure 3-1 for different levels of irradiance.

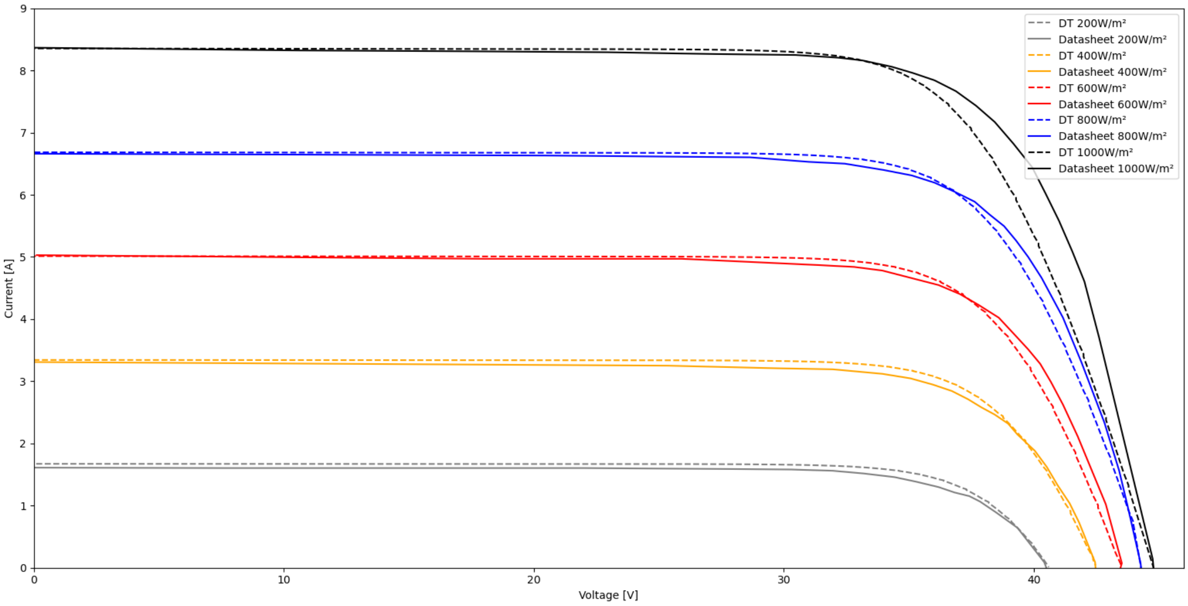


FIGURE 3-1: REPRODUCED AND EMPIRICAL I-V CURVES

To evaluate the effectiveness of the proposed DT, the RMSE between the empirical and the reproduced I-V curve is evaluated. The RMSE is calculated as per Equation 3-1.

$$RMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(I - I_i)^2}}{Isc}$$

EQUATION 3-1

Where:

- $I_i, \hat{I}_i$ are the real and modelled output current of the PV module.
- N is the number of samples of the empirical I-V curve
- Isc it's the short circuit current of the PV module

The RMSE value stands at 0.18, much lower than its target value (0.6).

3.2 KPI2: REDUCE SOILING LOSSES (RSL)

This KPI represents the ratio between the energy of the soiled PV panel and the cleaned one. The higher it is, the more cleaned the PV is for a long period of time. It considers losses due to both dust or organic soiling. This KPI is evaluated considering the actual energy produced by the PV farm, during the validation timeframe (from April 2022 to June 2023). This KPI is computed using Equation 3-2.

$$RSL = \frac{\int_0^T P_{PV\_soiled} dt}{\int_0^T P_{PV\_cleaned}\, dt}$$

EQUATION 3-2

Where:

- P_{PV_soiled}, represents the output power of the soiled module, thus it is the real output power of the PV park;
- $P_{PV\_cleaned}$, represents the ideal power output wherever the PV panels were always perfectly clean, thus when the performance ratio it is at its maximum value;
- T is the observation time, thus the whole validation frame.

