UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

João Lucas de Souza Silva

Analysis of Photovoltaic Systems with emphasis on Anomaly Classification under a Supervised Approach

Análise de Sistemas Fotovoltaicos com ênfase na Classificação de Anomalias sob Abordagem Supervisionada

Campinas

2024

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

João Lucas de Souza Silva

Analysis of Photovoltaic Systems with emphasis on Anomaly Classification under a Supervised Approach

Análise de Sistemas Fotovoltaicos com ênfase na Classificação de Anomalias sob Abordagem Supervisionada

Thesis presented to the Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas in partial fulfillment of the requirements for the degree of Doctor in Electrical Engineering, in the area of Electrical Energy.

Tese apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Engenharia Elétrica, na Área de Energia Elétrica.

Supervisor: Prof. Dr. Tárcio André dos Santos Barros

Este exemplar corresponde à versão final da tese defendida pelo aluno João Lucas de Souza Silva, e orientada pelo Prof. Dr. Tárcio André dos Santos Barros

---

Campinas

2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Si38a

Silva, João Lucas de Souza, 1994-
Analysis of photovoltaic systems with emphasis on anomaly classification under a supervised approach / João Lucas de Souza Silva. – Campinas, SP : [s.n.], 2024.

Orientador: Tárcio André dos Santos Barros.
Tese (doutorado) – Universidade Estadual de Campinas (UNICAMP), Faculdade de Engenharia Elétrica e de Computação.

1. Sistemas de energia solar fotovoltaica. 2. Classificação. 3. Aprendizado de máquina. 4. Aprendizagem Supervisionada (Aprendizado de máquina). I. Barros, Tárcio André dos Santos, 1987-. II. Universidade Estadual de Campinas (UNICAMP). Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações Complementares

Título em outro idioma: Análise de sistemas fotovoltaicos com ênfase na classificação de anomalias sob abordagem supervisionada
Palavras-chave em inglês:
Photovoltaic solar energy systems
Classification
Machine Learning
Supervised Learning (Machine Learning)
Área de concentração: Energia Elétrica
Titulação: Doutor em Engenharia Elétrica
Banca examinadora:
Tárcio André dos Santos Barros [Orientador]
Luiz Carlos Pereira da Silva
Denis Gustavo Fantinato
Ricardo Ruther
Leandro Michels
Data de defesa: 28-06-2024
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-3206-2241
- Currículo Lattes do autor: http://lattes.cnpq.br/0789910185774767

COMISSÃO JULGADORA – TESE DE DOUTORADO

Candidato: João Lucas de Souza Silva RA: 211497

Data da Defesa: 28 de junho de 2024

Título da Tese: “Analysis of Photovoltaic Systems with emphasis on Anomaly Classification under a Supervised Approach”.

Prof. Dr. Tárcio André dos Santos Barros (Presidente, FEEC/UNICAMP)
Prof. Dr. Luiz Carlos Pereira da Silva (Membro Interno, FEEC/UNICAMP)
Prof. Dr. Denis Gustavo Fantinato (Membro Interno, FEEC/UNICAMP)
Prof. Dr. Ricardo Ruther (Membro Externo, Universidade Federal de Santa Catarina)
Prof. Dr. Leandro Michels (Membro Externo, Universidade Federal de Santa Maria)

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de PósGraduação da Faculdade de Engenharia Elétrica e de Computação.

To my parents, Quitéria and Mozart,
for all the support in my education.

Acknowledgements

I express my gratitude to God for all the opportunities and individuals placed in my way that have contributed to the realization of this journey.

To my family, especially my parents Quitéria and Mozart, and my grandmother Rita Gomes, I extend my thanks for the love, support, and resources dedicated to my education.

To my brother Matheus, aunts, uncles, and cousins, I am thankful for your constant presence and support.

To my partner Michelle Melo, who has been with me since my arrival at Unicamp, I appreciate our collaboration in building this trajectory together.

I cannot forget to mention my beloved dogs, Raj and Chanel, who provided moments of joy and distraction throughout this journey.

To the late Prof. Marcelo Villalva, who supported my move to Campinas-SP during my master's and was instrumental in my solar energy education. He consistently sought all possible resources to fund our research, always believing that anything was possible despite the challenges.

To my advisor Prof. Tárcio Barros, who provided all the necessary support for the completion of my work and the continuation of the legacy of the Energy and Photovoltaic Systems Laboratory - Marcelo Villalva (LESF-MV). I am grateful for the trust in my work.

To the laboratory friends and everyone who contributed within UNICAMP during this time, especially those who engaged in discussions and shared valuable insights during this journey.

I extend my thanks to the University of Campinas (UNICAMP), the Faculty of Electrical and Computer Engineering, and the partnerships that aided my progress in the university: ANP (Brazilian National Oil, Natural Gas and Biofuels Agency) through the R&D levy regulation with Totalenergies; ANEEL and CPFL PD-00063-3032/2017 - PA3032. To the Federal Institute of São Paulo, Campus Hortolândia, for the experience as a professor.

Do not wait until the conditions are perfect to begin.
Beginning makes the conditions perfect. (Alan Cohen)

Abstract

Anomalies in photovoltaic (PV) systems are phenomena responsible for deviations from the normal behavior of a system. These deviations are natural in a PV system during its useful life, and constant monitoring is important to ensure efficiency and return on investment. Monitoring can be done, for example, by comparing simulations carried out before installing the PV system. However, there is a challenge in analyzing various data, understanding, and even classifying the problem, since it would be necessary to analyze data in a smaller range than just the monthly or annual energy generation provided by simulation software. Furthermore, simulations are also subject to variation depending on the solarimetric data used. One of the solutions for these analyses is the application of data science concepts, especially machine learning techniques. In this sense, this work aimed to investigate and propose a flow process to classify anomalies in PV systems with a supervised approach. To achieve this, the work went through the installation phases of a model PV system within the Sustainable Campus project, data collection, analysis of the installation with PV simulators, and finally, data science studies with supervised machine learning techniques. As results, we highlighted the evaluations of several nearby PV systems using software, creation of a methodology for exploratory analysis of PV data (macro analysis), creation of datasets for training/testing, and a new proposal flow process for an algorithm for classifying anomalies (micro analysis). The proposed flow process was a Random Forest ensemble with K-nearest neighbors (k-NN) and an inference machine for specific classes, tested on a real and synthetic basis. Testing indicated that the method classified anomalies achieved an AUC of 0.9815 for the synthetic dataset and an AUC of 0.9861 for the real dataset, and accuracy of 0.9647 for the real dataset. Thus, it was perceived that exploratory analysis is capable of providing clues to anomalies, and anomaly classification can be a viable and important step to limit the scope of action of the maintenance team.

Keywords: Photovoltaic Systems; Anomalies; Machine Learning; Anomalies in Photovoltaic Systems

Resumo

Anomalias em sistemas fotovoltaicos (FV) são fenômenos responsáveis por desvios do comportamento normal de um sistema. Esses desvios são naturais em um sistema FV durante sua vida útil, sendo importante o acompanhamento constante para garantir a eficiência e retorno do investimento. O acompanhamento pode ser feito, por exemplo, com a comparação de simulações realizadas antes da instalação do sistema FV. Porém, existe um desafio em analisar diversos dados, entender, e até classificar o problema, já que, seria preciso analisar dados em um intervalo menor do que somente a geração de energia mensal ou anual fornecida por *software* de simulação. Além disso, simulações também são passíveis de variação dependendo dos dados solarimétricos utilizados. Uma das soluções para essas análises é a aplicação dos conceitos de ciência de dados, sobretudo, técnicas de aprendizado de máquinas. Neste sentido, esse trabalho teve como objetivo investigar e propor um fluxo de processo para classificar anomalias em sistemas FV com abordagem supervisionada. Para isso, o trabalho passou pelas fases de instalação de um sistema FV modelo dentro do projeto Campus Sustentável, coleta de dados, análise da instalação com simuladores FV, e por fim, os estudos de ciência de dados com técnicas de aprendizado de máquina supervisionada. Como resultados, destacou-se as avaliações de diversos sistemas FV próximos por software, criação de metodologia para análise exploratória de dados FV (análise macro), criação de datasets para treino/teste, e uma proposta de processo de fluxo para um algoritmo de classificação de anomalias (análise micro). O processo de fluxo proposto utilizou um *ensemble* de *Random Forest* com *K-nearest neighbors*(k-NN) e uma máquina de inferência para classes específicas, testado em base real e sintética. Os testes indicaram que o método classificou anomalias alcançou uma AUC de 0,9815 para o conjunto de dados sintético e uma AUC de 0,9861 para o conjunto de dados real, e precisão de 0,9647 para o conjunto de dados real. Assim, percebeu-se que a análise exploratória é uma etapa capaz de mostrar indícios de anomalias e métodos para classificação de anomalias se mostraram viáveis e importantes para limitar o escopo de atuação da equipe de manutenção.

Palavras-chaves: Sistemas Fotovoltaicos; Anomalias; Aprendizagem de Máquina; Anomalias em Sistemas Fotovoltaicos.

List of Figures

List of Tables

List of Acronyms and Abbreviations

AC: Alternating Current

AE-LSTM: AutoEncoder Long Short-Term Memory

ANEEL: National Electricity Agency

AOI: Angle of Incidence

AUC: Area under the ROC Curve

CBR: Case-based reasoning

CEPETRO: Center for Energy and Petroleum Studies

CF: Capacity Factor

CG: Centralized Generation

COBEP: Brazilian Power Electronics Conference

CPFL: Local Electricity Distributor

DC: Direct Current

DG: Distributed Generation

DHI: Diffuse Horizontal Irradiance

DNI: Direct Normal Irradiance

DT: Decision Tree

FEC: School of Civil Engineering

FEEC: School of Electrical and Computer Engineering

FN: False Negative

FP: False Positive

G: Irradiance

GHI: Global Horizontal Irradiance

GTB: Gymnasium Test Base

HDKR: Hay–Davies–Klucher–Reindl

I: Photovoltaic Cell Current

I_0: Diode Reverse Saturation Current

I_{PV}: Current generated by the incidence of light

IEC: International Electrotechnical Commission

IRR: Internal Rate of Return

ISO: International Organization for Standardization

k-NN: K-nearest neighbors

K_I: Temperature coefficient of the short-circuit

K_V: Temperature coefficient of the open-circuit

LCOE: Levelized Cost of Energy

LESF: Laboratory of Energy and Photovoltaic Systems

LR: Logistic Regression

MEC: Maximum Entropy Classifier

ML: Machine Learning

MLP: Multilayer Perceptron

MPPT: Maximum Power Point Tracking

NB: Naïve Bayesian

NIPE: Interdisciplinary Energy Planning Center

NN: Neural Networks

NPV: Net Present Value

O&M: Operation and Maintenance

OC: Open-circuit

P_{ac}: Output power of the inverter

PCA: Principal Component Analysis

PID: Potential-induced degradation

POA: Plane of Array (Irradiance)

PV: Photovoltaic

R_{AVG}: Average Recall

R_P: Parallel Resistance

R_S: Series Resistance

RF: Random forest

ROC: Receiver Operating Characteristic Curve

SC: Short-circuit

SMLT: Supervised Machine Learning Techniques

SMOTE: Synthetic Minority Over-sampling TEchnique

STB: Synthetic Test Base

SVM: Support Vector Machine

TN: True Negative

TP: True Positive

UCP: Power Conditioning Unit

UMG: Unicamp's Multidisciplinary Gymnasium

UNICAMP: Universidade Estadual de Campinas

V_{dc1}: Voltage at the DC input of the inverter

V_T: Thermal Voltage

Y_n: Measured data

Contents

1 Introduction

1.1 Analysis of Photovoltaic Systems and Anomalies

Monitoring and analyzing photovoltaic (PV) systems is a necessary task to ensure financial return and extend the useful life of the equipment. With monitoring and maintenance it becomes possible to have more efficient systems, as the identification of anomalies helps to avoid low levels of energy production [1]. This task is often overlooked by integrators due to the costs involved, or even lack of knowledge.

The lack of monitoring and evaluation in PV systems can result in failures or interference in energy production. These deviations from conventional behavior are commonly called anomalies. When identified, it becomes imperative to take corrective actions. These interventions can take place on various components, such as PV modules, inverters, electrical cables, among others, as recommended in the maintenance guidelines proposed by Spertino and Corona [2].

Silva [3] summarized the types of losses/anomalies that exist in PV systems. These losses were manufacturing mismatch and operation mismatch [4], as a solution one must comply with manufacturing standards [5] or use different architectures [6, 7]; partial and total shading which is generally caused by dirt, clouds, constructions, among others [8–11], and as a solution after identification it is desired to optimize the project; Soiling [8, 11,12], which is also linked to shading; Variations in irradiance, which is a natural anomaly due to the aeration of gases in the atmosphere, the passage of clouds, reflection, among other phenomena [3]; Temperature, natural climate variations [13], soiling [11], defects in components and poor sizing; Hotspot [14]; failures in DC/DC Converters or DC/AC Converters, whether in switching [15, 16], temperature; parasitic resistances, inductances and capacitances; oversizing or undersizing; electromagnetic interference [17]; Quality of Electricity, disturbances in the electrical network that interfere with the system [18]; Connections and Cabling, these are generally failures caused during design and installation that worsen over time [3]; and aging and degradation of the system, which is normal during its useful life, as long as adequate maintenance is carried out.

The reality is that these faults can be identified through the data monitored by the PV converter [19]. PV converters perform a massive data collection, forming datasets. For the integrator, it is still a challenge to manage and understand these data. Thus, data science can be important for validating the integrity, analyzing, and classifying these data through machine learning (ML). Analyzing data-driven systems provides an opportunity

to comprehend underlying information and detect potential anomalies. Furthermore, it facilitates the recognition of patterns, features, and trends in the data.

1.2 Grid-connected Photovoltaic Systems

Grid-connected PV systems stand out for their ability to utilize the local utility grid, enhancing profitability and versatility in applications, which is the case for the majority of systems, particularly in Brazil. The Figure 1.1 illustrates the main types of systems that are purely connected to the electrical grid, i.e., without the use of batteries, diesel generators, or other forms of on-site energy.

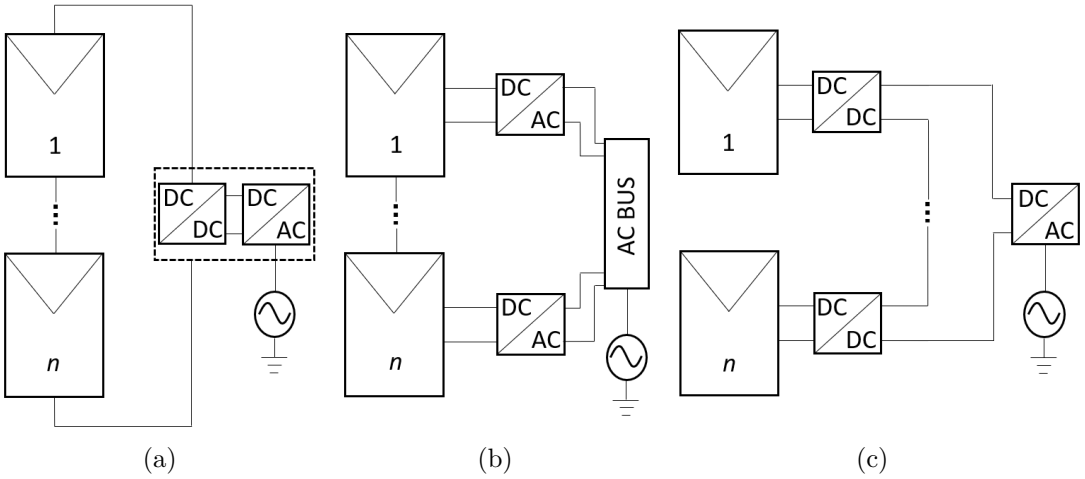


Figure 1.1 – Main types of PV systems connected to the electrical grid: (a) Conventional, (b) microinverters, and (c) Power Optimizers [20].

The grid-connected PV system is composed, in addition to protection elements, disconnection devices, and mechanical components (supports), of two basic elements: PV modules and converters. PV modules are responsible for capturing and converting energy from the sun through the PV effect. However, for the PV module to function, it is necessary to use a converter, which will process this energy from the PV module and enable the system to inject it into the electrical grid. All of this is done following local and international standards/guidelines, which ensure the safety and quality of the operation.

The connection method and type of converter determine the operation of a PV system. Conventional systems, as shown in Figure 1.1(a), use a single converter, commonly referred to as an inverter, which is known in Brazilian standards as a UCP (Power Conditioning Unit). In this type of system, the PV modules are directly connected to each other, forming a series and/or parallel connection, in one or more inputs. It is in this case that the concept of mismatch arises, which is imbalance exists between modules connected in series and/or parallel due to extrinsic and/or intrinsic factors related to the

PV cell [3, 4]. It is worth noting that there are inverters with more than one input that combine sets of PV modules internally, and others have more than one internal DC/DC converter (more than one MPPT), which improves the system's flexibility. It is important to highlight that, in this thesis, the systems covered are as shown in Figure 1.1(a), with one or several UCP.

This has led to the emergence of other forms of connection and the use of converters in PV applications. One of the objectives was to avoid direct connection between PV modules. The solution shown in Figure 1.1(b), which is the microinverters, works by being an inverter for a small group of PV modules. On the other hand, the solution shown in Figure 1.1(c), power optimizers, adds a DC/DC converter in addition to the inverter, which separates the PV modules and controls them individually or in small groups. Both solutions sought to mitigate mismatch; however, they added new safety and maintenance features since it is possible to monitor at the module level with these two latter solutions [20].

The fact is that all three solutions generate data that can aid in anomaly classification through ML. However, managing multiple systems and their data becomes challenging without the use of intelligent solutions. Many issues can go unnoticed by merely observing inverter data, such as unmonitored problems in the combine box (a box that combines multiple PV strings and protections) or a simple bypass diode failure. This work primarily focused on the conventional solution, as it is still widely used. However, the tests can be replicated for other PV architectures. Moreover, with module-level data, identifying potential anomalies becomes easier.

To underscore the importance of grid-connected PV systems in Brazil, one can observe the exponential growth of the technology. Currently, there are over 41.1 GW in operation, creating 1.2 million new jobs and avoiding 47.6 million tons of CO_2 emissions [21]. This growth is driving economic activity and enhancing the country's sustainability. Figure 1.2 depicts the evolution in recent years, categorized into Distributed Generation (DG) and Centralized Generation (CG). DG systems have an installed power level of up to 3 MW, while Centralized Generation exceeds this limit. Thus, with such growth, it becomes necessary to analyze the performance and anomaly detection on various fronts to avoid financial waste in the use of PV systems.

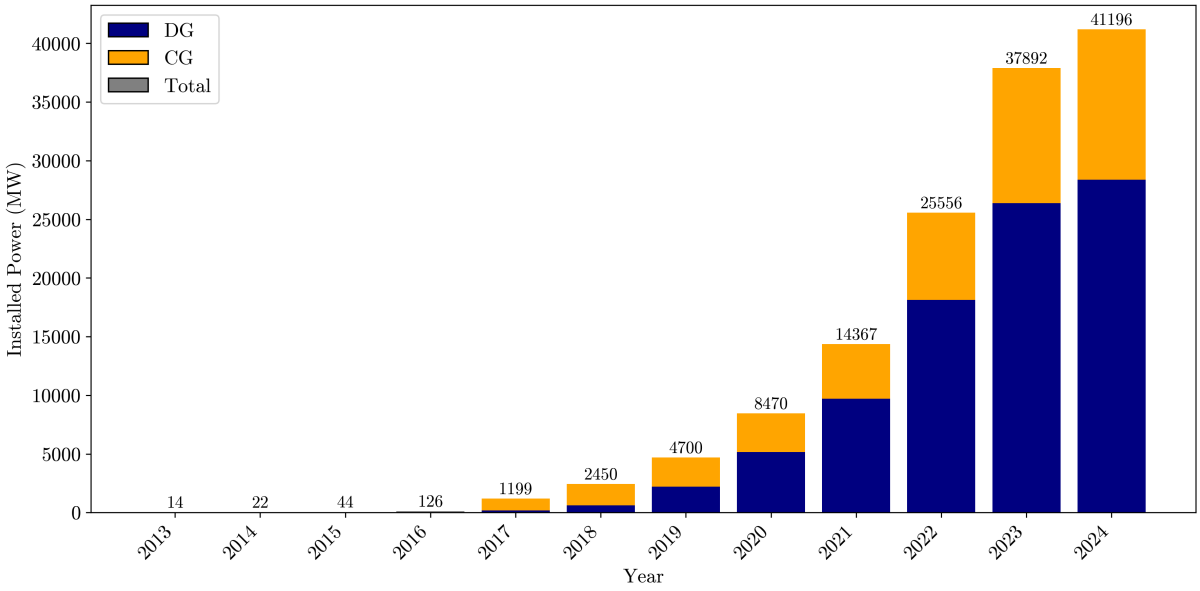


Figure 1.2 – Growth of DG and CG in Brazil until March 2024. Adapted from [21]

1.3 Machine Learning and Supervised Approach

ML is the science of programming computers to enable them to learn from data [22]. In this way, it is possible to escape the traditional programming approach in which it was necessary to create lists with several complex (conditional) rules to extract or analyze behavior from a set of data, and even so, sometimes it was not possible to extract all the information , as these were complex problems.

For this, as generally illustrated in Figure 1.3, a robust and reliable data set is necessary that will be subsequently prepared for analysis according to the preparation methodology adopted by the data scientist (*dataprep*). In the preparation stage, there is also the partitioning of data into training and testing sets, in some cases also validation; the percentage of data allocation is also a decision to be made. Subsequently, the data go through the training/testing process with the chosen algorithm, and the results are obtained. The result can be demonstrated through *storytelling* techniques for better use. With these steps completed, new data can be applied to the model.

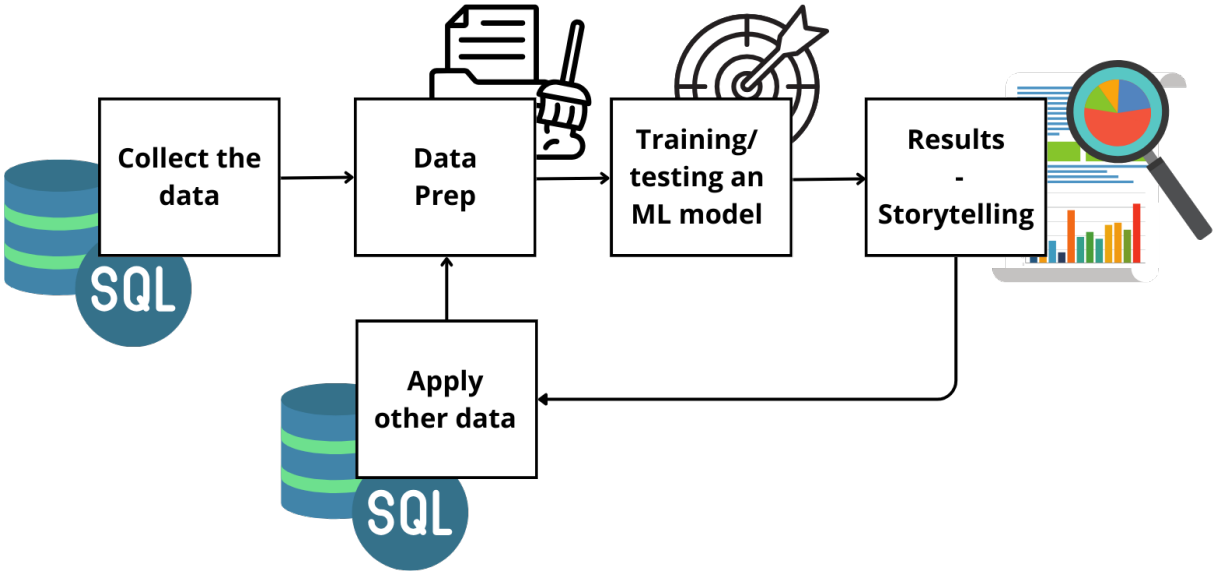


Figure 1.3 – Schematic of generic ML application.

In the data preparation stage, the application of *Data-Driven Analysis* can be essential. These initial analyses, using statistical methods, can aid in detecting noise and outliers. Often, a thorough data analysis using statistical methods is sufficient to identify issues or anomalies in a PV system, for example, and subsequently improve anomaly classification. These analyses also help in pattern recognition, feature selection (data correlation), and extracting meaningful insights from available data.

ML fields of application include computer vision, prediction, semantic analysis, natural language processing, and information retrieval [23]. Within these domains, ML models empower engineers, researchers, data scientists, and analysts to produce dependable and valid results and decisions [24].

In Géron (2019) [22], several criteria that can be met by ML algorithms are listed, with the recognition that an algorithm can fulfill more than one criterion. These criteria include: their capacity for human supervision during training (supervised, unsupervised, semi-supervised, and reinforcement learning methods); their speed of incremental learning (online and batch learning); and their capability to identify patterns in data, allowing for the creation of predictive models or simple comparison of new data points with existing ones (Instance-Based Versus Learning, and Model-Based Learning) [22]. Figure 1.4 shows the ML types defined by [22].

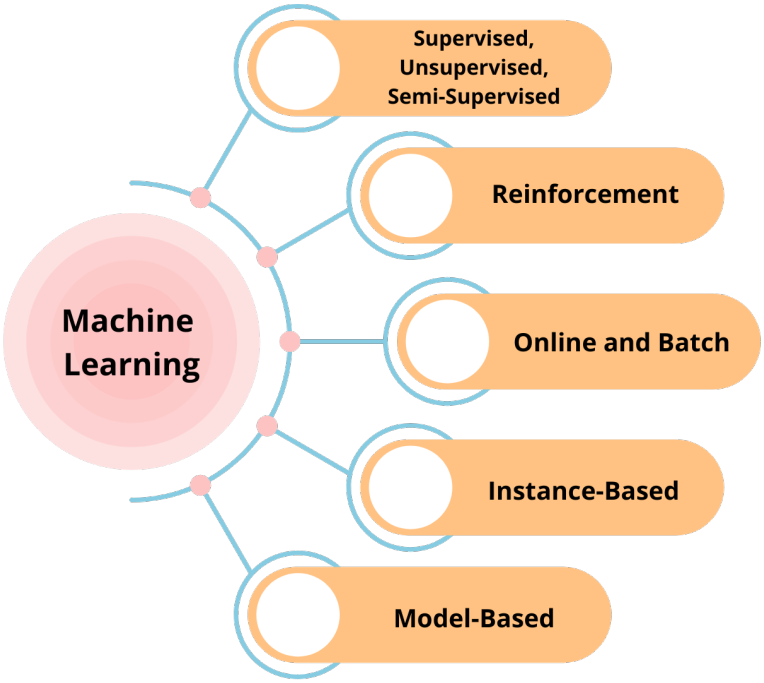


Figure 1.4 – Types of ML. Adapted from [22].

Supervised methods can classify a desired outcome through training with labeled datasets [25]. Labeling data can pose a challenge depending on the type of data being analyzed and the availability of experts. For PV projects, the labeling process needs to be done by experts or using labeling algorithms when synthetic data are available. The fact remains that with supervised methods, accuracy can be higher compared to other methods, as labeling facilitates training.

1.4 Challenges in Supervised Data Classification

The data classification process involves distinguishing and categorizing data into different relevant groups, making it a significant research topic due to its importance in human activities [26]. ML techniques have significantly enhanced the classification process; however, several challenges persist.

The formulation of the data classification problem is presented by Czarnowski and Jędrzejowicz [26]. When it comes to data classification, it is necessary to define a set to serve as an example to the algorithm (represented by U), and this set must be non-empty and finite.

A data $x \in U$ is described by a defined set of features, $A = \{a_1, a_2, \ldots, a_n\}$, where n denotes the number of features. For each feature a_i, there exists a value $a_i(x) \in V_{a_i}$, with V_{a_i} being the set of potential values for the feature a_i [26]. As a result, learning from multiple labeled data instances enables the determination of the dependency between these class labels and the features, facilitating the application of this relationship to new data.

The class labels are members of a finite set of predefined decision classes, $C = \{c_1, \ldots, c_k\}$, where k denotes the total number of these classes [26]. Consequently, the set U for the classification task is represented, according to Czarnowski and Jędrzejowicz [26], as follows in Equation 1.1:

$$U = \{[x_{ij}, d(x_j)] : i = 1, \ldots, n; j = 1, \ldots, N\} \tag{1.1}$$

In summary, a single data can be represented by $[x_{ij}, d(x_j)]$, where the set U consists of N such vectors, and $d(x)$ denotes the class label value for the example x, with $\forall_{x \in U} d(x) \in C$ [26].

When applying ML, it is feasible to generate a classifier $h \in H$ from the set U, where H denotes the hypothesis space. With a dataset U, a set of hypotheses H, and a performance criterion or criteria F, the learning algorithm produces a hypothesis $h \in H$. Consequently, learning from examples involves creating algorithms to find the best $h \in H$ relative to the performance values defined in F [26].

Thus, the output of the classifier is assessed using the performance measure $f \in F$. This process of learning from examples can be framed as maximizing the performance measure relative to the hypothesis h [26], as seen in Equation 1.2:

$$h = \arg\max_{h \in H} f(h) \tag{1.2}$$

In this structure, a classifier h is defined as a mathematical function that assigns examples from D to a predefined set of classes [26], as shown in Equation 1.3. It is important to highlight that the classification process involves different steps depending on the algorithm used and dataset.

$$h : U \rightarrow \{\emptyset, C_1, C_2, \ldots, C_k\} \tag{1.3}$$

However, data classification poses various challenges such as data preprocessing, handling imbalanced datasets, feature selection, data variability, classes with similar characteristics, and the volume of data.

The data preprocessing stage is important because it directly impacts the quality of the training, validation, and testing phases. It involves checking for missing values, dealing with noise depending on the application, and identifying outliers in the data collection process.

The challenge of dealing with imbalanced datasets exists mainly when it comes to anomalies within a dataset that is predominantly normal. This issue is particularly pronounced in the classification of data generated by PV systems, as these systems typically operate normally most of the time. Therefore, it is necessary to develop strategies to make anomalies more representative during model training. Imbalanced data can lead to overfitting, where the model replicates the behavior of the majority class [27], rather than learning to identify anomalies. Consequently, taxonomy strategies are often proposed in the literature to address this issue. Some of the strategies are performing resampling [28], synthetic data generation [29], using class weights [30], or even adopting a set of these.

Another challenge is Data Variability, which is particularly pronounced in PV systems. An example of data variability is when a single dataset exhibits changes in data patterns across different periods. PV systems, being exposed to varying weather conditions and seasons, often experience this phenomenon. To address this, various strategies can be employed, such as utilizing models that are more robust to data variability or dividing the dataset according to different periods.

Depending on the classification problem, there may also be the challenge of classes with similar characteristics (class overlap [31]). This occurs when the features defining the classes exhibit very similar behavior, making them difficult to distinguish. In such cases, it may be necessary to seek out new features. For instance, in PV systems, a damaged bypass diode in a PV module might produce behavior similar to a certain level of shading in the DC voltage and current features. This can be problematic when trying to separate faults but may not be an issue when the goal is merely to identify the anomaly and classify it into a broader category of faults (as proposed in this thesis).

The volume of data can also present a challenge during the classification process. Managing data quality becomes difficult with large volumes, and the data can overwhelm classification systems. While this is not typically an issue for PV systems at the inverter level, it becomes considerable when dealing with central PV power plants and multiple inverters together. This problem is addressed with models designed for Big Data Classification [32].

1.5 Objectives

The general objective of the work is to investigate and classify anomalies in PV systems through the application of a supervised approach. To achieve the general objectives, the following specific objectives will be considered:

- Analyze and compare different systems through simulations and, subsequently, analyze the behavior of energy generation between nearby PV installations;
- Develop a methodology for exploratory analysis of *dataset* with field PV data;
- Develop procedures for creating synthetic datasets with anomalies from solar irradiance field data;
- Implement a proposed process flow for a supervised method of type *ensemble* to identify anomalies in PV systems.

1.6 Contributions

This work has as scientific/technical contributions to literature the following points:

- Development and validation a new methodology for generating a *dataset* for training/testing ML algorithms for detecting/classifying anomalies using supervised approach;
- Proposal for a methodology for exploratory analysis of PV data;
- Proposal for a process flow for a supervised ML method of the *ensemble* type;
- Understand the variation in generation behavior for PV facilities nearby and how data from those facilities report.

1.7 Thesis Structure

This thesis was subdivided into four scientific papers, each of which is considered an integral chapter of the thesis. The selection of these papers was based on their degree of relevance to the project, although other works have been published and are listed at the end of the thesis.

In the second chapter, the scientific paper entitled "Case Study of Photovoltaic Power Plants in a Model of Sustainable University in Brazil" is presented, published in the journal Renewable Energy by Elsevier in August 2022. The present study contributed to the design, simulation and monitoring of PV installations at the *Universidade Estadual de Campinas (Unicamp)*, allowing the analysis of the variation in performance of several plants in an area of 5 km, being a difference in relation to works already existing in the literature. Developments regarding solar PV within the project were presented. Additionally, the economic evaluation of PV systems when operated together is addressed, within the context of the free energy market model.

The third chapter was composed of the scientific paper entitled "Evaluating the Significance of Solarimetric Data for Photovoltaic System Simulation in a Real-World Case" presented and published at the 2024 Brazilian Power Electronics Conference (COBEP). The contribution of the paper was to analyze the integration of real solarimetric data obtained close to the installation site (PV system of the Unicamp's Multidisciplinary Gymnasium) within the PVsyst software, compared with the use of data from solarimetric bases (Meteonorm) used in the installation project. In addition, statistical studies were carried out on the solarimetric database. The data collected from the station was analyzed for inclusion in other studies conducted in this thesis.

The fourth chapter consisted of the scientific paper entitled "Data-Driven Analysis of Solar Photovoltaic Systems: Correlation and Distribution Patterns" presented and published at the 2024 Brazilian Power Electronics Conference (COBEP). In the literature, a notable challenge has been identified in terms of outlining the appropriate approach for data analysis during the data preparation stage (dataprep). In response to this, a novel methodology has been formulated as a contribution, drawing upon analyzes employed in prior studies. This methodology proposes a distinct approach to the analysis of PV data.

The fifth chapter brings the article entitled "Classification of Anomalies in Photovoltaic Systems using Supervised Machine Learning Techniques and Real Data" is presented, published in the Energy Reports by Elsevier in June 2024. As a contribution in this paper, methods for classification of anomalies, developed a procedure for creating synthetic datasets with anomalies from solar irradiance field data, and finalized with a proposal process flow for a supervised method of type *ensemble* to identify anomalies.

The proposed method was compared with other methods on a synthetic and real dataset. Finally, the considerations obtained from the results of each chapter were synthesized in a discussion section and the conclusions.

2 Case Study of Photovoltaic Power Plants in a Model of Sustainable University in Brazil

Paper published in Renewable Energy [33] ©2022 Elsevier. Reprinted, with permission, from de Silva, J.L. S. Case Study of Photovoltaic Power Plants in a Model of Sustainable University in Brazil, Renewable Energy, August, 2022.

This manuscript, authored by João Lucas de Souza Silva, Karen Barbosa de Melo, Kaio Vieira dos Santos, Elson Yoiti Sakô, Michelle Kitayama da Silva, Hugo Soeiro Moreira, Giulianno Bolognesi Archilli, João Guilherme Ito Cypriano, Rafael Espino Campos, Luiz Carlos Pereira da Silva, and Marcelo Gradella Villalva. The paper is associated with the Digital Object Identifier (DOI): <https://doi.org/10.1016/j.renene.2022.06.103. >. This work was developed under the Electricity Sector Research and Development Program PD-00063-3032/2017 - PA3032: "Sustainable campus model at the University of Campinas - Brazil: An integrated living lab for renewable energy, electric mobility, energy efficiency, monitoring and energy demand management", regulated by the National Electricity Agency (ANEEL in Portuguese), in partnership with CPFL Brazil (Local Electricity Distributor). This work was also partially funded by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and BYD Energy Brazil with resources from the PADIS/MCTI/MDIC program.

Abstract: Universities play an important role in the search for a sustainable environment. The challenge for universities is to outline models of a sustainable society and show their benefits, seeking the engagement of all. In this opportunity, the University of Campinas (UNICAMP), with support from CPFL Brazil (Local Electricity Distributor), created a project called "Sustainable Campus" in August 2017. The Sustainable Campus project encompasses several sub-projects which aim to enhance the sustainability of the university. Among the subprojects (SP) of Sustainable Campus, the subproject photovoltaic (PV) was designed to follow up on the construction and operation of 6 PV power plants gathering a power equivalent of 535.26 kWp. In this way, this work explores the initiatives of the Sustainable Campus project with a focus on PV plants. Thus, it is possible to validate the simulations carried out with the PV plants and verify the impact of the system on the university. For this, the simulator data were compared with the measurement of the PV plants. Assembling the information, it was possible to verify different behaviours in PV plants in a range of 5 km. Quantitatively, for a year of data collected, an actual energy generation of 784.29 MWh was observed, while the simulated result showed a total of 759.04 MWh. PV generation resulted in 1.13% of the energy con-

sumed by the university due to the large size of the campus. Furthermore, as UNICAMP is a consumer that is in the free energy market model, the analysis of a possible financial return is presented. With an estimated payback of 7.65 years considering the average cost of R$ 4.64/Wp, which makes the system viable. Finally, the success of this project aims to bias other universities to deploy renewable energy sources, contributing to sustainable development and scaling the initiative to other sectors.

