GENERATION AND VALIDATION OF SYNTHETIC DATA FOR TRAINING MACHINE LEARNING MODELS IN PHOTOVOLTAIC SYSTEMS

1st Ernesto Manuel Distinto Ufuene
Polytechnic School
University of São Paulo (USP)
São Paulo, Brasil
ernestoufuene@usp.br

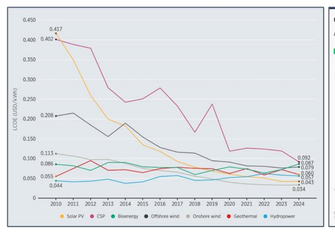
2nd Sergio Takeo Kofuji
Polytechnic School
University of São Paulo (USP)
São Paulo, Brasil
kofuji@usp.br

3rd Ricardo Queirós
Faculty of Engineering
Agostinho Neto University (UAN)
Luanda, Angola
ricardo.queiros@feuan.ao

INTRODUCTION

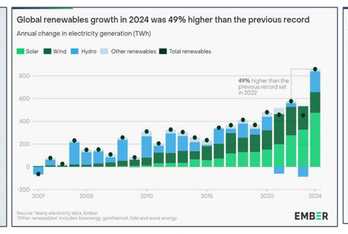
- **Photovoltaic solar energy** is the **fastest-growing source** in the global electricity mix.

a) Decline in PV costs (2010–2024) **PV**



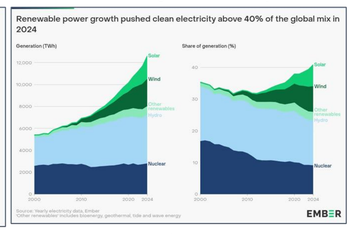
Source: IRENA (2025)

b) Growth of global PV generation **PV**



Source: Ember *(2025)*

c) PV share in the global electricity mix **PV**



Source: Ember *(2025)*

- However, this rapid growth also **introduces new challenges** related to system reliability, operation, and management.
- It is no longer sufficient to adopt only corrective or preventive maintenance strategies; intelligent approaches **(machine learning techniques)** are required to handle dynamic environmental conditions such as variations in irradiance, temperature, and, particularly, the occurrence of **partial shading**.
- The **effectiveness of any ML model** critically depends on the availability of representative and/or well-labeled data, which remains a practical challenge for its widespread adoption.
- Real data
- The collection and label process of real data faces both operational and economic challenges, such as high instrumentation costs, extended acquisition times, etc.
- **These limitations have driven the adoption of synthetic data as a viable alternative** (Huang & Zhao, 2024; Kolahi et al., 2024; Rashidi et al., 2024; Whang et al., 2023).

Objectives

General Objective

To develop and validate a synthetic, physics-informed data generation approach to train a machine learning model for detecting partial shading conditions in photovoltaic systems.

Specific Objectives

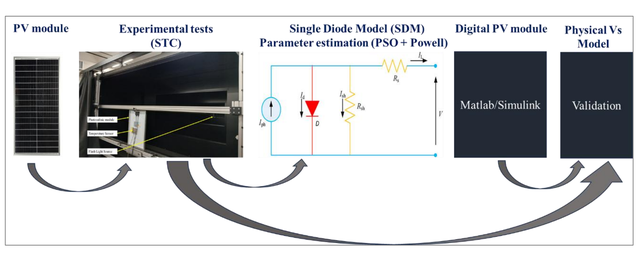
- To model and validate the PV module using the Single-Diode Model (SDM);
- To collect real irradiance and temperature data;
- To simulate the PV system under normal and shaded conditions;
- To generate a labeled synthetic dataset;
- To train and validate a Random Forest classifier;
- To compare performance using synthetic versus real data.

Key Contributions

- **A low-cost and efficient methodology** for generating and validating labeled synthetic data for partial shading detection in PV systems;
- **Use of experimentally validated SDM-based models** to produce physically consistent synthetic data;
- **Experimental testing of a Random Forest classifier**, trained with synthetic data and evaluated with real measurements;
- **Evidence that synthetic data reduces the cost and time** needed for labeled data acquisition, enabling scalable intelligent PV applications;
- **Publication of a real environmental dataset** (irradiance and temperature) on Kaggle for public scientific use;

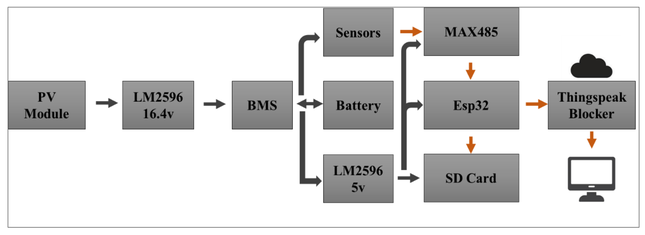
METHODOLOGY

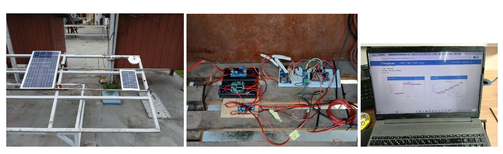
1. Modeling of a PV module using the SDM + PSO–Powell optimization.



Source: author

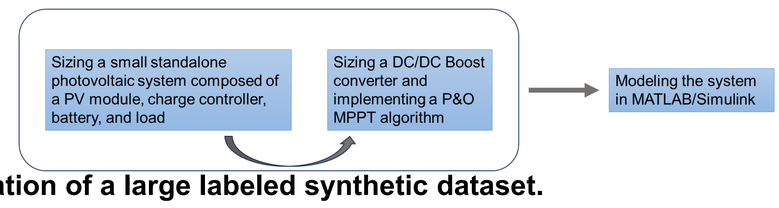
2. Real Data Acquisition, irradiance and temperature.