For confidentiality reasons, the amount of energy produced within the validation timeframe can not be disclosed. Nevertheless, it can be said that the average Performance Ratio (PR), as a result of the AI4PV policies stands at 82%. Having said that, comparing this value with the theoretical one (which is ideally, as it considers the PV panels being always cleaned and thus it is not economically viable as it would require huge CAPEX), the RSL can be calculated as per Equation 3-3.

$$RSL = \frac{PR_{average}}{PR_{ideal}} = 0.83$$

EQUATION 3-3

It can be noticed that this value is higher than the KPI defined in D4.2 (0.8) [17].

3.3 KPI3: NUMBER FAULTS AND/OR FAILURES DETECTED AUTOMATICALLY THROUGH DATA ANALYSIS

The inspection of the SCADA and sensor data of the inverter by AI, ML, algorithms will detect trending and deviations in the measurements that may indicate a fault or a failure in the PV plant. For confidentiality reasons, no absolute figures can be disclosed in terms of fault detected and fault registered by the O&M teams. Nevertheless, it can be said that 85% of the conditions detected by the AI4PV solutions, were considered as True Positive. This value is higher than the target (80%).

3.4 KPI4: FAULT DETECTION ACCURACY

This KPI describe the accuracy of the fault detection and classification algorithms developed within the AI4PV project. It is calculated via Equation 3-4.

$$FDA = \frac{N\_true\_positive\_state}{N\_true\_positive\_state + N\_false\_positive\_state}\%$$

EQUATION 3-4

A preliminary validation was already performed during the development phase, upon the training dataset (as explained in [1]) which showed promising results with the accuracy of the developed tool standing at 95-96%.

However, an additional validation was performed during the validation phase, on new data. The number of failures registered by the O&M team and recorded in O&M reports is used as reference to which the classification algorithms are compared. It is worth to mention, that the term "state" into the formula describe all the possible states that represent a particular component, which can be divided into two main categories: fault-free or faulty condition.

For what concern the fault detection algorithms for the Power transformer, 0 false positive states were registered during the validation phase, which lead to an FDA of 100%.

During the validation, for what concerns the inverter AI algorithms, 2502 false positive were registered. It is worth to mention that these values occur before sunrise and after sunset, with very low values of irradiance. Nevertheless, even considering these mispredictions, the FDA stands at 93%, much higher than the target value (80%).

3.5 KPI5: NUMBER OF MAINTENANCE ACTIONS AT VALIDATION SITE

Depending on the output of the recommendation system, predictive maintenance may be carried out to avoid failures. It is the number of interventions advised to the O&M team by AI4PV recommender system. However, these recommendations can also include the "do-nothing" option, whenever all the components are in normal conditions and the cleaning is not advisable. Having said that, the AI4PV task recommendation engine was able to define the best policy on a daily basis suggesting 445 different tasks, one for each day of the validation. The accuracy of these recommendation is evaluated in the following section.

3.6 KPI6: RECOMMENDATION ACCURACY (RA)

This KPI describes the number of correct recommendations. It can be evaluated through Equation 3-5.

$$RA = \frac{N\_good\_recommendation}{N\_tot\_recommendation}\%$$

EQUATION 3-5

The number of good recommendations is obtained by analysis in detail the different policy suggested, leveraging on the expertise and experience of an O&M technician and comparing the AI4PV results against the O&M team's plan. Having said that, it turns out that 378 out of the 445 recommended actions can be deemed as good recommendation. Thus, the RA during the validation stands at 85%.

This number is higher than the target (70%).

3.7 KPI7: PERCENTAGE OF LOSSES & DEGRADATION UNDERPERFORMANCE QUANTIFICATION (AEL_UD)

During the validation, no major degradation phenomena and underperformance were registered, thus this KPI is not applicable.

3.8 KPI8: AVOIDED ENERGY LOSSES DUE TO EARLY DETECTION PROBLEMS (AEL_ED)

During the validation, no major faults and failures were registered, thus this KPI is not applicable.