Keywords: Sustainable, Photovoltaic, Photovoltaic Plants, Sustainable University, Sustainable Campus.

2.1 Introduction

Several transformations are happening in the electric sector throughout the world, motivated by the search for a more sustainable, technological, and self-sufficient energy conversion, as well as the diversification of the energy matrix. However, the challenges and actions for sustainability may differ according to each country and its peculiarities [34]. All in pursuit of economic and social well-being, making possible a more sustainable use of environmental resources [35].

This could not be different within universities, given the challenges of seeking the best solutions for the electric sector. In mobilising for a more sustainable world, universities have the role of proposing solutions, testing ideas in laboratories, and training and educating new leaders to promote a sustainable future for the next generations. In the last decade, a number of universities have started engaging with sustainable practices and sustainable development within their activities and operation. These sustainable practices involved signing national and international declarations, and partnershis aiming sustainable commitments [36]. The foundations of a "sustainable university" exists when it has an administration that respects green environmental, economic and social practices, whilst providing research and encouraging the community outside the university to adopt practices aiming a sustainable civilization [37].

Therefore, the University of Campinas (UNICAMP), with support from CPFL Brazil (Local Electricity Distributor), created a project called "Sustainable Campus" in August 2017, with several sub-projects that seek to make the university sustainable [38]. The Project "Sustainable Campus" is composed by twelve subprojects, focusing in differents areas of research. In the following subsections, these subprojects will be presented briefly. Figure 2.1 presents a graphic summary of all the subprojects within the "Sustainable Campus" project.

One of the subprojects of the "Sustainable Campus" of UNICAMP and the

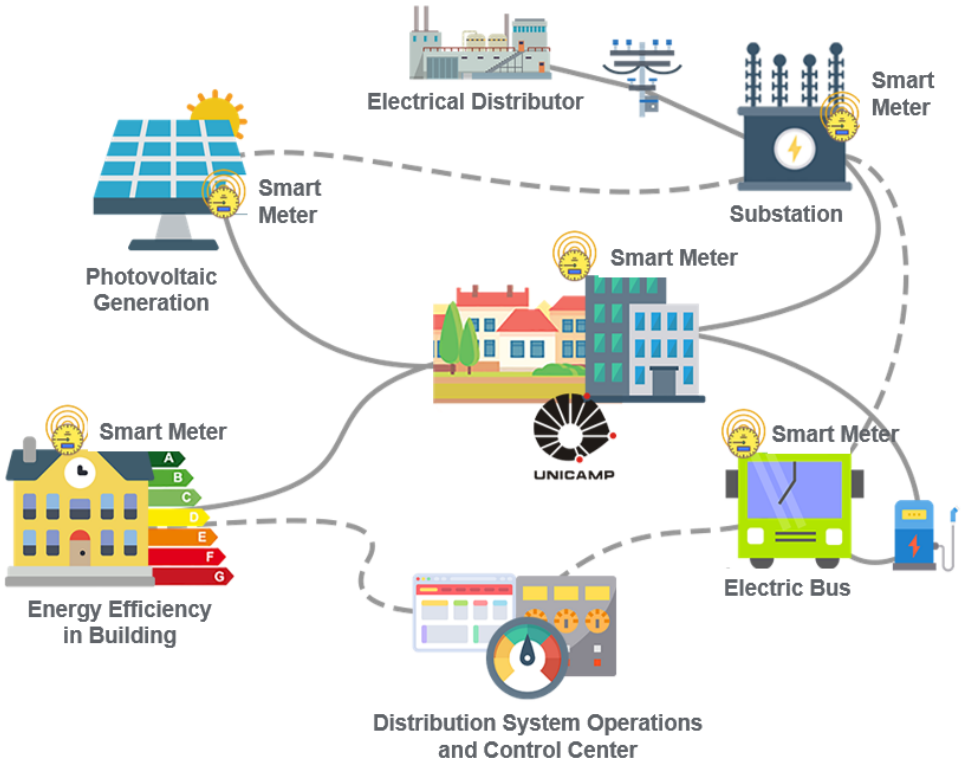


Figure 2.1 – Graphic summary of UNICAMP's "Sustainable Campus" project [38].

focus of this paper was the solar photovoltaic (PV) subgroup. Solar PV energy is a promissing energy source as the sun's energy is a clean, safe, and inexhaustible resource. Also, Brazil is one of the leading markets in South America for this technology [39]. Thus, PV solar energy deserves attention in the search for sustainable environments.

In parallel with the PV project, UNICAMP chose to integrate the free energy market contract model. This option in Brazil has been applied to consumers with large loads, who can freely negotiate the purchase of energy in the market, having the advantage of paying for a lower energy price [40]. In this model, the energy produced by a PV system must be used immediately by the loads or it will not bring profit, since it cannot participate in a later compensation model. Thus, there is a challenge in making PV installations viable in this scenario of lower electricity costs, and immediate use of energy. This results in few free-market customers with PV plants and little data in the literature.

In this way, this paper presents the case study of the six PV plants of the PV subproject of the "Sustainable Campus". The presentation of the plants, simulation, and comparison with 12 months of real data for each PV plant was performed. Thus, it was possible to analyze a year of data with a simulated year in the PVsyst software [41]. Besides, it was commented on several points of the "Sustainable Campus" project to present and encourage the scientific community to develop similar projects. The following contributions are obtained from the project:

- Check several PV installations of the "Sustainable Campus" of UNICAMP compared to simulation software;
- Understand the variation in generation behavior for PV facilities nearby and how data from those facilities relate;
- Create a sustainable university model and to demonstrate its importance to the community.
- Economic evaluation of PV systems for consumers in the free market model.

After this introduction, section 2 comments on some research initiatives involving sustainability, followed by section 3 that explores the PV subproject. Subsequently, the methodology of the present study was presented, followed by the results obtained and discussions. Finally, proposals for future PV installations and conclusions were presented.

2.2 Sustainable campus initiatives around the world

University campuses are excellent places to approach sustainable and smart initiatives because they are big enough for studies and small enough for implementation [42]. The smart/sustainable campus proposal must be scalable. Sustainable campuses are within the smart campus group. There are several initiatives within the smart campuses, such as smart building, smart environment, mobility, energy efficiency, renewable energy, energy management, among others.

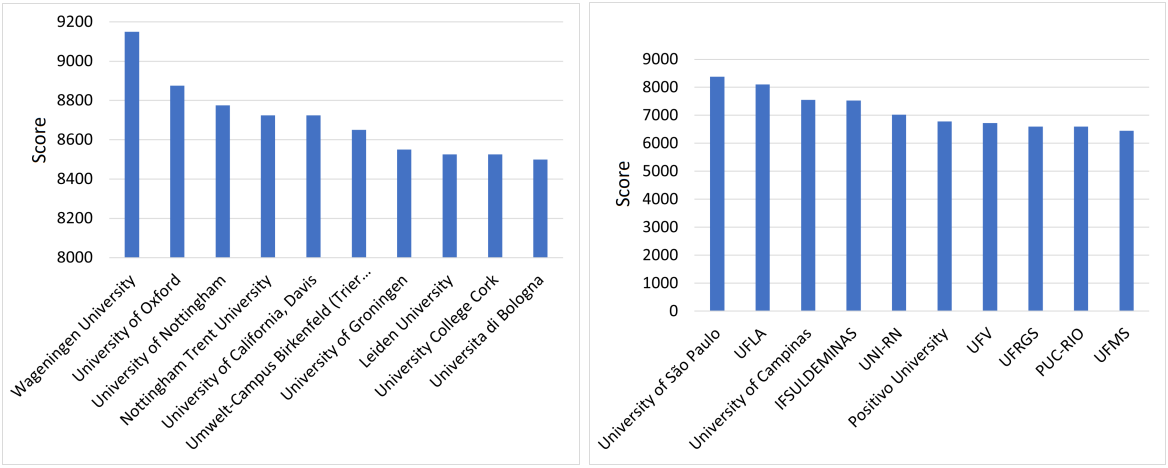
In terms of renewable energy, many campuses create microgrids as they end up studying different energy sources. The microgrid is important because it facilitates energy management. For example, at Hangzhou Dianzi University there is a hybrid system with a 120 kWp PV system, 120 kWp diesel generator, set of 100 kW super-capacitors (EDLC - Electric Double Layer Capacitor), and 50 kW/50 kWh lead-acid battery [43]. The highlight of Hangzhou Dianzi was the high penetration of renewable energy (50%), providing an interesting scenario for study.

The University of California-Irvine (USA) also stands out for having 1 MW of solar energy in 2014, with a planned expansion to 4 MW, reducing utility bills, with initiatives other than solar energy [44]. At University Politehnica of Bucharest (Romania), there are studies to optimize the operation of smart campuses in order to reduce costs [45, 46]. Off-grid PV systems and other initiatives were installed on the Auckland Park campus of the University of Johannesburg (South Africa) to meet the demand on the campus [47]. At the University of Genova (Italy) there are studies to explore microgrids with PV, storage and integration of electric vehicles [48].

With a proposal to produce buildings that reduce energy costs and waste, universities began to apply smart buildings. Angelis et al. [49] evaluated productions with renewable energies, with a PV system of 10-15 kW at the University of Brescia. Christensen et al. [50] evaluated buildings on a university campus in Denmark; Chalfoun [51] studied projects of smart buildings at the University of Arizona (USA). Escobedo et al. [52] evaluated issues related to energy consumption and greenhouse gases in constructing the National Autonomous University of Mexico (Mexico). Leon et. al. [53] assessed environmental impacts of the University of Basque Country and the city of Donostian-San Sebatián - Spain, where the university is located, since 205, in order to indentify reform scenarios to make the university more sustainable, enabling the university to align its efforts with the Convenant of Mayors of Climate and Energy.

To assess the interaction of the various research initiatives that make up a sustainable campus, the UI website GreenMetric World University Rankings was used in [54]. The score considers infrastructure, energy, water, transport, education and research. However, it is also linked to the size (area) of the Campus. In 2020, in the overall ranking, stood out Wageningen University & Research (Netherland), University of Oxford (United Kingdom), University of Nottingham (United Kingdom), as shown in Figure 2.2-(a). The universities of the United Kingdom and the Netherland stand out in the ranking.

In Brazil, the University of Sao Paulo, the Federal University of Lavras and the University of Campinas stand out in UI GreenMetric World University Rankings. Despite the high investment in sustainability studies at the University of Campinas, supported by CPFL Brazil (Local Electricity Distributor), more initiatives are needed to serve the entire campus due to the university's size in the area. Figure 2.2-(b) shows the 10 universities in Brazil in UI GreenMetric.



(a) (b)

Figure 2.2 – UI GreenMetric World University Rankings: (a) The top ten scores in the world, and (b) The ten best scores in Brazil [54].

2.3 Photovoltaic mini-generation subgroup on the Sustainable Campus

2.3.1 Photovoltaic Generation

The implementation of photovoltaic (PV) generation at UNICAMP is an important initiative to reduce the university's energy costs, and mainly, to establish a living laboratory for research, training, and education of technicians and specialists in PV energy generation.

Unicamp currently has six sites with PV installations of the Sustainable Campus project:

- Unicamp's Multidisciplinary Gymnasium (UMG) with 336.96 kWp;
- School of Electrical and Computer Engineering (FEEC) with 94.62 kWp, divided into FEEC 1 and FEEC 2;
- Interdisciplinary Energy Planning Center (NIPE) with 38.88 kWp;
- Exploratory Science Museum with 4.05 kWp;
- Extecamp with 22.95 kWp;
- School of Civil Engineering with (FEC) with 37.80 kWp.

The sum of all these systems results in a total of 535.26 kWp. PV systems are shown in Figure 2.3.

The data from these PV systems allows the validation of research results related to the different PV modules evaluation; solarimetric studies; solar irradiance modeling; modeling of PV modules; study of energy simulation methodology and performance evaluation of PV systems; development of simulation software for the performance evaluation of PV systems; development of I-V curve tracer equipment for commissioning PV systems.

(a) (b)

(c) (d)

(e) (f)

Figure 2.3 – Photovoltaic system of: (a) Unicamp's Multidisciplinary Gymnasium; (b) School of Electrical and Computer Engineering (FEEC); (c) Interdisciplinary Energy Planning Center (NIPE); (d) Exploratory Science Museum; (e) Extecamp; (f) School of Civil Engineering (FEC).

2.3.2 IV Curve Tracer

An I-V curve tracer with low-cost and high precision was developed as a solution for small PV systems installers [55]. This device was tested in partial shading conditions, which generates I-V curves that requires a complex algorithm to trace, and the obtained results were close to those of high-cost commercial tracers. The average errors of voltage and current in relation to the commercial IV curve tracer were 2.46% and 0.98%, respectively, for the knee points of IV curve and 2.35% and 1.3%, respectively, for the inflection points.

The I-V curves in partial shading occur for PV modules at different temperatures and/or irradiance during use. This can occur due to shading by objects, dirt, dust, bird droppings, and fog [56]. The Figure 2.4 shows examples of curves generated under partial shading conditions in modeling. The gray areas represent the different levels of shading, which give rise to the knees and inflection points on the curve.

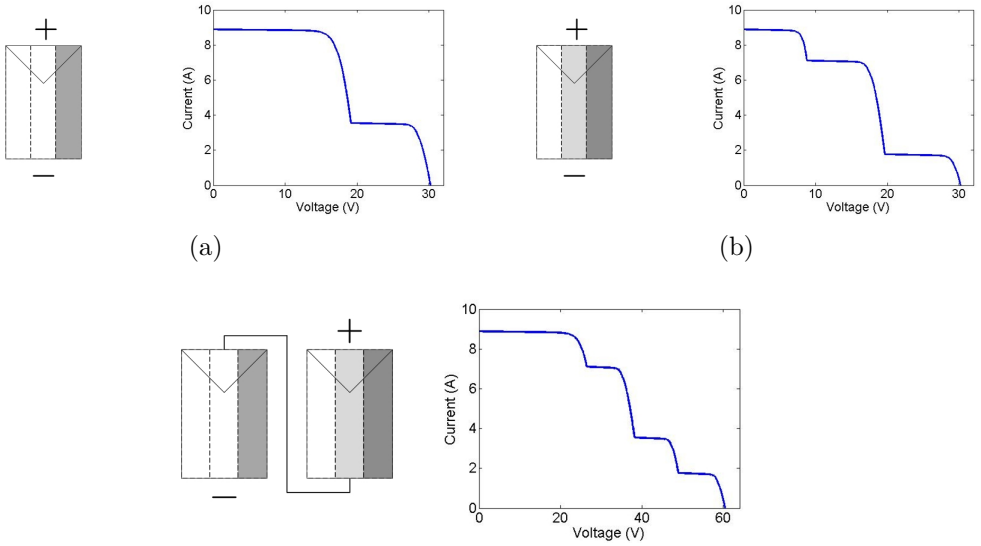


Figure 2.4 – Examples of PV modules and string under partial shading conditions [55].

The Figure 2.5 shows the developed equipment, and the Figure 2.6 shows curves obtained with the commercial tracer I-V and the developed equipment. The equipment developed has the potential for a much lower cost than the commercial one and is easier to configure.

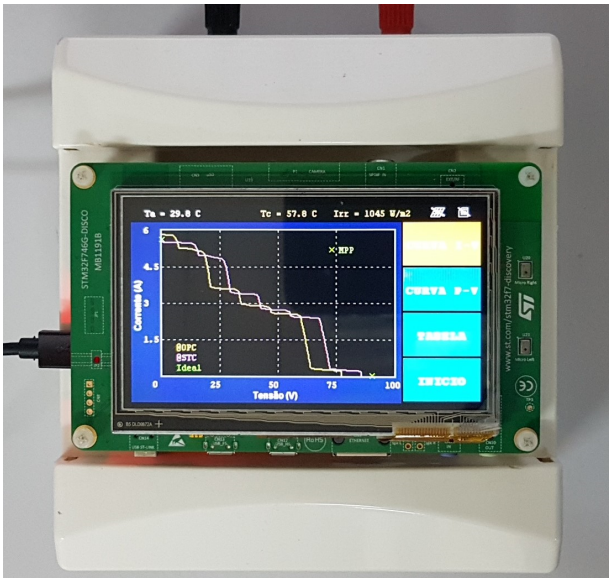


Figure 2.5 – I-V curve tracer developed at LESF (Laboratory of Energy and Photovoltaic Systems) at UNICAMP [55].

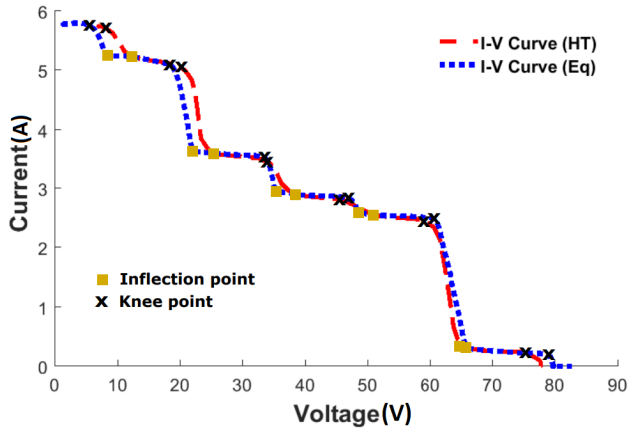


Figure 2.6 – Example of I-V curves for a string of PV modules, HT is the commercial tracer curve, and Eq the tracer curve developed [55].

2.3.3 Photovoltaic systems simulation software

The group responsible for the PV subproject is developing a software that simulates the aforementioned PV plants within the scope of research on solar irradiance and geometry modeling. This software uses the open-source tool pvlib-python [57] and will have an interface to present the simulation results of the installed PV systems. The main factors considered in this software are the solarimetric database it adopts, model of PV module and inverter. Shading is not taken into account in the developed software. The preliminary result has Figure 2.7 as an interface with the plans of the present project for simulation.

Figure 2.7 – Software interface with the project's pre-loaded PV plants.

2.3.4 Training courses

The PV subproject offered, since 2017, courses to disseminate the most relevant concepts of the PV universe, including academic and market approaches to the public. This initiative gathered around 3,000 students from all over Brazil, developing skills and enhancing the level of PV projects found in our country.

The first course, "Introduction to Solar Photovoltaic Energy - Isolated Systems and Connected to the Grid", aims to present fundamental concepts of PV solar energy, it shows the panorama of the Brazilian and world market, teach the basic dimensioning of PV systems, design of isolated systems and design of systems connected to the electric grid, present resolution 482 of ANEEL and other essential regulations of the PV sector.

The second course, "Design and Dimensioning of Solar Power Plants and Photovoltaic Systems with PVSyst" aims to teach how to design projects using the PVSyst software. With this software, it is possible to design PV systems connected to the grid, isolated and pumping. Two significant differentials are the bases of equipment cataloged by manufacturers of modules, inverters, optimizers, batteries, charge controllers, generators, pumps, and the 3D drawing tool. Thus, the user can obtain information about a

vast range of equipment on the market and make 3D simulations of where the PV installation will occur, for example, to measure the impact of shadows on the system's power generation.

The third course, "Installation and Integration of Photovoltaic Systems Connected to the Electric Grid" teaches all the steps of a complete installation. Students learn to climb on two types of roofs (ceramic and metallic), technical skill and safety measures to work at heights, handle tools, cables, and PV modules on top of the roof. Each group of students does all the electrical installation of modules, cables, protection devices, and inverters until the connection to the power grid. Finally, all groups commission their systems, conducting tests to ensure that the entire system can be safely connected to the power grid, and connect them. All students see their systems generating energy.

The courses are sought after by a wide range of people with different interests, from undergraduate and graduate students from various courses to engineers, architects, entrepreneurs, administrators, lawyers, landscapers, physicists, among other professions.

2.4 Methodology

The PV plants of UNICAMP were all simulated in PVsyst software to compare with the actual data and to check if the data are consistent or if there are problems with the installation. With the simulation, it is possible to predict the PV system's production, which makes it possible to take full advantage of a given PV plant from the design stage by positioning the modules in locations with less shading and choosing the best string arrangement. After the system is installed, it is possible to perceive possible problems if the system does not generate the expected amount of energy.

PVsyst software is dedicated to the study, design, and analysis of PV systems. PVsyst can simulate grid-connected, autonomous pumping systems and systems with direct current supply, using an extensive database of PV modules and inverters. It is a highly versatile software. Also, it presents a complete analysis of the system, including optimization of dimensioning, energy estimation over the years, losses due to near and far shadowing with 3D modeling, analysis of financial conditions, and return estimates. Allied to PVsyst, Meteonorm's 7.2 solarimetric base was used, which presents good results for the region [58]. It is worth noting that the simulation results can vary according to the parameters selected by the user, and the precision is not linked only to the software. The goal is to try to get a conservative forecast number for the project. In the simulations of this paper, the parameters and modeling were the PVsyst software standards, only choosing the components and performing the 3D design for shadow impact of each installation.

PVsyst is a commercial software, where the modeling used for the simulation is implicit. However, it is possible to mention some models used by the software. The models used in PVsyst were: the Erbs model [59] for irradiance decomposition, the Perez model [60] for the irradiance transposition, and the one diode model for the panel model described in [61].

After the PV plants were in operation and simulated, data for 12 months of power generation were collected. The data obtained are the values of energy injected into the electrical grid. As the plants were installed at different times, the data collected are not simultaneous. In other words, the 12 months were used, however, with a small difference between plants. The first plant to be activated was the GMU, followed by FEEC 2, both in April 2019. The last plant to be activated was the NIPE plant in October 2019. The PV plants are described in Tables 2.1 to 2.7.

For the analysis with simulation software, the important thing is that the analyzes are made with annual data (12 months) of power generation. The data were processed, and tables and graphical analyzes were produced for each PV plant, looking for problems and whether the expected energy was generated.

The uncertainty found in the literature for grid-connected PV systems must be less than ±30% monthly and less than ± 10% annually [62]. Generally, this uncertainty is linked to solarimetric data or errors in preparing the simulation. The modeling of PV modules and converters is already quite accurate [63].

Table 2.1 – Configuration of the PV system of the Unicamp's Multidisciplinary Gymnasium.

Equipment	**Model**	**Amount**
PV Modules	Canadian CS6K-270P	1248
Inverters	Ingecon Sun 55TL PRO	5

Table 2.2 – Configuration of the PV system of the School of Electrical and Computer Engineering (FEEC) 1.

Equipment	**Model**	**Amount**
PV Modules	BYD 320P6D-36	174
Inverters	Fronius Symo 15.0-3 208 PRO	3
Inverters	Fronius Primo 8.2-1 208 PRO	1

Table 2.3 – Configuration of the PV system of the School of Electrical and Computer Engineering (FEEC) 2.

Equipment	**Model**	**Amount**
PV Modules	BYD 320P6D-36	93
PV Modules	Canadian CS6K-270P	34
Inverters	Fronius Symo 15.0-3 208 PRO	2
Inverters	Fronius Primo 8.2-1 208 PRO	1
Inverters	Fronius Primo 4.0-1 208 PRO	1

Table 2.4 – Configuration of the PV system of the Interdisciplinary Energy Planning Center (NIPE).

Equipment	**Model**	**Amount**
PV Modules	Canadian CS6K-270P	144
Inverters	Fronius Symo 15.0-3 208 PRO	1
Inverters	Fronius Primo 8.2-1 208 PRO	2
Inverters	Fronius Primo 6.0-1 208 PRO	1

Table 2.5 – Configuration of the PV system of the Exploratory Science Museum.

Equipment	**Model**	**Amount**
PV Modules	Canadian CS6K-270P	15
Inverters	Fronius Primo 6.0-1	1

Table 2.6 – Configuration of the PV system of the Extecamp.

Equipment	**Model**	**Amount**
PV Modules	Canadian CS6K-270P	85
Inverters	Fronius Symo 12.0-3 208-240	2

Table 2.7 – Configuration of the PV system of the School of Civil Engineering (FEC).

Equipment	**Model**	**Amount**
PV Modules	Canadian CS6K-270P	140
Inverters	Fronius Symo 12.0-3 208-240	3

Eq. 2.1 was applied to analyze the error between the real and the simulated data, both for monthly and average annual errors. Thus, it was possible to analyze trends in the different PV plants' errors and non-standard data for possible corrections and studies. Positive errors indicate that the system generated more than expected in the simulations. Therefore, the simulation was conservative. Negative errors indicate that the energy generation predicted by the simulation was greater than the actual generation. Therefore, the simulation was optimistic. An optimistic simulation represents a problem for the consumer who has purchased or planned a PV system. Ideally, the actual system will generate the amount of energy promised, or a slightly larger amount. Therefore, PV software allows the use of probability indicators to make the result more conservative.

$$Error(\%) = \left(\frac{Energy_{Real} - Energy_{simulated}}{Energy_{Real}} \right) \cdot 100 \tag{2.1}$$

2.5 Results and discussion

2.5.1 Technical evaluation

The first installation analyzed and compared with the data for the generation of energy injected into the power grid was the Unicamp Multisport Gymnasium. This installation is the one with the highest installed PV power. Table 2.8 shows the comparison of simulated data with reals in one year of data.

The data presented showed a higher average annual energy generation for the actual data compared to the simulation. This type of result is interesting, as it shows that the PV system produced more than expected. As well, the PV system had an error within the expected by the literature. Thus, it can be said that PVsyst was more conservative in this installation and that the operation took place as expected for the first year.

Table 2.8 – Comparison of power generation for Unicamp's Multisport Gymnasium with simulated data in PVsyst

Month	**Energy (MWh)**		**Error (%)**
	Real	**PVsyst**	
January	48.62	42.60	12.38
February	38.92	41.98	-7.86
March	50.83	41.26	18.83
April	42.02	36.88	12.23
May	32.98	34.70	-5.22
June	33.59	31.28	6.88
July	35.74	35.87	-0.36
August	38.32	40.21	-4.93
September	41.76	40.41	3.23
October	52.67	42.98	18.40
November	48.70	47.94	1.56
December	49.47	45.05	8.93
Average	**42.80**	**40.10**	**6.32**
Total	**513.62**	**481.16**	**-**

Table 2.9 shows the result for the set of PV plants of FEEC 1. In this installation, a problem occurred during the first year of data collection, which was the circuit breaker trip. In February, for example, the circuit breaker was tripped several days, as it was a vacation month and the problem was not corrected at early stage. Thus, the real generation was smaller than the simulated generation. However, the result was still close. Without the problem, energy generation would probably be closer.

Table 2.9 – Comparison of power generation for FEEC 1 with simulated data in PVsyst

Month	**Energy (MWh)**		**Error (%)**
	Real	**PVsyst**	
January	6.80	7.36	-8.22
February	5.28	7.16	-35.53
March	7.55	6.91	8.52
April	6.54	6.05	7.49
May	5.20	5.59	-7.44
June	4.30	4.98	-15.81
July	5.48	5.71	-4.11
August	5.95	6.54	-9.88
September	6.45	6.72	-4.16
October	8.17	7.30	10.71
November	6.73	8.23	-22.23
December	6.82	7.77	-13.93
Average	**6.27**	**6.69**	**-6.67**
Total	**75.27**	**80.29**	**-**

Table 2.10 shows the result for the set of PV plants from FEEC 2. One detail is that FEEC 2 is composed of a set of PV plants that are used to collect results and tests from a microgrid and converters. In this case, some plants are constantly disconnected for short periods. However, sometimes there is a long disconnection period, an example is the month of June. Thus, as in FEEC 1, the actual generation was less than the simulated generation.

Table 2.10 – Comparison of power generation for FEEC 2 with simulated data in PVsyst

Month	**Energy (MWh)**		**Error (%)**
	Real	**PVsyst**	
January	4.68	4.83	-3.12
February	3.77	4.79	-27.14
March	5.18	4.79	7.49
April	4.44	4.35	2.09
May	3.84	4.12	-7.37
June	2.40	3.75	-56.13
July	3.73	4.30	-15.20
August	3.92	4.75	-21.15
September	4.25	4.72	-11.04
October	5.38	4.95	8.01
November	4.89	5.42	-10.80
December	4.93	5.10	-3.35
Average	**4.28**	**4.65**	**-8.65**
Total	**51.41**	**55.86**	**-**

Table 2.11 presents the result for NIPE. In this installation, the real average generation was much higher than the simulated, with an error in a few months greater than that indicated in the literature for software simulation. The justification may be that the installation has a scenario that can cause shadows, and perhaps the data added in the software did not faithfully replicate the shadow profile. Thus, the losses in the simulation were more conservative compared to the actual data. This can occur in real installations, and it is interesting and important that the real PV plant has a higher generation than the simulated one, observing the customer's perspective and values promised by the integrator.

Table 2.11 – Comparison of power generation for NIPE with simulated data in PVsyst

Month	**Energy (MWh)**		**Error (%)**
	Real	**PVsyst**	
January	5.14	4.92	4.30
February	4.20	4.66	-11.03
March	5.94	4.31	27.44
April	5.32	3.54	33.54
May	5.00	3.10	37.96
June	4.04	2.68	33.78
July	5.05	3.01	40.31
August	4.86	3.71	23.63
September	5.03	4.11	18.38
October	5.97	4.70	21.26
November	5.10	5.46	-7.00
December	5.13	5.22	-1.83
Average	**5.06**	**4.12**	**18.69**
Total	**60.78**	**49.42**	**-**

Table 2.12 shows data for the Unicamp exploratory museum. In this installation two months (April and May) the installation was shutdown during the period that the exploratory museum at Unicamp was closed. The actual data was smaller than the simulated ones, due to these months without operation. However, the behavior of the other months was adequate, with the exception of March, which probably also stayed disconnected for a few days.

Table 2.12 – Comparison of power generation for the Unicamp exploratory Museum with simulated data in PVsyst

Month	**Energy (MWh)**		**Error (%)**
	Real	**PVsyst**	
January	0.60	0.51	14.86
February	0.49	0.51	-4.95
March	0.35	0.51	-44.65
April	0.00	0.47	-
May	0.15	0.46	-
June	0.40	0.42	-4.28
July	0.51	0.48	6.01
August	0.52	0.52	-0.87
September	0.55	0.51	8.27
October	0.68	0.53	22.48
November	0.60	0.58	3.68
December	0.61	0.54	11.59
Average	**0.45**	**0.50**	**-10.39**
Total	**5.46**	**6.03**	**-**

For the installation of Extecamp, the results can be seen in Table 2.13. The result for Extecamp, considering the annual average, was close to the simulated and within the expected according to the literature indicators for PV simulation error. The estimated generation in PVsyst ended up being slightly higher than the real one. This result is normal and can occur, as there are variations in the actual irradiance data for that of the meteonorm used in the simulation.

Table 2.13 – Annual results for PV power plant of the School of the Extecamp

Month	**Energy (MWh)**		**Error (%)**
	Real	**PVsyst**	
January	3.03	2.97	2.13
February	2.39	2.89	-20.72
March	3.04	2.78	8.56
April	2.60	2.43	6.64
May	2.18	2.25	-3.18
June	1.84	2.01	-9.07
July	2.21	2.31	-4.62
August	2.17	2.62	-20.87
September	2.59	2.72	-5.10
October	3.30	2.94	10.80
November	3.03	3.32	-9.42
December	3.15	3.14	0.32
Average	**2.63**	**2.70**	**-2.67**
Total	**31.54**	**32.38**	**-**

The last PV plant analyzed was the FEC plant shown in Table 2.14. It can be seen with the average annual value that the real generation was -16.67% less than the simulated one. When realizing this result with the analysis, possible problems in the installation were studied. The diagnosis was that one of the inverters does not reach maximum power for much of the day, which could be a problem with the MPPT or inverter control. Thus, the importance of simulating, comparing results, and monitoring the system is perceived.

Table 2.14 – Annual results for PV power plant of the School of the FEC

Month	**Energy (MWh)**		**Error (%)**
	Real	**PVsyst**	
January	3.67	4.79	-30.57
February	3.03	4.70	-54.91
March	4.32	4.61	-6.81
April	3.99	4.13	-3.58
May	3.64	3.90	-7.18
June	2.86	3.53	-23.56
July	4.50	4.04	10.33
August	4.42	4.51	-1.95
September	3.75	4.52	-20.26
October	4.38	4.81	-9.88
November	3.92	5.35	-36.63
December	3.74	5.04	-34.70
Average	**3.85**	**4.49**	**-16.67**
Total	**46.21**	**53.92**	**-**

Subsequently, all tables were joined in a single one to compare the performance of all installations together and to analyze the behavior of the irradiance base, checking if there is any standard behavior between the simulations. Table 2.15 shows the result. The annual average of real energy generation was 3.22% higher than that simulated. Thus, the PV generation of the campus generated more than estimated, despite the problems that can be corrected, such as installing the FEC or the inevitable, such as the FEEC 2, which is a study facility and will undergo shutdowns during its useful life.

Table 2.15 – Comparison of power generation with all PV plants together with simulated data in PVsyst.

Month	**Energy (MWh)**		**Error (%)**
	Real	**PVsyst**	
January	72.55	67.98	6.30
February	58.09	66.69	-14.82
March	77.21	65.17	15.59
April	64.91	57.85	10.89
May	52.99	54.12	-2.13
June	49.43	48.64	1.60
July	57.22	55.71	2.63
August	60.15	62.85	-4.49
September	64.38	63.69	1.07
October	80.54	68.20	15.32
November	72.98	76.30	-4.55
December	73.85	71.85	2.70
Average	**65.36**	**63.25**	**3.22**
Total	**784.29**	**759.04**	**-**

Figure 2.8 presents each PV plant's errors to study the upward or downward trend of the set of errors calculated from the measured and simulated data. The result shows that the pattern of increasing or reducing monthly errors is similar for most PV plants.

The most divergent curve is that of the Nipe plant, where the error is positive most of the time, that is, the real energy generation is greater than expected from the simulations. The justification may be the fact that the installation has a scenario that can cause shading and, also, has high slopes. A possible explanation for this is the addition of factors that did not faithfully replicate the shadow profile in the simulation, generating errors concerning the real data. This situation can easily occur in installations, so it is interesting and important that the installed PV system has more energy generation than expected in the simulations. For the PV integrator to meet the customer's expectations and demands.

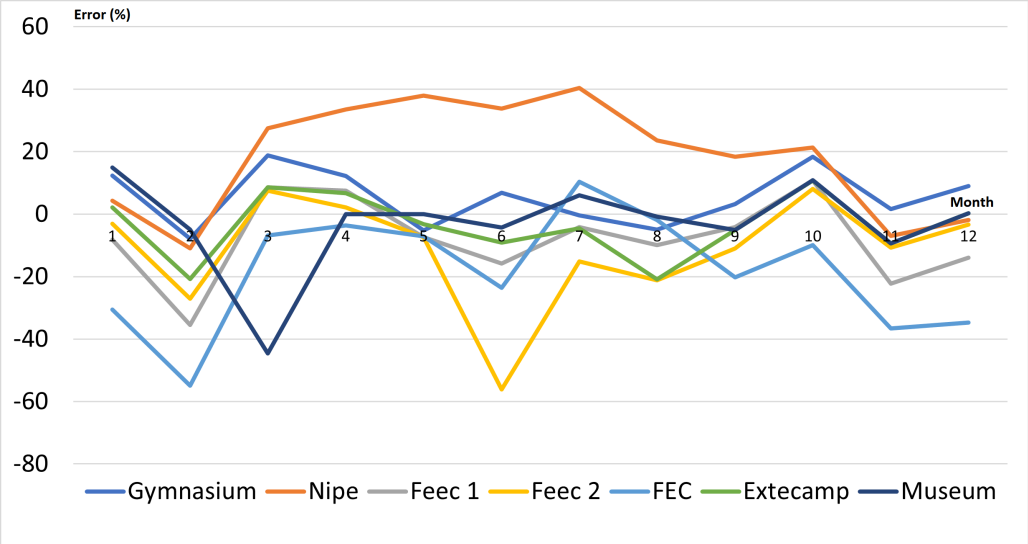


Figure 2.8 – Error up/down trend (%) for all PV plants.

Figure 2.9 presents a boxplot to graphically represent the pattern of errors, allowing to visualize the dispersion, symmetry, and discrepant data of a data set. Figure 2.9 shows that the Museum plan has less variability and a median close to zero, while the Unicamp's Multisport Gymnasium and Extecamp plants have similar variability. However, the Extecamp boxplot is more asymmetrical since the median is closer to the quartile 25%. Two outliers, which are nonstandard errors in the data, occurred. One at FEEC 2, possibly due to the plant's shutdown for studies carried out in a laboratory; and another at the Museum, due to the shutdown of the PV system by employees during a period when the museum was closed.