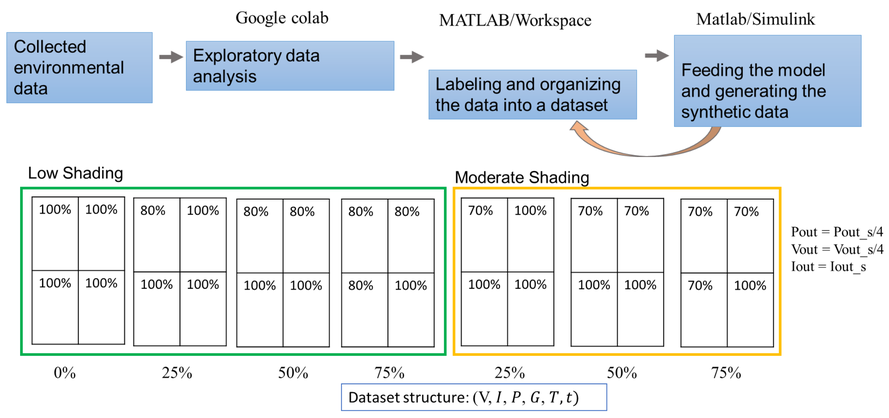


Source: author

3. Simulation of a complete PV system under normal and shaded conditions using the MATLAB/Simulink environment;



4. Generation of a large labeled synthetic dataset.

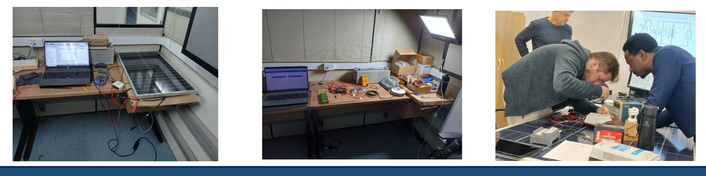


Note: The dataset comprised 3,844,900 samples.

5. Training of a Random Forest model and testing with real measurements.

A Random Forest model was trained using synthetic PV data (70/30 split, K-fold validation) with V, I, P, G, T, and t as inputs.

The trained model was then applied to real measurements from the 60 W PV module to verify compatibility between synthetic-data training and real-data operation.



Source: author

RESULTS

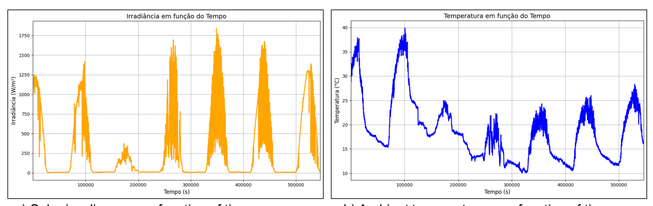
1. Validation of the Single-Diode

Source: author

Metrics	I–V		I–P	
	1000 W/m²	500 W/m²	1000 W/m²	500 W/m²
MAE (A|W)	0.0156	0.0061	0.2932	0.1021
RMSE (A|W)	0.0333	0.0094	0.7025	0.1809
R^2	0.9983	0.9993	0.9985	0.9996
MAPE (%)	1.36	0.90	1.39	0.93

Metrics	Good Range	References
MAE	< 0.05 A e < 0.5 W	*Silva et al., 2020*
RMSE	< 0.1 A e < 1 W	Villalva et al., 2019
R^2	0.995–0.999 V.E	*Silva et al.*, 2020
MAPE	1–3% V.E	Khan et al., 2019

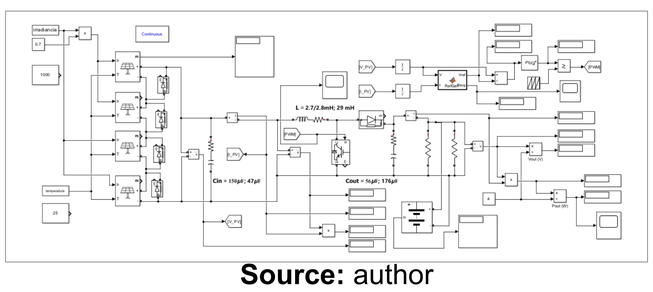
2. Real Data Acquisition



a) Solar irradiance as a function of time; b) Ambient temperature as a function of time

Source: author

3. Complete PV system under normal and shaded conditions in the MATLAB/Simulink environment



Source: author

4. Performance of the Machine Learning Model

The Random Forest model **achieved 78% accuracy during validation with synthetic data** and **81% accuracy when tested with real measurements**. The classifier reliably identified normal operating conditions and consistently detected partial shading, demonstrating robust generalization from synthetic to real data..

Conclusions

The proposed methodology shows that physics-informed synthetic data generated from an experimentally validated PV model can effectively train ML algorithms for partial shading detection. The Random Forest model trained solely on synthetic data performed consistently on real measurements, confirming the feasibility of using synthetic datasets to reduce the cost and effort of labeled data acquisition. Overall, the approach provides a scalable and low-cost strategy for intelligent monitoring and fault detection in small PV systems.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the **PDCT Angola program** through its Graduate Scholarship Program (Master's, Doctorate, and Postdoctoral Studies). The authors also thank the Instituto de Energia e Ambiente of the University of São Paulo **(IEE-USP)**, especially the Solar Photovoltaic Energy Division, for providing laboratory facilities, measurement equipment, and technical support essential for the experimental validation of this work.