3.9 KPI9: REDUCE UNEXPECTED OUTAGES (RUO) IN THE TRANSFORMER STATIONS

During the validation, no major transformer outages were registered, thus this KPI is not applicable.

3.10 KPI10: REDUCE RESPONSE TIME

It is the time between failure occurrence and detection. The objective of this KPI is to measure the promptness of the AI4PV faults detection algorithms and benchmark it against the conventional methods in terms of timings. It can be calculated via Equation 3-6.

$$RRT = \frac{RT_{AI4PV}}{RT_{conventional}} \%$$

EQUATION 3-6

Where:

- RT_{AI4PV} is the response time with AI4PV in place, for a particular failure;
- $RT_{conventional}$ is the conventional response time (without AI4PV) for a particular failure.

The AI4PV solutions are able to detect faults and failures at each Timestep. During the validation phase the granularity of the data was 15 minutes so is the AI4PV response time.

Traditional methods consist in creating reports on a hourly or daily basis and send alarms to O&M team. This said, assuming a response time of 1 hour for conventional method, we conclude that the RRT stands at 25%, which is in compliance with the target value (<90%) [17].

3.11 KPI11: PLANT AVAILABILITY INCREASE (PAI)

During the validation, no downtime or shutdown of the PV park was registered, thus this KPI is not applicable.

COMPETE 2020 PORTUGAL 2020 CDTI @CDTIoficial European Union European Regional Development Fund

4. CBA: SOILING USE CASE

In order to evaluate the effectiveness of the proposed solutions, particularly the cleaning optimiser, a Cost Benefit Analysis (CBA) is performed to assess potential benefits due to the AI4PV methodology.

The AI4PV cleaning module, its objectives and operations, are extensively described in [18] and [19].

In order to evaluate and quantify potential benefits, the AI4PV approach is compared against traditional methods. A common strategy is what is called "the threshold-based approach" (hereafter referred as TR-based policy), where PV panels are cleaned whenever the PR is below a certain threshold.

Two options are investigated:

- **Option A**: rain events are not taken into account in the cleaning schedule optimisation.
- **Option B**: rain events are considered and modelled in the optimisation.

4.1 CBA OPTION A

When performing this option, rain events are not modelled into the optimisation. The cleaning schedule is affected by two main parameters:

1. Cost of cleaning;
2. PR threshold for the TR-based policy.

Having said that, a sensitivity analysis is performed to see the changes in the potential benefits due to these parameters. The results obtained for this option are shown in Figure 4-1.