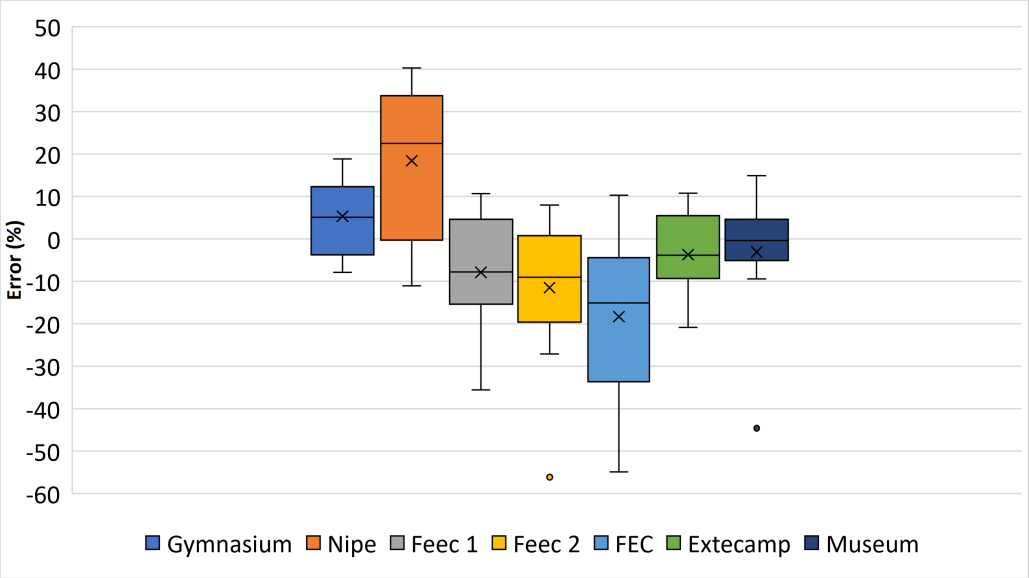


Figure 2.9 – Error up/down trend (%) for all PV plants.

Figure 2.10 shows measured data of energy injected into the electrical grid by the PV plants. It is observed that, in the first year of operation of UNICAMP's PV plants, the lowest generation of energy occurred in June, while the highest generation of energy occurred in October. A correlation of the data in Figure 2.10 with Figure 2.11 is shown, which shows the irradiance data in the PVsyst database. However, the differences are normal since the simulation follows the database, and the actual data may vary in certain months. Also, there are slight variations in the irradiance base due to the location used for each installation when generating the data on the meteonorm, or versions of meteonorm.

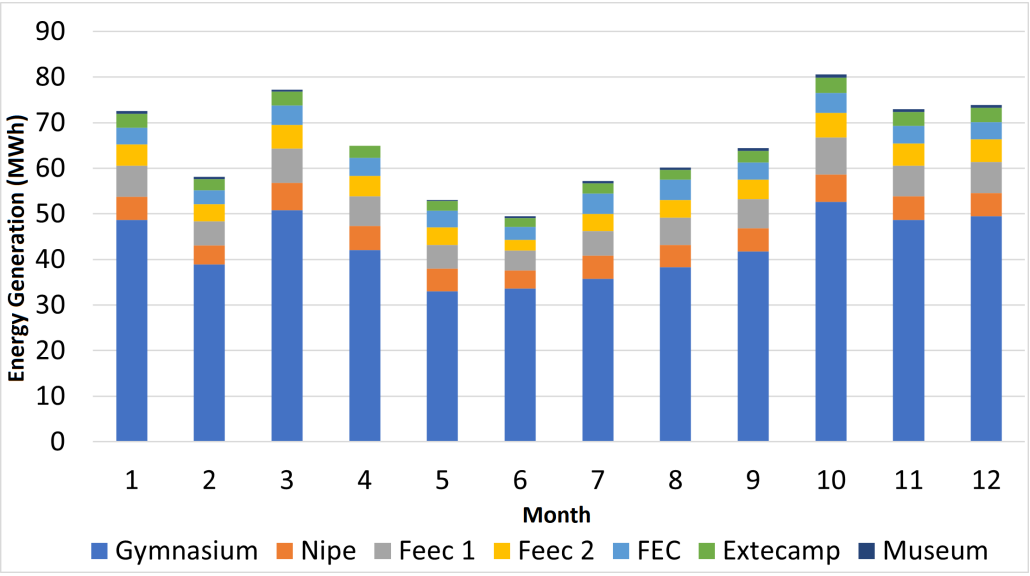


Figure 2.10 – Generation of UNICAMP PV plants in MWh.

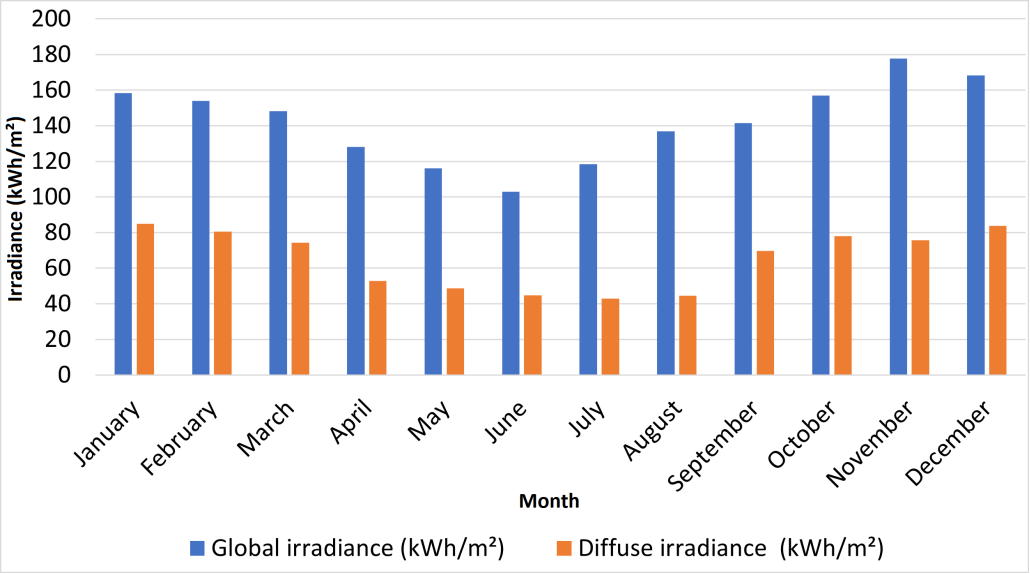


Figure 2.11 – Meteonorm irradiance data in PVsyst software in the region of Campinas-Brazil.

The real and simulated capacity factor (CF) was also calculated and plotted in Figure 2.12. The equation used was the 2.2 [64].It can be seen that the CF was in a close range, with the exception of the UMG, as it is an installation without shading by objects or buildings. The result shows that in total the CF was higher for the real data. This fact is due to the UMG that exerts greater weight, as it is a larger installation.

$$CF = \frac{Net\,AC\,energy\,(Wh)}{Maximum\,output\,power\,(W) \cdot 8760\,h} \tag{2.2}$$

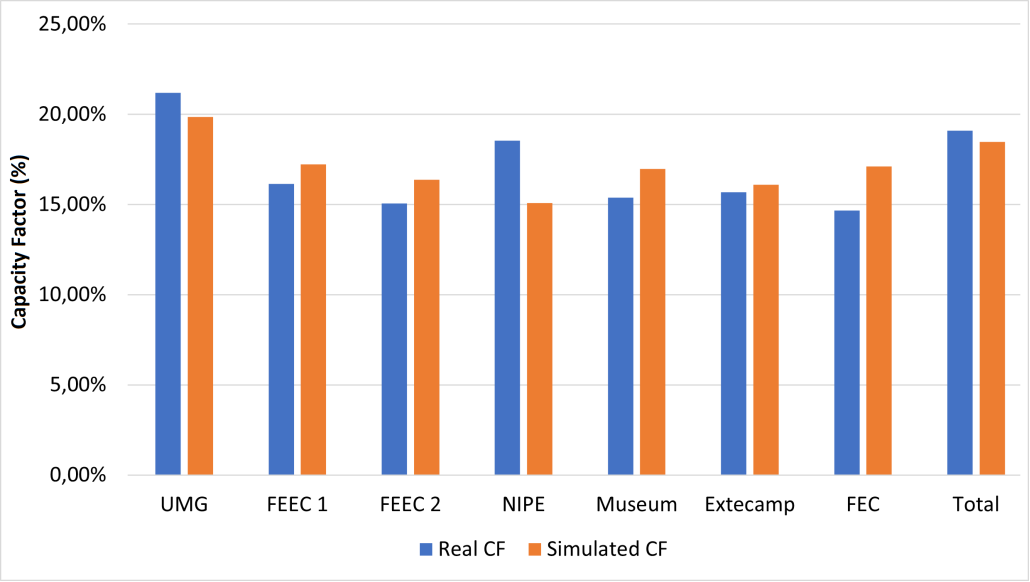


Figure 2.12 – Real and simulated capacity factor of PV installations.

To verify the impact of PV production on the university's energy consumption,

the annual consumption of 2019 was used and compared with the first twelve months of the PV plants. The result was presented in table 2.16, verifying that all PV plants produce about 1.13 % of the energy. The year 2019 was chosen since later, other energy bills were impacted by the quarantine during COVID-19, showing a large reduction in consumption. To meet UNICAMP's consumption, more than 40 MWp of PV power would be needed due to the size of the campus, also called the university city. This would pass from Brazilian resolution to distributed generation. Thus, the energy consumed began to be purchased in the free contracting environment model, originated by a renewable source. In Figure 2.13, the difference between consumption and energy generated by PV plants can be seen in a graph.

Table 2.16 – Comparison of energy consumed by UNICAMP in relation to the amount of energy converted by PV plants.

Months	Energy Consumption 2019 (MWh)	First months of PV energy generation (MWh)	Energy generated (%)
1	6,738.97	72.55	1.08
2	5,768.16	58.09	1.01
3	6,189.12	77.21	1.25
4	6,264.38	64.91	1.04
5	5,798.96	52.99	0.91
6	4,847.49	49.43	1.02
7	4,699.87	57.22	1.22
8	5,149.54	60.15	1.17
9	5,841.89	64.38	1.10
10	6,707.65	80.54	1.20
11	6,128.38	72.98	1.19
12	5,462.49	73.85	1.35
Total	**69,596.9**	**784.29**	**1.13**

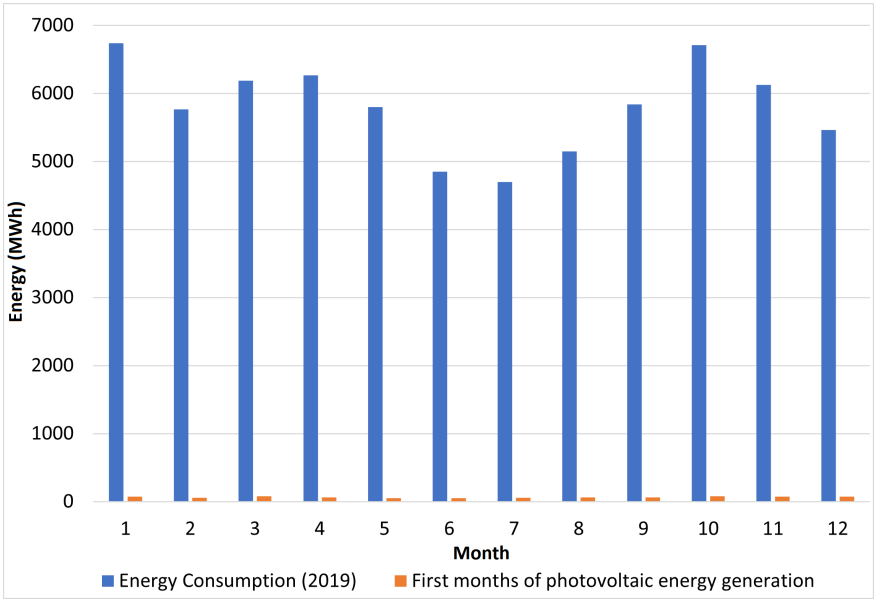


Figure 2.13 – UNICAMP's energy consumption in 2019 and energy converted by PV plants in the first months in MWh.

2.5.2 Economic evaluation

The UNICAMP is a consumer of the free energy market in Brazil. The Free energy market is a type of modality in which the consumer buys energy from other plants through contracts. Therefore, the cost of energy is lower than that of the local distributor. In the free energy market, the energy produced by the PV system is consumed by the load instantly or injected into the grid, in the second case, without financial return.

To show the economic evaluation of PV systems in this type of market, several figures of economic merit are presented in this section: Payback, Net Present Value (NPV), Internal Rate of Return (IRR), and Levelized Cost of Energy (LCOE).

Payback is the time to return the investment. NPV is the figure of merit that represents the present value of future payments discounted based on an interest rate (discount rate). The equation 2.3 presents the NPV calculation, where, C_t total costs of the solar plant during its useful life, d the discount rate, t number of years, and I is the initial investment. The IRR is the rate equivalent to the discount rate that a cash flow must have for the NPV to be equal to zero.

$$NPV = \left(\sum_{t=1}^{T} C_t / \left(1+d\right)^t \right) - I \tag{2.3}$$

Finally, LCOE, is the leveled value that costs the electrical energy produced by the PV system over its lifetime, calculated by the equation 2.4 [65]. For these calculations, the parameters in the Table 2.17 were used. The currency amount in *real* (R$) may be converted into dollars by applying a conversion rate of approximately R$ 5 per dollar. As there are taxes, and the direct conversion sometimes does not represent the correct value, the use in *real* (R$) currency was chosen for the paper.

$$LCOE = \frac{\sum_{T}^{t=0} C_t / \left(1+d\right)^t}{\sum_{T}^{t=0} E_t / \left(1+d\right)^t} \tag{2.4}$$

Table 2.17 – Parameters for economic evaluation

Parameter	**Value**
System Cost	R$/Wp 4.64
O&M annual	1% of the system cost
Energy inflation per year	8%
Annual Degradation Losses (PID)	0,5%
Discount rate	10 %

As a result, the values in the Table 2.18 were obtained. The financial return was higher than predicted in the simulations based on the first year of recorded data. The LCOE was presented considering the NPV and without using the discount rate.

Table 2.18 – Result of the economic evaluation

Figures of Merit	Simulation	Real Data
Payback	7.65 Years	7.44 Years
NPV	R$ 2,000,339.05	R$ 2,158,264.23
IRR	16.33 %	16.77 %
LCOE-NPV	R$ 0,414/kWh	R$ 0,4/kWh
LCOE	R$ 0,174/kWh	R$ 0,168/kWh

The payback based on data from the first measured year plus degradation resulted in 7.44 years, and simulation 7.65 years. In Figure 2.14 it is possible to verify the cash flow with this return period. The payback increased considerably because it is in the free energy market, unlike what it would be for consumers who buy energy from the distributor, and it has a higher energy rate, resulting in lower paybacks.

However, consumers who are in the free energy market have the benefit of acquiring energy from renewable sources at a low cost, being an option instead of using PV systems. Or make use of both (like the UNICAMP), even with a higher payback than it would be for a consumer of the local distributor.

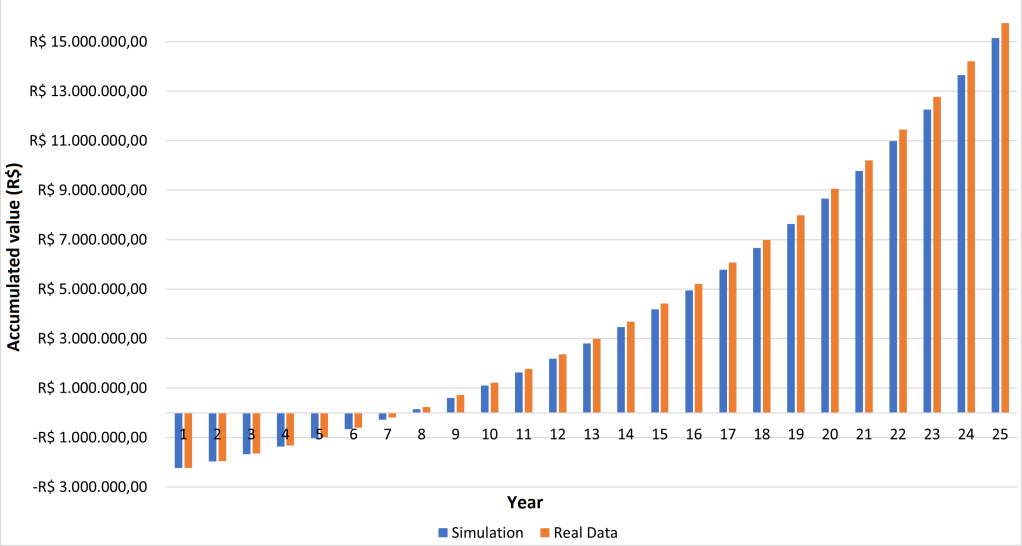


Figure 2.14 – Accumulated Cash Flow for both scenarios.

2.6 Conclusions and future works

This paper presented studies and information relevant to the insertion of PV solar energy at UNICAMP. For a year of data collected, a real generation of 784.29 MWh was observed, compared to simulation results of 759.04 MWh for the same system. This shows that the PV plants are generating in the expected value range, even with FEEC 2 system being turned off for studies in certain periods and some problems that occurred

during the 12 months that were commented above. With payback of less than 8 years and LCOE of R$ 0.4/kWh based on the first year of data.

The impact of current PV plants is 1.13 % of the consumption obtained in the year 2019. The result is a consequence of the university's high consumption, which would need more than 40 MWp of PV plant. However, as it is a consumer of the free energy market and has demand at night, it would not be possible to maintain itself with PV energy alone, without a storage system, which is not feasible at the moment.

The insertion of other PV plants in UNICAMP continues with an electro-post for electric buses with 19 kWp, Microgrids with 607 kWp, an energy efficiency program (PEE) in hospitals with 1,095 kWp and a 6 kWp PV kiosk. It is expected that, in the upcoming two years, the total power of PV systems at UNICAMP will be approximately 2.2 MWp. These new PV systems are from other projects linked to the "sustainable campus". In addition, new projects are underway to scale sustainable campus initiatives to other public buildings in São Paulo, Brazil.

With the studies carried out by the group, we hope to encourage the construction of other PV plants within UNICAMP, as we have a current success story and encourage other universities to use renewable energy sources, contributing to sustainable development and more studies for the sector. The project also served to study new PV module technologies; modeling and creating a tool for simulation of campus PV systems; creation of a low-cost PV curves tracer prototype to assist in the commissioning of small PV installers; and dissemination of knowledge through training and workshops. All of this content was of paramount importance for the success of the Sustainable Campus project and exchanges of experience with the market and academic community.

For future work related to PV plants, it is intended to analyze the mismatch of different types of PV modules in different PV plants; check the degradation of the different PV modules; evaluate the reliability of PV inverters; and impact of dirt in the region on energy conversion by PV plants.

Supplementary material

Update of Sustainable Campus project initiatives

The *Sustainable Campus* project remained active after the publication of the article presenting the results of its first cycle. By 2024, there are 7 projects within the Sustainable Campus with companies and/or government entities for improvements and studies on the Unicamp campus, with approximately 45 partners. Over 6 years, more than R$ 50 million in external funding and R$ 40 million from Unicamp's own resources have been secured for project initiatives. The result is an annual savings of over R$ 15 million. Through the project, 34 courses have been developed based on research findings, benefiting 932 students as of 2024.

Regarding energy efficiency, the project has replaced over 30,000 bulbs with LEDs and is currently replacing over 100,000 more internally, along with 2,615 street lights on campus. Additionally, more than 2,000 air conditioners have been upgraded to more efficient models. The PV systems have also been expanded; by 2024, there is an installed capacity of 2.5 MWp, with plans to expand to 5 MWp with new systems to be installed.

Within the *Sustainable Campus* project, the *CPTEn (São Paulo Center for Energy Transition Studies)* was established with support from FAPESP, aimed at experimenting with new energy transition solutions and technologies. The project now involves about 209 researchers and has 8 sources of funding to support students. Project partners include UNESP, UTFPR, USP, UFG, Mackenzie, PUC Minas, IFSP, PUC-Campinas, Uninove, CTI Renato Archer, TU Delft, and LUT University. The project has already published 221 articles.

Month of the first data obtained from real generation

Table 2.19 – Month of the first data obtained from real generation

PV Installation	**First data**
Gymnasium	May/2019
FEEC I	August/2019
FEEC II	September/2019
NIPE	October/2019
Museum	September/2019
Extecamp	September/2019
FEC	September/2019

Note on nomenclature

In scientific literature, the term "Error" is commonly used to compare real data with simulated data, as is done in this thesis. However, some authors prefer to reserve the term "Error" exclusively for the comparison between two real datasets or two simulations (with one serving as a reference). In such cases, the term "Discrepancy" may be more appropriate.

3 Evaluating the Significance of Solarimetric Data for Photovoltaic System Simulation in a Real-World Case

Paper published in COBEP [66] ©2023 IEEE 8th Southern Power Electronics Conference and 17th Brazilian Power Electronics Conference (SPEC/COBEP). Reprinted, with permission, from de Silva, J.L. S. Evaluating the Significance of Solarimetric Data for Photovoltaic System Simulation in a Real-World Case.

This manuscript, authored by João Lucas de Souza Silva, João Antonio Fernandes Gonçalves da Silva, João Frederico Souza de Paula, Eslam Mahmoudi, Tárcio André dos Santos Barros and Marcelo Gradella Villalva. The paper is associated with the Digital Object Identifier (DOI): <10.1109/SPEC56436.2023.10407437>. This work was developed with TotalEnergies financial support. In addition, we are grateful to all collaborators from University of Campinas (UNICAMP). We acknowledge the support of ANP (Brazilian National Oil, Natural Gas and Biofuels Agency) through the R&D levy regulation. Acknowledgements are extended to the Center for Energy and Petroleum Studies (CEPETRO) and School of Electrical and Computer Engineering (FEEC). The authors would like to thank the Campus Sustainable Project Unicamp for the data. Furthermore, the authors extend their gratitude to the late Professor Marcelo Villalva (1978-2023), foremost figures in the global solar energy scenario, for providing us with all the knowledge and opportunities during our professional engagement.

Abstract: Solarimetric stations in photovoltaic (PV) systems serve an essential role in identifying issues and evaluating performance by providing data for simulations. This study aims to illustrate the value of employing solarimetric station data through a case study of a PV installation. The methodology involved using solarimetric station data, comprising global, direct normal, and diffuse irradiance, to generate Plane of Array (POA) irradiance data. The POA data were input into the PVsyst software to simulate and compare with actual energy generation. A significantly lower discrepancy of 0.21% was discovered between the real and simulated data using POA, versus a 3.13% discrepancy with meteonorm data in PVsyst. This outcome emphasizes the enhancement in simulation precision when integrating real-world POA data, highlighting the necessity to leverage authentic irradiance data for accurate PV system simulations. Additionally, statistical analyses were conducted to comprehend the relation between POA and the generation data.

Keywords: Solar Power Plant, Monitoring, Solarimetric Data.

3.1 Introduction

A solarimetric station or solar weather station, is a set of instruments for measuring various aspects/characteristics of sunlight intensity and weather conditions [67, 68]. It typically consists of devices, such as pyranometers to measure global solar radiation, pyrheliometers for direct solar radiation, and sensors for other meteorological parameters like temperature, wind speed, and humidity. Some stations also include devices to measure diffuse solar radiation, albedometer [69], and sunlight incidence angle, mainly in studies of photovoltaic (PV) plants.

These stations provide comprehensive, accurate, and real-time solar irradiance and meteorological data, which are integral for precise planning, design, and performance assessment of PV power systems. For example, they can be used to perform simulations in PV software more accurately, since part of the error comes from the solarimetric bases of satellites. Simulation outcomes may vary with data used [70].

In this context, this paper presents a case study analyzing the energy generation of a PV system using data from a solarimetric station. A comprehensive dataset was collected from both the solarimetric station and PV inverters at the same location over the same year. The data were then processed to generate Plane of Array (POA) irradiance [71] information and used in PVsyst simulations to validate against real-world outcomes. This process allowed for an assessment of the similarity between the expected generation, obtained through simulations, and the real production when using local data in PVsyst.

The results were compared with previous simulation data created for the design of the PV plant using PVsyst and Meteonorm. Additionally, a statistical evaluation was conducted to understand the relationship between the POA data and PV power, adding further depth to the analysis. The scientific and technical contributions were as follows:

- Integration of Real-World Data: This work incorporates real-world data from a solarimetric station and PV inverters, emphasizing the importance of using actual data in the analysis and simulation of PV systems.
- Validation of Simulation Models: The study validates the accuracy of simulations performed in PVsyst using data generated by the conversion model for the POA Irradiance. This contributes to the reliability of simulation tools in predicting real-world outcomes.

- Comparison with Existing Simulation Data: The paper compares the results with previous simulation data created using PVsyst and Meteonorm. This provides insights into the consistency and accuracy of different simulation methods, guiding best practices for PV system design.
- Statistical Analysis of Irradiance and PV Power: It includes statistical evaluation to understand the relationship between POA data and PV power production. This contributes to a scientific understanding of the factors influencing PV system performance in real-world scenarios.

3.2 Solarimetric Data

3.2.1 Components of solar irradiance

The Fig. 1 showcases the components of solar irradiance. The segment of irradiance that hits the Earth's surface along the line from the observer to the center of the sun, untouched by external factors such as dust, gasses, clouds, or other particles, is referred to as Direct Normal Irradiance (DNI) [72]. There is also a portion of irradiance that journeys through the atmosphere, undergoing scattering events, for instance by a cloud, which is termed Diffuse Horizontal Irradiance (DHI). The amalgamation of these two, the direct and diffuse horizontal irradiance, results in what is known as the global horizontal irradiance [73].

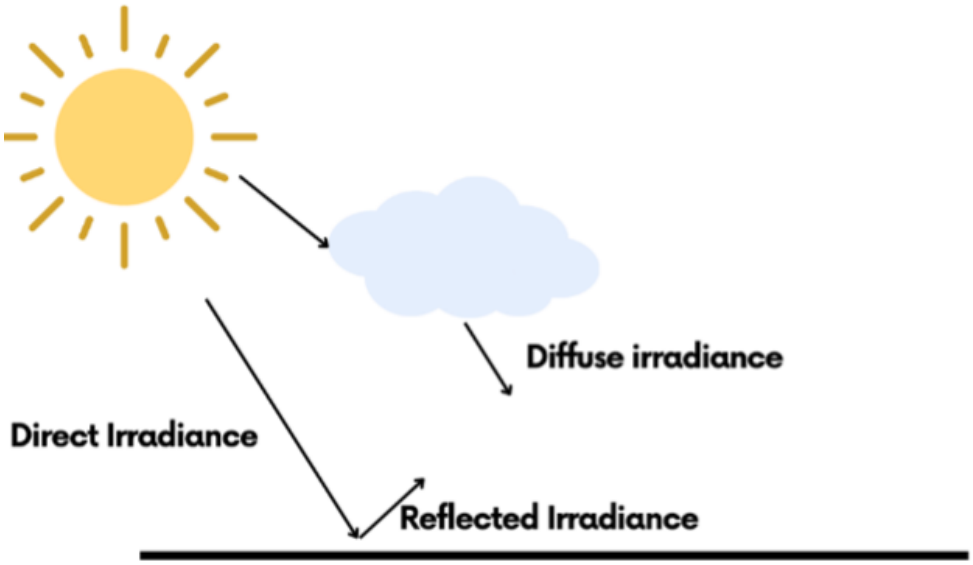


Figure 3.1 – Components of solar irradiance.

Knowing this, the global horizontal irradiance can be defined as the aggregate of solar energy that reaches the Earth's surface, as expressed in Eq. (3.1). Knowledge of the

global (GHI), direct (DNI) and diffuse irradiance (DHI) spectrum incident on the earth's surface is crucial to understand and analyze the power generation of PV systems [73].

$$GHI = DNI \cdot cos\,(\theta) + DHI \tag{3.1}$$

3.2.2 Plane of Array (POA) irradiance data

In the context of PV systems, it's crucial to convert irradiance to the POA, which corresponds to the inclined plane of the PV module. This is because the system is usually tilted at a certain angle to optimize the usage of irradiance, particularly when considering the annual average for fixed systems. Transposition models ascertain the total POA irradiance by computing the individual contributions from direct, ground-reflected diffuse, and sky diffuse components incident on the POA [74]. One of the classic examples for POA irradiance modeling is the Perez model [75].

3.2.3 Solarimetric Station

The solarimetric station used in the tests is shown in the Fig. 3.2. This station provided the capability for meticulous data collection. Since 2020, it has been possible to collect a full year's worth of data on PV energy generation as well as solarimetric information.

A solarimetric station is composed by sensors that measure solarimetric and meteorological parameters of the environment, as shown in the Table I. This data provides important information regarding the performance assessment of PV power plants. The Unicamp Solarimetric Station has two pyranometers, that measure GHI and DHI (measured with a shading plate that follows the sun track and block the DNI), and one pyrheliometer (provides the DNI component from GHI).

Table 3.1 – Solarimetric station sensors and measured environmental factors

Sensors	**Measured Parameters**	**Unit**
Pyranometer	**Global Solar Irradiance**	**W/m^2**
Thermohygrometer	**Ambient Air Temperature and Humidity**	**°C and %**
Anemometer	**Wind speed and direction**	**m/s and**
Pluviometer	**Rainfall**	**cm**
Albedometer	**Albedo**	**Dimensionless**

The IEC 61724-1 (Photovoltaic System Performance Part 1: Monitoring) [76] is the standard that provides informations about which parameters are needed and how to measure them in a solarimetric station and throughout the PV power plant. This standard classifies the PV monitoring systems between two classes: Class A and B. Both of them

Figure 3.2 – Unicamp Solarimetric Station.

measure irradiance, environmental factors and electrical output data, the difference being the types of parameters that need to be measured, the samples and records interval, and the sensors classifications.

The main sensor used in a solarimetric station is the pyranometer, which measures the global solar irradiance on different situations, depending on tilt and position in the field. The pyranometer has three classifications, according to the ISO 9060:2018 [77]: Class A, B and C, going from highest to lowest. The pyranometer used on this study has a Class B classification.

3.2.4 POA irradiance modeling according to the PV Software

In this work, a pyranometer wasn't installed in the same tilt as the PV modules, because the solarimetric station wasn't set up for this PV study. Thus, the POA parameter was calculated, where the horizontal values from irradiance components collected on the solarimetric station were applied. To find the POA value it's necessary to sum the three components of irradiance that reach the PV module's surface [78]: incident beam irradiance (Ib), incident sky diffuse irradiance (Id), and incident ground-reflected

irradiance (Ir), as can be seen in Eq. (3.2).

$$POA = Ib + Id + Ir \tag{3.2}$$

The software calculated each one of the irradiance components in Eq. (3.2), according to physical modules, to provide the POA value. To find the Ib, the Eq. (3.3) was used to transpose the DNI from horizontal plane to PV module surface [78].

$$Ib = Eb \cdot cos\,(AOI) \tag{3.3}$$

AOI is the "Angle of Incidence", which is the sun incidence angle defined as the angle between beam irradiance and the normal line considering the subarray surface. Eb is the Direct Normal Irradiance (W/m^2) [78]. For the Id value, the software has three options of physical models, but only one can be choosen to make the estimative. For this study, the equation defined to calculate the incident sky diffuse irradiance was from the Perez 1990 model [75], as it can be seen in Eq. (3.4).

$$Id = Di + Dc + Dh \tag{3.4}$$

The Perez model is a more complex computacional method then the other two (Isotropic and HDKR), due to the fact that it considers the isotropic (Di), circumsolar (Dc) and horizon brightening (Dh) components of incident diffuse irradiance [78]. Furthermore, the SAM's Perez sky diffuse irradiance model has a modification in the implementation that treats diffuse irradiance as isotropic for 87,5° <= Z <= 90°, as Z is the solar zenith angle [78]. Also, the Perez model differs from the isotropic and HDKR in the use of empirical coefficients present on the table from [78], which is derived from measurements over a range of sky conditions and locations instead of mathematical representations of the sky diffuse components. Each component from the diffuse irradiance on the Perez model has your modelling showed in [78].

Finally, to provide the Ir, the software applied Eq. (3.5). The equation is a function of the beam normal irradiance and sun zenith angle, sky diffuse irradiance, and albedo (ground reflectance) [79]:

$$Ir = \rho \cdot (Eb \cdot cosZ + Ed) \cdot \left(\frac{1 - cos\beta}{2}\right) \tag{3.5}$$

According to Eq. (3.5), the Albedo (ρ) is considered, being the reflectance property of the material, which makes up the surface through which light is reflected and

reaches the module's surface, E_d is the diffuse irradiance and β is the subarray surface's tilt. For more details about albedo selection, the [78] report provides these information.

3.3 Simulation Software: PVsyst

To effectively design PV systems and gather comprehensive data for future analyses, simulations are highly recommended. One prevalent software in the industry that aids in such simulations is PVsyst [80]. This software offers the ability to specify equipment, examine potential shading effects, assess various losses, and perform other crucial studies. To facilitate these simulations, PVsyst incorporates a solarimetric database, usually Meteonorm. However, to enhance the accuracy of these simulations, incorporating site-specific real data, when available, is considered beneficial and advantageous.

3.4 Methodology

Firstly, we collected data from the solarimetric station for the year 2020, which included global, direct normal, and diffuse irradiance measurements. The data and installation are part of the Unicamp Sustainable Campus project - *"Projeto Campus Sustentável"* (see Fig. 3.3). The collected data was then converted into POA irradiance data format.

Table 3.2 – Comparison between real and simulated data

Month	Real (MWh)	PVsyst Projected (MWh)	PVsyst with POA (MWh)	Error - Real x Projected	Error - Real x POA
January	48.61	42.60	44.92	12.36%	7.59%
February	38.92	41.98	33.47	-7.86%	14.02%
March	50.85	41.26	48.21	18.86%	5.20%
April	42.02	36.88	39.39	12.23%	6.26%
May	35.00	34.70	39.08	0.86%	-11.66%
June	29.28	31.28	33.08	-6.83%	-12.97%
July	34.38	35.87	38.29	-4.33%	-11.37%
August	35.50	40.21	39.08	-13.27%	-10.08%
September	40.59	40.41	44.28	0.44%	-9.08%
October	42.55	42.98	44.45	-1.01%	-4.46%
November	52.49	47.94	44.30	8.67%	15.61%
December	46.53	45.05	47.14	3.18%	-1.32%
Total	**496.72**	**481.16**	**495.67**	**3.13%**	**0.21%**

In later stage, the converted POA data was imported into PVsyst for simulation purposes. This data was compared with the original project's simulation data, which had been modeled using Meteonorm data. The primary aim of this comparison was to analyze and quantify the discrepancies between the simulated and actual energy production of the PV system, thus enabling a better understanding of the influence and importance of utilizing accurate irradiance data for PV system simulations. The findings from this study stress the significance of harnessing real-world solarimetric station data to boost the precision of PV system simulations and performance evaluations. The flowchart of the steps described in the methodology can be seen in the Fig. 3.4.

Figure 3.3 – PV plant installed at UNICAMP.

Equation (3.6) was employed for assessing the discrepancy between actual (Y_i) and simulated data ($\hat{Y}_i$), both on a monthly and annual average scale [80]. This facilitates a comparison between the simulated power generation using either meteorological or POA irradiance data and the actual data.

$$Error(\%) = \left(\frac{Y_i - \hat{Y}_i}{Y_i} \right) \cdot 100 \tag{3.6}$$

3.5 Results and Discussions

The Table 3.2 showcases a comparative analysis of the generated energy, the projected output in PVsyst prior to acquisition of the solarimetric data, and the simulation in PVsyst post incorporation of local solarimetric data. When compared to the real data, the simulation with Meteonorm data exhibited an error of 3.13%, whereas the simulation incorporating local solarimetric data exhibited a significantly lower error of 0.21%. While both simulations were close to the real data, the performance noticeably improved when solarimetric data collected from an onsite station was utilized in the simulation.

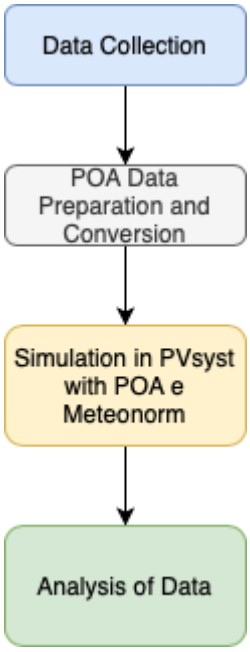


Figure 3.4 – Flowchart of methodology.

Another insightful visualization, generated using Tableau 2023, is the comparative analysis of the daily curves of the PV system and the POA irradiance (Fig. 3.5). Visually, these curves exhibit similar characteristics, differing in the few delays of occurrence, likely attributable to the data recording process and averaging techniques.

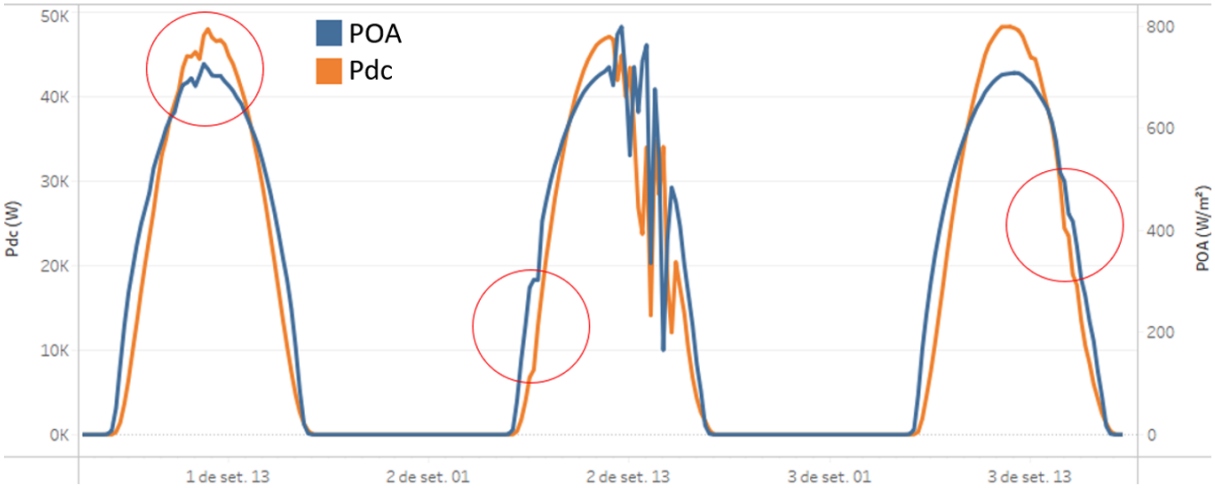


Figure 3.5 – Graph of selected days for DC power (Pdc) and POA irradiance.

In Fig. 3.6, histograms were plotted for POA irradiance and *Pdc*, highlighting a similarity in the behavior of the curves, particularly in the blue-shaded region. This indicates that the measured solarimetric data closely follows the data for energy generation. The histograms provide a visual confirmation of the correlation between solar irradiance (POA) and power output (*Pdc*), demonstrating their interdependence in the context of PV system performance analysis. Some additional PV installation data can be found on the platform www.profjl.com/projetoif.

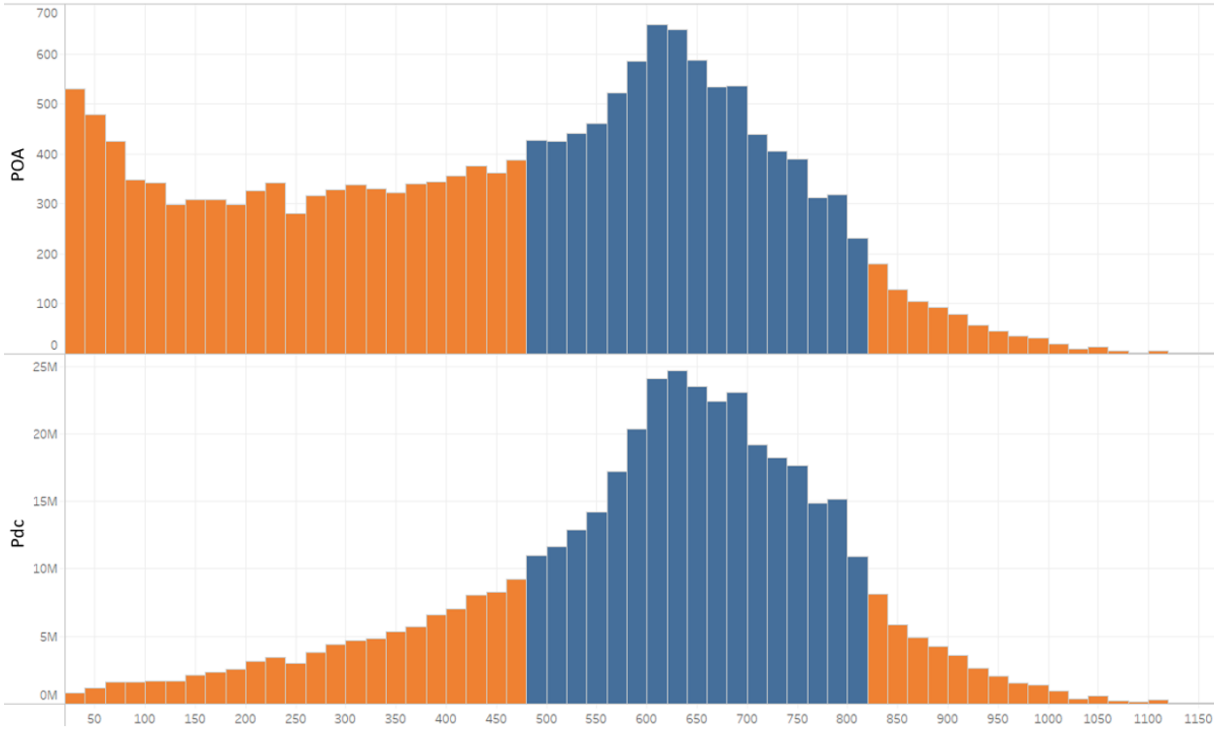


Figure 3.6 – Year 2020 histogram for DC power (Pdc) and POA irradiance.

Finally, Pearson correlation was applied between the POA irradiance and the data from the PV inverters at the Unicamp Gymnasium. The matrix was constructed using a 15-minute step for the data. Pearson correlation yields a correlation coefficient that ranges from -1 to 1 [81, 82]. A value of 1 indicates a perfect positive correlation, signifying that both variables increase in perfect proportion. Conversely, a value of -1 indicates a perfect negative correlation, implying that one variable increases as the other decreases. A value close to 0 suggests a weak linear relationship between the variables. The matrix obtained is illustrated in Fig 3.7. As a result, it was observed that there is a strong positive correlation between POA irradiance and the installation data, confirming the importance of using local data in simulations.

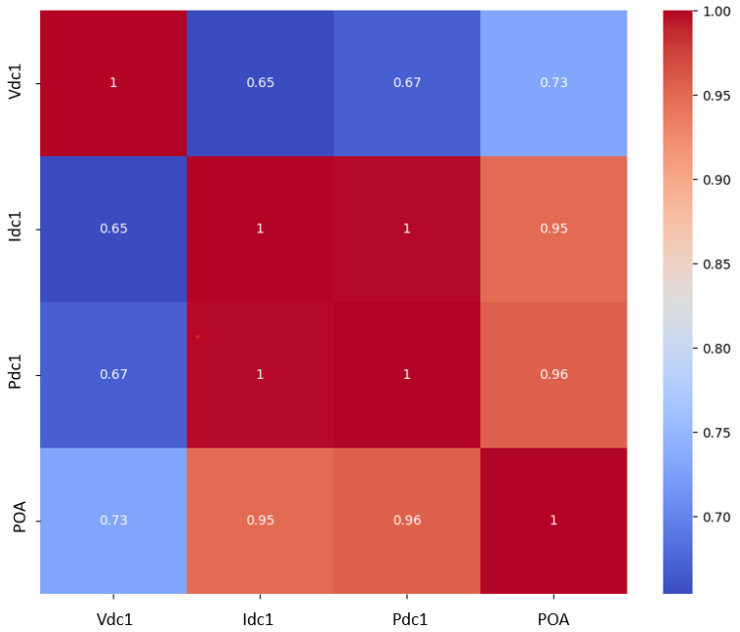


Figure 3.7 – Correlation matrix for POA irradiance and inverters data.

3.6 Conclusion

This paper stressed the significance of incorporating solarimetric data in PV systems simulations, demonstrated through a case study. The use of local solarimetric data was found to enhance the performance of PVsyst simulations, yielding a discrepancy of a mere 0.21% when juxtaposed with real energy generation. Furthermore, the analysis included graphical comparisons and an examination of the data distribution between the POA irradiance and the power output from the inverter, details of which will be elaborated upon in the full version of this paper. Looking towards future investigations, the intention is to apply the same methodology to other PV installations. This is motivated by the aspiration to reinforce the utilization of solarimetric data in performance evaluations and thereby further optimize the simulation's reliability. The results from this study serve as a compelling testament to the critical role of solarimetric data in advancing the precision of PV system's simulations. In addition, it is interesting to conduct further tests with POA irradiance data from pyranometers of different classes without using data conversion in the horizontal plane.

4 Data-Driven Analysis of Solar Photovoltaic Systems

Paper published in COBEP [83] ©2023 IEEE 8th Southern Power Electronics Conference and 17th Brazilian Power Electronics Conference (SPEC/COBEP). Reprinted, with permission, from de Silva, J.L. S. Data-Driven Analysis of Solar Photovoltaic Systems: Correlation and Distribution Patterns.

This manuscript, authored by João Lucas de Souza Silva, Michelle Melo Cavalcante, Samuel Botter Martins, Everton Josué da Silva, Tárcio André dos Santos Barros e Marcelo Gradella Villalva. The paper is associated with the Digital Object Identifier (DOI): <10.1109/SPEC56436.2023.10407565>. This work was developed with TotalEnergies financial support. In addition, we are grateful to all collaborators from University of Campinas (UNICAMP). We acknowledge the support of ANP (Brazilian National Oil, Natural Gas and Biofuels Agency) through the R&D levy regulation. Acknowledgements are extended to the Center for Energy and Petroleum Studies (CEPETRO) and School of Electrical and Computer Engineering (FEEC). The authors would like to thank the Campus Sustainable Project Unicamp for the data. Furthermore, the authors extend their gratitude to the late Professor Marcelo Villalva (1978-2023), foremost figures in the global solar energy scenario, for providing us with all the knowledge and opportunities during our professional engagement.

Abstract: With the increasing number of photovoltaic (PV) systems, it has become increasingly important to understand the behavior of all variables that permeate a system and their relationships to detect anomalies and optimize system performance. This work proposes applying Data Science concepts to three years' worth of data from a 336.96 kWp system located in Campinas, Brazil. One of the challenges seen in the literature was defining what to apply to analyze the data. To this end, a new methodology was devised based on analyzes applied in other works, proposing a way of analyzing PV data. For the PV plant studied, the results showed that there was a reduction in energy generation over the years, possibly due to degradation or soiling in the modules, and the correlation of variables for the inverter model studied. The application of data science techniques can provide valuable information to optimize system performance, increase energy efficiency, and reduce maintenance costs, especially when combined with PV inverters power electronic data.

Keywords: Photovoltaic systems, data science, pearson correlation, histogram.

4.1 Introduction

With the aim of diversifying energy sources, promoting sustainability and reducing costs, the use of photovoltaic (PV) systems has grown all over the world [20]. However, it is crucial to regularly assess the performance of these systems to ensure that technical and economic forecasts are sustained. In this sense, data science can be a valuable ally.

Analyzing data-driven systems is crucial, as it provides an opportunity to comprehend the underlying data and identify potential anomalies or latent issues. Conducting this preliminary analysis not only contributes to the early detection of any irregularities but also enables an understanding of patterns, characteristics, and trends within the data. Furthermore, it can assist in the appropriate selection of machine learning algorithms and types, thereby facilitating the process of model implementation and optimization, if necessary.

Exploratory data analysis consists of clarifying the structure of the data and, when necessary, enabling transformation into a format that enables the extraction of the necessary information [84]. Depending on the case, methods can be applied to filter data, reduce dimensionality, clean noise, for example, Principal Component Analysis (PCA) [85].

Several works in the literature utilize datasets for various analyses of PV systems [86–91]. However, there is a certain difficulty in finding a standardized methodology for conducting exploratory analyses on PV systems. Each study adopts a different approach to process and perform the necessary tests. This variability may be attributed to the diverse nature of these systems and data acquisition. In this study, the methodology for exploratory analysis is grounded in datasets extracted from PV inverters, containing basic input and output information, such as voltages and currents.

Thus, this study presents a novel methodology for assessing data extracted from PV systems. To achieve this, data from a 336.96 kWp PV system located at the University of Campinas, Brazil, as part of the Campus Sustentavel project [33], were utilized. The performance of the PV system was evaluated using data science techniques, encompassing data collection and preprocessing, Pearson's correlation analysis, histogram analysis, and forecasting employing the Weibull distribution. With the analyzes carried out and tests, this work resulted in the following contributions:

- Creation of a methodology for exploratory analysis of PV systems.
- Examination of three years' worth of authentic PV data through data-science approaches.

- Provision of insights for optimizing PV system output and identifying potential issues.
- Facilitation of data-driven decision-making for future PV installations.

4.2 Data Science and PV systems

4.2.1 Grid-connected PV power system

A grid-connected PV system is made up of several components that work together to produce electricity from sunlight. Fig. 4.1 illustrates a conventional PV architecture when connected to the electrical grid [20].

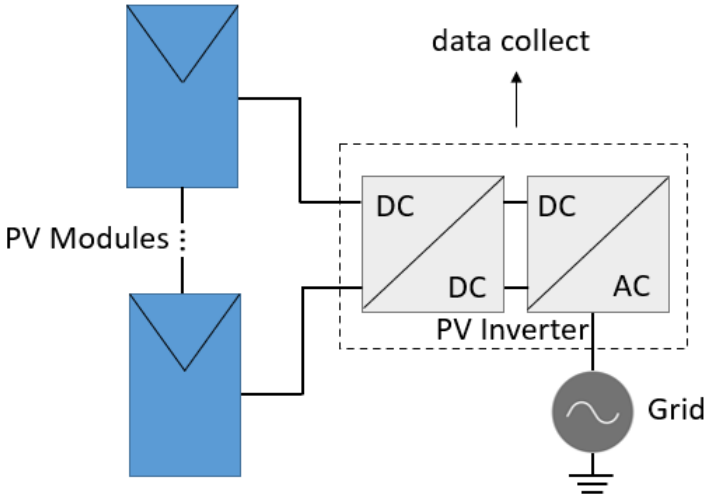


Figure 4.1 – Conventional PV Architecture. Adapt [20].

PV modules are responsible for transforming sunlight into electrical energy. The PV inverter converts the energy produced by the set of PV modules into energy usable by the electrical grid. Data collection in a PV system connected to the electrical grid is done by the PV inverter [19], which as a rule, records information about energy generation, such as the amount produced and the efficiency of the system.

4.2.2 Pearson's Correlation

Pearson's correlation is a widely used statistical tool that measures the linear association between two quantitative variables. This correlation coefficient can be computed only if data regarding at least two quantitative variables are available. The Pearson's correlation coefficient ranges from -1 to +1, where values near +1 imply a strong positive correlation between the variables, values near -1 suggest a strong negative correlation, and values close to zero indicate either a weak correlation or no correlation

at all [82]. Numerous studies employ Pearson correlation as a well-established measure in the literature, including applications within the field of solar research [92–94].

4.2.3 Histograms and Data Distributions

Histograms represent a powerful statistical technique employed for depicting the distribution of data within a dataset [95]. This visualization tool works by breaking down the data range into smaller intervals or classes and then counting the number of observations present in each class. The resulting bar chart illustrates the frequency for each class, enabling the identification of patterns, including skewness and kurtosis. Additionally, histograms can be instrumental in identifying outliers within a dataset, such as data gathered from a PV inverter.

When it comes to data distributions, it is basically a mathematical function capable of returning the probability that a random variable assumes a certain value. Therefore, it is useful for modeling the distribution of a random variable in terms of its mean, standard deviation, and other statistical properties. A very common example in renewable energy applications is the Weibull distribution, being applied in works like [96].

Numerous studies employ histograms in conjunction with data science in the realm of solar energy research. In the study by [97], histograms played a pivotal role in gaining insights into the degradation of PV cells, providing a valuable tool for comprehending the dataset employed by the authors. In contrast, [68] noted in their concluding remarks that histograms were utilized to demonstrate the departure from a normal distribution in solar irradiance. In a distinct approach, [98] harnessed histograms to scrutinize the error associated with training, validation, and testing of machine learning models employing neural networks for solar resource assessment, ultimately aiding in the determination of suitable locations for solar system installations.

4.2.4 Boxplot

The boxplot technique is a graphical method employed to detect outliers. This method is notable for not including extremely outlier values when calculating a measure of dispersion. Internal and external boundaries are defined based on quartiles, preventing extreme values from distorting the analysis [99].

In [68], the boxplot was employed to assess the extent to which a dataset adheres to a normal distribution when dealing with solar irradiance. In [100], it was used to evaluate the detection of potential faults, integrated into the methodology for analyzing the quality of PV systems with descriptive statistics. In another study [101], the boxplot technique was also applied to detect faults in PV system strings.

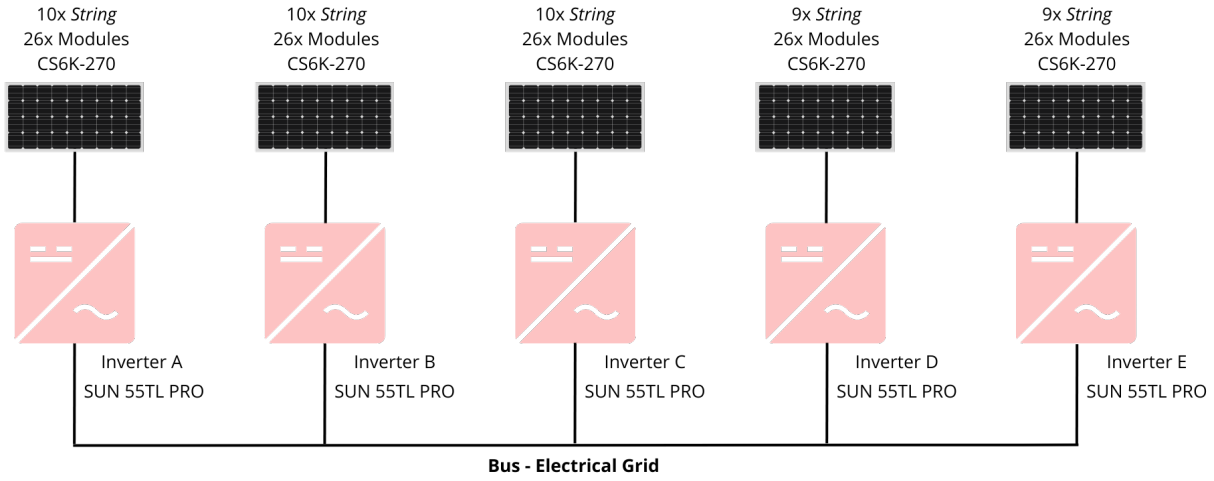


Figure 4.2 – Description of the PV Project at the Unicamp Gymnasium.

4.3 Sustainable Campus Project

The University of Campinas (UNICAMP), in collaboration with CPFL Brazil (the local electricity distributor), initiated the "Sustainable Campus" project in August 2017, aimed at promoting sustainability across the university [38]. This comprehensive project encompasses twelve subprojects, each concentrating on various research areas. The data used was from the PV solar energy subproject, especially in the largest installation, which was the Gymnasium. Fig. 4.3 shows the installation used for the research.

Figure 4.3 – PV Installation at the Unicamp Gymnasium, Brazil.

The installation, with a capacity of 336.96 kWp, consists of five inverters, each with the same power rating. Consequently, each inverter can be regarded as an individual

system since the data is obtained from each inverter. Figure 4.2 provides a description of the project, including the number of modules in each inverter and their respective models.

4.4 Methodology

First, Pearson's correlation analysis is performed to investigate the relationship between variables, enabling the identification of potential interdependencies among the data. Next, analyses of data distributions are conducted to understand their variability and statistical characteristics. To visualize these comparisons, histogram plots are used, providing a clear and concise representation of data distribution. Additionally, boxplots are employed, highlighting the median, quartiles, and outliers of the data for each year or season (due to page limitations). This allows for easy comparison and identification of any variations. Finally, an annual energy generation study is conducted, and the results are analyzed to internally construct a storytelling narrative.

All the discussed implementations were based on statistical methods found in the literature [68, 92–94, 96–98, 100, 101]. In this work, we propose the use of the flowchart in Fig. 4.4 as a data-driven methodology for standardizing such analyses or as a starting point.

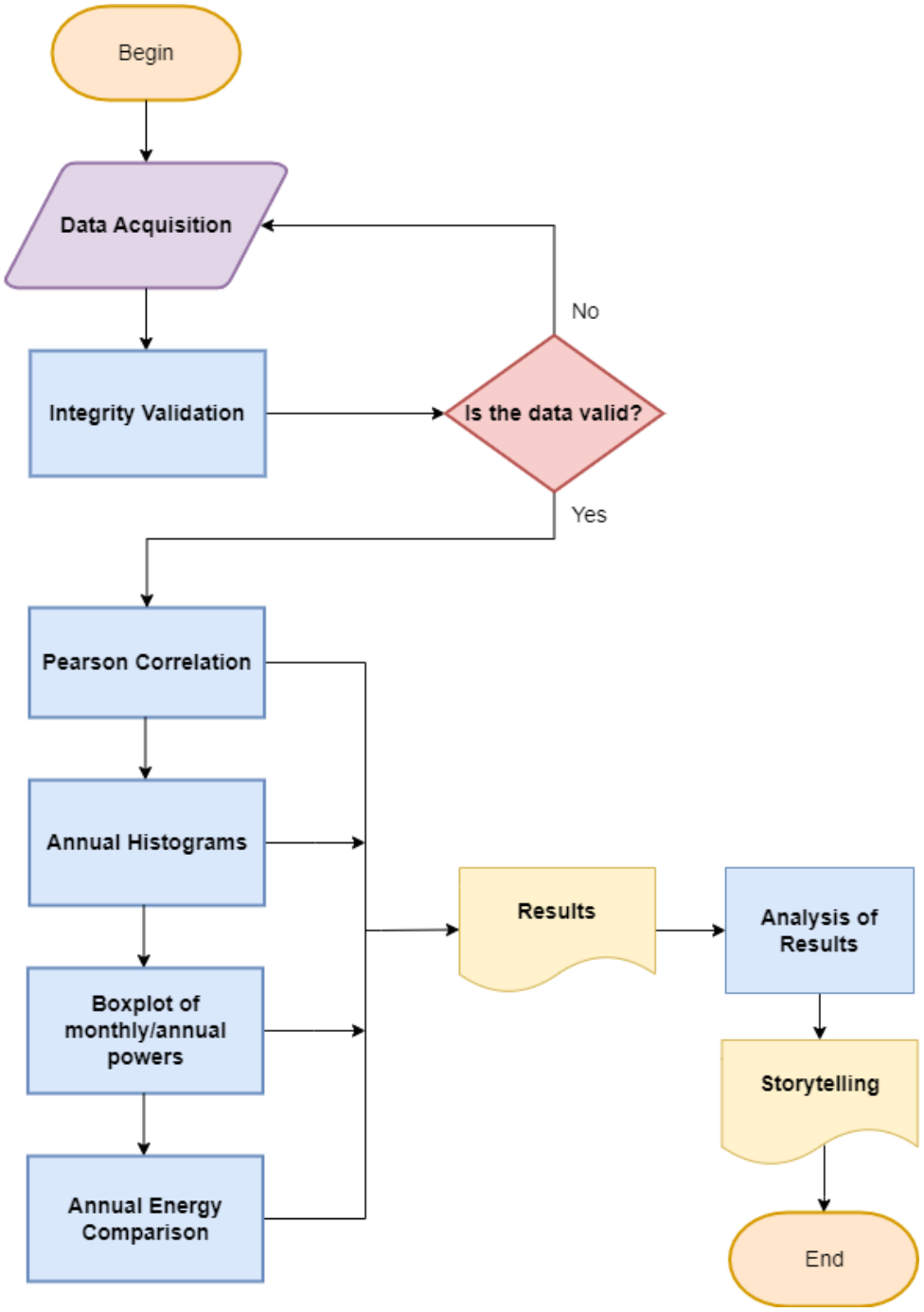


Figure 4.4 – Flowchart for Exploratory Analysis of data collected from PV systems.

Lastly, the Weibull distribution is applied to calculate the probability of the inverter operating at maximum power, providing valuable insights for system sizing and efficiency. This information can be used to optimize operation and maintenance strategies, maximizing energy generation and reducing costs in the PV system. Overall, our methodology enables a comprehensive analysis of PV systems, facilitating the identification of patterns, correlations, and optimization opportunities for improved system performance.

4.5 Results and Discussions

After the data had been processed, the statistical analyses outlined in Fig. 4.4 were conducted. The initial analysis involved Pearson's correlation, which was computed for the entire system as a whole and separately for individual inverters. As illustrated in Fig. 4.5, the correlation analysis was performed on the three years of data for all inverters combined. Fig. 4.6 further presents the correlation results separated by year for Inverter A, given its resemblance to the patterns observed in the other inverters, although these are not presented in this paper.

As part of the discussion, it was observed that current exhibited a perfect positive correlation with power, indicating their proportionality. In contrast, voltage and frequency demonstrated weak or negligible correlations with power and current. These findings align with the inherent complexities of PV inverters, stemming from intricate interactions between the electrical characteristics of PV modules and the control processes within the PV inverter. In the event of a non-positive correlation between current and power, it could potentially signify the presence of an anomaly.

The next step involved annual histograms for each inverter, along with annual comparisons. The results are presented in Fig. 4.7, 4.8, 4.9, 4.10, and 4.7. Two noteworthy observations emerged from the histogram analysis. Firstly, it was evident that in the first year, the inverter exhibited higher power values compared to the subsequent years, indicating visible degradation—a phenomenon expected in a PV system. However, as a second observation, it became apparent that Inverters D and E experienced a significantly more substantial reduction in power compared to the other inverters, hinting at an anomaly, despite having fewer PV modules.

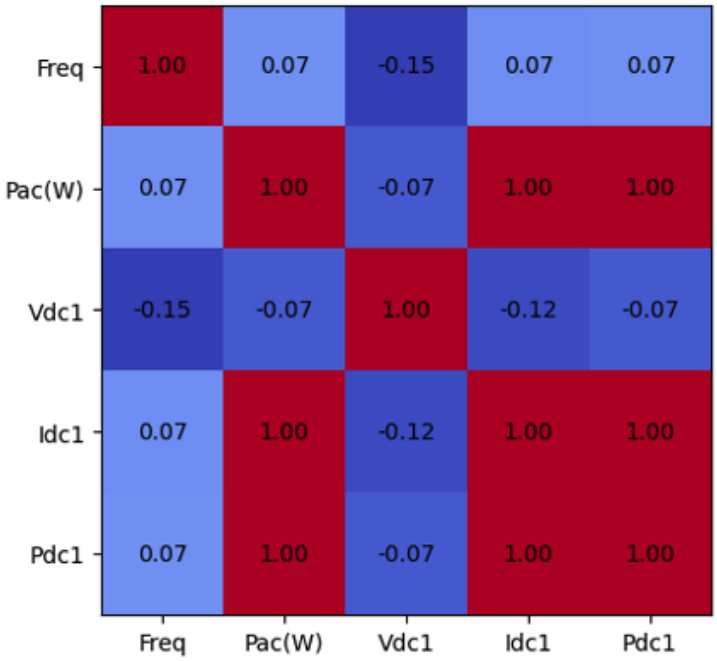


Figure 4.5 – Correlation Matrix of the Unicamp Gymnasium dataset for three years.

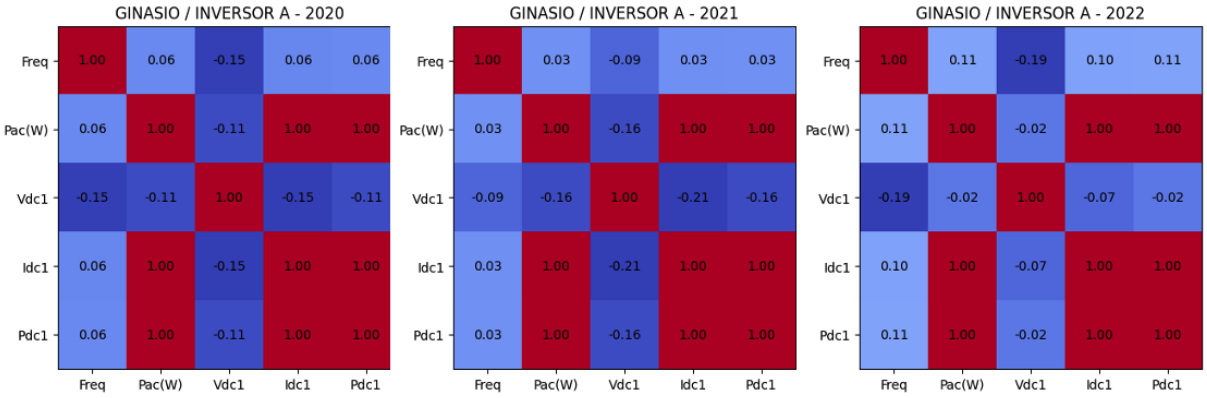


Figure 4.6 – Correlation Matrices for Inverter A at the Unicamp Gymnasium.

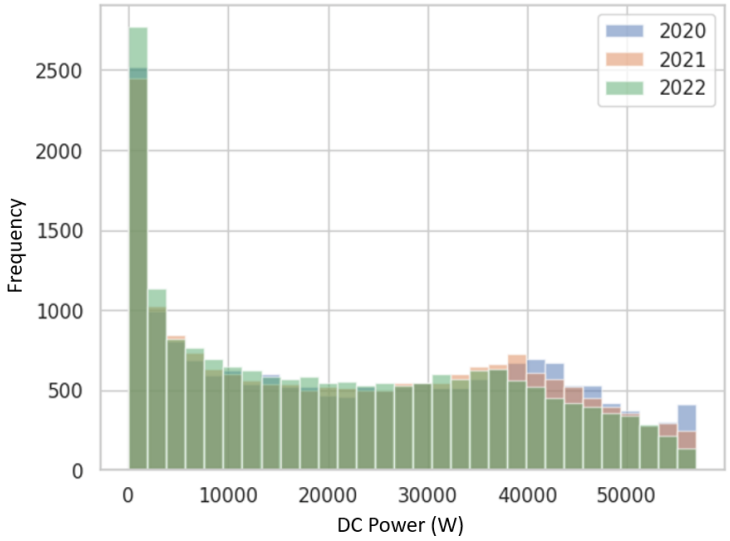


Figure 4.7 – Histogram for Inverter A at the Unicamp Gymnasium.

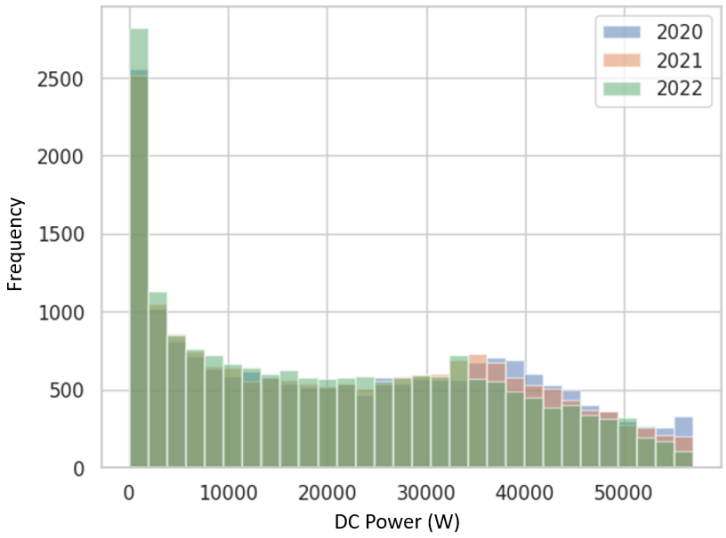


Figure 4.8 – Histogram for Inverter B at the Unicamp Gymnasium.

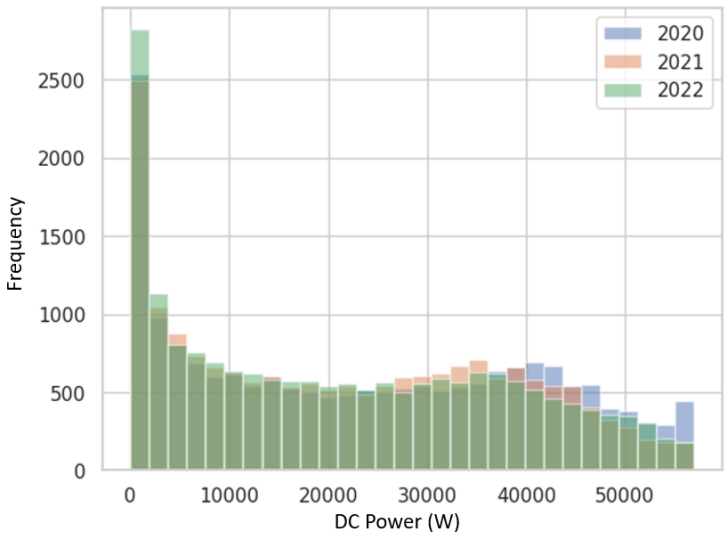


Figure 4.9 – Histogram for Inverter C at the Unicamp Gymnasium.

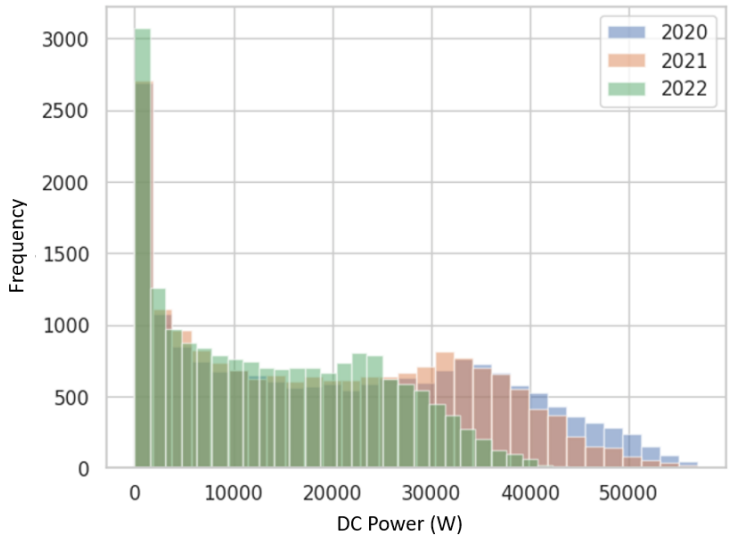


Figure 4.10 – Histogram for Inverter D at the Unicamp Gymnasium.

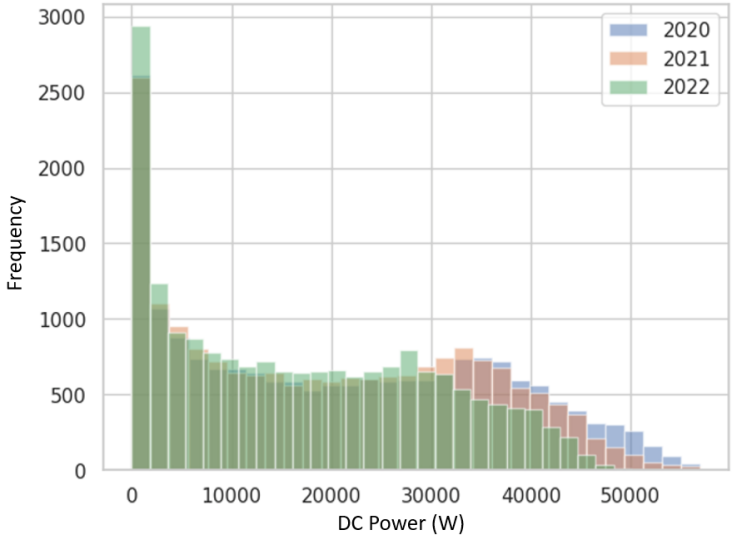


Figure 4.11 – Histogram for Inverter E at the Unicamp Gymnasium.

The Weibull distribution also was applied and an attempt was made to verify the probability of the inverter working at its maximum in a data set for each year. The result obtained showed that the probability has been reducing each year (8.23%; 7.25%; 5.87%), proving the degradation of the system or soiling. Thus, it is necessary to evaluate and define new maintenance routines.

Subsequently, boxplots were generated with the initial focus on assessing the power output of all PV inverters throughout the years 2020, 2021, and 2022. Notably, inverter A exhibited a consistent performance profile (see Fig. 4.12), indicative of normalcy, while in the case of inverter D (see Fig. 4.13), significant disparities emerged across the years. This underscores the efficacy of the boxplot tool within this context. Notably,

within inverter D, the boxplot for the year 2022 exhibited distinct behavior compared to the others, standing out due to a significant reduction in power. This reduction also led to a notable deviation of the median from the mean; however, in the case of inverter D, the boxplot was considerably more compressed. This observation raises a flag for potential anomalies. While a decrease in irradiance levels could conceivably lead to a reduction in the boxplot values for 2022, the magnitude of reduction in inverter D appears to deviate from the patterns observed in the other inverters.

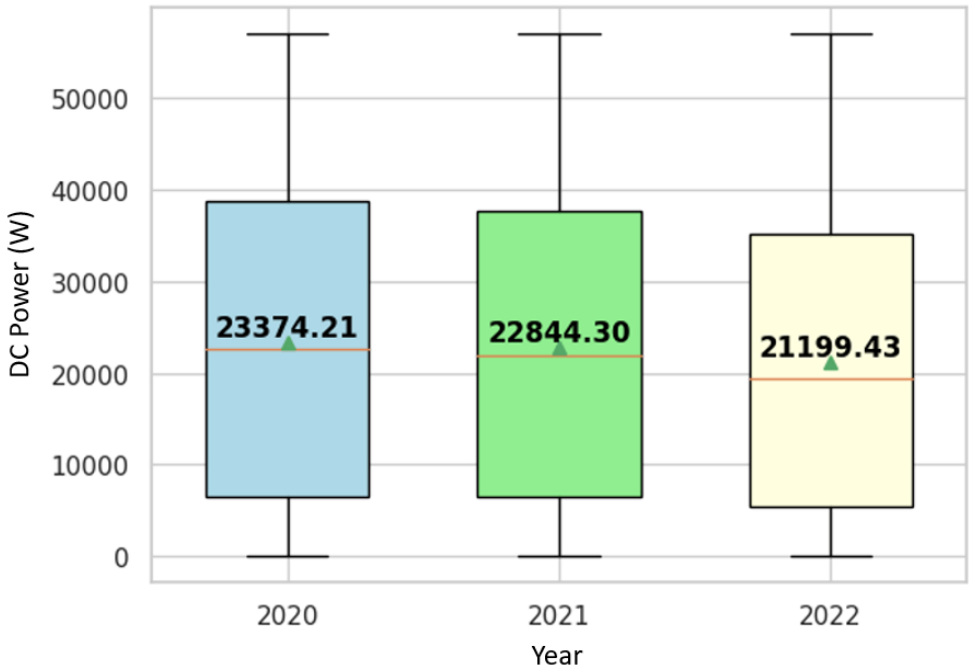


Figure 4.12 – Annual boxplot of Inverter A at the Unicamp Gymnasium.