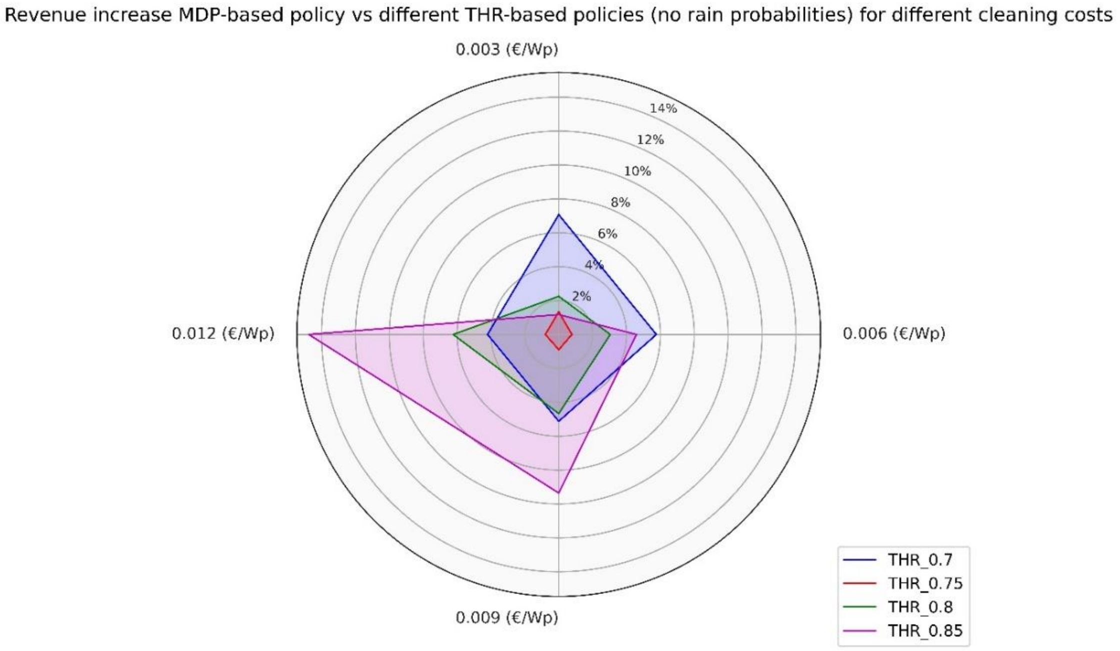


FIGURE 4-1: OPTION A - CBA AI4PV APPROACH VS TR BASED POLICY

As it can be seen, the AI4PV approach brings additional revenues regardless of the cost of cleaning and threshold value, when compared to the TR-based policy. However, the magnitude of such increase varies according to the level of the cost of cleaning and threshold value. The AI4PV benefits are summarised in Table 4-1.

TABLE 4-1: OPTION A CBA SUMMARY

TR value / Cost of Cleaning (€/Wp)	TR=0.7	TR=0.75	TR=0.8	TR=0.85
0.003	7%	1%	2%	1%
0.006	6%	1%	3%	4%
0.009	5%	1%	2%	9%
0.012	2%	1%	6%	14%

4.2 CBA OPTION B

Option B include rain events, and their impact on the PR, into the model. The results obtained for this option are shown in Figure 4-2.

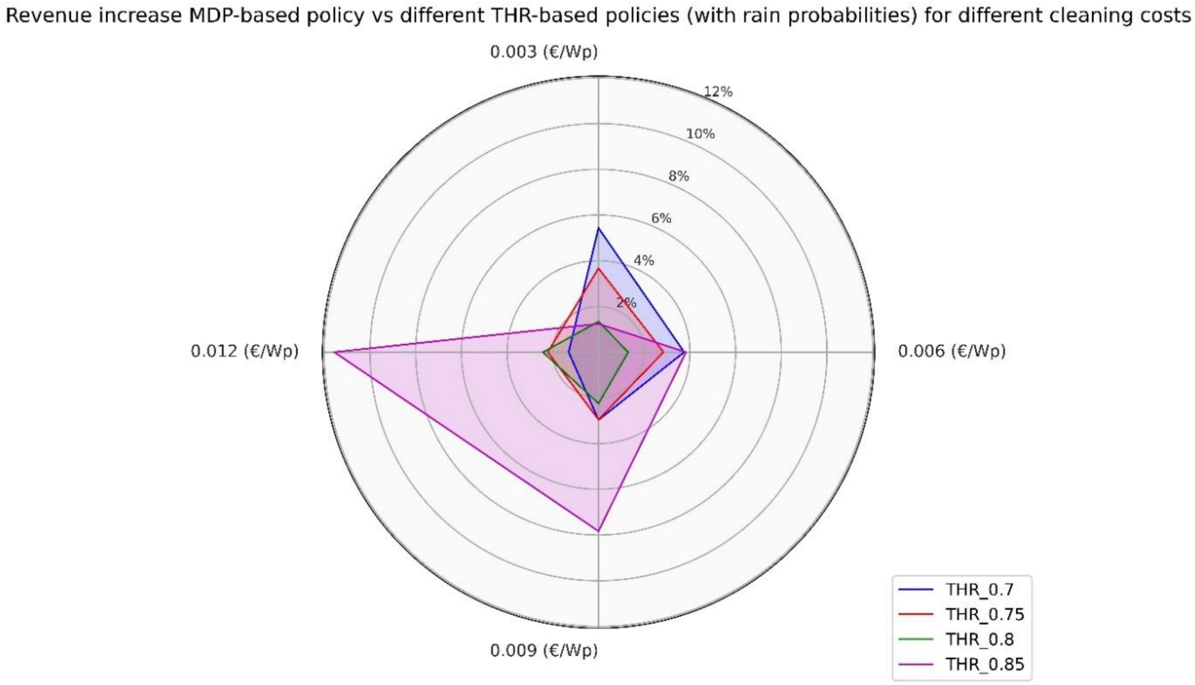


FIGURE 4-2: OPTION B - CBA AI4PV APPROACH VS TR BASED POLICY

As it can be seen, even in this case, the AI4PV approach brings additional revenues regardless of the cost of cleaning and threshold value, when compared to the TR-based policy. However, the magnitude of such increase varies according to the level of the cost of cleaning and threshold value. The AI4PV benefits are summarised in Table 4-2.

TABLE 4-2: OPTION B CBA SUMMARY

TR value / Cost of Cleaning (€/Wp)	TR=0.7	TR=0.75	TR=0.8	TR=0.85
0.003	5%	4%	2%	2%
0.006	4%	3%	1%	4%
0.009	3%	3%	2%	8%
0.012	1%	2%	2%	12%

4.3 CBA AI4PV CLEANING MODULE WITH AND WITHOUT RAIN

Finally, a CBA between the two AI4PV method, with and without considering rain events into the model, was performed. The results are shown in Figure 4-3. As it can be seen, including rain into the definition of cleaning schedule might bring additional revenues, that varies from 0.6% to 1% depending on the cost of cleaning.

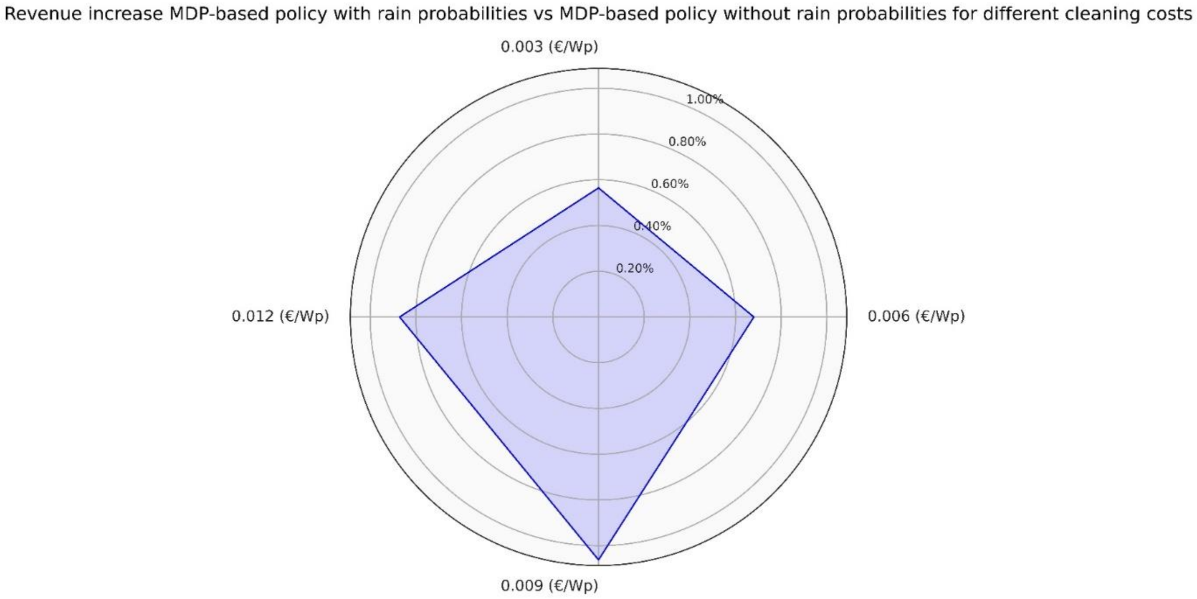


FIGURE 4-3: CBA AI4PV MODEL WITH RAIN VS AI4PV MODEL WITHOUT RAIN

5. CONCLUSIONS

This deliverable reported all the results and key finding of the AI4PV project collected during the validation phase.