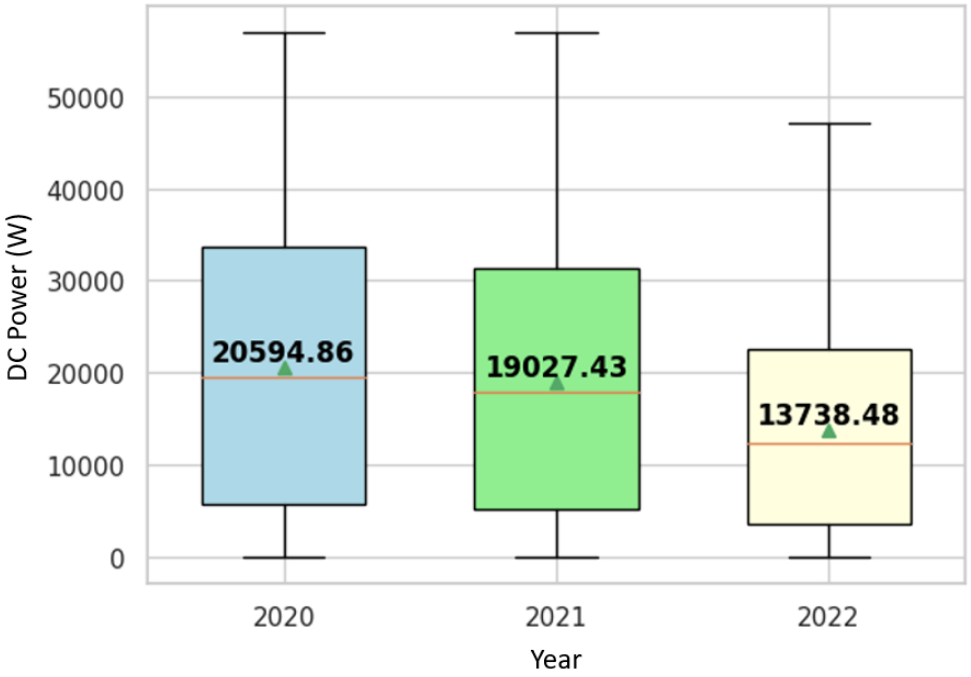


Figure 4.13 – Annual boxplot of Inverter D at the Unicamp Gymnasium.

Subsequently, an investigation into PV generation was conducted. To this end, graphical representations were generated depicting the total PV generation for the year (see Fig. 4.14), as well as the generation attributed to each inverter over the span of three years (see Fig. 4.15). The graph presented in Fig. 4.14 highlights a discernible decrease of 10.90 % between the years 2021 and 2022. In Fig. 4.15, a reflection of the findings from the previously examined boxplot analysis is evident, with inverter D exhibiting a notably substantial reduction in generation, prominently during the year 2022. These results substantiate the imperative for an exhaustive examination of the installation, with a specific focus on inverter D. Drawing upon these insights, it becomes feasible to craft an easily comprehensible narrative, a strategy that aligns seamlessly with the art of storytelling (ex. https://www.profjl.com/projetoif).

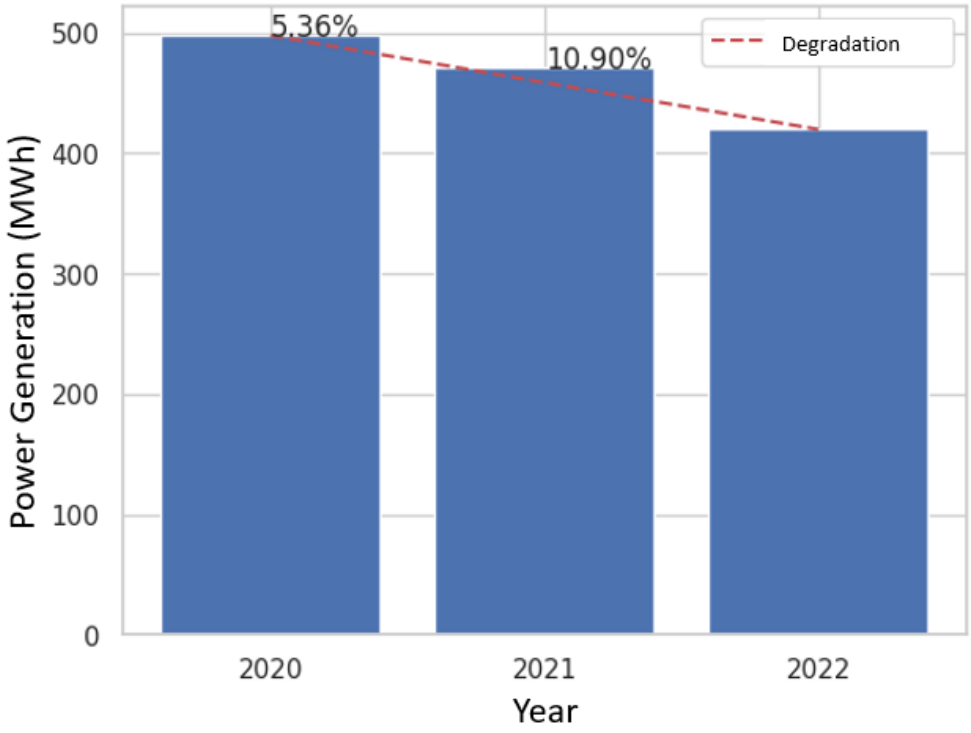


Figure 4.14 – Total energy generation for the PV installation of the Unicamp Gymnasium.

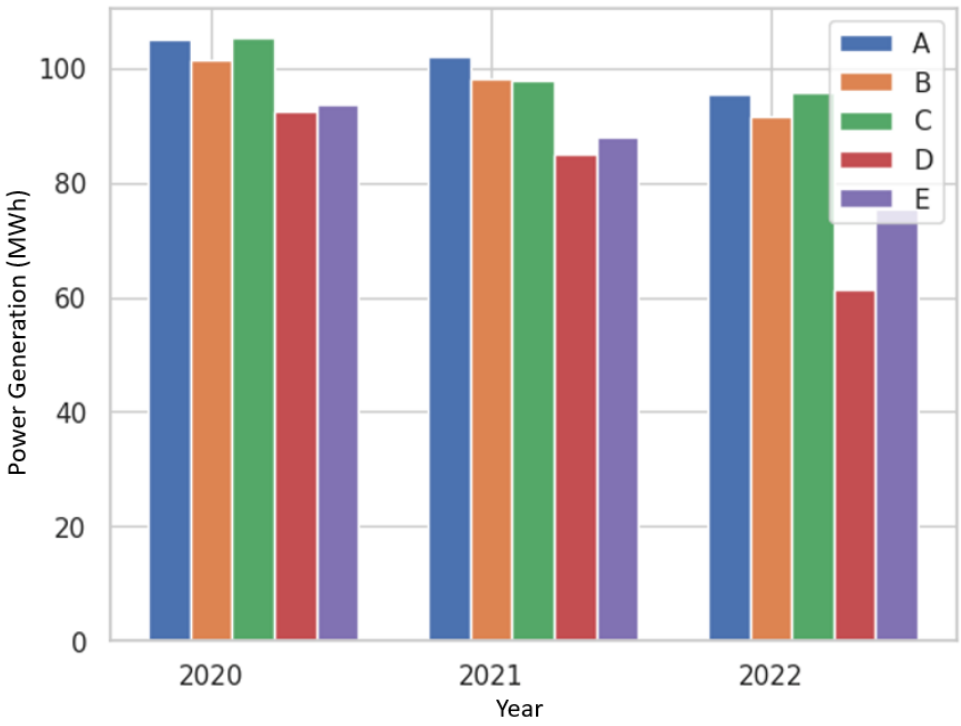


Figure 4.15 – Total energy generation of each PV Inverter at the Unicamp Gymnasium.

4.6 Conclusion

The presented data underscores a notable downward trend in the occurrence of higher power outputs from the PV system over the years, indicative of a decrease in its overall production. To safeguard against further declines, it is imperative to delve into the root causes driving this trend. Furthermore, the investigation unveiled that, within this system, DC current exerts a more substantial influence on power output than voltage. As a primary outcome, this study accentuates the system's potential for early detection of performance deviations, a phenomenon notably exemplified in the case of inverter D.

Therefore, the application of data science techniques proves indispensable for comprehending the PV system's dynamics, particularly with regard to the PV power converter. As well, such analytical approaches aid in elucidating the dataset's intricacies and serve as a preparatory step for more intricate analyses involving artificial intelligence.

5 Classification of Anomalies in Photovoltaic Systems using Supervised Machine Learning Techniques and Real Data

Paper published in Energy Reports [102] ©2024 Elsevier. Reprinted, with permission, from de Silva, J.L. S. Classification of Anomalies in Photovoltaic Systems using Supervised Machine Learning Techniques and Real Data, June, 2024.

This manuscript, authored by João Lucas de Souza Silva, Eslam Mahmoud, Rômullo Randell Macedo Carvalho, Tárcio André dos Santos Barros. The paper is associated with the Digital Object Identifier (DOI): <10.1016/j.egyr.2024.04.040>. This work was developed with TotalEnergies financial support. In addition, we are grateful to all collaborators from University of Campinas (UNICAMP). We acknowledge the support of ANP (Brazilian National Oil, Natural Gas and Biofuels Agency) through the R&D levy regulation. Acknowledgements are extended to the Center for Energy and Petroleum Studies (CEPETRO) and School of Electrical and Computer Engineering (FEEC). The authors would like to thank the Campus Sustainable Project Unicamp for the data. Furthermore, the authors extend their gratitude to the late Professor Marcelo Villalva (1978-2023), foremost figures in the global solar energy scenario, for providing us with all the knowledge and opportunities during our professional engagement.

Abstract: In a photovoltaic (PV) plant, various types of anomalies can lead to a decrease in energy conversion efficiency. These anomalies represent deviations from the normal behavior of the plant. Understanding the system's behavior and identifying deviations from normality are crucial for implementing preventive and corrective measures to ensure the expected economic return on investment in the plant. One effective approach for anomaly detection and classification is the utilization of Supervised Machine Learning Techniques (SMLT). However, this approach comes with several challenges, including the classification of the training dataset, temporal variations, uncertainties, model interpretability, data scarcity, and generalization of the model to different systems and locations, among others. In this way, this study aims to explore the application of SMLT in PV systems and compare different methods. To achieve this, a methodology was developed to create a training and testing synthetic dataset based on real irradiance data, followed by a new process flow for ensemble SMLT. Finally, the new process flow for ensemble SMLT was tested on a real PV plant and synthetic dataset. The methodology for

the proposed algorithm was an ensemble of Random Forest with K-nearest neighbors (k-NN) and an inference machine for specific classes. The results indicate that the algorithm successfully classified anomalies, achieving an AUC of 0.9815 for the synthetic dataset and an AUC of 0.9861 for the real dataset, as well as an accuracy of 0.9647 for the real dataset. It is evident that SMLT can serve as valuable tools for detecting and addressing issues, particularly due to the abundance of data and the diverse characteristics inherent to PV plants.

Keywords: Anomalies, Photovoltaic, Photovoltaic Plants, Photovoltaic Anomalies, Machine Learning.

5.1 Introduction

Evaluating and monitoring photovoltaic (PV) plants has become important to ensure that technical and economic forecasts are sustained, especially with the constant expansion of PV systems with an interest in sustainability and reducing energy costs [20]. This expansion created a market with products of dubious origin and unqualified labor, which makes the existence of tools for the regular evaluation of plants even more important.

During the evaluation of a PV plant, anomalies can be found at all stages of the energy generation chain, and this directly reflects on the data provided by the PV converter. Anomalies in the literature regarding PV modules are subdivided into irreversible failures and temporary failures [103], both of which result in a reduction in generation capacity, generally added to losses due to mismatch [4]. Irreversible failures are caused by physical damage, whether mechanical, electrical, or environmental, as well as natural degradation. Temporary failures are generally characterized by shadowing, and this can be corrected depending on the nature of the cause [56].

Likewise, it is important to categorize PV converter failures as irreversible and temporary. Irreversible failures are generally linked to problems in the power electronics, resulting from a variety of factors. These include temperature variations, fluctuations in the electrical network, construction errors, switching losses, parasitic resistances, inductances, capacitances, incorrect sizing, and electromagnetic interference, among other challenges. Temporary failures in PV converters, on the other hand, can be induced by factors such as temperature variations, where the inverter adjusts its operational power [64].

The fact is that PV converter monitoring data can capture these failures [19]. PV converters carry out a massive collection of data, and this makes it difficult for PV integrators to work in various installations. In this sense, data science can be a valuable

ally for integrity verification and analysis. Data-driven systems analysis provides an opportunity to understand the underlying data and identify potential anomalies or latent problems. As well, it also allows an understanding of patterns, characteristics, and trends in the data. Along these lines, it is interesting to carry out exploratory analyzes on PV plants and search for anomalies using machine learning (ML) methods.

Regarding exploratory analyses, works in the literature prepare datasets or perform analyzes with these data. For example, in [86], data analysis was conducted using Python on a real 500 kW plant in India. In [87], an important dataset was developed with 42 systems deployed in five campuses of La Trobe University, Victoria, Australia. The work mainly described data cleaning (preparation). In [89], the authors proposed a methodology to assess the maximum power of a PV plant through data analysis, mainly based on the characterization of irradiance sensors. Zhang [90] used data to verify the spatial correlation of different PV systems in different locations. Finally, Sundararajan [91] created a roadmap to prepare data for the performance analysis of PV plants. However, there is some difficulty in finding a standardized methodology for carrying out exploratory analyzes or data preparation on PV plants. Each study takes a different approach to processing and performing the necessary tests. This variability can be attributed to the diverse nature of these systems and data acquisition.

In the case of applying machine learning algorithms, it is clear in the literature that due to the difficulty in classifying and finding some specific failures, authors concentrate on parts of the system or just some failures. In [104] the authors studied some types of ML applications, such as *AutoEncoder Long Short-Term Memory* (AE-LSTM), *Facebook-Prophet* and *Isolation Forest*, in addition to citing other similar works. In [105] the authors applied ML to classify possible anomalies in the PV module. In [106] a compilation of possible ML applications in PV plants was carried out. In [107], the identification of faults in PV modules was studied, however neural networks were trained for each network for a type of fault, not being a flexible alternative.

In this scenario, this paper proposes contributions to training/testing anomaly classification in PV dataset. To achieve this, it was proposed to create a synthetic dataset and another with real data. The general objective was to investigate and classify anomalies in PV systems through the Supervised Machine Learning Techniques (SMLT). Contributions include:

- Novel approach compared to the literature, employing SMOTE and an ensemble of SMLT for anomaly classification;
- Proposal for a methodology for generating a dataset for training/testing ML algorithms for detecting/classifying anomalies;

- Tested classic supervised ML methods applied to datasets;
- Used, monitored, and tested a minigeneration PV installation in the University of Campinas (Unicamp) Sustainable Campus project to validate the studies.

After this introduction, Section 5.2 provides the background of the work, beginning with insights into anomalies in PV systems and concluding with an exploration of ML methods. Section 5.3 outlines the methodology employed in this study, followed by the presentation of results and discussions, including the proposal of synthetic datasets, analyses of real datasets, and the application of SMLT.

5.2 Background

5.2.1 Anomalies in Photovoltaic Systems

Anomalies in PV systems have the potential to influence the efficiency of electrical energy production, a factor that plays a crucial role in the financial return of an enterprise. It is important to highlight that such deviations can occur due to problems ranging from the grouping of PV modules to the integration with the electricity grid [3], as illustrated in Figure 5.1 examples.

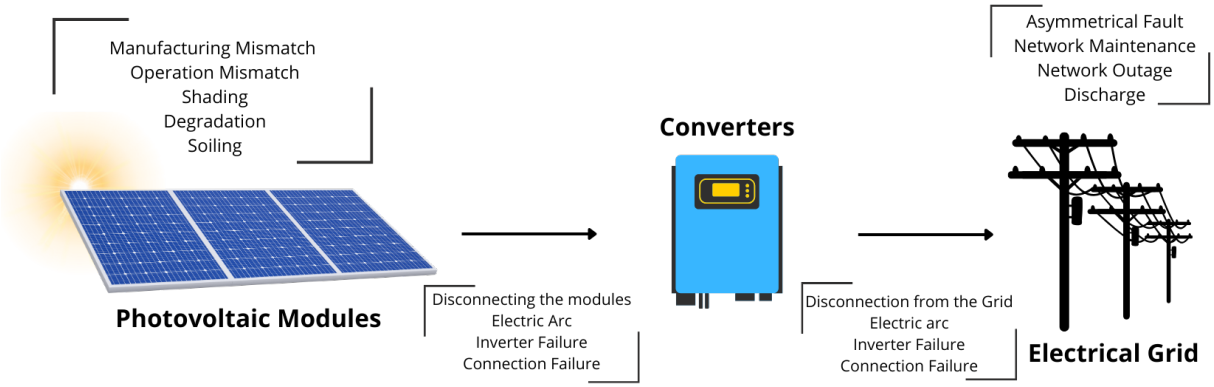


Figure 5.1 – Various problems that cause anomalies in PV systems.

However, most of the problems that appear in PV plants will inevitably be reflected, directly or indirectly, in the behavior of PV modules. This factor, in turn, makes the application of detection algorithms a more accessible approach. As an example, a failure in a converter can cause the modules to enter a short-circuit (SC) or open-circuit (OC) state.

The origins of anomalies include several factors, manifesting themselves, especially when the designer does not pay attention to crucial aspects of the PV project. An example of this is the insufficient understanding of how to avoid the occurrence of

anomalies in response to the specific environmental conditions of the system's location or the choice of lower-quality components. It is important to highlight that the presence of anomalies is inherent to any installation. This arises from the existence of unavoidable elements, such as the interruption of the electrical supply by the grid or the progressive degradation of system components, for example, [3].

As many types of anomalies are correlated, each author deals with the types of failures/faults differently. For example, Vieira [107] divides faults and faults in modules into a mismatch, bypass diode, module disconnection, asymmetric faults, electric arc, atmospheric discharges and grounding faults. Hong and Pula [108] divided them into a mismatch, SC, OC, bypass diode problems, line fault, grounding faults, junction box problems, electric arc, bridge failure (low resistance in the PV module cables), problems in the maximum power location algorithm, inverter problem, network problem and natural disaster.

Many works restrict it to fewer classes or types of failures, which makes it difficult to apply in practice. Garoudja observed the DC side to detect SC and OC in PV modules with Direct Propagation Neural Network [109]. Harrou [110] used the unsupervised Support Vector Machine (SVM) to detect SC, OC and shadows. Chen [111] worked more concerned with degradation and open circuit failure; for this, he applied Random Forest, and also used the measurement of the DC part. Li [112] worked based on the I-V curve of the PV module, which makes it difficult to apply in real-time, but he tested methods such as K-Nearest Neighbor (k-NN), Decision Tree (DT), Regression forest (RF), and Naïve Bayesian (NB). In Hong [108], a bibliographical review of several works involving detection/classification was presented, and it is possible to verify that a large part uses data from the DC part and there is no consensus on how to classify the types of anomalies. Thus, each research acts on a type of anomaly.

For this work, the focus was on simplifying the classification of anomalies, accompanied by the assignment of a number to each class, in order to assist with labeling during ML training. This approach is due to the fact that anomalies in PV plants impact the behavior of the PV module, and the data available for standard analysis in a PV plant is limited (*features*). Generally, the available data are current, voltage, and power extracted from the PV inverter.

For example, works like [113] perform this class simplification. In [114], the elbow method was applied to define the optimal number of clusters (which serves as a basis for the number of classes) in real data from PV datasets, in this case, the ideal number of clusters is found by the point of "elbow", or inflection point of the graph obtained. With the elbow method, the Brazil dataset used resulted in a graph similar (in cluster numbers) to the work of [114] presented in Figure 5.2. Thus, Table 5.1 shows the

classes studied in the present work.

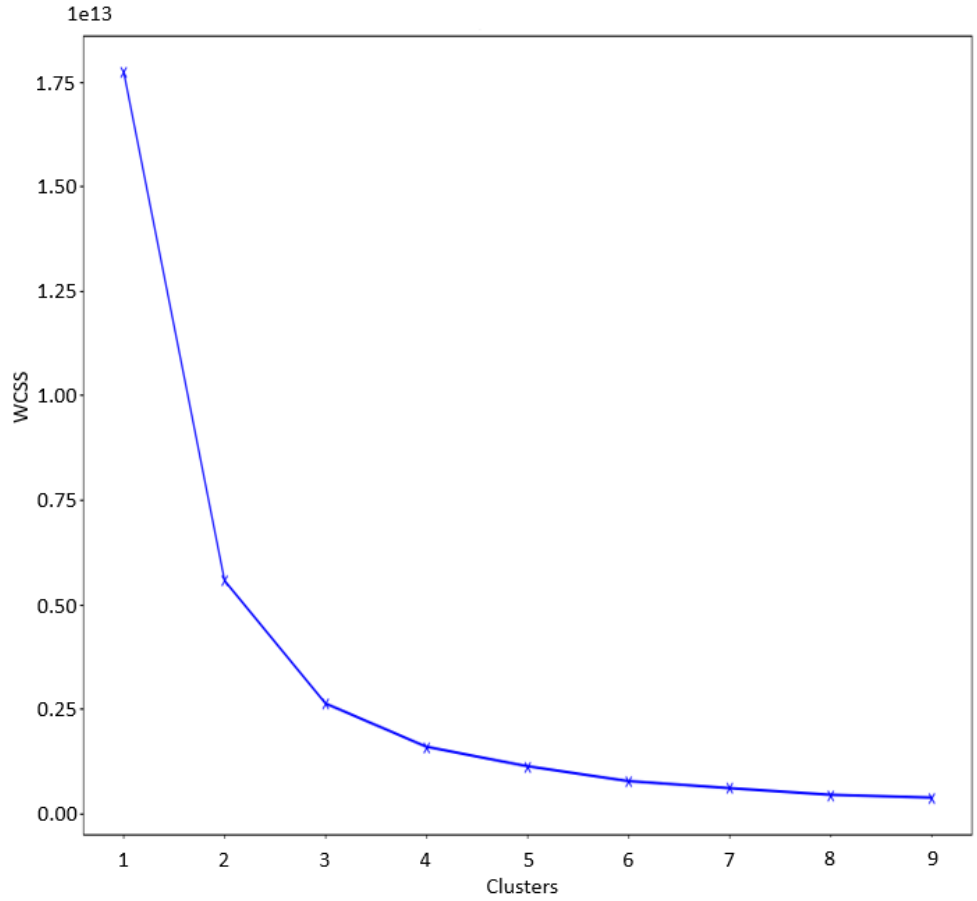


Figure 5.2 – Application of the elbow method on the Unicamp Photovoltaic Plant dataset.

Table 5.1 – Classes assigned for each type of anomaly.

Class	Anomaly	Description
0	Normal Behavior	-
1	Attenuation	Reduction in working power
2	Open-Circuit	Record of the PV module voltage reading in open-circuit
3	Short-Circuit	Record of the PV module current reading in short-circuit
4	N/A	Other unlisted problems, such as monitoring failure

The positive point of classifying into groups is that when the algorithm identifies an anomaly within a specific class, the scope for correcting the failure is already delimited. With a limited scope, maintenance and detailed searches in large PV plants are easier.

5.2.2 Machine Learning Methods

Machine learning can be defined as a science that, through computer programming, is capable of learning from data [22]. To this end, there are different types of

machine learning, which are classified in different ways depending on the author. In [22], machine learning systems are separated into three criteria: ability to be trained or not with human supervision (supervised, unsupervised, semi-supervised, and reinforcement learning methods); Incremental learning speed (online and batch learning); or whether they are able to detect patterns in data with the ability to create a predictive model or just compare new points with known data points (Instance-Based Versus Learning, and Model-Based Learning).

Supervised methods are highlighted by the possibility of classifying a result of interest [25]. However, the data used in training must be classified for learning. This can present a challenge in solar PV related projects, where manual classification of anomalies can be complex. However, when performing this classification and training the model, identification accuracy tends to be higher than other methods, as data labeling facilitates learning.

Supervised methods are categorized into probabilistic classifiers, linear classifiers, and other classifications [115]. Among the algorithms applicable to supervised learning are: *Naive Bayes* (NB) [116], *Bayesian Network (BN)* [117], *Maximum Entropy Classifier* (MEC) [118], *Support Vector Machine* (SVM) [119], *Multilayer Perceptron* (MLP) [120], *Logistic Regression* (LR) [121], *Rule-Based Classifiers* [122], *Decision Tree* (DT) [123], *Random Forest* [124], *Neural Networks* (NN) [125], *CBR* [126], *Boosting* [127], and *Quadratic Classifiers* [128].

5.3 Metodology

In the first part, synthetic datasets were prepared based on PV modeling with real irradiance data and anomalies were inserted. In the second part, a PV system installed at Unicamp was studied and a methodology for exploratory data analysis was devised with the aim of checking anomalies and preparing the dataset for analysis with ML algorithms. And, the project will end with the study/elaboration of ML algorithms applied to the available datasets.

The synthetic dataset is important to train/test the algorithm on something known, but then testing with real data is necessary to validate. Based on the work of [113] in which the authors proposed a dataset with I-V curves for detecting faults in PV systems, here the creation of the dataset with measured irradiance data was proposed. Different from the aforementioned work, the objective was, based on the irradiance measured by a solarimetric station, to create a synthetic energy generation scenario for a PV module, obtaining data from the PV inverter, not the I-V curve.

In the second part, exploratory analysis was necessary to evaluate the data that would be worked on later with SMLT. In this work, the methodology for exploratory analysis is based on datasets extracted from PV inverters, with basic input and output information (voltages and currents). The purpose of developing an exploratory methodology is to raise questions such as: (1) What is the correlation between the inverter data? (2) Are there any unusual patterns in energy generation when compared to other months or years? (3) Can possible anomalies be identified through this analysis? Furthermore, with the exploratory approach, it is feasible to provide the user with a deeper understanding of the behavior of the PV plant in question. The methodology used was proposed for this work in [83].

Finally, in the third part, the application of conventional algorithms from the literature, such as those mentioned in the previous section, is proposed, in addition to a new methodology for applying the SMLT in the scenario with PV data. The algorithms will be applied to synthetic and real bases. The real base had some problems added to the purpose of the study.

To evaluate the results, a confusion matrix will be obtained for each test, and performance indicators will be calculated. The confusion matrix is essentially a table that shows how often classifications were correct or incorrect in different classes. In the matrix, rows represent the actual data (true labels), while columns represent the predicted labels.

With the matrix, it is possible to obtain the precision, an indicator for the accuracy of positive predictions. In other words, it represents how often the algorithm was correct, whether it was detecting an anomaly or not [22]. This indicator is good for balanced datasets, but not for datasets where anomalies are sparse compared to non-anomalies. This is because you can have high precision in detecting non-anomalies and fail to detect the actual anomalies. Precision is calculated using Eq. 5.1, where TP is True Positive, and FP is False Positive.

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

Another indicator is Recall or sensitivity [22], which is usually used in conjunction with precision to assist in evaluating the model's performance and estimate the positive examples in a dataset. Thus, it provides a reference for false negative errors, and this helps to define how well the model identifies the class of interest. Recall is calculated using Eq. 5.2, where FN is False Negative.

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

A combination of Precision and Recall yields an indicator called F1-Score [22]. This indicator helps strike a balance with respect to the previous metrics and is calculated using Equation 5.3.

$$F1score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5.3}$$

The accuracy provides the proportion of all correct predictions relative to the total number of predictions made by the model. Thus, it can be said that it is a global indicator of the model. Accuracy is calculated using Equation 5.4, where TN is True Negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.4}$$

Lastly, the area under the curve (AUC) is a widely-used metric to evaluate the overall performance of a binary classification model, especially in situations where the data may be imbalanced. The AUC metric calculates the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate. The higher the AUC, the better the model's ability to distinguish between positive and negative classes.

5.4 Results and discussions

5.4.1 Synthetic dataset proposal

To evaluate SMLT, the creation of a synthetic dataset was proposed, and subsequently the dataset from the PV installation at the Unicamp Gymnasium was used. The dataset in this study stands out for being generated from real solarimetric data, based on data augmentation strategies. However, it is still referred to as synthetic because it is created by mathematical models, and the anomalies are artificially induced.

In [113], the authors proposed a dataset with I-V curves for detecting faults in PV systems. In practice, acquiring the I-V curve of a single PV module can be done with an I-V curve tracer, but when referring to a PV system, plotting the I-V curve of each module is more complicated. As the I-V curve test is generally performed in the manufacture of the PV module, an application such as [113] is valid in this scenario. However, the project aims to utilize data from solarimetric stations and the inverter, as this data is more accessible and applicable in field settings.

For this work, a synthetic dataset was proposed, obtained from real solarimetric data. The objective was to generate a dataset closer to the real one. To this end,

solarimetric data from the Unicamp station seen in Figure 5.3 were collected, the data was processed, and PV modeling was applied with the PVlib library [129].

Figure 5.3 – Unicamp Solarimetric Station.

The modeling of a PV cell is carried out using a simple diode model [130]. Figure 5.4 shows the representation of the model of a PV cell with one resistance in series and one in parallel. The resistance R_P essentially represents the leakage current present at the P-N junction of the PV cell, while R_S is an equivalent resistance that encompasses the internal resistances of the cell itself [131]. To equate the model according to [132], the thermal voltage (V_T), the number of PV cells that make up the module to be modeled (N_C), where K_I is the temperature coefficient of the short-circuit current, K_V is the temperature coefficient of the open-circuit voltage, and ΔT is the difference between the nominal temperature and the operating temperature. Thus, we obtain the equations from 5.5 to 5.7. Finally, resistances are found through interactions based on data entered from the PV modules.

$$I = I_{PV} - I_0 \left[\exp\left(\frac{V + IR_S}{nV_T}\right) - 1\right] - \frac{V + R_S I}{R_P} \tag{5.5}$$

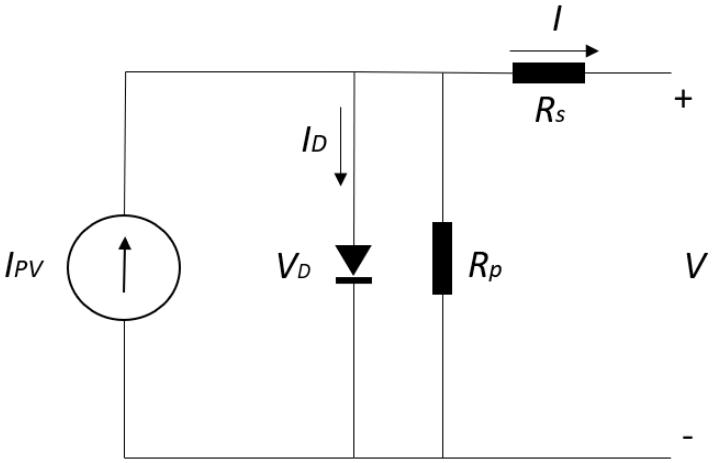


Figure 5.4 – Model of a PV cell with one resistance in series and one in parallel [3].

$$I_{PV} = \left(\frac{G}{G_{(STC)}}\right)\left(I_{PV(STC)} + K_I \Delta T\right) \tag{5.6}$$

$$I_{PV(STC)} = I_{SC(STC)} \frac{R_S + R_P}{R_P} \tag{5.7}$$

$$I_0 = \frac{I_{SC(STC)} + K_I \Delta T}{\exp\left(\frac{V_{OC(STC)} + K_V \Delta T}{n V_T}\right) - 1} \tag{5.8}$$

$$V_T = \frac{N_C k T}{q} \tag{5.9}$$

For the synthetic dataset, the proposal was to build with a single module, a scenario similar to systems with power optimizers [20] or microinverters [133] in which the modules are individualized. The faults inserted were based on the work of [113], with the addition of current reduced for mitigations. Table 5.2 shows the parameters adopted, and the dataset can be summarized in the flowchart in Figure 5.5.

Type of Anomaly	**Condition**
Attenuation	I_real <I_model
Short-Circuit	I_real >0.9*I_model and V_real <0.1*V_model
Open-Circuit	V_real >0.9*V_model and I_real <0.1*I_model

Table 5.2 – Conditions inserted in the data to represent failures.

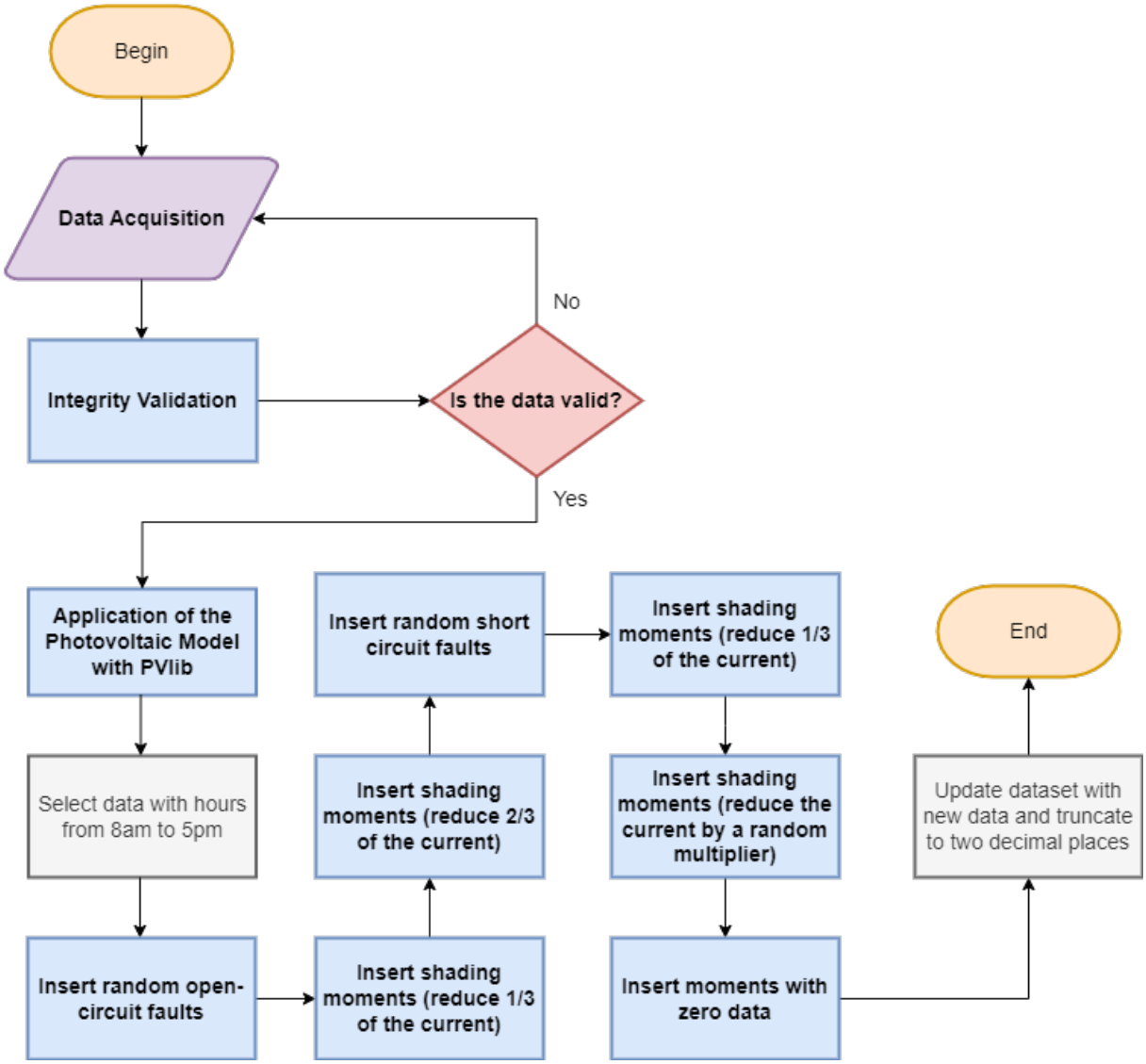


Figure 5.5 – Flowchart for PV dataset generation.

5.4.2 Exploratory Data Analysis based on a photovoltaic system from UNICAMP

The installation chosen for this project was the Unicamp Gymnasium, as it is the largest available with one year of data already obtained (336.96 kWp of installed power). The location of the installation on the roof ensured the absence of objects causing shading in the immediate vicinity, as the height of the roof is a mitigating factor. The installation's component elements are described in Figure 5.6.