As pointed out, all the solutions developed within the project, successfully passed the test phase, since all the KPIs defined were met.

Furthermore, as proved via the CBA, the AI4PV solutions might bring huge benefits to PV plant operators and owners, in the form of additional revenues due to optimal policy recommendation and early detection of faults and failures.

European Union
European Regional
Development Fund

6. REFERENCES

[1] L. Costa, A. Silva and C. Verrecchia, "D3.1 - Models for root-cause analysis with data analytics," 2023.

[2] M. Grieves, "Virtually intelligent product systems: Digital and physical twins," *Complex Systems Engineering: Theory and Practice*, 2019.

[3] Louelson Costa (INESC TEC), Christian Verrecchia (EDP NEW), Ruben Gonzalez Bernal (ISOTROL), "D1.1 - Use cases for O&M of solar power plants," 2021.

[4] M. G. Villalva, J. R. Gazoli and E. R. Filho, "Comprehensive Approach to Modeling and Simulation of Photovoltaic Arrays," *IEEE Transactions on Power Electronics*, 2009.

[5] 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," 2023.

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

[7] Gianfranco Di Lorenzo, Rodolfo Araneo, Massimo Mitolo, Alessandro Niccolai, and Francesco Grimaccia, "Review of O&M Practices in PV Plants: Failures, Solutions, Remote Control, and Monitoring Tools," *IEEE Journal of Photovoltaics*, 2020.

[8] "pandas.DataFrame.interpolate," 2023. [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.interpolate.html.

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

[10] "skopt.BayesSearchCV," 2023. [Online]. Available: https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html.

[11] "Bayesian optimization," 2023. [Online]. Available: https://en.wikipedia.org/wiki/Bayesian_optimization.

[12] "Bayes' theorem," 2023. [Online]. Available: https://en.wikipedia.org/wiki/Bayes'_theorem.

[13] R. Isermann, Fault-Diagnosis Applications: Model-Based Condition Monitoring: Actuators, Drives, Machinery, Plants, Sensors, and Fault-tolerant Systems, Springer, 2011.

European Union
European Regional
Development Fund

[14] "pandas.DataFrame.clip," 2023. [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.clip.html .

[15] J. Castro, D. Gómez and J. Tejada, "Polynomial calculation of the Shapley value based on sampling," *Computers & Operations Research*, vol. 36, no. 5, pp. 1726-1730, May 2009.

[16] P. Jain, J. Poon, J. P. Singh, C. Spanos, S. R. Sanders and S. K. Panda, "A Digital Twin Approach for Fault Diagnosis in Distributed Photovoltaic Systems," *IEEE Transactions on Power Electronics*, vol. 35, no. 1, pp. 940 - 956, January 2020.

[17] Christian Verrecchia (EDP NEW), Louelson Costa (INESC TEC), Ruben Gonzalez Bernal (ISOTROL), "D4.2 - Demonstration Plan," *AI4PV Project*, 2022.

[18] Christian Verrecchia (EDP NEW), Catarina Mendes Martins (EDP NEW), Miguel Chousal (EDP NEW), Flávia Barbosa (INESCTEC), "D3.2 - Method for return-on-investment prediction," *AI4PV project*, 2023.

[19] Miguel Angel Delgado (ISOTROL), Sergio Raigón (ISOTROL), Ricardo Morales (ISOTROL), Jose Garcia Franquelo (ISOTROL), Flávia Barbosa, Luis Guimarães (INESC TEC), Christian Verrecchia (EDP NEW), "D4.3 - Method for cost-optimized predictive maintenance," *AI4PV Project*, 2023.

COMPETE 2020 PORTUGAL 2020 CDTI @CDTIoficial European Union European Regional Development Fund