Figure 5.6 – PV installation at the Unicamp Gymnasium [33].

The distribution of *strings* in the inverters was shown in Figure 5.7. The inverters were named *A* to *E*, with *D* to *E* having a small difference in the number of PV modules.

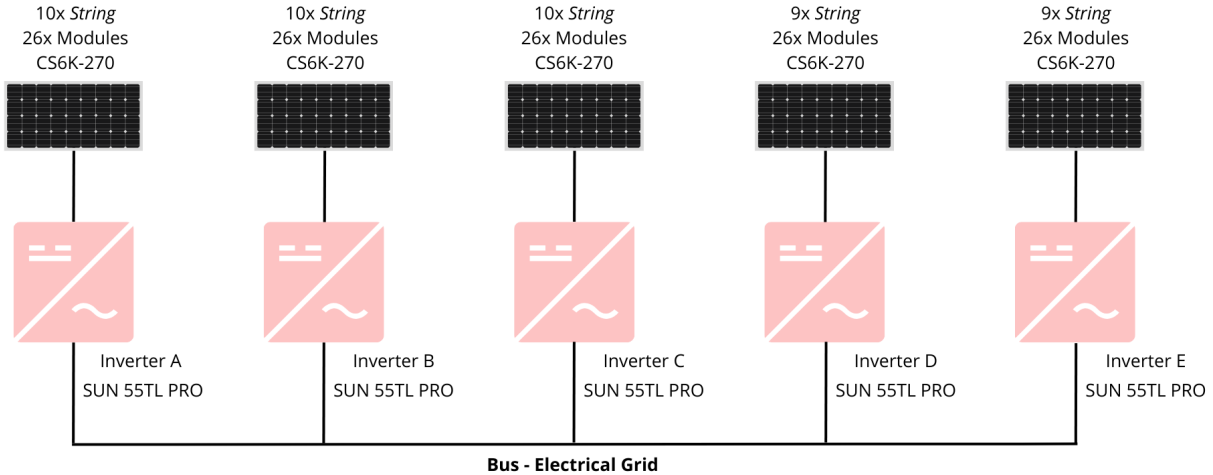


Figure 5.7 – Distribution of PV modules in inverters at the Unicamp Gymnasium.

Before using the data, in order to verify whether the PV system works as designed, the first twelve complete months of energy generation were collected and compared with simulation in the PVsyst software. Through simulation, it becomes feasible to study the production of the PV system, which allows the full optimization of a given plant from

the design phase, through the strategic arrangement of modules in areas with a lower incidence of shading, as well as the selection of more efficient arrangement of strings PV. After installing the system, it is possible to identify possible setbacks if energy generation deviates from the expected amount of energy. To this end, the PVsyst [33, 41] software was used.

It is important to highlight that the uncertainty identified in the literature for grid-connected PV systems must be less than ±30% monthly and less than ±10% annually [62]. Generally, this uncertainty is associated with solarimetric data or errors in the preparation of the simulation. The modeling of PV modules and converters already demonstrates considerable precision [63].

To check the discrepancy between the measured values and those predicted in simulation, the idea of the Estimated Error Matrix (EEM) [134] was adopted. The EEM will generate significant value in situations in which the measured data does not follow the simulated pattern. The equation 5.10 was used, where $\hat{Y_n}$ would be the simulated values and Y_n the measured ones.

$$EEM = \begin{bmatrix} \hat{Y_1} - Y_1 \\ \vdots \\ \hat{Y_n} - Y_n \end{bmatrix}^T \cdot \begin{bmatrix} Y_1^{-1} \\ \vdots \\ Y_n^{-1} \end{bmatrix} \tag{5.10}$$

And Eq. 5.11 was also used to analyze the error between real and simulated data, both for monthly errors and for annual average errors. Positive errors denote that the system generated more energy than predicted in the simulations, which indicates a conservative approach in the simulation. Negative errors indicate that the energy generation predicted by the simulation exceeded the actual generation, which denotes, in this case, an optimistic approach to the simulation. An optimistic simulation has implications for the consumer who has purchased or planned a PV system. The ideal scenario consists of generating energy as promised or slightly above this value. Therefore, the PV software makes it possible to use probability indicators to make the result more conservative [33].

$$Erro(\%) = \left(\frac{Y_i - \hat{Y_i}}{Y_i} \right) \cdot 100 \tag{5.11}$$

Therefore, the result found was presented in Table 5.3, and published in [33]. The data presented demonstrated a higher average annual energy generation in the measured data compared to the simulation. This type of result is interesting, as it indicates that the PV system generated more energy than expected in its first 12 months. Furthermore, the PV system presented an error as expected in the literature. The EEM was close to zero, as would be appropriate, with the exception of the month of March with

the highest EEM (-0.19), but the measured data was higher. Therefore, it is possible to state that PVsyst was more conservative in this installation and that the operation went according to plan during the first year, thus validating the installation and adopting it for the next phases of this work.

Table 5.3 – Comparison of measured data with simulations in PVsyst for the Unicamp Gymnasium [33].

Month	**Energy (MWh)**		**Error (%)**	**EEM**
	Measured	**PVsyst**		
January	48.62	42.60	12.38	-0.12
February	38.92	41.98	-7.86	0.08
March	50.83	41.26	18.83	-0.19
April	42.02	36.88	12.23	-0.12
May	32.98	34.70	-5.22	0.05
June	33.59	31.28	6.88	-0.07
July	35.74	35.87	-0.36	0.00
August	38.32	40.21	-4.93	0.05
September	41.76	40.41	3.23	0.03
October	52.67	42.98	18.40	-0.18
November	48.70	47.94	1.56	-0.02
December	49.47	45.05	8.93	-0.09
Average	**42.80**	**40.10**	**6.32**	-
Total	**513.62**	**481.16**	**-**	-

We proposed a methodology for exploratory analysis of PV datasets in [83], explained here for its importance, and presented in 5.8. The objective was to offer an exploratory analysis capable of addressing the questions previously presented, and contributing to a more in-depth understanding of the data to be used in the detection of anomalies.

To this end, the following key stages of the proposal can be considered: 1) data acquisition; 2) integrity validation (data processing); 3) application of Pearson correlation; 4) annual histograms; 5) boxplot of annual powers; 6) annual energy comparison; and, 7) analysis of results.

Initially, data integrity validation was carried out, which involves checking missing data, inserting zeros in missing data, adjusting times and joining data from different sources and time intervals. In this case, the years 2020, 2021 and 2022 were used. In the case of *dataset*, *GId* are the inverter models available in the base, *Freq* the frequency of the inverter synchronized with that of the electrical network, *Pac(W)* is the output power of the inverter, *Vdc1* is the voltage at the DC input of the inverter, *Idc* is the DC current, *Pdc1* the power in the DC part of the inverter, and *DateTime* is the time at which the data was collected.

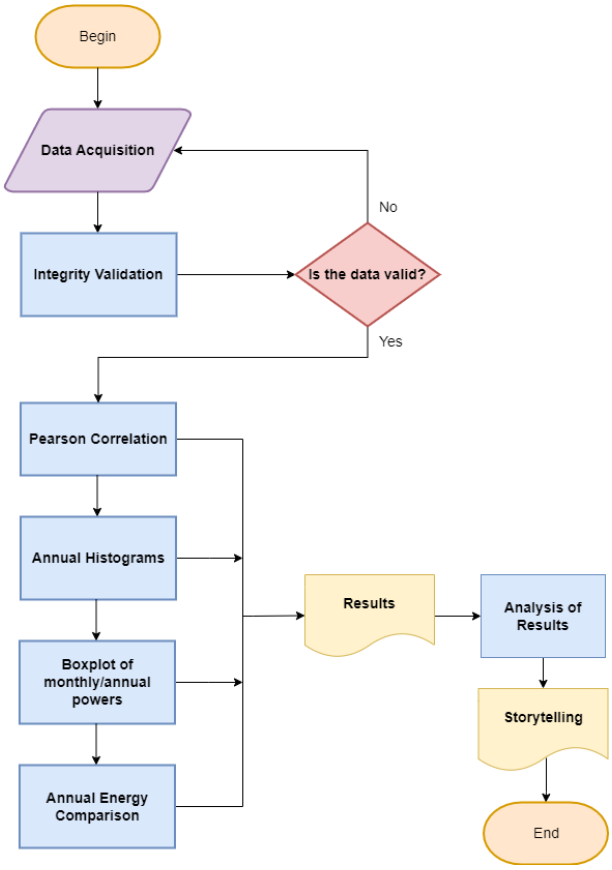


Figure 5.8 – Flowchart for Exploratory Analysis [83].

Subsequently, the first analysis of interest was to apply the *Pearson* correlation. The Pearson correlation was chosen for its ability to establish a measure of dependence between two variables [81]. The Pearson correlation coefficient ranges from -1 to +1, where values close to +1 imply a strong positive correlation between the variables, values close to -1 suggest a strong negative correlation, and values close to zero indicate a weak correlation or no correlation [82].

There are several works that use Pearson correlation as it is a measure already well regarded in the literature, including applications in the area of solar [92]. In [92], for example, correlation helped understand PV modeling and improve the model. The equation 5.12 was applied to find a matrix with the relationship between each variable in a *dataset* with PV data.

$$rcorr = \frac{[n \cdot (\sum x_1 x_2)] - [(\sum x_1) \cdot (\sum x_2)]}{\sqrt{\left[n \sum x_1^2 - (\sum x_1)^2\right] \cdot \left[n \sum x_2^2 - (\sum x_2)^2\right]}} \tag{5.12}$$

By applying the equation 5.12, it was possible to plot the correlation for all inverters together, shown in Figure 5.9. With this matrix, conclusions can now be drawn from the data, that the current has a perfect positive correlation with the DC or AC power; thus, they are proportional. Voltage and frequency have weak or no correlation

with power and current. This is compatible with how the PV inverter should work due to the complex interactions between the electrical characteristics of the PV modules and the PV inverter control process. If the correlation between current and power was not positive, it could be evidence of an anomaly. The correlation matrices for the years 2020, 2021, and 2022 for each inverter were also plotted, but as there were no considerable differences, they were not added to this document.

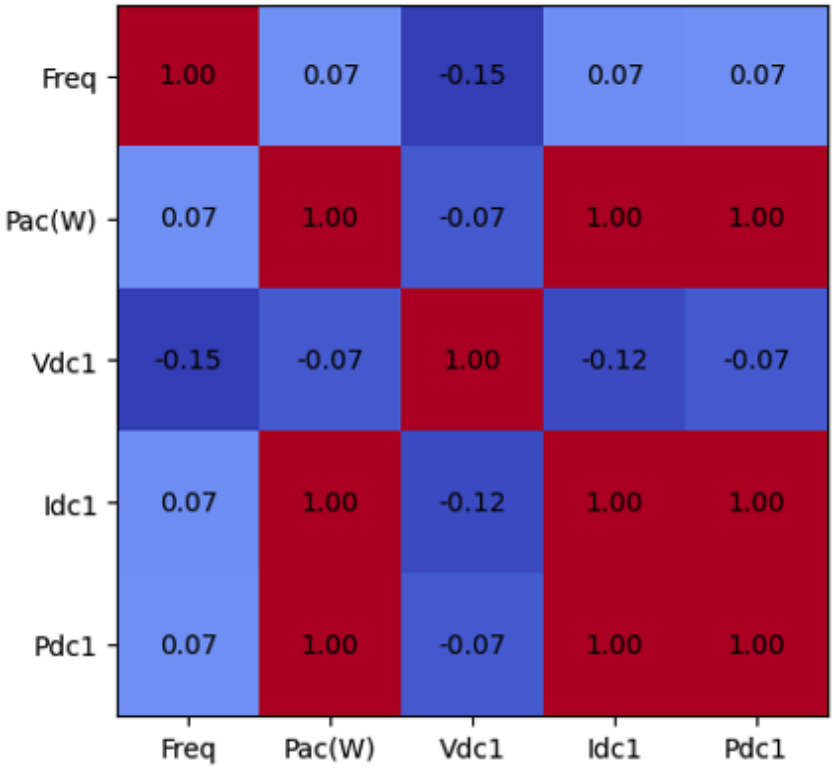


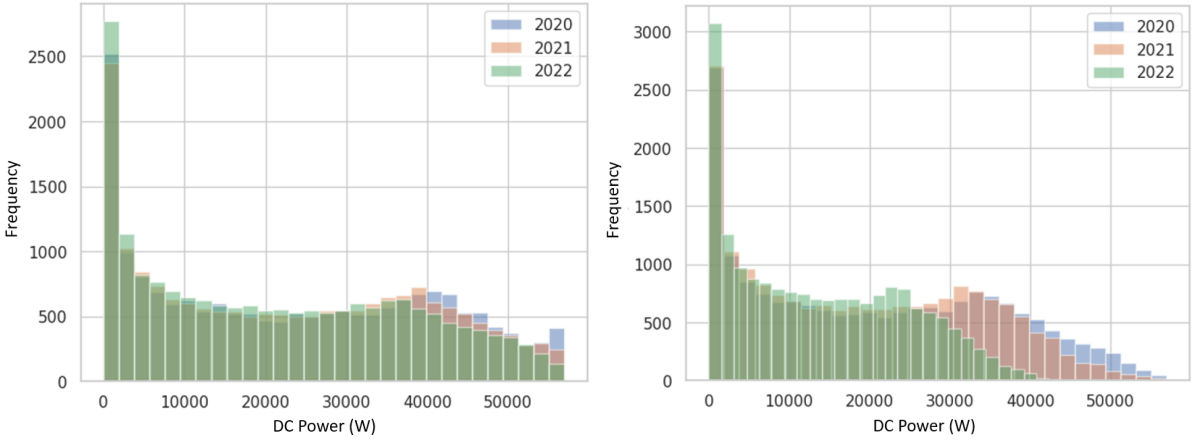
Figure 5.9 – Correlation Matrix of the Unicamp Gymnasium Dataset [83].

The next step was to obtain the annual histograms. The choice to obtain histograms is due to the fact that it is a statistical technique capable of representing the distribution of data within a dataset [95]. This way, one can divide the range of data into smaller ranges or classes and then count the number of observations present in each class. As a result, you have a resulting bar graph that illustrates the frequency of each class, which helps identify patterns, including skewness and kurtosis. This is a good tool for identifying outliers within a data set, such as data collected from a PV inverter.

Several works make use of histograms in PV energy, in addition to data science. In [97], the histogram played an important role in understanding the degradation of PV cells, being a tool to understand the *dataset* used by the authors. In [68], it was mentioned in the final considerations that, through histograms, it was verified that solar irradiance did not follow a normal distribution. With a different approach, in [98], the histogram was used to study the error in training, validation, and testing of ML with neural networks for evaluating solar resources to study the location where to install systems.

In this result for the applied methodology, the histogram was obtained for each inverter with the annual comparison. The results were presented in Figures 5.10 for inverter A (normal) and Inverter D (different behavior from the others). Two important observations were made when analyzing the histograms. Firstly, in the first year, the inverter reached higher power values than in subsequent years, indicating visible degradation, which is expected for a PV system. However, as a second observation, it was clear that inverter D showed a much more considerable reduction in power than the other inverters, being an indication of an abnormality.

Inverters D and E had fewer PV modules than the other inverters; this is a fact to be considered in the investigation. As inverters A to C had many more PV modules than the power of the PV inverter, there was a portion of energy that was not produced throughout the year, known as clipping [64]. This was much lower in inverters D and E, and with normal degradation, the incidence of powers close to the inverter's maximum or greater is reduced. On the other hand, the power drop in Inverter D is still greater than E and needs to be investigated. Other factors, such as shadowing, which did not exist in 2020, may exist in other years.



(a) (b)

Figure 5.10 – Histogram for Inverters A and D of the Unicamp Gymnasium [83].

The next step was to obtain the boxplot of the power. The boxplot technique is a graphical method used to identify atypical values (outliers). This method is notable for not considering extremely atypical values when calculating a measure of dispersion. Internal and external limitations are defined based on quartiles, which prevents extreme values from distorting the analysis [99].

In [68], a boxplot was used to evaluate how much a set of data obeys the normal distribution when dealing with solar irradiance. In [100] it was used to evaluate the detection of possible faults, being allocated to the PV system quality analysis methodology with descriptive statistics. Another work [101] also applied a boxplot to detect faults in

strings of PV systems.

The first boxplots aimed to analyze the power values of all PV inverters throughout the years 2020, 2021, and 2022. As a result, the Figures 5.11 were generated. The box corresponding to 2022 stood out for a significant reduction, which also resulted in the median moving away from the average in all inverters, however, in inverter D, the box was still much smaller. This observation raises an alert for possible anomalies. A drop in irradiance rates may occur, which would cause a reduction in the box in 2022, but that of inverter D reduced in a different proportion to the other inverters.

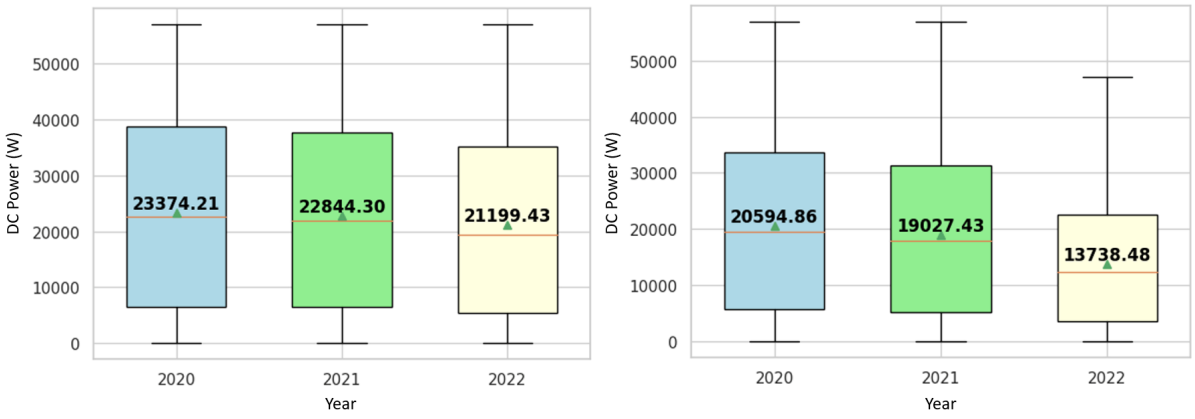


Figure 5.11 – Boxplot of Inverters A and D of the Unicamp Gymnasium [83].

An interesting strategy to deepen the analysis is to plot the monthly *boxplot*. For this, the results of inverter D were selected, which presented a reduced box in the year, thus obtaining Figure 5.12. It can be seen that the reduction in power limits occurred throughout 2022, indicating some anomaly in this inverter D.

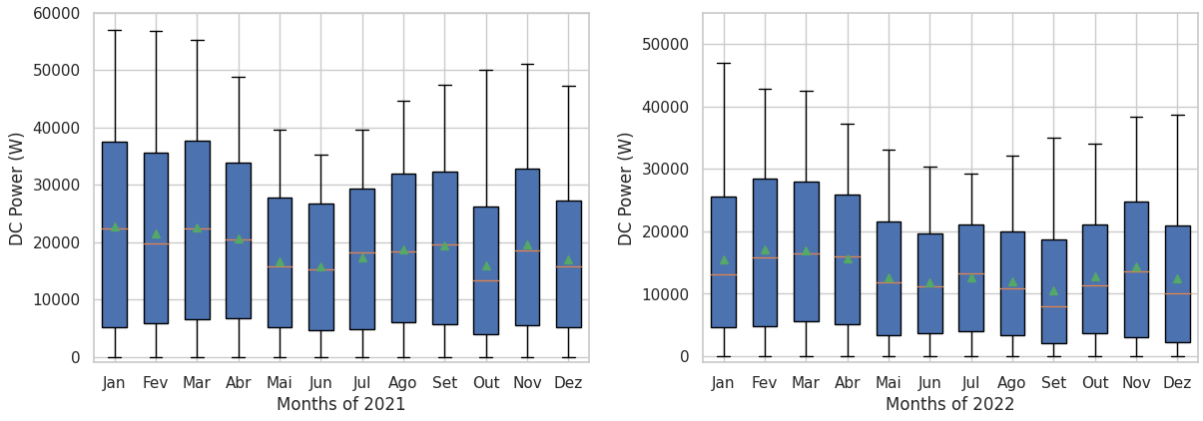
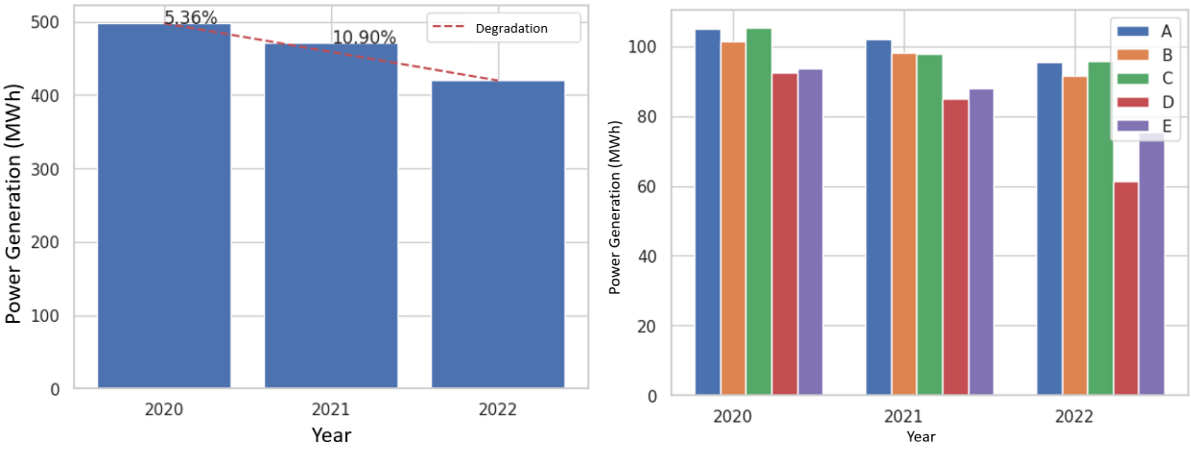


Figure 5.12 – Monthly boxplot of Inverter D for the years 2021 and 2022, respectively.

Finally, it is important to check the PV generation. For this, graphs were

created with the total PV generation in the year (See Figure 5.13-a) and the generation per inverter in the three years (See Figure 5.13-b). In the graph in Figure 5.13-a, a drop of 10.90 % from 2021 to 2022 can be seen. In Figure 5.13-b, again what was seen in *boxplot* is Therefore, inverter D showed a high generation drop (visibly) in 2022. The results show a problem in the PV system connected to inverter D, requiring an on-site analysis.



(a) (b)

Figure 5.13 – (a) Total Energy Generation for PV installation at the Unicamp Gymnasium and (b) Total Energy Generation for each PV Inverter at the Unicamp Gymnasium [83].

5.5 Proposal of a new process flow for a Supervised Machine Learning Technique

Detecting and classifying anomalies in PV systems presents several challenges. For example, there is the challenge of the nature of the collected data, since PV systems often generate a large amount of information, with measurements on the inverter throughout the year, at frequent intervals, such as every 15 min in this case. This results in the creation of unbalanced datasets, as most observations do not constitute anomalies. Thus, a challenge is to find algorithms capable of dealing with this inequality in class distribution, so that anomalies are not obscured by the volume of non-anomalous data.

On the other hand, it is important to highlight that failures often exhibit behavior that can be identified through graphical analysis and data studies by experienced professionals. This is feasible due to the presence of discernible patterns in the data, when this happens the performance of the algorithm will be less affected, as the minority classes (anomalies) are linearly separable, even with unbalanced data [135, 136].

Therefore, it is necessary to evaluate several algorithms in the literature for classification and propose a type of algorithm that meets the purpose. Some requirements

that are important are the accuracy in detecting/classifying anomalies; interpretability of the model, as it is necessary to understand the reasons for choosing the classification; mitigate the occurrence of false positives; and future real-time operation is possible.

For this work, it was proposed to use the synthetic test base of Figure 5.5 explored in the methodology (STB - Synthetic Test Base), and the Unicamp base (Inverter A) with short-circuit and open circuit faults inserted on the training/test base (GTB - Gymnasium Test Base). The *Logistic Regression* [121], *Decision Tree* [123], *Random Forest* [137], and a new process flow was proposed, and algorithms were studied.

The methodology for the new process flow included an *ensemble* system (technique that combines results from several models) of *Random Forest* with *K-nearest neighbors* (k-NN) finished with inference machine. The *ensemble* system implemented was VotingClassifier [138] to compensate for the individual weaknesses of each model. This system takes the average of the probabilities and selects the class with the highest average probability for each data point.

The choice for *Random Forest* with k-NN occurred after several tests. K-NN stands out for its sensitivity in identifying anomalies, as it can distinguish data samples, in addition to being very simple to interpret. Random Forest stands out for its ability to be robust to *overfitting* and can adapt to different classification tasks due to decision trees.

The methodology is illustrated in Figure 5.14. Basically, after filtering the data, a *Synthetic Minority Oversampling TEchnique* (SMOTE) is applied [139], a technique to balance the data from the minority classes with the majority, this is of paramount importance due to the nature of the data from PV systems that There is a class imbalance. Subsequently, the models that will be used in *VotingClassifier* are defined and trained, after which they are applied to the classes. Then, the inference machine comes into action, applying the conditions of OC, SC, and Undefined (or N/A), to eliminate incorrectly classified data in these classes since these are classes that have pre-defined criteria. Finally, there is the classification report.

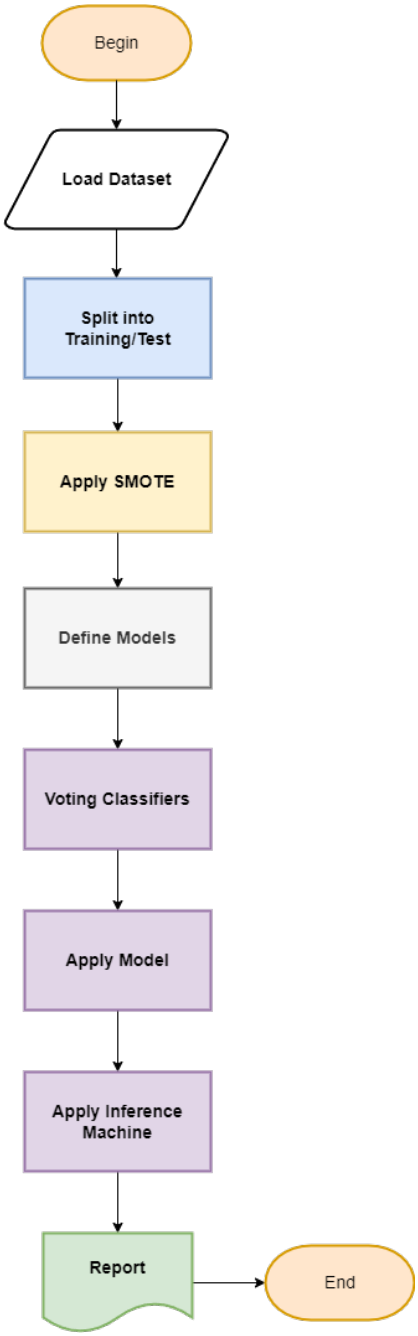


Figure 5.14 – New process flow proposal for an SMLT applied to anomaly classification for PV systems.

The algorithms then use data from each PV inverter to facilitate the identification/classification of anomalies, since in PV installations it is possible to obtain data from each inverter. However, evaluating the methods is a challenge since anomalies are minority classes compared to normal data (in real datasets), and algorithms that get more correct results on normals can score high in performance metrics. Therefore, class 1, 2, 3, 4 anomalies with recall were analyzed separately. Furthermore, the AUC stands out, which when higher indicates that the model is capable of more effectively distinguishing between instances of positive and negative classes.

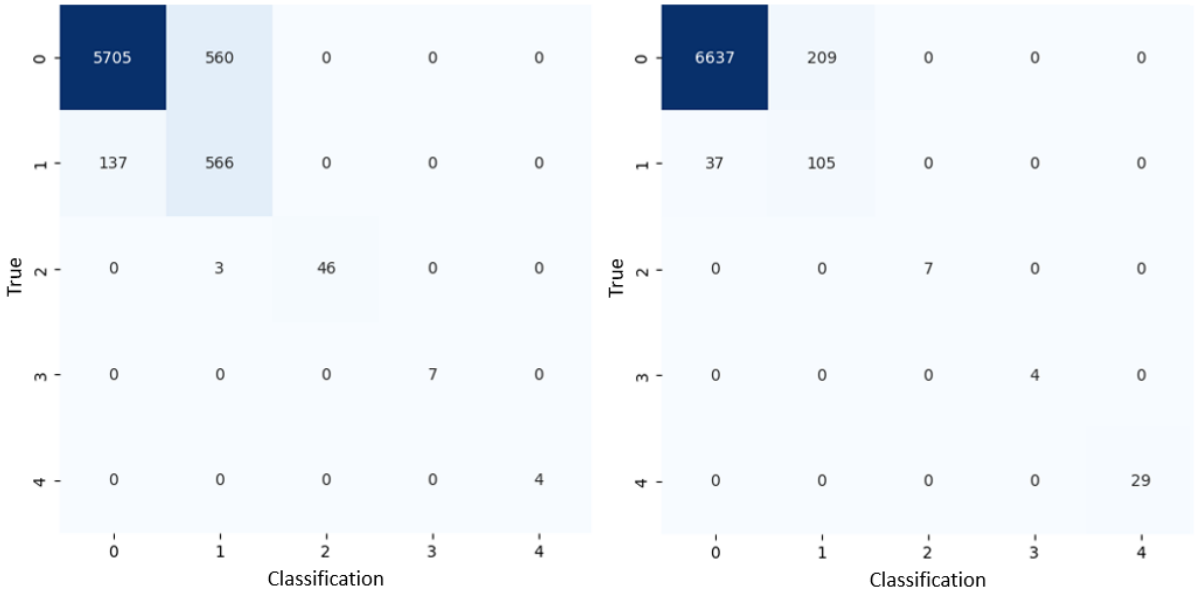
In addition, during the accuracy assessment, the cross-validation technique was employed to test the algorithm's behavior in classification [22], mainly the model's generalization capability. In this method, the two datasets, STB and GTB, are divided into k parts (folds), and the model is trained k times, each time using *k-1* parts for training and the remaining part for validation. Subsequently, the accuracy is calculated along with the average difference obtained in each of the k iterations; in this case, k was set to five.

As a result, one can check the metrics presented in Table 5.4 and confusion matrices in Figure 5.15. The proposed process flow was able to detect class 2, 3, and 4 failures in all cases, in addition to being better in class 1 than standard algorithms in the literature (in real dataset). The AUC of the proposed model outperformed other algorithms in all tests. The accuracy obtained with the cross-validation technique was also higher than the other methods in GTB, in addition to lower deviation. For STB, the accuracy was slightly lower than that proposed in relation to the Random Forest model. This difference in accuracy between the two predictions is related to the distribution of classes and the way in which the model is making trade-offs between FN and FP. In some cases, it is more important to minimize FN, while in others, minimizing FP is the priority, depending on the problem requirements. In this case, the synthetic-based model had more FP.

It is worth mentioning that the Logistic Regression (LR) model in GTB assigned zero values to two classes, as evident in R1 and R2. Nevertheless, it achieved a remarkably high level of accuracy by correctly identifying a significant portion of the normal data. This underscores the importance of using multiple metrics when working with anomaly classification. In the literature, it is quite common to rely solely on accuracy. In this sense, an Average Recall (R_{AVG}) for the anomalous classes was also presented in the Table 5.4. This average recall showed that the proposed process flow outperformed the others, presenting an average recall of 0.9360 and 0.9349 for STB and GTB respectively.

Algorithm	Dataset	Accuracy	Precision	Recall	F1-score	AUC	R1	R2	R3	R4	R_{AVG}
LR	STB	0.9083±0.0118	0.8949	0.9084	0.8989	0.9382	0.3570	1.0000	1.0000	0.7500	0.7768
	GTB	0.9542±0.0007	0.9324	0.9511	0.9275	0.7045	0.0000	0.0000	1.0000	0.0051	0.2513
DT	STB	0.9080±0.0031	0.8942	0.9104	0.8965	0.9282	0.3001	0.9796	1.0000	1.0000	0.8199
	GTB	0.9521±0.0057	0.9059	0.9488	0.9266	0.8046	0.0000	1.0000	0.0000	0.0000	0.2500
RF	STB	0.9188±0.0125	0.9202	0.9227	0.9213	0.9498	0.5747	1.0000	1.0000	1.0000	0.8937
	GTB	0.9512±0.0135	0.9430	0.9552	0.9416	0.8114	0.4577	1.0000	1.0000	0.0205	0.6196
Proposed	STB	0.9001±0.0005	0.9292	0.9004	0.9102	0.9815	0.8051	0.9388	1.0000	1.0000	0.9360
	GTB	0.9647±0.0004	0.9812	0.9650	0.9714	0.9861	0.7394	1.0000	1.0000	1.0000	0.9349

Table 5.4 – Metrics obtained from the tests carried out, with accuracy with cross-validation.



(a) (b)

Figure 5.15 – (a) Confusion Matrix for Synthetic Base and (b) Confusion Matrix for Real Base.

After classifying anomalies, it is possible to generate a file indicating these anomalies, as well as a graph to visualize anomalous points during energy generation. Figure 5.16 presents an example with four days of the STB, illustrating the behavior of this dataset. According to the definitions of anomaly types in Table 5.1, there are several instances of type 1 anomalies, which are typically caused by shading, type 2 anomalies indicating open circuits, and type 3 anomalies representing short-circuit faults at a specific moment on 26th April 2020.

In Table 5.5, the proposed process flow is compared with results from the literature that involve fault (anomaly) detection in PV systems. The studies vary in the algorithm analyzed and the dataset used. Some employ the I-V curve while others use data from the PV inverters. The I-V curve data is more difficult to obtain in practice for a PV plant, and these algorithms are more suitable for use during the PV module manufacturing process. Moreover, each study investigates different types of classes. Overall, it is observed that the result of the proposed algorithm is consistent with the literature.

The proposed process flow for ensemble SMLT, as well as those studied, can be applied to data downloaded from measurement equipment in PV plants during a scheduled period. As the application is not real-time, time and processing limitations are not a problem for the model. Furthermore, the algorithms have a fast execution time of just over a minute and can operate on contracted remote servers, such as Amazon and Google environments, the latter used in the present work. In the proposed process flow, the longest training/testing time was 96.32 s, and considering only the testing phase, it

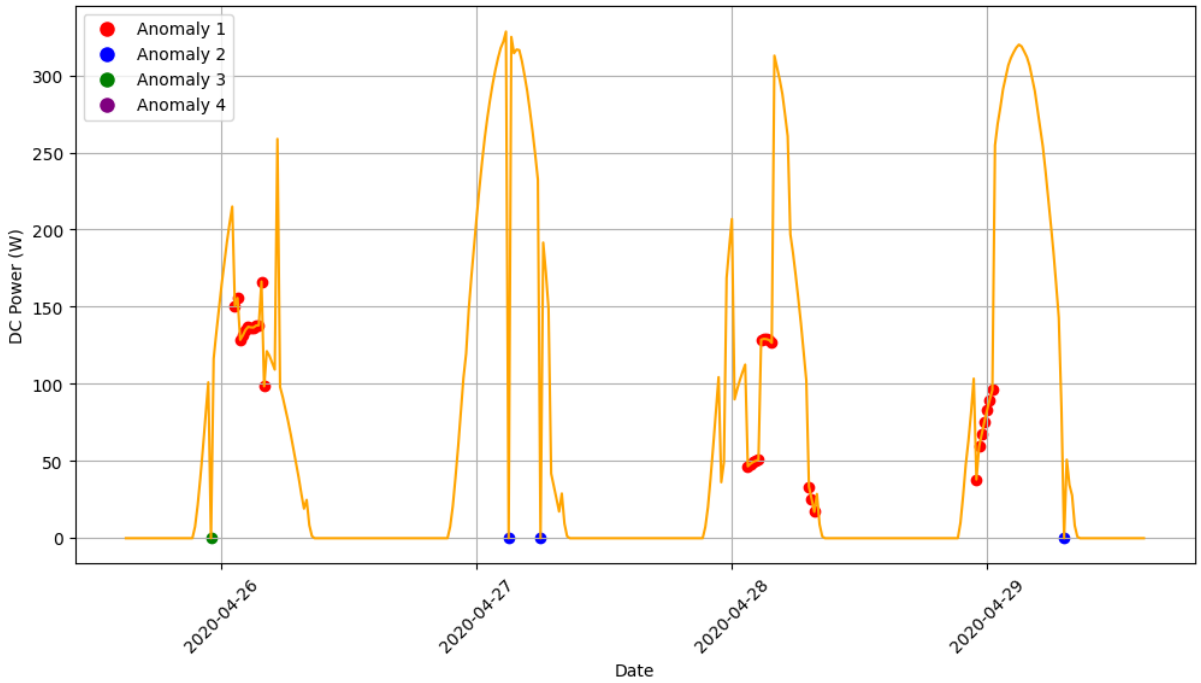


Figure 5.16 – Example of the four-day power curve after anomaly classification in the STB.

was 3.88 s, as shown in Table 5.5.

Table 5.5 – Comparison of the proposed method with literature works. Adapted from [113].

Method	**Anomalies**	**Accuracy**	**Notes**	**Ex. time**
Proposed	Normal, Attenuation, SC, OC, Undefined	0.9647±0.0004	With Real Dataset	3.88 s
Voutsinas et al. [113]	Normal, SC, OC, Mismatch, Undefined	0.9711	I-V Curve Data	8 ms
Voutsinas et al. [140]	Normal, SC, OC, Mismatch, Undefined	0.934	I-V Curve Data, NN	44 ms
Chen et al. [141]	Normal, SC, OC, Mismatch	0.9880	I-V Curve Data	-
Harrou et al. [142]	Normal, SC, OC, Mismatch	MAPE: 2.83%	DC Current Data, no use ML	-
Yi and Etemadi [143]	Line-to-Line Fault Detection	0.9474	Uses SVM	-
Xia et al. [144]	Arc fault Detection	0.96	Uses SVM	-
Harrou et al. [110]	SC, OC, Intermittent Faults (IF)	Minor Class 0.896	Unsupervised, One-Class SVM	-
Wang et al. [145]	Mismatch, SC	0.9778	Multiclass SVM	-
Winston et al. [146]	Hotspot, Microcrack	M1: 0.87 - M2: 0.99	NN	-
Yi and Etemadi [147]	SC	97.69 to 100%	Multi-Resolution Signal / Fuzzy	-
Memon et al. [148]	SC, OC, Mismatch	0.952	Convolutional NN	160–70 ms
Jia et al. [149]	Arc fault Detection	1.00	Logistic regression	-
Fadhel et al. [150]	Shadings	Minor Class 0.88	I-V Curve Data, Uses PCA	-
Dai et al. [151]	Aging, Dust, SC, OC, Inverter protection	0.966	Deep Reinforcement Learning	-

One of the disadvantages and challenges of the proposed process flow, primarily due to its supervised nature, is the need for a classification training dataset from a location near the PV installation. This is because climatic differences affect the PV system, consequently influencing the behavior of the data. This is one reason why it is interesting to have a synthetic dataset; however, even to generate a synthetic dataset, a solarimeter station near the analyzed PV system is necessary. Therefore, the disadvantages to be observed include the need for data, difficulty in generalizing the model training, and constant updating of training to maintain performance. Once these challenges are overcome, the supervised model will have a higher success rate than other models, if well designed.

5.6 Conclusions and future works

In this paper, a synthetic PV power generation dataset was proposed using real irradiance data collected from a solarimetric station. To this end, it was necessary to apply mathematical models of irradiance and PV cells. The use of real irradiance data contributed to a more reliable representation of the data. And, the use of synthetic datasets proved to have an important role, as PV data are often sensitive or difficult to obtain.

For real data, exploratory collection and analysis was performed before use in the SMLT model. The data were from a facility located at UNICAMP and a flowchart was proposed for exploratory analysis. Inverter D presented a significantly different energy generation than the others. This fact probably contributed to the drop in overall generation of 10.90% from 2021 to 2022.

With the two datasets, classic SMTL methods were applied, in addition to the proposed process flow, the main objective of the work. The proposed process flow for ensemble SMLT was noted for achieving an AUC of 0.9815 for the synthetic dataset and an AUC of 0.9861 for the real dataset, with an accuracy of 0.9647, the best result among algorithms for the real dataset. Furthermore, it demonstrated the significance of comprehending the purpose of implementing the algorithm and, consequently, examining various metrics to make an good selection.

Thus SMTL proves to be effective in classifying anomalies, being a possibility for application in PV plant analysis tools (framework). As future work, a pilot plant will be implemented for further testing with SMLT. This plant will allow the incorporation of other unconventional sensors to generate more features to improve anomaly classification and potentially create additional classes. One of these sensors is the rear-side measurement of PV modules at different points on the module. Additionally, the new process flow will be implemented for analyzing real-time data downloaded from PV systems. However, the classification analysis is conducted programmatically and takes place after data collection.

Supplementary material

Information of interest of the models and datasets applied

Table 5.6 – Information of interest of the models and datasets applied

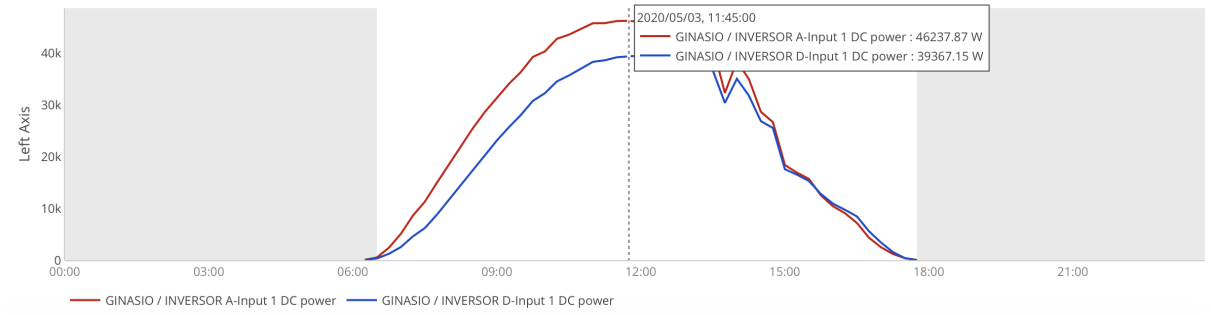
Information	**Definition/Numbers**
Number of rows in Real DataSet	35136
Number of rows in Synthetic DataSet	35136
Step	15 min
Imblearn (SMOTE)	Default
New data with Synthetic SMOTE	124960
Data Split	80/20
Number of Cross Validation	5
n_estimators (Random Forest)	100
Criterion (Random Forest)	Gini
Number of Neighbors (kNN)	5
Weight Function (kNN)	Uniform
Metric (kNN)	minkowski
VotingClassifier Type	Soft

Anomalies in the PV Inverter D of the Unicamp Gymnasium

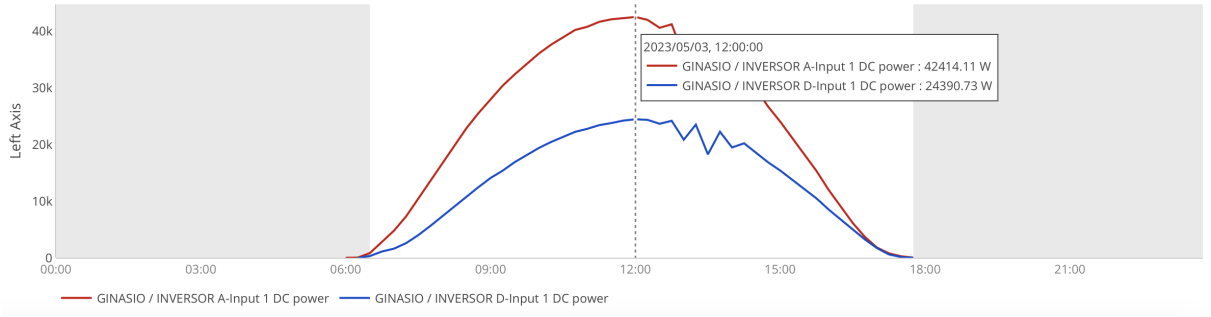
The inverter D exhibited a very significant attenuation in the exploratory analyses of this study. This showed that, solely through exploratory analysis, it is possible to identify anomaly behaviors in PV inverters. Prior to this work, the issue was not noticed, as the analysis is generally done on total generation, since there are many systems on site. This is also a behavior adopted in the PV solar energy market.

With a possible issue identified, we also sought to plot power versus time graphs for some days in 2020, 2023, and 2024, thus comparing them. It is possible to notice a significant difference that did not exist in the year 2020 from inverter A to inverter D (Figure 5.17 and 5.18). Since this inverter model has a single MPPT, this difference indicates a possible reduction due to issues in the PV module strings. Thus, this work indicates that the strings of inverter D should be evaluated through an I-V Curve test.

In this scenario, the proposed process flow was not applied for the year 2023 and for inverter D, as there were no POA data available for other years.

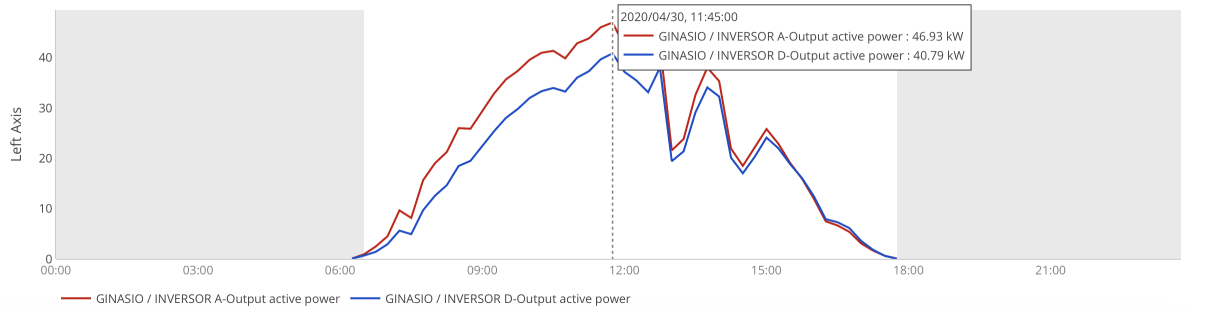


(a)

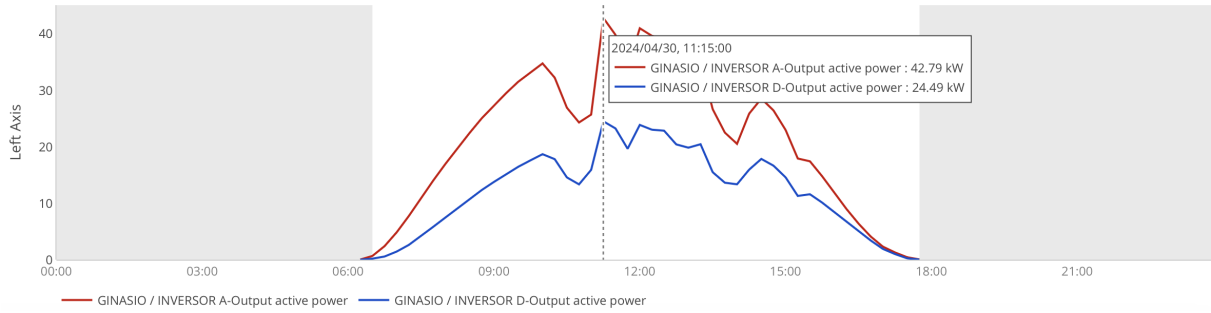


(b)

Figure 5.17 – Power generation data with good irradiance to the Unicamp Gymnasium: (a) 2020 and (b) 2023.



(a)



(b)

Figure 5.18 – Power generation data with shadows from clouds to the Unicamp Gymnasium: (a) 2020 and (b) 2024.

6 Discussion

It is evident the interest in the PV solar energy theme, mainly due to the growth in the last decade in Brazil and worldwide. Thus, several research studies have been proposed by private and governmental sectors. One of the challenges faced in the PV solar energy market is to analyze and identify issues. This thesis presented a flow of various analyses and contributions relevant to the theme. The articles presented, above all, encourage the routine of data analysis of PV installations and provide insights that can aid in these analyses.

In Chapter 2, the case study of the "Sustainable Campus" project at Unicamp was presented. This project enabled a living laboratory within the university for analyses and generated various contributions within the university and also for the external community. It was possible to inspect several PV installations and compare them with simulation software, studying the software's behavior and the energy generation of the plant. Subsequently, with all the data, it was possible to understand the behavior of each PV facility nearby and how data from those facilities relate. Finally, concluding with an economic analysis of the PV system facing a client from the free market. In Brazil, there were few customers in the year of the study (and only one Brazilian university), mainly in the public sphere, making it a potential study for comparisons.

All the study presented in Chapter 2 serves as a reference for universities in Brazil and worldwide that wish to be more sustainable. With the results and contributions, including those from the PV solar energy group, Unicamp stood out in the UI GreenMetric World University ranking. Additionally, throughout the study, it was possible to teach PV solar energy to the external community of Unicamp, disseminating the project and knowledge.

Recognizing the importance of using actual solarimetric data, Chapter 3 conducted a study utilizing solarimetric data from the "Sustainable Campus" station to simulate in PVsyst and compare it with the simulation result using Meteonorm data (available in PVsyst). This approach revealed that incorporating local solarimetric data significantly enhanced the accuracy of PVsyst simulations, resulting in a discrepancy of just 0.21% when compared to real energy generation.

It was also possible to perform some statistical analyses, such as the correlation of POA irradiance with plant data. From this, it was possible to contribute by demonstrating that using POA irradiance is necessary for PV system studies and that it would be important to use POA irradiance as a feature in models. Additionally, it is important

whenever possible to use real solarimetric data in simulations to analyze performance and, consequently, simulate the return on investment in PV systems.

In Chapter 4, an exploratory methodology was proposed to standardize initial studies when dealing with PV data. This methodology serves as a macro analysis of the PV system. Several papers were analyzed, and interesting statistical methods for the theme were collected. A flowchart was proposed and applied. Various insights were observed with the proposed exploratory methodology, culminating in the detection of a power attenuation problem in two inverters, mainly in Inverter D of the Unicamp gymnasium. Thus, it was demonstrated that exploratory analysis is a necessary step that should be carried out in PV systems and monitoring programs. The methodology will be implemented in a future framework.

With the importance of the POA irradiance feature detected in the study and the exploratory analysis, the study proceeded to anomaly classification. Chapter 5 tested algorithms from the literature and proposed a process flow for anomaly classification. The model was an ensemble of Random Forest with kNN and an inference machine. The proposal performed well for both synthetic annual datasets and real data. A key highlight is the ability to classify data with a single algorithm, as there are works that apply methods at different stages. However, future scalability tests are still necessary, not only for the proposed process flow but also for the algorithms from the literature.

It is also worth noting that in Chapter 5, the creation of a synthetic dataset was proposed, which was generated with real solarimetric data, but with randomly assigned faulty inverter data. This is important for training algorithms when data is not available, as well as for studying PV system data.

Therefore, this work, with its composition of papers, brought together various information, insights, and warnings about the analysis of PV systems, focusing on anomaly classification in PV systems. As a recent and hot topic, there are several future projects needed and gaps to be filled with new work.

7 Conclusion

The main objective of this research was to examine and classify anomalies in PV systems using a supervised approach. However, to achieve this objective, it was necessary to carry out advanced studies on PV systems, which involved simulations, implementations and data acquisition. This step was necessary to formulate a data set and develop the necessary specialized knowledge on the subject.

The project's first notable results were a long study based on the PV systems integrated at Unicamp with the "Sustainable Campus". Over the course of a year, actual energy generation reached 784.29 MWh, surpassing simulation results of 759.04 MWh for the same system. PV installations have helped provide datasets for these and other projects. Furthermore, the technologies developed, the licensed simulation tool and the spreading knowledge via training and workshops establish the necessary foundations to promote the expansion of PV plants at Unicamp and to conduct in-depth research on the topic.

Subsequently, it became essential to evaluate the available solarimetric data, which would be used to classify anomalies as a possible characteristic. One of the challenges was also to examine the influence of real data close to the installation on simulations, comparing them with meteornorm data and effective generation information. Utilizing solarimetric data from the local environment proved to improve the accuracy of PVsyst simulations, resulting in a minimal discrepancy of only 0.21% when compared to actual energy generation. Thus, this step showed the importance of using real solarimetric data to carry out simulations before implementing the PV system.

In the subsequent phase, with the data available, a gap was found in the literature regarding the identification of specific methodologies for the preparation and analysis of data from PV systems. Given this lack, a methodology was proposed based on studies in specialized literature. With the implementation of this methodology, the results obtained for the PV plant under analysis revealed a reduction in energy generation over the years, possibly attributed to degradation or accumulation of dirt in the modules, as well as the correlation between variables in the inverter model studied.

At the end of the process, the properly prepared and analyzed data were used to apply supervised anomaly classification methods. Additionally, a synthetic dataset, based on the PV cell model and real solarimetric data, was created to be tested in parallel. These datasets were then used in the different testing and training procedures. The final model was a Random Forest set with K-nearest neighbors (k-NN) and an inference machine to

specific classes tested compared with methods from the literature. The results suggest that the proposed methodology effectively categorized the anomalies, achieving an AUC of 0.9815 for the synthetic dataset and an AUC of 0.9861 for the real dataset. Furthermore, it demonstrated an accuracy of 0.9647 for the real dataset.

Therefore, this work made it possible to demonstrate that the preparation of datasets and analytical approaches are fundamental to understanding information related to PV systems. These are important steps before applying ML algorithms. The data exploration stage alone is capable of offering insights and checking for anomalies. In this way, it is evident that applying a thorough macro analysis, and potentially making automatic decisions based on these analyses in the future, has the potential to assist PV installations in detecting anomalies, even without the use of ML methods in cases where training algorithms is challenging.

Furthermore, in micro analysis, it showed that the classification of anomalies in PV systems is viable with the proposed process flow, identifying faults and restricting the scope of action of the maintenance team. However, for this type of classification the model must always be trained with labeled data. It is necessary to train the algorithm for systems with characteristics different from those studied, as well as to evaluate other regions and technologies. Consequently, this study provides valuable insights and underscores the importance of data science as an ally in monitoring PV systems throughout their operational life.

7.1 Future perspectives

Some of the future perspectives for continuing contributions related to this work are:

- Creating a framework with an inference machine capable of, upon inserting a dataset, generating a report with the proposed exploratory (statistical) approach;
- Test the proposed process flow in system with high shadow intensity;
- Test the proposed process flow for classifying anomalies in Unicamp's new PV systems after this work;
- Test the proposed process flow anomaly classification, as well as literature methods in PV systems with optimizers and microinverters;
- Test the proposed process flow in systems with bifacial modules and trackers;
- Test the proposed process flow in system with high shadow intensity;

- Test the proposed process flow for classifying anomalies in systems in other regions of Brazil;
- Implement and evaluate the proposed flow process for anomaly classification in environments such as Amazon SageMaker with large datasets;
- Evaluate the proposed process flow by training on systems of different power levels and locations from the test scenario, verifying the model's potential to work with diverse data;
- Discover and analyze new classifications within PV system datasets by adding new features (such as new sensors in PV plants);
- Serve as a reference for future implementations of PV systems in universities seeking sustainability.

7.2 Scientific Publications and Licenses

Papers Published in Journals

- **de Souza Silva, J. L.**; Mahmoudi, E.; Carvalho, R. R. M.; Barros, T. A. S. Classification of anomalies in photovoltaic systems using supervised machine learning techniques and real data. Energy Reports, vol. 11, pp. 4642-4656, Jun. 2024. [102]
- **de Souza Silva, J. L.**; Barbosa de Melo, K.; dos Santos, K. V.; Yoiti Sakô, E.; Kitayama da Silva, M.; Soeiro Moreira, H.; Bolognesi Archilli, G.; Ito Cypriano, J. G.; CAMPOS, R. E.; Pereira da Silva, L. C.; Gradella Villalva, M. Case study of photovoltaic power plants in a model of sustainable university in brazil. Renewable Energy, v. 196, p. 247–260, 2022. ISSN 0960-1481. [33]
- **de Souza Silva, J. L.**; Moreira, H. S.; dos Reis, M. V. G.; Barros, T. A. dos S.; Villalva, M. G. Theoretical and behavioral analysis of power optimizers for grid-connected photovoltaic systems. Energy Reports, v. 8, p. 10154–10167, 2022. ISSN 2352-4847. [20]
- Moreira, H. S.; **de Souza Silva, J. L.**; Gomes dos Reis, M. V.; de Bastos Mesquita, D.; Kikumoto de Paula, B. H.; Villalva, M. G. Experimental comparative study of photovoltaic models for uniform and partially shading conditions. Renewable Energy, v. 164, p. 58–73, 2021. ISSN 0960-1481. [56]
- Barbosa de Melo, K.; Kitayama da Silva, M.; **de Souza Silva, J. L.**; Costa, T. S.; Villalva, M. G. Study of energy improvement with the insertion of bifacial modules and solar trackers in photovoltaic installations in brazil. Renewable Energy Focus, v. 41, p. 179–187, 2022. ISSN 1755-0084. [152]
- Reis, M. V. G. d.; Narváez, D. I.; Moreira, H. S.; **de Souza Silva, J. L.**; Barros, T. A. S.; Villalva, M. G. A robust islanding detection method for inverter-based distributed generation systems using dc-link voltage perturbation. IEEE Journal of Photovoltaics, v. 12, n. 6, p. 1559–1566, 2022. [153]

Book chapters published:

- **de Souza Silva, J. L.**; Melo, K. B. ; Carneiro, R. K. ; Santos, K. V. ; Vieira, F. C. ; Villalva, M. G. Inserção de Energia Solar Fotovoltaica na UNICAMP. Campus Sustentável: um modelo de inovação em gestão energética para a América Latina e o Caribe. 1ed.Rio de Janeiro: Synergia Editora, 2021, v. 1, p. 124-137. [154]

Software Registration (license):

- Melo, K. B. de; Kitayama, M. S.; **de Souza Silva, J. L.**; Moreira, H. S; Sakô, E. Y.; Villalva, M. G. PVCAMPUS. Universidade Estadual de Campinas e Companhia Paulista de Força e Luz. BR 51 2022 000374 0.

Papers Published in Proceedings:

- **de Souza Silva, J. L.**; Paula, J. F. S. de; Silva, J. A. F. G. da; Barros, T. A. S.; Mahmoudi, E.; Villalva, M. G. Evaluating the significance of solarimetric data for photovoltaic system simulation in a real-world case. In: Proceedings of IEEE 8th Southern Power Electronics Conference (SPEC), 2023. [66]
- **de Souza Silva, J. L.**; Cavalcante, M. M.; Martins, S. B.; Silva, E. J.; Barros, T. A.; Villalva, M. G. Data-driven analysis of solar photovoltaic systems: Correlation and distribution patterns. In: Proceedings of IEEE 8th Southern Power Electronics Conference (SPEC), 2023. [83]
- Cavalcante, M. M.; **de Souza Silva, J. L.**; Martins, S. B.; Nunes, I. F. S.; Ribeiro, A. C.; Barros, T. A. S. Comparison and Application of Data Science Techniques for Anomaly Detection in Photovoltaic Systems. In: Proceedings of IEEE 8th Southern Power Electronics Conference (SPEC), 2023. [155]
- **de Souza Silva, J. L.**; Melo, K. B. de; Costa, T. S.; Machado, G. M. V.; Moreira, H. S.; Villalva, M. G. Impact of bifacial modules on the inverter clipping in distributed generation photovoltaic systems in brazil. In: 2021 Brazilian Power Electronics Conference (COBEP). [S.l.: s.n.], 2021. p. 1–6. [64]
- Lima, G. P. de; Prym, G. C. S.; Melo, K. B. de; Moreira, H. S.; **de Souza Silva, J. L.**; Filho, E. R.; Villalva, M. G. Thermal mathematical modeling of photovoltaic inverters and experimental validation. In: 2021 Brazilian Power Electronics Conference (COBEP). [S.l.: s.n.], 2021. p. 1–7. [156]
- Costa, T. S.; Rosolem, M. de F.; **de Souza Silva, J. L.**; Villalva, M. G. An overview of electrochemical batteries for ess applied to pv systems connected to the grid. In: 2021 14th IEEE International Conference on Industry Applications (INDUSCON). [S.l.: s.n.], 2021. p. 1392–1399. [157]
- Souza De Paula, J. F.; Pinheiro De Lima, G.; CERBATTO, G.; PRYM, S.; **de Souza Silva, J. L.**; Barbosa De Melo, K.; GRADELLA, M. Análise Comparativa De Desempenho De Um Sistema Fotovoltaico Simulado Com As Ferramentas Pvsyst

E Sam (System Advisor Model). IX Congresso Brasileiro de Energia Solar, 2022. [158]

- Prym, G. C. S.; Lima, G. P. de; **de Souza Silva, J. L.**; Neves, M. R. M.; Barros, T. A. d. S.; Villalva, M. G. Estudo Da Corrente De Fuga E Seus Efeitos Em Inversores Baseados Na Topologia Full-Bridge Com E Sem Transformador Para Aplicações Fotovoltaicas. IX Congresso Brasileiro de Energia Solar, p. 10, 2022. [159]

Bibliography

1 HERNANDEZ-CALLEJO, L.; GALLARDO-SAAVEDRA, S.; ALONSO-GOMEZ, V. A review of photovoltaic systems: Design, operation and maintenance. *Solar Energy*, v. 185, p. 401–415, 2019. Available from Internet: <https://doi.org/10.1016/j.solener.2019.06.017>.

2 SPERTINO, F.; CORONA, F. Monitoring and checking of performance in photovoltaic plants: A tool for design, installation and maintenance of grid-connected systems. *Renewable Energy*, v. 60, p. 722–732, December 2013. Available from Internet: <https://doi.org/10.1016/j.renene.2013.06.011>.

3 SILVA, J. L. de S. *Estudo e Desenvolvimento Experimental de Otimizador de Potência para Sistemas Fotovoltaicos Conectados à Rede Elétrica*. Dissertação (Mestrado) — Universidade Estadual de Campinas, Campinas, Brasil, 2020. 122 p. (in Portuguese).

4 SAKO, E. Y.; SILVA, J. Lucas de S.; MESQUITA, D. d. B.; CAMPOS, R. E.; MOREIRA, H. S.; VILLALVA, M. G. Concepts and case study of mismatch losses in photovoltaic modules. In: *2019 IEEE 15th Brazilian Power Electronics Conference and 5th IEEE Southern Power Electronics Conference (COBEP/SPEC)*. [S.l.: s.n.], 2019. p. 1–6.

5 WOHLGEMUTH, J. H. Standards for pv modules and components – Recent developments and challenges. *27th European Photovoltaic Solar Energy Conference and Exhibition*, n. October, p. 2976–2980, 2012.

6 ALLUHAYBI, K.; BATARSEH, I. Review and comparison of single-phase grid-tied photovoltaic microinverters. In: *2018 IEEE Energy Conversion Congress and Exposition (ECCE)*. [S.l.: s.n.], 2018. p. 7101–7108.

7 SILVA, J. L. de S.; MOREIRA, H. S.; MESQUITA, D. de B.; VILLALVA, M. G. Analysis of power optimizers in photovoltaic power plant. In: *13 th IEEE/IAS International Conference on Industry Applications, INDUSCON*. [S.l.]: São Paulo-SP, 2018.

8 FOUAD, M. M.; SHIHATA, L. A.; MORGAN, E. S. I. An integrated review of factors influencing the performance of photovoltaic panels. *Renewable and Sustainable Energy Reviews*, Elsevier Ltd, v. 80, n. July 2016, p. 1499–1511, 2017. ISSN 18790690. Available from Internet: <http://dx.doi.org/10.1016/j.rser.2017.05.141>.

9 MOREIRA, H. S.; SILVA, J. L. de S.; MESQUITA, D. de B.; PAULA, B. H. K. de; CARVALHO, J. L. de; KRETLY, L. C.; VILLALVA, M. G. Experimental analysis of photovoltaic modeling for partially shading conditions. In: *13 th IEEE/IAS International Conference on Industry Applications, INDUSCON*. [S.l.]: São Paulo-SP, 2018.

10 MAGHAMI, M. R.; HIZAM, H.; GOMES, C.; RADZI, M. A.; REZADAD, M. I.; HAJIGHORBANI, S. Power loss due to soiling on solar panel: A review. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 59, p. 1307–1316, 2016. ISSN 18790690. Available from Internet: <http://dx.doi.org/10.1016/j.rser.2016.01.044>.

11 SILVA, J. L. de S.; CAVALCANTE, M. M.; MOREIRA, H. S.; VILLALVA, M. G.; DELGADO, D. B. de M. Automated cleaning system for photovoltaic panels. In: *13 th IEEE/IAS International Conference on Industry Applications, INDUSCON*. [S.l.]: São Paulo-SP, 2018.

12 BARBOSA, E. R.; FARIA, M. d. S. F. de; GONTIJO, F. d. B. *Influência Da Sujeira Na Geração Fotovoltaica*. 2018. Anais do Congresso Brasileiro de Energia Solar. (in Portuguese).

13 PANAGEA, I. S.; TSANIS, I. K.; KOUTROULIS, A. G.; GRILLAKIS, M. G. Climate change impact on photovoltaic energy output: The case of Greece. *Advances in Meteorology*, v. 2014, 2014. ISSN 16879317.

14 BELHADJ, C. A.; BANAT, I. H.; DERICHE, M. A Detailed Analysis of Photovoltaic Panel Hot Spot Phenomena based on the Bishop Model. *2017 14th International Multi-Conference on Systems, Signals & Devices (SSD)*, p. 222–227, 2017.

15 KIM, J. H.; JUNG, D. Y.; KIM, J. H.; LEE, S. W.; JUNG, Y. C.; WON, C. Y. Soft switching interleaved boost converter for photovoltaic power generation system. *2008 IEEE International Conference on Sustainable Energy Technologies, ICSET 2008*, p. 257–262, 2008.

16 LUJARA, N. K.; WYK, J. D. V.; MATERU, P. N. Power Electronic Loss Models of dc-dc Converters in Photovoltaic Applications. *IEEE International Symposium on Industrial Electronics. Proceedings. ISIE'98*, p. 1–6, 1998.

17 CHAN, P.-W.; MASRI, S. DC-DC Boost Converter with Constant Output Voltage for Grid Connected Photovoltaic Application System. *International Conference on Intelligent and Advanced Systems*, v. 21, n. 4, p. 67–99, 2010.

18 ELTAWIL, M. A.; ZHAO, Z. Grid-connected photovoltaic power systems: Technical and potential problems-A review. *Renewable and Sustainable Energy Reviews*, v. 14, n. 1, p. 112–129, 2010. ISSN 13640321.

19 JONES, C. B.; ELLIS, B. H.; STEIN, J. S.; WALTERS, J. Comparative review of high resolution monitoring versus standard inverter data acquisition for a single photovoltaic power plant. In: *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC)*. Waikoloa, HI, USA: [s.n.], 2018. p. 0715–0720.

20 SILVA, J. L. de S.; MOREIRA, H. S.; REIS, M. V. G. dos; BARROS, T. A. dos S.; VILLALVA, M. G. Theoretical and behavioral analysis of power optimizers for grid-connected photovoltaic systems. *Energy Reports*, v. 8, p. 10154–10167, 2022. ISSN 2352-4847. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S2352484722014196>.

21 Associação Brasileira de Energia Solar Fotovoltaica. *Infográfico do Mercado de Energia Solar Fotovoltaica no Brasil*. S/D. Online. (in Portuguese). Available from Internet: <https://www.absolar.org.br/mercado/infografico/>.

22 GéRON, A. *Mãos à Obra Aprendizado de Máquina com Scikit-Learn & TensorFlow: Conceitos, Ferramentas e Técnicas Para a Construção de Sistemas Inteligentes*. [S.l.]: Alta Books, 2019. 576 p. (in Portuguese). ISBN 978-8550803814.

23 SHINDE, P. P.; SHAH, S. A Review of Machine Learning and Deep Learning Applications. *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, IEEE, p. 1–6, 2018.

24 ANGRA, S.; AHUJA, S. Machine learning and its applications: A review. In: *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. [S.l.: s.n.], 2017. p. 57–60.

25 JIANG, T.; GRADUS, J. L.; ROSELLINI, A. J. Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, Elsevier Ltd, v. 51, n. 5, p. 675–687, 2020. ISSN 18781888. Available from Internet: <https://doi.org/10.1016/j.beth.2020.05.002>.

26 CZARNOWSKI, I.; JęDRZEJOWICZ, P. Supervised classification problems–taxonomy of dimensions and notation for problems identification. *IEEE Access*, v. 9, p. 151386–151400, 2021.

27 ABOKADR, S.; AZMAN, A.; HAMDAN, H.; AMELINA, N. Handling imbalanced data for improved classification performance: Methods and challenges. In: *2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*. [S.l.: s.n.], 2023. p. 1–8.

28 KHUSHI, M.; SHAUKAT, K.; ALAM, T. M.; HAMEED, I. A.; UDDIN, S.; LUO, S.; YANG, X.; REYES, M. C. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, v. 9, p. 109960–109975, 2021.

29 LIM, P.; GOH, C. K.; TAN, K. C. Evolutionary cluster-based synthetic oversampling ensemble (eco-ensemble) for imbalance learning. *IEEE Transactions on Cybernetics*, v. 47, n. 9, p. 2850–2861, 2017.

30 DIALAMEH, M.; JAHROMI, M. Z. Dynamic feature weighting for imbalanced data sets. In: *2015 Signal Processing and Intelligent Systems Conference (SPIS)*. [S.l.: s.n.], 2015. p. 31–36.

31 VUTTIPITTAYAMONGKOL, P.; ELYAN, E.; PETROVSKI, A. On the class overlap problem in imbalanced data classification. *Knowledge-Based Systems*, v. 212, p. 106631, 2021. ISSN 0950-7051. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0950705120307607>.

32 MAILLO, J.; TRIGUERO, I.; HERRERA, F. Redundancy and complexity metrics for big data classification: Towards smart data. *IEEE Access*, v. 8, p. 87918–87928, 2020.

33 SILVA, J. L. de S.; MELO, K. B. de; SANTOS, K. V. dos; SAKO, E. Y.; SILVA, M. K. da; MOREIRA, H. S.; ARCHILLI, G. B.; CYPRIANO, J. G. I.; CAMPOS, R. E.; SILVA, L. C. P. da; VILLALVA, M. G. Case study of photovoltaic power plants in a model of sustainable university in brazil. *Renewable Energy*, v. 196, p. 247–260, 2022. ISSN 0960-1481. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0960148122009430>.

34 GUNNARSDOTTIR, I.; DAVIDSDOTTIR, B.; WORRELL, E.; SIGURGEIRSDOTTIR, S. Review of indicators for sustainable energy development. *Renewable and Sustainable Energy Reviews*, Elsevier Ltd, v. 133, n. July, p. 110294, 2020. ISSN 18790690. Available from Internet: <https://doi.org/10.1016/j.rser.2020.110294>.

35 International Energy Agency (IEA), INTERNATIONAL ATOMIC ENERGY AGENCY. Energy Indicators for Sustainable Development: Guidelines and. p. 7 – 140, 2007. Available from Internet: <http://www-pub.iaea.org/MTCD/publications/PDF/Pub1222_web.\pdf>.

36 LOZANO, R.; LUKMAN, R.; LOZANO, F. J.; HUISINGH, D.; LAMBRECHTS, W. Declarations for sustainability in higher education: Becoming better leaders, through addressing the university system. *Journal of Cleaner Production*, Elsevier Ltd, v. 48, p. 10–19, 2013. ISSN 09596526. Available from Internet: <http://dx.doi.org/10.1016/j.jclepro.2011.10.006>.

37 VELAZQUEZ, L.; MUNGUIA, N.; PLATT, A.; TADDEI, J. Sustainable university: what can be the matter? *Journal of Cleaner Production*, v. 14, n. 9-11, p. 810–819, 2006. ISSN 09596526.

38 UNICAMP. *About Campus Sustentável Project*. 2021. <http://www.campus-sustentavel.unicamp.br/en/\about-campus-sustentavel/>.

39 GIL, G. M. V.; CUNHA, R. B. A.; SANTO, S. G. D.; MONARO, R. M.; COSTA, F. F.; FILHO, A. J. S. Photovoltaic energy in South America: Current state and grid regulation for large-scale and distributed photovoltaic systems. *Renewable Energy*, Elsevier Ltd, v. 162, p. 1307–1320, 2020. ISSN 18790682. Available from Internet: <https://doi.org/10.1016/j.renene.2020.08.022>.

40 LIMA, D. A.; PAULA, D. N. T. Free contract environment for big electricity consumer in Brazil considering correlated scenarios of energy, power demand and spot prices. *Electric Power Systems Research*, Elsevier, v. 190, n. July 2020, p. 106828, 2021. ISSN 03787796. Available from Internet: <https://doi.org/10.1016/j.epsr.2020.106828>.

41 PVSYST. *PVSyst- Photovoltaic Software*. 2021. <https://www.pvsyst.com/>. Available from Internet: <https://www.pvsyst.com/>.

42 CHAGNON-LESSARD, N.; GOSSELIN, L.; BARNABé, S.; BELLO-OCHENDE, T.; FENDT, S.; GOERS, S.; SILVA, L. C. P. D.; SCHWEIGER, B.; SIMMONS, R.; VANDERSICKEL, A.; ZHANG, P. Smart campuses: Extensive review of the last decade of research and current challenges. *IEEE Access*, v. 9, p. 124200–124234, 2021.

43 SYMPOSIUMS, M. *Microgrid of hangzhou dianzi university*. 2021. <https://microgrid-symposiums.org/microgrid-examples-and\-demonstrations/hangzhou-dianzi-university-microgrid/>. Access at: 14 feb. 2021.

44 MCLARTY, D.; SABATE, C. C.; BROUWER, J.; JABBARI, F. Micro-grid energy dispatch optimization and predictive control algorithms; A UC Irvine case study. *International Journal of Electrical Power and Energy Systems*, Elsevier Ltd, v. 65, p. 179–190, 2015. ISSN 01420615. Available from Internet: <http://dx.doi.org/10.1016/j.ijepes.2014.09.039>.

45 LAZAROIU, G. C.; DUMBRAVA, V.; COSTOIU, M.; TELICEANU, M.; ROSCIA, M. Energy-informatic-centric smart campus. *EEEIC 2016 - International Conference on Environment and Electrical Engineering*, 2016.

46 LAZAROIU, G. C.; DUMBRAVA, V.; COSTOIU, M.; TELICEANU, M.; ROSCIA, M. Smart campus-an energy integrated approach. *2015 International Conference on Renewable Energy Research and Applications, ICRERA 2015*, v. 5, p. 1497–1501, 2015.

47 MALATJI, E. M.; NTSALUBA, S. B. K. Smart energy generation for a smart campus. *2018 International Conference on Intelligent and Innovative Computing Applications, ICONIC 2018*, IEEE, p. 1–5, 2019.

48 BRACCO, S.; CANCEMI, C.; CAUSA, F.; LONGO, M.; SIRI, S. Optimization model for the design of a smart energy infrastructure with electric mobility. *IFAC-PapersOnLine*, Elsevier B.V., v. 51, n. 9, p. 200–205, 2018. ISSN 24058963. Available from Internet: <https://doi.org/10.1016/j.ifacol.2018.07.033>.

49 ANGELIS, E. D.; CIRIBINI, A. L.; TAGLIABUE, L. C.; PANERONI, M. The Brescia Smart Campus Demonstrator. Renovation toward a zero Energy Classroom Building. *Procedia Engineering*, Elsevier B.V., v. 118, p. 735–743, 2015. ISSN 18777058. Available from Internet: <http://dx.doi.org/10.1016/j.proeng.2015.08.508>.

50 CHRISTENSEN, K.; MA, Z.; KORSGAARD, J.; JØRGENSEN, B. N. Location-based energy efficiency and flexibility strategies for smart campuses : Consideration of different levels of building intelligence and typologies. *2019 IEEE PES Conference on Innovative Smart Grid Technologies, ISGT Latin America 2019*, 2019.

51 CHALFOUN, N. Greening University Campus Buildings to Reduce Consumption and Emission while Fostering Hands-on Inquiry-based Education. *Procedia Environmental Sciences*, Elsevier B.V., v. 20, p. 288–297, 2014. ISSN 18780296. Available from Internet: <http://dx.doi.org/10.1016/j.proenv.2014.03.036>.

52 ESCOBEDO, A.; BRICEÑO, S.; JUÁREZ, H.; CASTILLO, D.; IMAZ, M.; SHEINBAUM, C. Energy consumption and GHG emission scenarios of a university campus in Mexico. *Energy for Sustainable Development*, International Energy Initiative. Published by Elsevier Inc. All rights reserved., v. 18, n. 1, p. 49–57, 2014. ISSN 09730826. Available from Internet: <http://dx.doi.org/10.1016/j.esd.2013.10.005>.

53 LEON, I.; OREGI, X.; MARIETA, C. Contribution of university to environmental energy sustainability in the city. *Sustainability*, v. 12, n. 3, 2020. ISSN 2071-1050. Available from Internet: <https://www.mdpi.com/2071-1050/12/3/774>.

54 UI. *UI GreenMetric World University Rankings*. 2021. <https://greenmetric.ui.ac.id/>. Access at: 07 Sep. 2021.

55 CAMPOS, R. E.; SAKÔ, E. Y.; MOREIRA, H. S.; SILVA, J. S. L. D.; VILLALVA, M. G. Experimental Analysis of a Developed I-V Curve Tracer under Partially Shading Conditions. *2019 IEEE PES Conference on Innovative Smart Grid Technologies, ISGT Latin America*, p. 1–5, 2019.

56 MOREIRA, H. S.; SILVA, J. L. d. S.; REIS, M. V. G. d.; MESQUITA, D. d. B.; KIKUMOTO, B. H. d. P.; VILLALVA, M. G. Experimental comparative study of photovoltaic models for uniform and partially shading conditions. *Renewable Energy*, v. 164, p. 58–73, 2021. ISSN 18790682.

57 PVLIB. *pvlib python*. 2021. <https://pvlib-python.readthedocs.io/en/\stable/>. Access at: 14 feb. 2021.

58 SILVA, M. K. da; NARVAEZ, D. I.; MELO, K. B. de; COSTA, T. S.; SIQUEIRA, T. G. de; VILLALVA, M. G. Comparative Analysis of Meteorological Databases and Transposition Models Applied To Photovoltaic Systems. *Proceedings XXII Congresso Brasileiro de Automática*, p. 237–241, 2018.

59 ERBS, D. G.; KLEIN, S. A.; DUFFIE, J. A. Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation. *Solar Energy*, v. 28, p. 293–302, 1982. ISSN 0038092X.

60 PEREZ, R.; SEALS, R.; INEICHEN, P.; STEWART, R.; MENICUCCI, D. A new simplified version of the perez diffuse irradiance model for tilted surfaces. *Solar Energy*, v. 39, p. 221–231, 1987. ISSN 0038092X.

61 DUFFIE, J. A.; BECKMAN, W. A. *Design of Photovoltaic Systems. Solar Engineering of Thermal Processes*. 2. ed. John Wiley Sons, Inc., 2013. 928 p. Available from Internet: <http://doi.wiley.com/10.1002/9781118671603.ch23>.

62 LORENZO, E. *Energy Collected and Delivered by PV Modules*. [S.l.]: Handbook of Photovoltaic Science and Engineering, John Wiley & Sons, 2003. 984–1042 p. ISBN 0-471-49196-9.

63 SILVA, J. L. D. S.; COSTA, T. S.; MELO, K. B. D.; SAKO, E. Y.; MOREIRA, H. S.; VILLALVA, M. G. A comparative performance of PV power simulation software with an installed PV plant. *Proceedings of the IEEE International Conference on Industrial Technology*, v. 2020-February, p. 531–535, 2020.

64 SILVA, J. L. de S.; MELO, K. B. de; COSTA, T. S.; MACHADO, G. M. V.; MOREIRA, H. S.; VILLALVA, M. G. Impact of bifacial modules on the inverter clipping in distributed generation photovoltaic systems in brazil. In: *2021 Brazilian Power Electronics Conference (COBEP)*. [S.l.: s.n.], 2021. p. 1–6.

65 BRANKER, K.; PATHAK, M. J.; PEARCE, J. M. A review of solar photovoltaic levelized cost of electricity. *Renewable and Sustainable Energy Reviews*, Elsevier Ltd, v. 15, n. 9, p. 4470–4482, 2011. ISSN 13640321. Available from Internet: <http://dx.doi.org/10.1016/j.rser.2011.07.104>.

66 SILVA, J. L. D. S.; SILVA, J. A. F. G. D.; MAHMOUDI, E.; PAULA, J. F. S. D.; BARROS, T. A. D. S.; VILLALVA, M. G. Evaluating the significance of solarimetric data for photovoltaic system simulation in a real-world case. In: *2023 IEEE 8th Southern Power Electronics Conference and 17th Brazilian Power Electronics Conference (SPEC/COBEP)*. [S.l.: s.n.], 2023. p. 1–6.

67 BASTE, P. A.; JADKAR, S. R.; PATHAK, A. M. Weather station for solar pv power plant using arduino mega. In: *2021 International Conference on Computer Communication and Informatics (ICCCI)*. [S.l.: s.n.], 2021. p. 1–6.

68 RODRIGUES, B. K. F.; GOMES, M.; SANTANNA, A. M. O.; BARBOSA, D.; MARTINEZ, L. Modelling and forecasting for solar irradiance from solarimetric station. *IEEE Latin America Transactions*, v. 20, n. 2, p. 250–258, 2022.

69 RIEDEL-LYNGSKæR, N.; RIBACONKA, M.; PO, M.; THORSTEINSSON, S.; THORSETH, A.; DAM-HANSEN, C.; JAKOBSEN, M. L. Spectral albedo in bifacial photovoltaic modeling: What can be learned from onsite measurements? In: *2021 IEEE 48th Photovoltaic Specialists Conference (PVSC)*. [S.l.: s.n.], 2021. p. 0942–0949.

70 PHAM, M.-H.; PHAP, V. M.; TRUNG, N. N.; SON, T. T.; KIEN, D. T.; THO, V. T. A. A study on the impact of various meteorological data on the design performance of rooftop solar power projects in vietnam: A case study of electric power

university. *Energies*, v. 15, n. 19, 2022. ISSN 1996-1073. Available from Internet: <https://www.mdpi.com/1996-1073/15/19/7149>.

71 LAVE, M.; HAYES, W.; POHL, A.; HANSEN, C. W. Evaluation of global horizontal irradiance to plane-of-array irradiance models at locations across the united states. *IEEE Journal of Photovoltaics*, v. 5, n. 2, p. 597–606, 2015.

72 ŞEN, Z. *Solar Energy Fundamentals and Modeling Techniques: Atmosphere, Environment, Climate Change and Renewable Energy*. London, UK: Springer, 2008.

73 SILVA, M. K. da. *Estudo de modelos matemáticos para análise da radiação solar e desenvolvimento de ferramenta para modelagem e simulação de sistemas fotovoltaicos*. Dissertação (Mestrado) — Fac. de Eng. Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, SP, Brazil, 2019. (in Portuguese).

74 LAVE, M.; HAYES, W.; POHL, A.; HANSEN, C. W. Evaluation of global horizontal irradiance to plane-of-array irradiance models at locations across the united states. *IEEE Journal of Photovoltaics*, v. 5, n. 2, p. 597–606, March 2015.

75 PEREZ, R.; INEICHEN, P.; SEALS, R.; MICHALSKY, J.; STEWART, R. Modeling daylight availability and irradiance components from direct and global irradiance. *Solar Energy*, v. 44, n. 5, p. 271–289, 1990.

76 COMMISSION, I. E. *IEC 61724-1: Photovoltaic system performance monitoring - Guidelines for measurement, data exchange and analysis - Part 1: Grid-connected systems*. Geneva, Switzerland, 2017. Available from Internet: <https://www.iec.ch/>.

77 STANDARDIZATION, I. O. for. *ISO 9060:2018 - Solar energy - Specification and classification of instruments for measuring hemispherical solar and direct solar radiation*. Geneva, Switzerland, 2018. Available from Internet: <https://www.iso.org/standard/>.

78 GILMAN, P. *SAM Photovoltaic Model Technical Reference*. [S.l.], 2015. Available from Internet: <https://www.nrel.gov/docs/fy15osti/64102.pdf>.

79 LIU, B.; JORDAN, R. A rational procedure for predicting the long-term average performance of flat-plate solar-energy collectors. *Solar Energy*, v. 7, n. 3, p. 53–74, 1963.

80 SILVA, J. L. de S.; MELO, K. B. de; SANTOS, K. V. dos; SAKO, E. Y.; SILVA, M. K. da; MOREIRA, H. S.; ARCHILLI, G. B.; CYPRIANO, J. G. I.; CAMPOS, R. E.; SILVA, L. C. P. da; VILLALVA, M. G. Case study of photovoltaic power plants in a model of sustainable university in brazil. *Renewable Energy*, v. 196, p. 247–260, 2022.

81 HARDLE, W.; SIMAR, L. *Applied Multivariate Statistical Analysis*. 2nd. ed. [S.l.]: Springer, 2007.

82 WU, W. J.; XU, Y. Correlation analysis of visual verbs' subcategorization based on Pearson's correlation coefficient. *2010 International Conference on Machine Learning and Cybernetics, ICMLC 2010*, IEEE, v. 4, n. July, p. 2042–2046, 2010.

83 SILVA, J. L. D. S.; CAVALCANTE, M. M.; MARTINS, S. B.; SILVA, E. J. D.; BARROS, T. A. D. S.; VILLALVA, M. G. Data-driven analysis of solar photovoltaic systems: Correlation and distribution patterns. In: *2023 IEEE 8th Southern Power Electronics Conference and 17th Brazilian Power Electronics Conference (SPEC/COBEP)*. [S.l.: s.n.], 2023. p. 1–7.

84 UZHGA-REBROV, O.; GRABUSTS, P. Comparative Evaluation of Four Methods for Exploratory Data Analysis. *ITMS 2021 - 2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University, Proceedings*, IEEE, v. 1, n. 2, p. 1–5, 2021.

85 LIU, W. M.; CHANG, C. I. Variants of principal components analysis. *International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, p. 1083–1086, 2007.

86 JOSE, S.; ITAGI, R. L. Data Analytics in Solar Photovoltaics Power Forecasting for Smart Grid Applications. *2021 International Conference on Intelligent Technologies, CONIT 2021*, IEEE, p. 1–5, 2021.

87 WIMALARATNE, S.; HAPUTHANTHRI, D.; KAHAWALA, S.; GAMAGE, G.; ALAHAKOON, D.; JENNINGS, A. UNISOLAR: An Open Dataset of Photovoltaic Solar Energy Generation in a Large Multi-Campus University Setting. *International Conference on Human System Interaction, HSI*, IEEE, v. 2022-July, p. 1–5, 2022. ISSN 21582254.

88 TANNAHILL, B. K.; MAUTE, C. E.; YETIS, Y.; EZELL, M. N.; JAIMES, A.; ROSAS, R.; MOTAGHI, A.; KAPLAN, H.; JAMSHIDI, M. Modeling of system of systems via data analytics - Case for "big data" in SoS. *Proceedings of 2013 8th International Conference on System of Systems Engineering: SoSE in Cloud Computing and Emerging Information Technology Applications, SoSE 2013*, IEEE, p. 177–183, 2013.

89 CAPASSO, C.; RUBINO, L.; RUBINO, G.; VENERI, O. Data Analytics for Performance Modelling of Photovoltaic Systems in the Internet of Energy Scenario. *2021 IEEE 15th International Conference on Compatibility, Power Electronics and Power Engineering, CPE-POWERENG 2021*, IEEE, p. 1–6, 2021.

90 ZHANG, R.; MA, H.; HUA, W.; SAHA, T. K.; ZHOU, X. Data-Driven Photovoltaic Generation Forecasting Based on a Bayesian Network with Spatial-Temporal Correlation Analysis. *IEEE Transactions on Industrial Informatics*, v. 16, n. 3, p. 1635–1644, 2020. ISSN 19410050.

91 SUNDARARAJAN, A.; SARWAT, A. I. Roadmap to prepare distribution grid-tied photovoltaic site data for performance monitoring. *2017 International Conference on Big Data, IoT and Data Science, BID 2017*, v. 2018-Janua, p. 110–115, 2018.

92 KHANNA, V.; DAS, B. K.; VANDANA; SINGH, P. K.; SHARMA, P.; JAIN, S. K. Statistical analysis and engineering fit models for two-diode model parameters of large area silicon solar cells. *Solar Energy*, Elsevier Ltd, v. 136, p. 401–411, 2016. ISSN 0038092X. Available from Internet: <http://dx.doi.org/10.1016/j.solener.2016.07.018>.

93 VARGA, Z.; RACZ, E. Influence of the Temperature on the Open-Circuit Voltage and Short-Circuit Current of a yellow colorized Dye Sensitized Solar Cell using Correlation Approach. *SACI 2021 - IEEE 15th International Symposium on Applied Computational Intelligence and Informatics, Proceedings*, p. 253–258, 2021.

94 ZHANG, G.; YUAN, J.; MAO, Y.; HUANG, Y. Two-dimensional Janus material MoS2(1-x)Se2x ($0 < x < 1$) for photovoltaic applications: A machine learning and density functional study. *Computational Materials Science*, v. 186, n. August 2020, p. 1–7, 2021. ISSN 09270256.

95 MEZEI, J.; LUUKKA, P.; COLLAN, M. Similarity of histograms and circular histograms from interval and fuzzy data. In: IEEE. *2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS)*. Otsu, Japan, 2017. p. 1–6.

96 AFZAAL, M.; SAJJAD, I.; AWAN, A.; PARACHA, K.; KHAN, M.; BHATTI, A.; ZUBAIR, M.; REHMAN, W.; AMIN, S.; HAROON, S.; LIAQAT, R.; HDIDI, W.; TLILI, I. Probabilistic generation model of solar irradiance for grid connected photovoltaic systems using weibull distribution. *Sustainability*, v. 12, p. 2241, 2020.

97 GOLIVE, Y. R.; KOTTANTHARAYIL, A.; VASI, J.; SHIRADKAR, N.; ZACHARIAH, S.; DUBEY, R.; CHATTOPADHYAY, S.; BHADURI, S.; SINGH, H. K.; BORA, B.; KUMAR, S.; TRIPATHI, A. K. Analysis of Field Degradation Rates Observed in All-India Survey of Photovoltaic Module Reliability 2018. *IEEE Journal of Photovoltaics*, IEEE, v. 10, n. 2, p. 560–567, 2020. ISSN 21563403.

98 SANTIAGO, R. M. C.; BANDALA, A. A.; DADIOS, E. P. Artificial neural network model for solar resource assessment: An application to efficient design of photovoltaic system. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, v. 2017-December, p. 2672–2676, 2017. ISSN 21593450.

99 SCHWERTMAN, N. C.; OWENS, M. A.; ADNAN, R. A simple more general boxplot method for identifying outliers. *Computational Statistics and Data Analysis*, v. 47, n. 1, p. 165–174, 2004. ISSN 01679473.

100 KIM, G. G.; HYUN, J. H.; CHOI, J. H.; AHN, S. H.; BHANG, B. G.; AHN, H. K. Quality Analysis of Photovoltaic System Using Descriptive Statistics of Power Performance Index. *IEEE Access*, IEEE, v. 11, n. March, p. 28427–28438, 2023. ISSN 21693536.

101 ZHAO, Y.; LEHMAN, B.; BALL, R.; MOSESIAN, J.; PALMA, J.-F. de. Outlier detection rules for fault detection in solar photovoltaic arrays. In: *2013 Twenty-Eighth Annual IEEE Applied Power Electronics Conference and Exposition (APEC)*. [S.l.: s.n.], 2013. p. 2913–2920.

102 SILVA, J. L. de S.; MAHMOUDI, E.; CARVALHO, R. R. M.; BARROS, T. A. dos S. Classification of anomalies in photovoltaic systems using supervised machine learning techniques and real data. *Energy Reports*, v. 11, p. 4642–4656, 2024. ISSN 2352-4847. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S2352484724002488>.

103 URBANETZ, I. V. *Diagnóstico de falhas em módulos fotovoltaicos*. Dissertação (Mestrado) — Escola Superior de Tecnologia e Gestão, Bragança, Portugal, 2019. (In Portuguese).

104 IBRAHIM, M.; ALSHEIKH, A.; AWAYSHEH, F. M.; ALSHEHRI, M. D. Machine learning schemes for anomaly detection in solar power plants. *Energies*, v. 15, n. 3, p. 1–17, 2022.

105 BABASAKI, T.; HIGUCHI, Y. Using pv string data to diagnose failure of solar panels in a solar power plant. In: *2018 IEEE International Telecommunications Energy Conference (INTELEC)*. [S.l.: s.n.], 2018. p. 1–4.

106 KHELIFI, B.; ZDIRI, M. A.; SALEM, F. B. Machine learning for solar power systems - a short tour. In: *2021 12th International Renewable Energy Congress (IREC)*. Hammamet, Tunisia: [s.n.], 2021. p. 1–6.

107 VIEIRA, R. G. *Aplicação de Técnicas de Inteligência Artificial para Identificação de Falhas em Módulos Fotovoltaicos*. Tese (Doutorado) — Universidade Federal do Rio Grande do Norte, Natal, Brazil, 2021. (In Portuguese).

108 HONG, Y. Y.; PULA, R. A. Methods of photovoltaic fault detection and classification: A review. *Energy Reports*, Elsevier Ltd, v. 8, p. 5898–5929, 2022. ISSN 23524847. Available from Internet: <https://doi.org/10.1016/j.egyr.2022.04.043>.

109 GAROUDJA, E.; CHOUDER, A.; KARA, K.; SILVESTRE, S. An enhanced machine learning based approach for failures detection and diagnosis of PV systems. *Energy Conversion and Management*, v. 151, p. 496–513, 2017. ISSN 01968904.

110 HARROU, F.; DAIRI, A.; TAGHEZOUIT, B.; SUN, Y. An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class Support Vector Machine. *Solar Energy*, Elsevier, v. 179, n. December 2018, p. 48–58, 2019. ISSN 0038092X. Available from Internet: <https://doi.org/10.1016/j.solener.2018.12.045>.

111 CHEN, Z.; HAN, F.; WU, L.; YU, J.; CHENG, S.; LIN, P.; CHEN, H. Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. *Energy Conversion and Management*, Elsevier, v. 178, n. August, p. 250–264, 2018. ISSN 01968904. Available from Internet: <https://doi.org/10.1016/j.enconman.2018.10.040>.

112 LI, B.; DELPHA, C.; MIGAN-DUBOIS, A.; DIALLO, D. Fault diagnosis of photovoltaic panels using full I–V characteristics and machine learning techniques. *Energy Conversion and Management*, v. 248, 2021. ISSN 01968904.

113 VOUTSINAS, S.; KAROLIDIS, D.; VOYIATZIS, I.; SAMARAKOU, M. Development of a machine-learning-based method for early fault detection in photovoltaic systems. *Journal of Engineering and Applied Science*, Springer Berlin Heidelberg, v. 70, n. 1, p. 0–17, 2023. ISSN 25369512. Available from Internet: <https://doi.org/10.1186/s44147-023-00200-0>.

114 ZULFAUZI, I. A.; DAHLAN, N. Y.; SINTUYA, H.; SETTHAPUN, W. Anomaly detection using k-means and long-short term memory for predictive maintenance of large-scale solar (lss) photovoltaic plant. *Energy Reports*, v. 9, p. 154–158, 2023. ISSN 2352-4847.

115 SARAVANAN, R.; SUJATHA, P. A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, n. Iciccs, p. 945–949, 2018. Available from Internet: <https://ieeexplore.ieee.org/abstract/document/8663155>.

116 LI, L.; ZHANG, Y.; CHEN, W.; BOSE, S. K.; ZUKERMAN, M.; SHEN, G. Naïve bayes classifier-assisted least loaded routing for circuit-switched networks. *IEEE Access*, v. 7, p. 11854–11867, 2019.

117 KHANAFER, R. M.; SOLANA, B.; TRIOLA, J.; BARCO, R.; MOLTSEN, L.; ALTMAN, Z.; LAZARO, P. Automated diagnosis for umts networks using bayesian

network approach. *IEEE Transactions on Vehicular Technology*, v. 57, n. 4, p. 2451–2461, 2008.

118 MAZUELAS, S.; SHEN, Y.; PEREZ, A. Generalized maximum entropy for supervised classification. *IEEE Transactions on Information Theory*, v. 68, n. 4, p. 2530–2550, 2022.

119 JANG, H. S.; BAE, K. Y.; PARK, H.-S.; SUNG, D. K. Solar power prediction based on satellite images and support vector machine. *IEEE Transactions on Sustainable Energy*, v. 7, n. 3, p. 1255–1263, 2016.

120 OBIORA, C. N.; ALI, A.; HASAN, A. N. Using the multilayer perceptron (mlp) model in predicting the patterns of solar irradiance at several time intervals. In: *2023 31st Southern African Universities Power Engineering Conference (SAUPEC)*. [S.l.: s.n.], 2023. p. 1–6.

121 CHEN, C. C.; SCHWENDER, H.; KEITH, J.; NUNKESSER, R.; MENGERSEN, K.; MACROSSAN, P. Methods for identifying snp interactions: A review on variations of logic regression, random forest and bayesian logistic regression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, v. 8, n. 6, p. 1580–1591, 2011.

122 CHRISTOPHER, J. The science of rule-based classifiers. In: *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. [S.l.: s.n.], 2019. p. 299–303.

123 PANIGRAHI, B. k.; PARIJA, B.; PATTANAYAK, R.; TRIPATHY, S. K. Faults classification in a microgrid using decision tree technique and support vector machine. In: *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*. [S.l.: s.n.], 2018. p. 145–148.

124 RANGELOV, D.; BOERGER, M.; TCHOLTCHEV, N.; LäMMEL, P.; HAUSWIRTH, M. Design and development of a short-term photovoltaic power output forecasting method based on random forest, deep neural network and lstm using readily available weather features. *IEEE Access*, v. 11, p. 41578–41595, 2023.

125 EVANS, M.; ELLACOTT, S.; HAND, C. A multiresolution neural network classifier for machine vision. In: *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*. [S.l.: s.n.], 1991. p. 2594–2599 vol.3.

126 LI, Y.; SHIU, S.; PAL, S. Combining feature reduction and case selection in building cbr classifiers. *IEEE Transactions on Knowledge and Data Engineering*, v. 18, n. 3, p. 415–429, 2006.

127 BOUREL, M.; GHATTAS, B. Direct multiclass boosting using base classifiers' posterior probabilities estimates. In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.: s.n.], 2017. p. 228–233.

128 SETIAWAN, R. Quadratic classifier from discriminant analysis for classification of multiple attributes data: (case study: Fertility data set). In: *2017 International Conference on Information Management and Technology (ICIMTech)*. [S.l.: s.n.], 2017. p. 112–115.

129 DEVELOPERS, P. P. *pvlib-python Documentation*. 2023. Acesso em: 02 de agosto de 2023. Available from Internet: <https://pvlib-python.readthedocs.io/en/stable/>.

130 VILLALVA, M.; GAZOLI, J.; FILHO, E. Comprehensive Approach to Modeling and Simulation of Photovoltaic Arrays. *IEEE Transactions on Power Electronics*, v. 24, n. 5, p. 1198–1208, 2009. ISSN 08858993.

131 VILLALVA, M. G. *Conversor Eletrônico de Potência Trifásico para Sistema Fotovoltaico Conectado à Rede Elétrica*. 292 p. Tese (Doutorado) — Universidade Estadual de Campinas (UNICAMP), São Paulo, Brasil, 2010. (in Portuguese).

132 MOREIRA, H. S. *Estudo de Técnicas de Rastreamento de Máxima Potência Tolerantes a Sombras para Sistemas Fotovoltaicos*. 2018. Universidade Estadual de Campinas (UNICAMP), Dissertação de Mestrado em Engenharia Elétrica, Campinas, Brasil. (in Portuguese).

133 HARB, S.; KEDIA, M.; ZHANG, H.; BALOG, R. S. Microinverter and string inverter grid-connected photovoltaic system - a comprehensive study. *Conf. Record of the IEEE Photovoltaic Specialists Conference*, p. 2885–2890, 2013. ISSN 01608371.

134 SHIN, J. H.; KIM, J. O. On-line diagnosis and fault state classification method of photovoltaic plant. *Energies*, v. 13, n. 17, 2020. ISSN 19961073.

135 KALID, S. N.; NG, K.-H.; TONG, G.-K.; KHOR, K.-C. A multiple classifiers system for anomaly detection in credit card data with unbalanced and overlapped classes. *IEEE Access*, v. 8, p. 28210–28221, 2020.

136 JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, v. 6, n. 5, p. 429–449, 2002.

137 WANG, W.; KEEN, J.; BANK, J.; GIRALDEZ, J.; MONTANO-MARTINEZ, K. An automated approach for screening residential pv applications using a random forest model. *IEEE Open Access Journal of Power and Energy*, v. 10, p. 327–334, 2023.

138 THAKUR, K.; LAL, K.; KUMAR, V. Ensemble method to predict impact of student intelligent quotient and academic achievement on placement. In: *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*. [S.l.: s.n.], 2021. p. 249–253.

139 AZHAR, N. A.; POZI, M. S. M.; DIN, A. M.; JATOWT, A. An investigation of smote based methods for imbalanced datasets with data complexity analysis. *IEEE Transactions on Knowledge and Data Engineering*, v. 35, n. 7, p. 6651–6672, 2023.

140 VOUTSINAS, S.; KAROLIDIS, D.; VOYIATZIS, I.; SAMARAKOU, M. Development of a multi-output feed-forward neural network for fault detection in photovoltaic systems. *Energy Reports*, Elsevier, v. 8, n. May, p. 33–42, 2022. Available from Internet: <https://doi.org/10.1016/j.egyr.2022.06.107>.

141 CHEN, Z.; WU, L.; CHENG, S.; LIN, P.; WU, Y.; LIN, W. Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and i-v characteristics. *Applied Energy*, Elsevier, v. 204, p. 912–931, 2017. Available from Internet: <https://doi.org/10.1016/j.apenergy.2017.05.034>.

142 HARROU, F.; SUN, Y.; TAGHEZOUIT, B.; SAIDI, A.; HAMLATI, M. Reliable fault detection and diagnosis of photovoltaic systems based on statistical monitoring approaches. *Renewable Energy*, Elsevier, v. 116, p. 22–37, 2018. Available from Internet: <https://doi.org/10.1016/j.renene.2017.09.048>.

143 YI, Z.; ETEMADI, A. Line-to-line fault detection for photovoltaic arrays based on multiresolution signal decomposition and two-stage support vector machine. *IEEE Transactions on Industrial Electronics*, IEEE, v. 64, n. 11, p. 8546–8556, 2017. Available from Internet: <https://doi.org/10.1109/TIE.2017.2703681>.

144 XIA, K.; HE, S.; TAN, Y.; JIANG, Q.; XU, J.; YU, W. Wavelet packet and support vector machine analysis of series dc arc fault detection in photovoltaic system. *IEEJ Transactions on Electrical and Electronic Engineering*, Wiley, v. 14, n. 2, p. 192–200, 2019. Available from Internet: <https://doi.org/10.1002/tee.22797>.

145 WANG, L.; LIU, J.; GUO, X.; YANG, Q.; YAN, W. Online fault diagnosis of photovoltaic modules based on multi-class support vector machine. In: IEEE. *Proceedings of the 2017 Chinese Automation Congress (CAC)*. 2017. p. 4569–4574. Available from Internet: <https://doi.org/10.1109/CAC.2017.8243586>.

146 WINSTON, D. *et al.* Solar pv's micro crack and hotspots detection technique using nn and svm. *IEEE Access*, IEEE, v. 9, p. 127259–127269, 2021. Available from Internet: <https://doi.org/10.1109/ACCESS.2021.3111904>.

147 YI, Z.; ETEMADI, A. Fault detection for photovoltaic systems based on multi-resolution signal decomposition and fuzzy inference systems. *IEEE Transactions on Smart Grid*, IEEE, v. 8, n. 3, p. 1274–1283, 2017. Available from Internet: <https://doi.org/10.1109/TSG.2016.2587244>.

148 MEMON, S.; JAVED, Q.; KIM, W.; MAHMOOD, Z.; KHAN, U.; SHAHZAD, M. A machine-learning-based robust classification method for pv panel faults. *Sensors*, MDPI, v. 22, n. 21, p. 1–14, 2022. Available from Internet: <https://doi.org/10.3390/s22218515>.

149 JIA, F.; LUO, L.; GAO, S.; YE, J. Logistic regression based arc fault detection in photovoltaic systems under different conditions. *Journal of Shanghai Jiaotong University*, Springer, v. 24, n. 4, p. 459–470, 2019. Available from Internet: <https://doi.org/10.1007/s12204-019-2095-1>.

150 FADHEL, S. *et al.* Data-driven approach for isolated PV shading fault diagnosis based on experimental I-V curves analysis. In: IEEE. *Proceedings of IEEE International Conference on Industrial Technology (ICIT)*. 2018. p. 927–932. Available from Internet: <https://doi.org/10.1109/ICIT.2018.8352302>.

151 DAI, S.; WANG, D.; LI, W.; ZHOU, Q.; TIAN, G.; DONG, H. Fault diagnosis of data-driven photovoltaic power generation system based on deep reinforcement learning. *Mathematical Problems in Engineering*, v. 2021, 2021. Available from Internet: <https://doi.org/10.1155/2021/2506286>.

152 MELO, K. B. de; SILVA, M. K. da; SILVA, J. L. de S.; COSTA, T. S.; VILLALVA, M. G. Study of energy improvement with the insertion of bifacial modules and solar trackers in photovoltaic installations in brazil. *Renewable Energy Focus*, v. 41, p. 179–187, 2022. ISSN 1755-0084. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S1755008422000084>.

153 REIS, M. V. G. d.; NARVAEZ, D. I.; MOREIRA, H. S.; SILVA, J. L. d. S.; BARROS, T. A. S.; VILLALVA, M. G. A robust islanding detection method for inverter-based distributed generation systems using dc-link voltage perturbation. *IEEE Journal of Photovoltaics*, v. 12, n. 6, p. 1559–1566, 2022.

154 SILVA, L. da. *Um modelo de inovação em gestão energética para a América Latina e o Caribe*. 1. ed. Rio de Janeiro: 2021, 2021. –373 p. (in Portuguese). ISBN 978-65-86214-61-1.

155 CAVALCANTE, M. M.; SILVA, J. L. D. S.; MARTINS, S. B.; NUNES, I. F. S.; RIBEIRO, A. C.; BARROS, T. A. D. S. Comparison and application of data science techniques for anomaly detection in photovoltaic systems. In: *2023 IEEE 8th Southern Power Electronics Conference and 17th Brazilian Power Electronics Conference (SPEC/COBEP)*. [S.l.: s.n.], 2023. p. 1–5.

156 LIMA, G. P. de; PRYM, G. C. S.; MELO, K. B. de; MOREIRA, H. S.; SILVA, J. L. de S.; FILHO, E. R.; VILLALVA, M. G. Thermal mathematical modeling of photovoltaic inverters and experimental validation. In: *2021 Brazilian Power Electronics Conference (COBEP)*. [S.l.: s.n.], 2021. p. 1–7.

157 COSTA, T. S.; ROSOLEM, M. de F.; SILVA, J. L. d. S.; VILLALVA, M. G. An Overview of Electrochemical Batteries for ESS Applied to PV Systems Connected to the Grid. In: *2021 14th IEEE International Conference on Industry Applications (INDUSCON)*. [S.l.: s.n.], 2021. p. 1392–1399.

158 PAULA, J. F. S. D.; LIMA, G. P. D.; CERBATTO, G.; PRYM, S.; SILVA, J. L. D. S.; MELO, K. B. D.; GRADELLA, M. Análise comparativa de desempenho de um sistema fotovoltaico simulado com as ferramentas Pvsyst e SAM (System Advisor Model). *IX Congresso Brasileiro de Energia Solar*, 2022. (in Portuguese).

159 PRYM, G. C. S.; LIMA, G. P. de; SILVA, J. L. d. S.; NEVES, M. R. M.; BARROS, T. A. d. S.; VILLALVA, M. G. Estudo da corrente de fuga e seus efeitos em inversores baseados na topologia Full-Bridge com e sem transformador para aplicações fotovoltaicas. *IX Congresso Brasileiro de Energia Solar*, 2022. (in Portuguese).

APPENDIX A – Usage licenses for copyrighted papers

This Appendix provides the licensing information for the use of IEEE and Elsevier copyrighted papers in this thesis.

Permissions

As a general rule, permission should be sought from the rights holder to reproduce any substantial part of a copyrighted work. This includes any text, illustrations, charts, tables, photographs, or other material from previously published sources. Obtaining permission to re-use content published by Elsevier is simple. Follow the guide below for a quick and easy route to permission.

Permission guidelines

For further guidelines about obtaining permission, please review our Frequently Asked Questions below:

When is permission required?

Permission is required for any material that is not original. As a rule, written permission must be obtained from the rightsholder to re-use any copyrighted material. Typically, the rightsholder of published material is the publisher unless it is explicitly indicated otherwise. Copyrighted material can include figures, illustrations, charts, tables, photographs, and text excerpts. Re-use of any borrowed material must be properly acknowledged, even if it is determined that written permission is not necessary.

For any further clarifications, you can submit your query via our online form

When is permission not required?

From whom do I need permission? +

How do I obtain permission to use photographs or illustrations? +

Do I need to obtain permission to use material posted on a website such as Blogs/Google images/e-commerce websites? +

What rights does Elsevier require when requesting permission? +

How do I obtain permission from another publisher? +

What is RightsLink/CCC? +

What should I do if I am not able to locate the copyright owner? +

Can I obtain permission from a Reproduction Rights Organization (RRO)? +

Is Elsevier an STM signatory publisher? +

Do I need to request permission to re-use work from another STM publisher? +

Do I need to request permission to text mine Elsevier content? +

ermission guidelines ScienceDirect content ClinicalKey content Tutorial videos Help and suppoı

Yes. Authors can include their articles in full or in part in a thesis or dissertation for non-commercial purposes.

For any further clarifications, you can submit your query via our online form ↗

Which uses of a work does Elsevier view as a form of 'prior publication'? +

RightsLink

Sign in/Register

IEEE Requesting permission to reuse content from an IEEE publication

Evaluating the Significance of Solarimetric Data for Photovoltaic System Simulation in a Real-World Case

Conference Proceedings: 2023 IEEE 8th Southern Power Electronics Conference (SPEC)

Author: João Lucas De Souza Silva

Publisher: IEEE

Date: 26 November 2023

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

RightsLink

Sign in/Register

IEEE Requesting permission to reuse content from an IEEE publication

Data-Driven Analysis of Solar Photovoltaic Systems: Correlation and Distribution Patterns

Conference Proceedings: 2023 IEEE 8th Southern Power Electronics Conference (SPEC)

Author: João Lucas De Souza Silva

Publisher: IEEE

Date: 26 November 2023

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW