



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

João Vitor Rosa Negri

**Estimation of Missing Strings Based on Tracker and Stringbox Operational Data
in Utility-Scale Photovoltaic Power Plants**

Florianópolis
2025

João Vitor Rosa Negri

**Estimation of Missing Strings Based on Tracker and Stringbox Operational Data
in Utility-Scale Photovoltaic Power Plants**

Dissertação submetida ao Programa de Pós-Graduação
em Engenharia Elétrica da Universidade Federal de
Santa Catarina para a obtenção do título de mestre
em Engenharia Elétrica.

Orientador: Prof. Richard Demo Souza, Dr.

Coorientador: Prof. Arliones Stevert Hoeller Jr, Dr.

Florianópolis

2025

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Negri, João Vitor Rosa

Estimation of Missing Strings Based on Tracker and Stringbox Operational Data in Utility-Scale Photovoltaic Power Plants / João Vitor Rosa Negri ; orientador, Richard Demo Souza, coorientador, Arliones Stevert Hoeller Junior, 2025.

43 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia Elétrica, Florianópolis, 2025.

Inclui referências.

1. Engenharia Elétrica. 2. Fotovoltaica. 3. Usinas fotovoltaica de grande escala. 4. Detecção de falhas. 5. Dados operacionais. I. Souza, Richard Demo. II. Hoeller Junior, Arliones Stevert. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia Elétrica. IV. Título.

João Vitor Rosa Negri

Estimation of Missing Strings Based on Tracker and Stringbox Operational Data in Utility-Scale Photovoltaic Power Plants

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca
examinadora composta pelos seguintes membros:

Roberto Francisco Coelho
Universidade Federal de Santa Catarina

Guilherme Luiz Moritz
Universidade Tecnológica Federal do Paraná

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi
julgado adequado para obtenção do título de mestre em Engenharia Elétrica.



Documento assinado digitalmente

Carlos Renato Rambo

Data: 01/12/2025 10:15:23-0300

CPF: ***.748.198-**

Verifique as assinaturas em <https://v.ufsc.br>

Prof. Carlos Renato Rambo, Dr.
Coordenador do Programa



Documento assinado digitalmente

Richard Demo Souza

Data: 28/11/2025 16:42:46-0300

CPF: ***.267.379-**

Verifique as assinaturas em <https://v.ufsc.br>

Prof. Richard Demo Souza, Dr.
Orientador

Florianópolis, 2025.

AGRADECIMENTOS

Agradeço à minha família pelo apoio incondicional ao longo da minha vida. Desde a infância, estiveram presentes não apenas no incentivo aos estudos e no compartilhamento de conhecimento, mas também nos momentos simples de convivência, que se tornaram fundamentais para minha formação pessoal e acadêmica.

Expresso também minha gratidão à minha namorada, Monisse, por seu companheirismo constante. Sua presença nos momentos de alegria e nas fases mais desafiadoras me deu forças para seguir em frente. Obrigado por acreditar em mim, me motivar diariamente e compartilhar comigo esta caminhada.

Agradeço aos amigos e colegas de trabalho do Inova IFSC, em especial ao Gabriel e ao Yago, cuja parceria foi essencial durante o desenvolvimento desta pesquisa. Além das contribuições técnicas e intelectuais, sua amizade tornou esta jornada mais leve e repleta de boas lembranças.

Registro também minha gratidão aos profissionais da ENGIE, que colaboraram de forma proativa em diferentes etapas deste trabalho, assim como aos professores envolvidos no projeto de pesquisa, cujas contribuições enriqueceram a construção deste estudo.

Este trabalho contou com apoio financeiro parcial da CAPES (Finance Code 001), do CNPq (Processo 305021/2021-4) e do Programa de Pesquisa e Desenvolvimento da ENGIE Brasil Energia S.A., regulado pela ANEEL (Projeto PD-00403-0056/2023).

Deixo um agradecimento especial ao meu orientador, Richard, pela dedicação, paciência e disposição em sempre trazer novas ideias e perspectivas, além de oferecer motivação nos momentos em que mais precisei. Agradeço igualmente ao meu coorientador, Arliones, pelo companheirismo e apoio constante desde a graduação, e ao professor Mario, cuja orientação e sugestões foram valiosas para o amadurecimento deste trabalho.

Por fim, estendo meus agradecimentos a todos que, direta ou indiretamente, contribuíram para a realização desta dissertação. Cada gesto de apoio, incentivo ou colaboração foi essencial para que este objetivo pudesse ser alcançado.

*"All we have to decide is what to do with the time that is given us."
J. R. R. Tolkien, The Fellowship of the Ring*

RESUMO

Este trabalho propõe uma nova metodologia para detecção de *strings* ausentes, ou seja, *strings* que se tornam eletricamente desconectadas devido a falhas, em usinas fotovoltaicas em grande escala (USPVPPs), por meio da integração de dados operacionais de sistemas de controle de supervisão e aquisição de dados (SCADA) nos níveis de *tracker* e *stringbox*. Falhas nos *trackers* são inicialmente identificadas por meio da análise de desvio angular e taxa de variação. Um modelo de regressão polinomial treinado com cenários simulados de falha estima as perdas de potência correspondentes, ajustando dinamicamente o número esperado de *strings* por *stringbox*. As medições de corrente das *stringboxes* dentro das subunidades (SUs) da usina são agrupadas com o algoritmo k-means para estabelecer linhas de base de comportamento saudável, e desvios em relação a elas permitem a estimativa de *strings* ausentes. Além disso, um critério de recorrência é aplicado para reduzir alertas esporádicos isolados. Validada em uma usina fotovoltaica de 30 MW, a metodologia resultou em uma redução de 71% nos falsos alertas e alcançou 83% de acurácia na distinção entre falhas relacionadas a *strings* e a *trackers*. A abordagem não requer sensores adicionais, aproveitando a infraestrutura SCADA existente para aumentar a confiabilidade operacional.

Palavras-chave: Fotovoltaica, Usinas fotovoltaica de grande escala, Detecção de falhas, Dados operacionais.

ABSTRACT

This work proposes a novel methodology for detecting missing strings, i.e., strings that become electrically disconnected due to faults, in utility-scale photovoltaic power plants (USPVPPs) by integrating operational supervisory control and data acquisitions (SCADAs) system data from both tracker and stringbox levels. Tracker faults are first identified via angular deviation and change-rate analysis. A polynomial regression model trained on simulated fault scenarios estimates the corresponding power losses, dynamically adjusting the expected number of strings per stringbox. The current measurements from stringboxes within plant subunits (SUs) are clustered using k-means to establish healthy behavior baselines, and deviations from them enable missing string estimation. Additionally, a recurrence criterion is applied to reduce spurious onetime alerts. Validated on a 30 MW photovoltaic (PV) plant, the method led to a 71% reduction in false alerts and achieved 83% accuracy in distinguishing between string-related and tracker-related faults. It requires no additional sensors, leveraging existing SCADA infrastructure to enhance operational reliability.

Keywords: Photovoltaic, Utility-scale PV power plants, Fault detection, Operational data.

RESUMO EXPANDIDO

Introdução

A energia solar vem apresentando rápida expansão mundial, impulsionada por fatores como a redução no preço dos módulos e incentivos políticos à transição energética. Esses fatores tornam a energia fotovoltaica uma das principais fontes de geração da atualidade (IEA, 2024). Devido ao aumento na oferta e à natureza imediata da energia solar, usinas fotovoltaicas geralmente operam com baixas margens de lucro, o que torna essencial o monitoramento da degradação e das falhas nos equipamentos.

O monitoramento e a manutenção de usinas solares fotovoltaicas de grande escala (USPVPPs) apresentam desafios significativos devido à sua extensão física. Como essas usinas são compostas por milhares de módulos e centenas de *stringboxes*, o acompanhamento individual de cada equipamento torna-se inviável pelo alto custo de instalação, manutenção e armazenamento de dados. Por esse motivo, os sensores costumam ser instalados em pontos estratégicos, como inversores e *stringboxes*, o que reduz custos, mas dificulta a detecção de falhas localizadas.

Diversas técnicas têm sido propostas para mitigar perdas em USPVPPs, incluindo modelos baseados em aprendizado de máquina, simulações físicas da usina e inspeções visuais por drones ou câmeras terrestres. Apesar de seus avanços, muitas dessas soluções dependem de instrumentação adicional, alta resolução temporal, alta granularidade ou infraestrutura dedicada à aquisição de imagens, o que nem sempre é viável na operação contínua de usinas já em funcionamento. As metodologias que utilizam apenas dados operacionais apresentam maior compatibilidade com a realidade do monitoramento, mas geralmente analisam as falhas de forma isolada, sem considerar efeitos combinados entre diferentes equipamentos.

A ausência dessa integração limita a precisão dos diagnósticos, especialmente em situações em que o mau funcionamento de um componente afeta o desempenho de outros, como ocorre em casos de curtos em módulos, desconexão de *strings* ou falhas em *trackers*. Esses problemas podem gerar alarmes falsos, deslocamentos desnecessários de equipes e desperdício de recursos. Para enfrentar essa lacuna, este trabalho propõe uma metodologia que expande o que foi desenvolvido por (Simão et al., 2024), identificando *strings* ausentes e considerando o efeito combinado das perdas causadas por falhas de *trackers*, utilizando exclusivamente a infraestrutura SCADA já disponível em uma USPVPP.

A proposta integra dados de *trackers* e *stringboxes* para a detecção de *strings* ausentes em USPVPs, incorporando um modelo de estimativa de perdas por desalinhamento e um critério de recorrência temporal para reduzir alarmes esporádicos. A abordagem foi concebida para operar a partir da infraestrutura de dados existente, que incluem dados operacionais como corrente da *stringbox* e posição angular do *tracker*, além de dados meteorológicos como irradiância global inclinada (GTI) e temperatura ambiente, dispensando a instalação de sensores adicionais. Essa integração permite não apenas melhorar a precisão da detecção, mas também distinguir perdas atribuídas a *strings* daquelas associadas a *trackers*, oferecendo subsídios mais precisos para a tomada de decisão em manutenção e operação. Assim, o trabalho contribui com uma solução prática, de baixo custo e alinhada às necessidades reais de monitoramento de usinas solares fotovoltaicas de grande escala.

Objetivos

O objetivo deste trabalho é desenvolver uma metodologia capaz de distinguir falhas em *strings* de falhas em *trackers* em USPVPs.

A partir desse objetivo geral, foram definidos os seguintes objetivos específicos:

- Desenvolver um algoritmo para detecção de falha em *trackers*.
- Obter um modelo que traduz a falha em *trackers* para a sua perda relacionada.
- Ampliar a metodologia de detecção de *strings* ausentes desenvolvida por (Simão et al., 2024) para que seja considerada a correlação temporal dos alertas.
- Permitir a análise de perda energética de uma USPVP, separando perdas causadas por *trackers* e por *strings* ausentes.

Metodologia

A metodologia proposta compreende quatro etapas principais:

- Limpeza e filtragem dos dados, removendo leituras inválidas e períodos com forte *curtailment*.
- Detecção de desalinhamento de *trackers* por meio da análise da taxa de variação angular e do desvio em relação à trajetória ideal calculada pelo PVLIB (Holmgren et al., 2018).
- Estimativa das perdas de potência associadas ao desalinhamento, utilizando um

modelo de regressão polinomial treinado a partir de simulações com dados meteorológicos reais.

- Detecção de *strings* ausentes baseada na metodologia apresentada em (Simão et al., 2024), que consiste na análise de correntes das *stringboxes* via agrupamento k-means, ajustando o número esperado de *strings* com base nas perdas estimadas dos *trackers* e aplicando um critério de recorrência para reduzir alarmes isolados.

Resultados

A metodologia foi validada em uma USPVP de 30 MW no Nordeste do Brasil, utilizando seis meses de dados operacionais com resolução de 10 minutos. Os resultados mostraram que a introdução do critério de recorrência reduziu em até 71% o número de alarmes falsos, enquanto o ajuste para perdas de *trackers* permitiu reclassificar corretamente 83% dos casos antes atribuídos incorretamente a falhas de *strings*. Além disso, a abordagem proposta possibilitou estimar separadamente as perdas de energia associadas a falhas de *trackers* e de *strings*, oferecendo informações mais precisas para a priorização de ações de manutenção e otimização da operação.

Considerações finais

A metodologia desenvolvida mostrou-se eficaz na detecção de *strings* ausentes e na redução de alarmes falsos, distinguindo também as causas das perdas de energia. Por utilizar exclusivamente dados já disponíveis no SCADA, ela pode ser implementada sem custos adicionais com instrumentação, ampliando a aplicabilidade prática. Como trabalhos futuros, sugere-se a inclusão do efeito de *clipping* na análise e a extensão do método para diferentes configurações de usinas e arquiteturas de monitoramento.

Palavras-chave: Fotovoltaica, Usinas Solares de Grande Porte, Detecção de falhas, Dados Operacionais.

LIST OF FIGURES

Figure 1 – Centralized inverter architecture, with N stringboxes connected to an inverter. Strings are placed over trackers, which support one or more strings.	23
Figure 2 – Sample data from a clear-sky day illustrating normalized AC power output, tracker angle, and GTI measurements. A distinct curtailment event is clearly visible in the power profile.	25
Figure 3 – Comparison of power output under three tracker behaviors: ideal tracking, fixed at 0° (stowed position), and fixed at -55° . The deviation from the ideal path results in reduced energy production.	26
Figure 4 – Overview of the proposed methodology for missing string detection. The process comprises four main stages: (1) data cleaning and filtering, (2) tracker misalignment detection, (3) power loss estimation, and (4) estimation of missing strings through cluster-based current deviation analysis.	27
Figure 5 – (a) Ideal tracker path compared with healthy, stowed (-5°), and faulty trackers. (b) Change rate of tracker angle, highlighting the initial deviation triggering fault detection. (c) Angular deviation from the ideal path, emphasizing the first threshold exceedance, indicating a fault.	29
Figure 6 – Illustration of the energy produced by an ideal tracking path and a fixed angle tracker fault, highlighting the energy lost in the proposed period of the day due to the tracker malfunctioning.	30
Figure 7 – Flowchart of the proposed missing strings algorithm, accounting for tracker power loss and fault recurrence.	32
Figure 8 – Comparison of the total number of alerts generated under different recurrence settings, with and without tracker-related loss correction.	35
Figure 9 – Comparison of estimated energy losses using the baseline method (Simão et al., 2024) and the proposed method, which separates tracker-related and string-related losses.	36

LIST OF ACRONYMS

ASPP	atrous spatial pyramid pooling
CSI	clear sky index
CUSUM	cumulative sum
DCNN	deep convolutional neural network
DNN	deep neural network
GHI	global horizontal irradiance
GTI	global tilted irradiance
PV	photovoltaic
R-CNN	region-based convolutional neural network
RLS	recursive least squares
SCADA	supervisory control and data acquisition
SU	subunit
UAV	unmanned aerial vehicle
USPVPP	utility-scale photovoltaic power plant

CONTENTS

1	INTRODUCTION	15
1.1	OBJECTIVES	16
1.1.1	Specific objectives	16
2	RELATED WORKS	18
3	SYSTEM MODEL	22
3.1	DATASET	23
4	PROPOSED METHODOLOGY	26
4.1	DATA CLEANING	27
4.2	TRACKER MISALIGNMENT DETECTION	28
4.3	TRACKER LOSSES ESTIMATION	28
4.4	MISSING STRINGS ESTIMATION	31
5	RESULTS	34
5.1	LOSSES ASSESSMENT	35
6	FINAL REMARKS	37
6.1	FUTURE WORK	38
	REFERENCES	39

1 INTRODUCTION

Solar energy is projected to be the main driver of global renewable energy expansion, accounting for nearly 80% of new capacity additions in the coming years (IEA, 2024). Political incentives, financial viability, and environmental concerns support this growth. Notably, the cost of photovoltaic (PV) modules dropped by 60% between 2022 and 2024, making utility-scale photovoltaic power plant (USPVPP), which are grid-connected installations typically exceeding 5 MW of capacity, together with onshore wind energy, the current most cost-effective sources of electricity generation in most of the world (IEA, 2024).

Despite this rapid growth, USPVPP systems still face significant challenges related to energy losses. Operational inefficiencies, such as soiling, module mismatch, shading, and equipment failures, can lead to substantial energy losses over time (Iftikhar et al., 2021; Alhmoud, 2023; Raptor Maps, 2025). The study in (Raptor Maps, 2025) reports that system underperformance has tripled over the past five years, with estimated losses reaching \$10 billion in centralized solar power systems.

In addition to the reported financial impact, power loss in USPVPP plants has been quantified at a global average of 5.77% of annual production (Raptor Maps, 2025). While some of the issues, such as heavy soiling or inverter failures, may be detected relatively easily through conventional monitoring or scheduled maintenance, others are more subtle. For example, a single disconnected string or a tracker misalignment may only cause localized losses that remain hidden at the plant level, accumulating into significant underperformance over time.

Industry practices such as infrared drone inspections, electroluminescence testing, and IV-curve tracing can provide detailed diagnostics, but they are costly, time-consuming, and often performed only a few times per year. As a result, many faults remain undetected between inspection cycles, emphasizing the need for monitoring solutions that can leverage existing supervisory control and data acquisition (SCADA) data to provide continuous, low-cost visibility into system performance.

Thus, in light of this recent increase in system underperformance (Raptor Maps, 2025), improving monitoring strategies has become essential to identify and address performance issues. In USPVPPs, maintaining detailed visibility into the health of individual components is a logistical and financial challenge, and detailed monitoring systems are still not widely implemented in many installations (Moreno-Garcia et al., 2016). High equipment counts often limit the feasibility of deploying fine-grained monitoring systems. As a result, many installations rely on supervision at the inverter or stringbox level only (Simão et al., 2024; Zhao et al., 2019), which may not detect localized minor issues. While the literature presents a variety of techniques for performance analysis and fault detection in PV systems, to the best of our knowledge, no prior studies have explored

the correlation between tracker and stringbox faults to accurately distinguish faulty or disconnected strings from misaligned or stuck trackers while using only operational data already existing in the USPVP.

This study addresses this gap by proposing a methodology to detect missing strings through the combined analysis of tracker and stringbox data. In this context, missing strings refer to strings that are electrically disconnected from the rest of the photovoltaic array, resulting in an open-circuit condition that prevents the entire string from contributing to power generation. The approach builds upon and expands the method proposed in (Simão et al., 2024), by incorporating tracker-induced power losses and an alert recurrence mechanism. Considering the temporal recurrence of alerts has proven to be an important aspect, as it reduced up to 71% of identified isolated alerts. Moreover, by explicitly modeling tracker-induced power losses, the methodology was able to correctly reclassify 83% of the tracker-related faults that were previously misidentified as string failures. As a result, the proposed method not only improves fault attribution between string and tracker components but also enhances energy loss estimates, all while requiring no additional instrumentation beyond the existing SCADA infrastructure.

The contributions of this work can be summarized as follows: (i) development of a tracker misalignment detection algorithm, (ii) design of an AI-based model that estimates power loss from tracker malfunctions, (iii) integration of tracker data into an existing missing-string detection methodology (Simão et al., 2024), (iv) introduction of a temporal alert recurrence mechanism to reduce false positives, and (v) development of an energy loss estimation model.

1.1 OBJECTIVES

The main goal of this work is to distinguish missing strings faults from tracker malfunctions in USPVP.

1.1.1 Specific objectives

- Develop a tracker misalignment detection algorithm.
- Obtain a model that takes a tracker fault as input and outputs a power loss.
- Improve the existing missing string methodology in the literature to consider a temporal correlation.
- Enable energy loss analysis separating tracker-induced from the string-induced.

The remainder of this work is organized as follows: Section 2 discusses related works. Section 3 describes the architecture of the USPVP system used in this study.

Section 4 details the proposed methodology, including data preprocessing and alert generation. Section 5 discusses the experimental results. Finally, Section 6 presents the conclusions and outlines directions for future research.

2 RELATED WORKS

Fault detection in USPVPP systems has been widely studied in the literature (Simão et al., 2024; Zhao et al., 2019; Ren et al., 2025; Wang, D. et al., 2023; Silva et al., 2025; Wang, Y. et al., 2024; Moreno et al., 2024; Qorbani et al., 2025; Chang et al., 2024; Anderson et al., 2022; Skomedal et al., 2021; Zhao et al., 2020; Gaaloul et al., 2025; Sheppard et al., 2022; Rao et al., 2021; Klyve et al., 2024; Li et al., 2020; Awedat et al., 2025; Rocha et al., 2023; Oliveira et al., 2023; Ramirez et al., 2024; Aarseth et al., 2022), with methodologies varying significantly depending on the available monitoring infrastructure. This section categorizes existing works according to data granularity and detection strategy.

Several studies focus on scalable and cost-effective fault detection by relying exclusively on existing SCADA infrastructure, thereby avoiding the need for additional instrumentation. Zhao et al. (Zhao et al., 2019) proposed a hierarchical, unsupervised framework for anomaly detection and classification using only inverter-level data. Wang et al. (Wang, D. et al., 2023) introduced a self-supervised pre-training approach to extract meaningful representations from large volumes of unlabeled SCADA data, enhancing downstream diagnostic tasks. Silva et al. (Silva et al., 2025) employed a Random Forest regressor with dynamic thresholds to detect power anomalies at the inverter level, achieving high accuracy with minimal data requirements. Wang et al. (Wang, Y. et al., 2024) utilized stringbox-level data to create a fault description library using unsupervised learning and then employed a deep neural network (DNN) to estimate module parameters and classify module-level faults. Moreno et al. (Moreno et al., 2024) proposed a methodology using recursive least squares (RLS) to detect abnormal operations in PV systems. A simplified model was used to identify real-time faults and anomalies and a complex model was used to estimate the plant's long-term degradation. Qorbani et al. (Qorbani et al., 2025) utilizes signal processing techniques to identify long-term degradation, separating it from other performance factors like seasonality, recoverable soiling, and abrupt system changes. Chang et al. (Chang et al., 2024) proposes an AI-assisted methodology for fault detection and failure mode diagnosis, it uses machine learning models to predict inverter power output, and a decision tree algorithm to diagnose specific failure modes like shading or fuse issues. Similarly, Anderson et al. (Anderson et al., 2022) estimated energy losses from tracker faults using only time-series power data, while Simão et al. (Simão et al., 2024) applied clustering techniques to stringbox current and spatial information to identify missing strings.

To improve fault localization and enable more precise detection, several works leveraged string-level, module-level measurements or higher resolution data, allowing finer-grained analysis of system behavior. Ren et al. (Ren et al., 2025) use features extracted from I–V curves obtained under controlled conditions to train a similarity le-

arning neural network, which accurately identifies degradation, partial shading, and short-circuit faults in a 50 MW USPVPP. Skomedal et al. (Skomedal et al., 2021) used cumulative sum (CUSUM) control charts applied to the relative specific yield of strings to detect persistent power losses as low as 1%. Zhao et al. (Zhao et al., 2020) proposed a collaborative filtering technique that estimated string current based on neighboring string behavior. Gaaloul et al. (Gaaloul et al., 2025) employed Random Forest and KNN models to classify healthy vs. faulty strings and modules using current and voltage signals. Instead of relying purely on measured data, Sheppard et al. (Sheppard et al., 2022) simulated I-V curves using the PVLIB library (Holmgren et al., 2018) and compared them with field measurements to identify string and tracker anomalies, applying quality filters to ensure data reliability. Rao et al. (Rao et al., 2021) combined autoencoders with traditional classifiers, achieving high accuracy in string fault classification. Klyve et al. (Klyve et al., 2024) analyzed how different temporal resolutions impact the modeling of hybrid PV–wind power plants. Using one year of 5-second resolution data, the study compares curtailment loss estimations obtained with high-resolution data (5 s, 1 min) against those derived from hourly data. The results show that relying on hourly data leads to an underestimation of actual curtailment losses.

In contrast to signal-based methods, imaging techniques offer a complementary approach for detailed module-level diagnostics. Li et al. (Li et al., 2020) used unmanned aerial vehicle (UAV) to capture images and employed a deep convolutional neural network (DCNN) for real-time fault identification, with uncertain detections sent to a ground station for retraining. Awedat et al. (Awedat et al., 2025) used thermal imaging with an enhanced U-Net architecture integrating Residual Blocks, atrous spatial pyramid pooling (ASPP), and Attention Mechanisms. This approach enhances feature extraction, contextual understanding, and fault localization, thereby reducing the number of false positives. Rocha et al. (Rocha et al., 2023) compared different neural network models on UAV-based infrared images, with region-based convolutional neural network (R-CNN) showing superior performance in defect classification. Oliveira et al. (Oliveira et al., 2023) proposed a framework for the automatic detection and localization of faults in USPVPPs. The approach integrates Mask R-CNN to identify failures from aerial thermography images, with the results combined using orthomosaicking techniques to generate a complete map of the PV plant, highlighting the detected faults. The system demonstrated high accuracy in detecting hotspots and disconnected sub-strings, thereby facilitating rapid maintenance. Ramirez et al. (Ramirez et al., 2024) developed an image-processing pipeline to segment PV modules from ground-level infrared thermography and identify hotspots at the module level. Aarseth et al. (Aarseth et al., 2022) estimated power output steps between parallel strings and validated the results through infrared thermography.

To support comparison and interpretation, Table 1 summarizes key aspects of the

reviewed literature, including data type, sampling resolution, fault targets, and modeling strategy. This classification highlights the diversity of approaches, ranging from high-resolution simulations and image-based inspections to coarse-grained SCADA-based diagnostics.

Although some studies target key objectives like string-level fault detection (Simão et al., 2024; Zhao et al., 2020), tracker fault identification (Anderson et al., 2022), or both (Sheppard et al., 2022), their methodologies either neglect the critical interaction between tracker and string behaviors or rely on high-frequency data typically unavailable in standard USPVPP SCADA systems. The methodology proposed in this work stands out by leveraging only low-sampling-rate (10-minute), stringbox-level operational data to jointly analyze missing strings and tracker misalignments. This integrated perspective enables accurate fault attribution and loss estimation while remaining compatible with the SCADA infrastructure typically deployed in real-world USPVPP.

Table 1 – Summary of Related Work in PV Fault Detection

Publication	Instrumentation	Sampling Time	Main Data Source	Fault Target	Model Type
(Simão et al., 2024)	Sensor	10 minutes	Stringbox	String	Algorithmic
(Zhao et al., 2019)	Sensor	1/5/10 minutes	String	String	Machine Learning
(Ren et al., 2025)	Sensor	–	String	String	Machine Learning
(Wang, D. et al., 2023)	Sensor	5 minutes	String	–	Machine Learning
(Silva et al., 2025)	Sensor	15 minutes	Inverter	Inverter	Machine Learning
(Wang, Y. et al., 2024)	Sensor	2 minutes	Stringbox	Module	Machine Learning
(Moreno et al., 2024)	Sensor	1 hour	Inverter	Inverter	Machine Learning
(Qorbani et al., 2025)	Sensor	–	Inverter	Degradation	Algorithmic
(Chang et al., 2024)	Sensor	5 minutes	Inverter	Degradation	Machine Learning
(Anderson et al., 2022)	Sensor	6 minutes	Stringbox	Tracker	Algorithmic
(Skomedal et al., 2021)	Sensor	1 minute	String	Module/String	Algorithmic
(Zhao et al., 2020)	Sensor	10 minutes	Stringbox	String	Algorithmic
(Gaaloul et al., 2025)	Sensor	15 minutes	Stringbox	Module/String	Machine Learning
(Sheppard et al., 2022)	Sensor	1 minute	Stringbox	String/Tracker	Algorithmic
(Rao et al., 2021)	Sensor	1 hour	Module	String	Machine Learning
(Klyve et al., 2024)	Sensor	5 sec/ 1 min/ 1 hour	Inverter	Curtailment	Algorithmic
(Li et al., 2020)	Image	–	String	Module	Machine Learning
(Awedat et al., 2025)	Image	–	String	Module	Machine Learning
(Rocha et al., 2023)	Image	–	String	Module	Machine Learning
(Oliveira et al., 2023)	Image	–	String	Module/String	Machine Learning
(Ramirez et al., 2024)	Image	–	String	Hotspot	Machine Learning
(Aarseth et al., 2022)	Sensor/Image	10 minutes	String	Bypass diode	Algorithmic

3 SYSTEM MODEL

The basis of a PV system is the photovoltaic cell, which is responsible for converting electromagnetic radiation into electrical energy (Zidane et al., 2023). Sets of these cells form modules (Zidane et al., 2023), which are typically connected in series to create a string. As stated on (Ali Khan et al., 2020), various string arranges may be utilized based on the particular application. These configurations range from simple series or parallel connections to more advanced topologies, such as total-cross-tied or honeycomb-like structures. Although complex arrangements can improve performance under shading and module mismatch conditions, the series–parallel configuration remains the most commonly adopted in USPVPP systems due to its simplicity, cost-effectiveness, and ease of maintenance (Awan et al., 2022).

Among the various system configurations, the most common in large PV systems are the centralized inverter, string inverter, and multi-string inverter (Kolantla et al., 2020). Each of these offers distinct trade-offs in terms of efficiency, reliability, and maintenance complexity. USPVPP systems typically use the centralized inverter configuration due to its greater ease of maintenance, which is a critical factor in large-scale deployments (Kolantla et al., 2020).

In centralized configurations, multiple strings are typically connected via a string-box, also known as a combiner box, which consolidates electrical outputs while providing overcurrent protection and enabling monitoring functionalities (Anderson et al., 2022; Zhao et al., 2019; Simão et al., 2024). Additionally, many USPVPP systems employ solar trackers, which are mechanical structures that adjust the modules' orientation throughout the day using either a predetermined path or an irradiance sensor, to considerably enhance energy yield (Rodríguez-Gallegos et al., 2020).

In such configurations, the inverter is not only responsible for the conversion from DC to AC power, but also performs maximum power point tracking (MPPT), which dynamically adjusts the voltage-current operating point of the connected strings to maximize energy extraction. In centralized inverter architectures, a single inverter typically manages one or two MPPT channels, each regulating a group of stringboxes. During curtailment events, this mechanism is intentionally altered to limit power output according to system constraints.

The overall system topology is illustrated in Fig. 1, which shows the typical component connections in a centralized USPVPP system. PV modules are connected in series to form strings, which are then grouped in parallel through stringboxes before delivering their output to a centralized inverter. In systems equipped with solar trackers, each tracker structure typically supports multiple strings that are electrically connected to the same stringbox.

As previously discussed, data sensing plays a critical role in the operation and

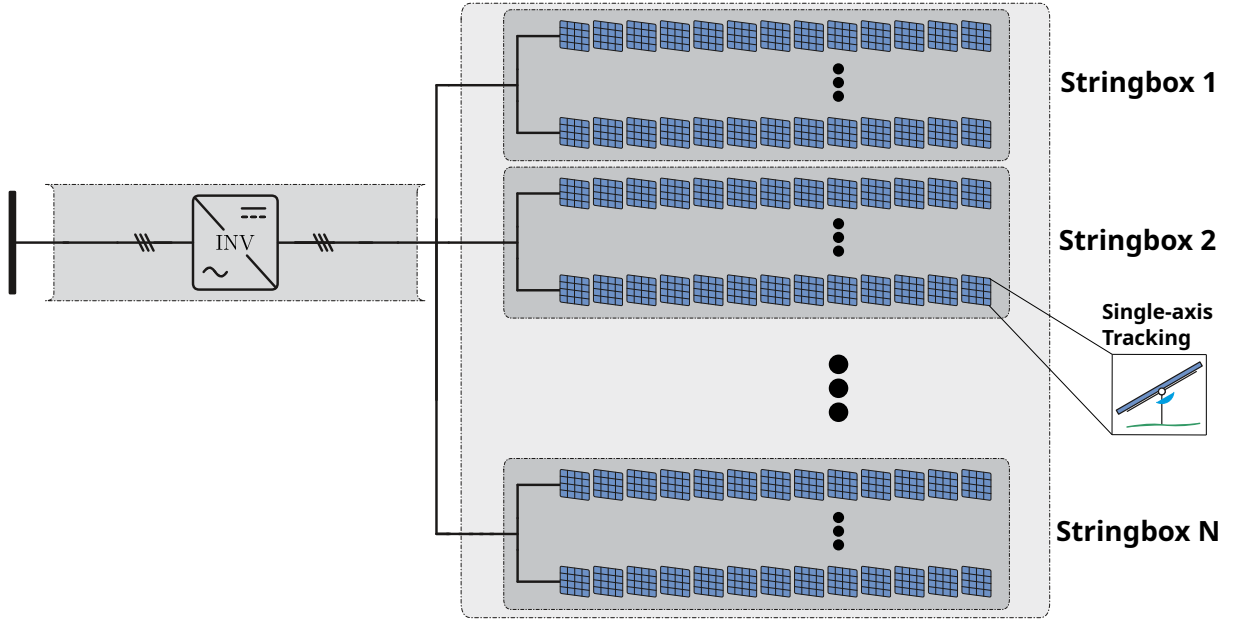


Figure 1 – Centralized inverter architecture, with N stringboxes connected to an inverter. Strings are placed over trackers, which support one or more strings.

maintenance of USPVPs. Due to logistical constraints, large-scale systems often rely on limited instrumentation, concentrating measurements primarily at stringboxes, trackers, and inverters, key components that aggregate data from multiple pieces of equipment (Anderson et al., 2022; Zhao et al., 2019; Simão et al., 2024).

3.1 DATASET

The dataset utilized in this work originates from a USPVP located in the Northeast of Brazil, with an installed capacity of approximately 30 MW, comprising approximately 2660 strings, organized into 180 stringboxes, which feed into 30 centralized inverters. Data acquisition is performed by a SCADA system and processed to produce measurements at a 10-minute sampling interval. The dataset includes both operational and environmental variables, measured at different hierarchical levels of the system over an 18-month period. An important aspect of the dataset is the presence of curtailment events, defined as intentional reductions in power output due to grid or operational constraints. The frequency of such events has increased alongside the growing penetration of decentralized renewable generation in the country (Operador Nacional do Sistema Elétrico (ONS), 2024).

The plant adopts the centralized inverter configuration. Internally, it is organized into subunits (SUs), each composed of three inverters and a step-up transformer that increases the voltage level. Multiple SUs are then connected to form circuits, and the full plant consists of an arrangement of these circuits. This hierarchical structure is utilized for data grouping and processing along this work.

Table 2 lists the variables used in this study, along with their respective measu-

Table 2 – Monitored variables used in the analysis.

Variable	Data granularity level
Direct current (DC)	Stringbox
Alternating current (AC) power	Inverter
Tracker angle	Tracker
Curtailement	Subunit
Global tilted irradiance (GTI)	Subunit
Ambient temperature	Plant
Wind speed	Plant
Global horizontal irradiance (GHI)	Plant

rement granularity. Environmental variables are collected by a meteorological station located near the power plant and are also integrated into the SCADA system with the same sampling interval.

To partially illustrate the data used in this study, Fig. 2 presents a sample from a cloudless day, including the global tilted irradiance (GTI), tracker angle, and inverter alternating current (AC) power output. A pronounced curtailement event is observed between 07:00 and 14:00 in the AC power curve, significantly limiting energy generation despite favorable irradiance and an optimal tracker trajectory. In the absence of curtailement, as the modules are continuously adjusted by the tracker, the GTI and AC power output exhibit strong correlation, apart from minor deviations caused by transient shadowing or system-level inefficiencies.

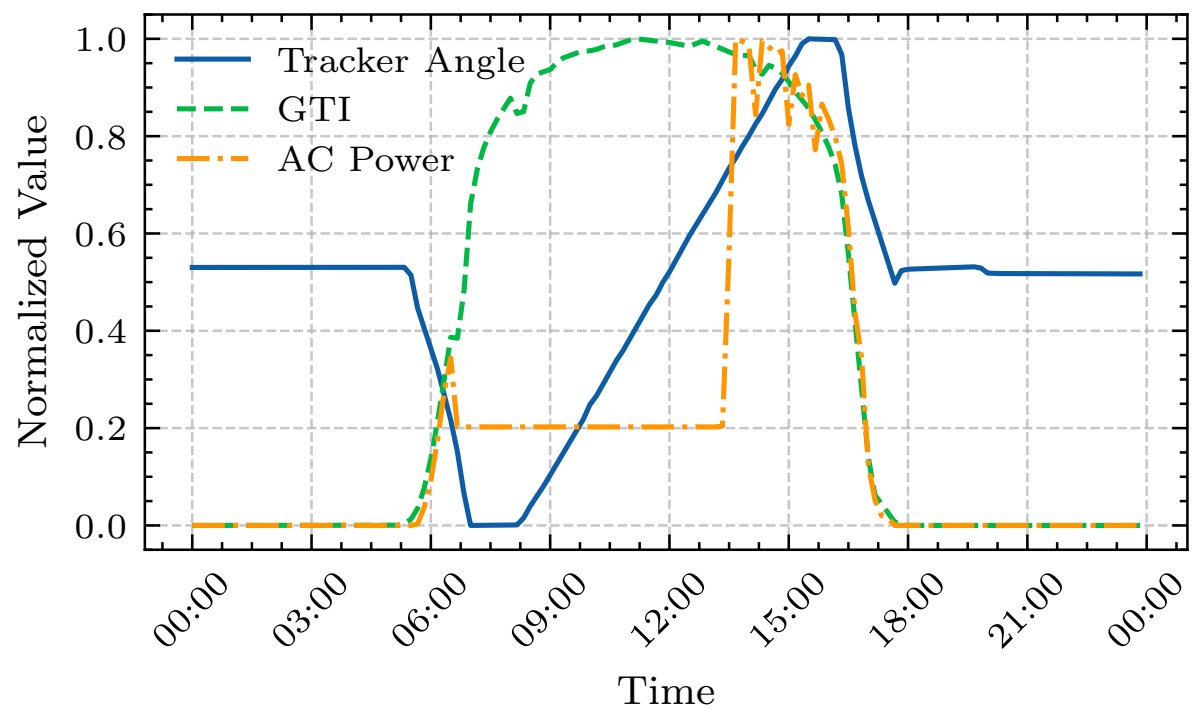


Figure 2 – Sample data from a clear-sky day illustrating normalized AC power output, tracker angle, and GTI measurements. A distinct curtailment event is clearly visible in the power profile.

4 PROPOSED METHODOLOGY

The proposed methodology extends (Simão et al., 2024) to detect missing strings by incorporating tracker data into the analysis. Unlike (Simão et al., 2024), which assumes faulty trackers always go to the horizontal wind-safe (stow) position and therefore lead to minimal energy loss, we consider all tracker failure scenarios, including both stowed trackers and those stuck at arbitrary angles for any period of time. Under the second condition, directly applying the original method would overestimate the number of missing strings and overlook the real fault, which could mislead maintenance prioritization.

Fig. 3 shows the power output for three tracker behaviors. The “Tracking” curve corresponds to the ideal trajectory that keeps the modules aligned with the sun. “Stowed 0°” represents the horizontal resting position assumed in (Simão et al., 2024). “Fixed 55°” illustrates a fault in which the tracker is stuck at -55° .

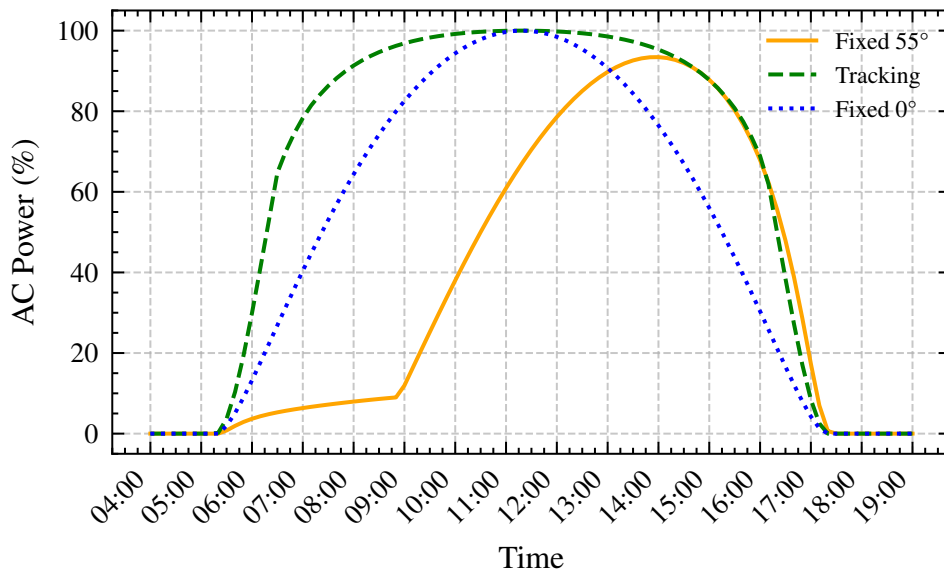


Figure 3 – Comparison of power output under three tracker behaviors: ideal tracking, fixed at 0° (stowed position), and fixed at -55° . The deviation from the ideal path results in reduced energy production.

Figure 3 shows that tracker faults at angles other than the horizontal stow position can cause substantial power losses. If only string-box current is analyzed, these losses may be misinterpreted as missing-string faults, yielding false positives. To avoid this problem, we introduce a methodology that combines additional contextual data with the string-box currents. Besides the measured currents, the method uses tracker angle readings and a recurrence analysis. An AI-based model then estimates the power loss attributable to tracker misalignment, allowing more reliable fault identification.

The complete workflow appears in the flowchart of Figure 4. The next subsections detail each stage: data cleaning, tracker-misalignment detection, tracker-loss esti-

mation, and the final calculation of missing strings.

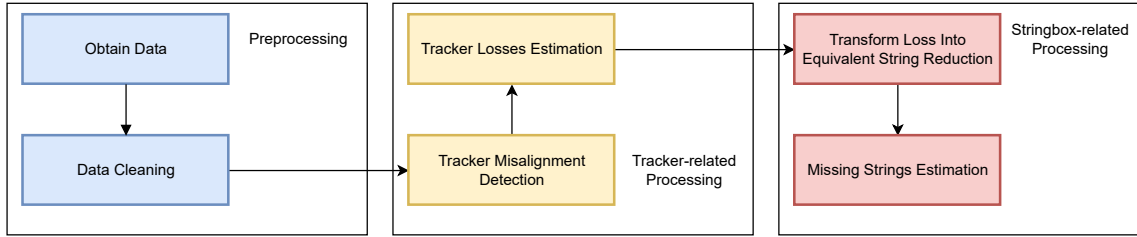


Figure 4 – Overview of the proposed methodology for missing string detection. The process comprises four main stages: (1) data cleaning and filtering, (2) tracker misalignment detection, (3) power loss estimation, and (4) estimation of missing strings through cluster-based current deviation analysis.

4.1 DATA CLEANING

Ensuring that the provided data is consistent and free from anomalies plays a crucial role in guaranteeing that the proposed method performs as intended. As described in Section 3.1, the dataset consists of 10-minute sampled data collected over one and a half years. The first year of data is employed to train and test the AI model responsible for estimating power losses due to tracker faults, while the remaining period is used as test data for the proposed methodology.

Given that the training data for the AI model and the data used in the overall methodology differ in nature and objectives, distinct preprocessing procedures are applied to each. The preprocessing of the dataset used on the AI model begins with the generation of an ideal global horizontal irradiance (GHI) curve for the plant's location, using the PVLIB library (Holmgren et al., 2018). This theoretical irradiance profile serves as a reference to filter invalid power data, excluding samples corresponding to periods when the ideal GHI is zero, which typically indicate physically implausible or invalid measurements. Additionally, to ensure data quality, outliers were removed using the interquartile range (IQR) method, where values falling below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR were excluded from the analysis (Tukey, 1977), and any day with curtailment was removed.

Regarding the data used for the overall methodology, all samples are filtered to correspond to the midday period (11:00–13:00), as it is most likely to contain the peak operating conditions, thus maximizing the contrast in performance between healthy and faulty strings, as well as avoiding inter-string shading. Additionally, negative readings of DC current or AC power are removed, as they stem from sensor errors and do not reflect valid operating behavior. Furthermore, days exhibiting curtailment intensity over 80% during the analysis period are excluded, since the artificial current limitation imposed by the MPPT under such conditions reduces the contrast needed for accurate fault detection.

4.2 TRACKER MISALIGNMENT DETECTION

The proposed method to identify faults consists of evaluating the change rate of the tracker angle, defined as the difference between consecutive samples, and its deviation from the ideal tracking path during the analysis period. The first metric detects abnormal behavior by assessing the consistency of movement among nearby trackers, while the second evaluates their conformity with the expected operation across the USPVP.

The change rate limit is obtained by grouping the trackers associated with the same SU, followed by estimation of the mean and standard deviation. Outliers are then identified as values exceeding 1.5 standard deviations from the mean and are classified as faults. For the second criterion, the ideal tracker path is generated using the PVLIB library (Holmgren et al., 2018), which takes the USPVP location and the data timestamp as inputs. A tracker is classified as faulty when its angle deviates from the ideal path by more than six degrees. This threshold accounts for a consistent offset observed across all trackers between the PVLIB-generated ideal path and the actual angles measured at the plant during specific periods of the year. A similar threshold is also adopted in (Anderson et al., 2022). If either of the two criteria is not satisfied, the corresponding data sample is flagged as faulty, and the timestamp and angular position are stored for further power loss estimation.

To illustrate the tracker's faulty behavior and the detection method, Fig. 5(a) presents the angular trajectories of an ideal unit, a healthy one, a stowed one, and a faulty one. The two identification metrics are analyzed in Figs. 5(b) and 5(c). Fig. 5(b) displays the change rate; this value remains relatively constant in the analysis period for both healthy and faulty trackers, but at the red point around noon, it deviates significantly from the ideal behavior, triggering detection by this criterion. Fig. 5(c) shows the angular deviation from the ideal path, highlighting where the faulty trackers exceed the predefined threshold.

4.3 TRACKER LOSSES ESTIMATION

Accurately estimating energy losses from tracker malfunctions in USPVPs is challenging because many interacting factors affect system performance, including irradiance variability, ambient conditions, tracker mechanics, and inter-string shading. The USPVP analyzed in this study features a high-fidelity simulator capable of quantifying these factors in detail. However, its computational demands make it impractical for real-time or large-scale operational use. To enable fast and reliable loss estimation during plant operation, we therefore developed an AI-based model that emulates the simulator's behavior while substantially reducing computational cost.

To develop the tracker-angle-based power loss estimation model, several simu-

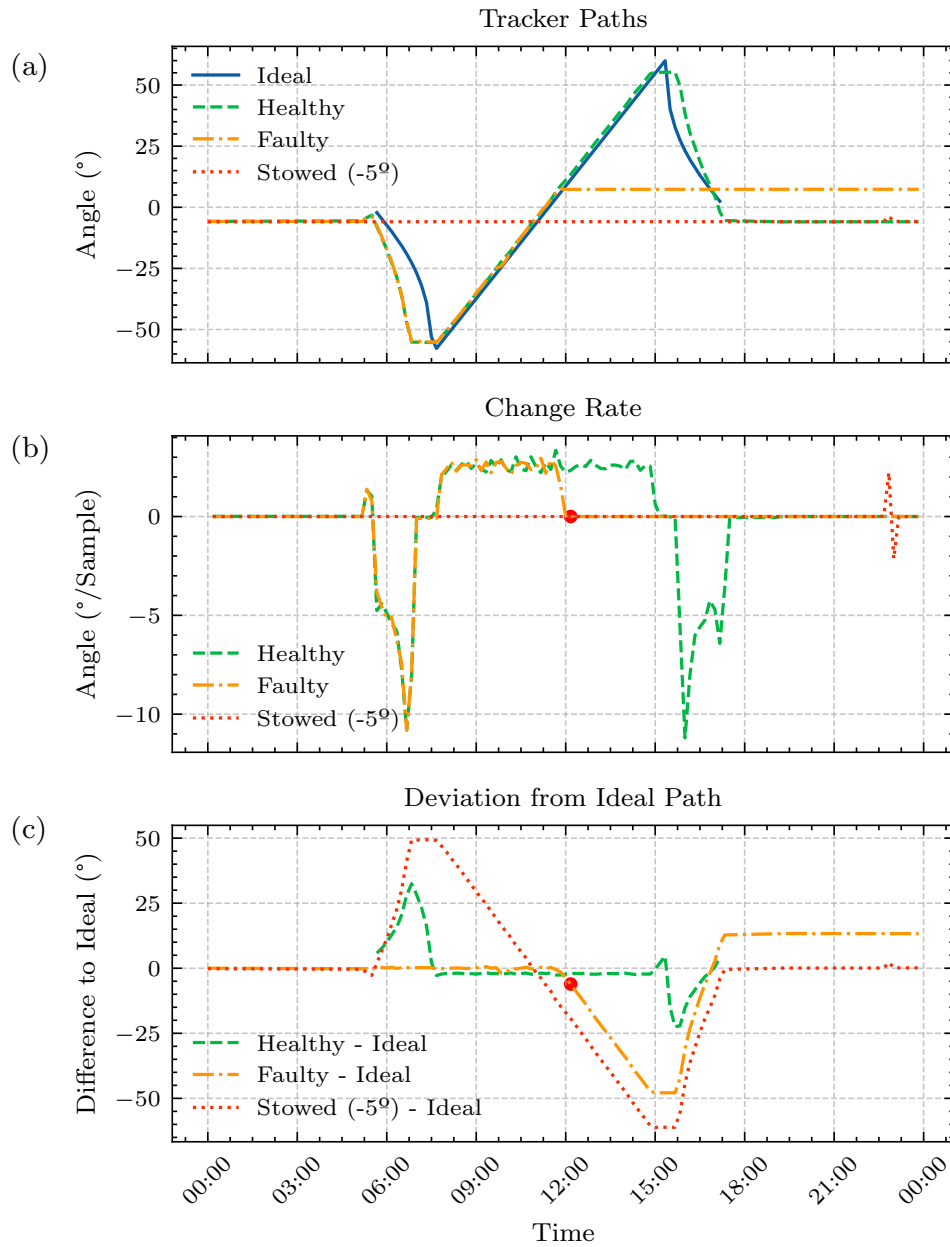


Figure 5 – (a) Ideal tracker path compared with healthy, stowed (-5°), and faulty trackers. (b) Change rate of tracker angle, highlighting the initial deviation triggering fault detection. (c) Angular deviation from the ideal path, emphasizing the first threshold exceedance, indicating a fault.

lations of the PV system were performed using actual meteorological data, including both ideal tracking scenarios and multiple faulty tracker scenarios, characterized by different malfunction angles. The simulator combines measured GTI, wind speed, and ambient temperature with synthetic tracker trajectories to generate power output curves. Fig. 6 illustrates the energy loss caused by a stuck tracker over the selected analysis period. The primary objective of this step is to express this loss as a percentage so that subsequent missing-string estimations can properly account for it.

The USPVP simulator was used to generate accurate power output data under

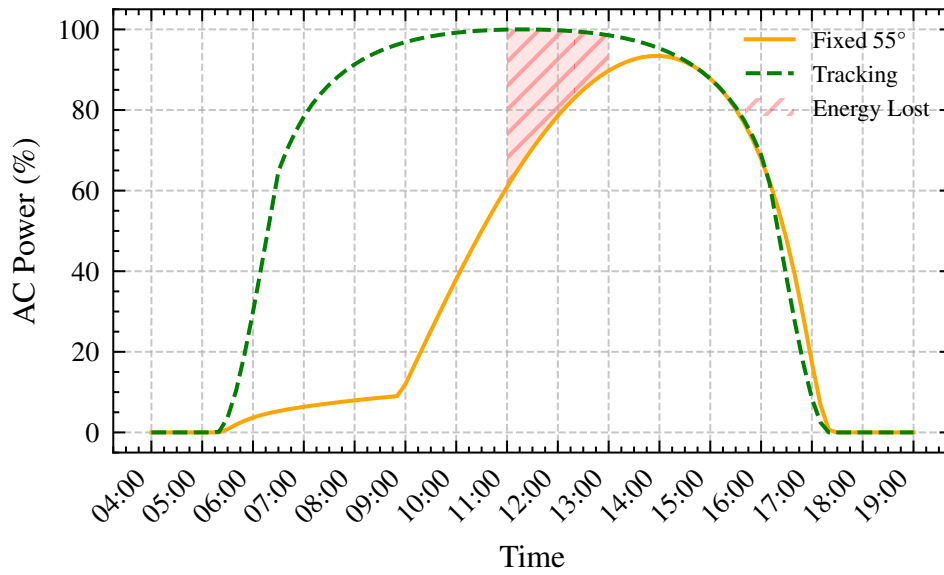


Figure 6 – Illustration of the energy produced by an ideal tracking path and a fixed angle tracker fault, highlighting the energy lost in the proposed period of the day due to the tracker malfunctioning.

multiple tracker fault scenarios over one year of real ambient data. These simulations served as the training basis for the AI model, which estimates power losses directly from operational conditions. To avoid inter-string shading and maintain a high operating point, the midday period was chosen as the analysis period. In this configuration, the AI model uses only two input features: (i) the clear sky index (CSI) (Marty et al., 2000), defined as

$$CSI = \frac{GHI_{\text{real}}}{GHI_{\text{theoretical}}} \quad (1)$$

where GHI_{real} is the measured horizontal irradiance, and $GHI_{\text{theoretical}}$ is the ideal irradiance profile generated by PVLIB; and ii) the tracker malfunction angle.

The model's output is the estimated power loss, expressed as a percentage. This loss is computed by comparing the energy produced in the ideal scenario to that produced under each fault condition within the midday period previously mentioned, according to

$$\Delta P \% = \frac{E_{\text{ideal}} - E_{\text{faulty}}}{E_{\text{ideal}}} \times 100, \quad (2)$$

where E_{ideal} is the total energy generated during the ideal tracking scenario, and E_{faulty} is the energy under a specific fault condition.

As stated in Section 3.1, one year of operational data was utilized to train and test the model. Considering the seasonal variability inherent to PV data, the test dataset was constructed by randomly selecting one representative month from each season, and the remaining data were used for training.

Table 3 – Performance metrics for the evaluated models.

Model	R^2	RMSE
Linear Regression	0.47	12.66
Polynomial Regression	0.96	3.55
Gradient Boosting Regressor	0.96	3.53

To balance predictive performance and model interpretability, three AI approaches were tested: a Gradient Boosting Regressor automatically selected via the TPOT AutoML library (Le et al., 2020) as the best performing method, and two simpler models, Linear and Polynomial Regression, chosen for their transparency and ease of interpretation.

The performance of the power loss estimation models was evaluated using two standard metrics, the coefficient of determination (R^2) and the root mean squared error (RMSE). The results are summarized in Table 3.

As shown in Table 3, both the Polynomial Regression and the Gradient Boosting Regressor models demonstrated superior predictive performance, achieving an R^2 of 0.96 and a significantly lower RMSE when compared to the Linear Regression model. The Polynomial Regression was then selected due to its greater interpretability and reduced complexity with respect to the Gradient Boosting Regressor.

4.4 MISSING STRINGS ESTIMATION

This step is based on the methodology proposed in (Simão et al., 2024), but with important modifications to account for the tracker malfunctions and temporal contextual data. The core concept involves clustering geographically proximate stringboxes and subsequently identifying distinct performance clusters within these groups. The median performance of the healthiest cluster is then established as a reference, enabling the detection of underperforming stringboxes that may indicate missing strings.

The first modification is the adoption of stringbox grouping by SU instead of clustering based on geographical coordinates as used in (Simão et al., 2024). This change was motivated by the heterogeneous module distribution of the analyzed USPVPP and the absence of a detailed module mapping, both of which undermine the reliability of the geographic clustering approach from (Simão et al., 2024). Two measures were therefore implemented to address this limitation: grouping by SU, and validation of the current-to-string ratio within each SU, ensuring that the nominal current of strings in the same group are equal.

The second modification was the incorporation of a recurrence criterion into the detection process: an alert is only triggered when underperformance is consistently observed over multiple consecutive days. This new criterion for alert generation showed a

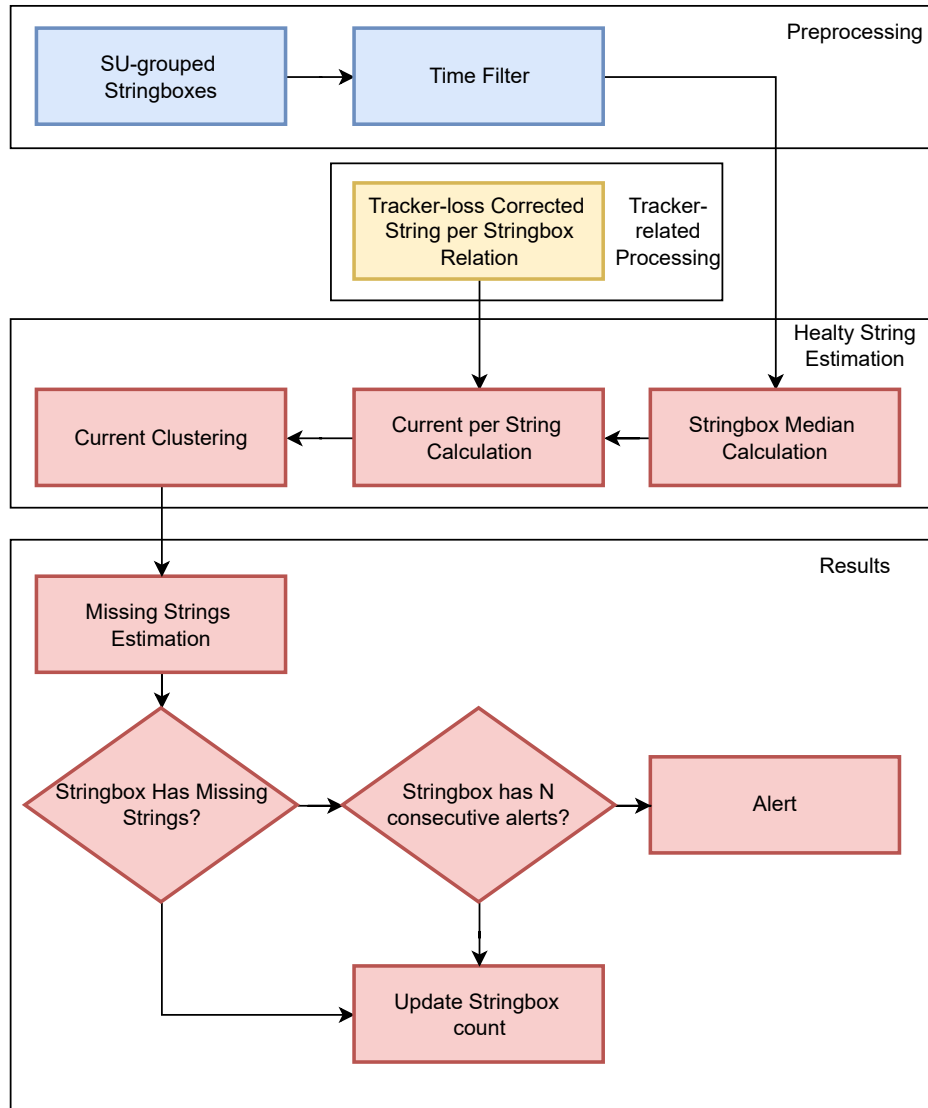


Figure 7 – Flowchart of the proposed missing strings algorithm, accounting for tracker power loss and fault recurrence.

great reduction in the false alarm ratio. While the recurrence criterion enhances reliability, it also introduces latency in fault detection, as N consecutive days of monitoring are required to confirm an alert.

As a third and larger modification, the proposed methodology integrates the tracker power loss estimation model from the tracker malfunction detection process. This integration allows for the adjustment of the expected number of strings per stringbox, effectively compensating for the equivalent power loss induced by tracker faults. The equivalent number of strings is calculated as

$$S_{eq,i} = S_i - (\Delta P \times S_i) \quad (3)$$

where S_i is the nominal number of strings connected to the stringbox i , and ΔP is the percentage of estimated power loss due to tracker malfunctioning.

Using the revised equivalent number of strings, faulty string identification is then carried out. Stringboxes from the same SU are grouped, midday current values are filtered, and the median current of each stringbox is calculated. The per-string current for each stringbox is then derived from its median current and the corresponding equivalent number of strings as

$$I_i^{\text{string}} = \frac{I_i^{\text{med}}}{S_{\text{eq},i}}, \quad (4)$$

where I_i^{med} is the median current of stringbox i , and $S_{\text{eq},i}$ is its equivalent number of strings as defined in (3).

Within each group, these per-string currents are clustered using k-means (MacQueen, 1967). The cluster with the highest current is considered healthy, and its median current is used as a reference. The number of missing strings for each stringbox is then computed as

$$M_i = S_{\text{eq},i} \cdot \left(1 - \frac{I_i^{\text{string}}}{I_{\text{ref}}^{\text{string}}} \right), \quad (5)$$

where I_i^{string} is the measured per-string median current in stringbox i , and $I_{\text{ref}}^{\text{string}}$ is the reference per-string median current obtained from the healthy cluster.

Then, if a stringbox is found to have one or more missing strings, it is flagged as a potential alert. If this condition persists for N consecutive days, the alert is confirmed as a true fault. To summarize the proposed methodology, Fig. 7 contains a flowchart with the key aspects of the procedure.

5 RESULTS

As previously described, the proposed methodology was evaluated using six months of operational data from a USPVPP with a capacity exceeding 30 MW. Three key modifications were made to the baseline method proposed in (Simão et al., 2024): (i) geographic clustering was replaced by SU-based clustering, (ii) a recurrence criterion was introduced to reduce spurious alerts, and (iii) tracker-related power losses were incorporated into the missing string estimation procedure.

To assess the impact of recurrence filtering and tracker loss estimation, Fig. 8 presents a comparison of the total number of alerts generated under different recurrence configurations, using two string clusters. The recurrence filter notably reduced the volume of alerts: applying a three-day recurrence criterion resulted in a 45% reduction, while a seven-day criterion led to a 71% reduction. These results indicate that many alerts identified by the baseline method were isolated events, likely false positives, given that string-level issues are typically resolved within one week or more. Fig. 8 also shows the tracker-related alert reduction, being around 2% with no recurrence, 1% with three-day recurrence, and 0.1% in the seven-day recurrence.

These findings also highlight that the choice of recurrence window should be application-dependent. Increasing the recurrence threshold substantially reduces the alert volume, but at the cost of delaying the detection of newly emerging issues. Thus, the recurrence configuration inherently represents a trade-off between detection efficiency and operational responsiveness, since larger windows slow down the activation of corrective actions. In practice, the optimal recurrence period should balance these competing objectives according to the specific requirements and dynamics of each USPVPP context.

To further evaluate the impact of accounting for tracker-related losses in the fault detection process, a detailed analysis was conducted using the configuration with two string clusters and a seven-day recurrence filter over the six-month test period. This assessment compared the baseline and the tracker-loss-corrected methodologies. Two main criteria were used to evaluate the differences: (i) whether the per-string current remained within its cluster-defined range, and (ii) whether alerts originally triggered by tracker faults were correctly removed by the updated estimation procedure.

The revised methodology correctly identified tracker-induced losses in 83% of the analyzed cases. These cases involved alerts initially classified as string failures by the baseline method, but later correctly attributed to tracker faults based on the newly proposed current estimation approach. By excluding tracker-induced deviations from the string current imbalance calculation, the proposed method reduced the likelihood of wrongly classifying tracker faults as string-level outages.

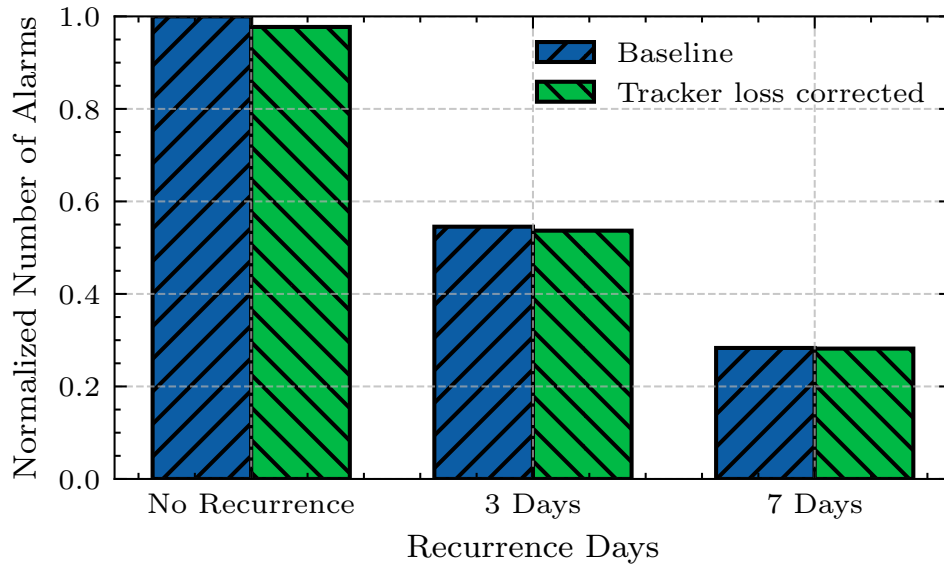


Figure 8 – Comparison of the total number of alerts generated under different recurrence settings, with and without tracker-related loss correction.

5.1 LOSSES ASSESSMENT

An additional advantage of the proposed methodology is its ability to separately estimate tracker- and string-related energy losses. The latter can be computed directly using the healthy current reference from each SU cluster, along with the number of missing strings detected. Tracker-related losses require two additional steps. First, since the original loss estimation model was trained using data from the midday period, suitable for string-related faults, a separate instance of the same model was trained using full-day data to properly estimate tracker-related losses, which occur throughout the day. Second, the misalignment detection algorithm was adapted by changing its analysis window to cover the entire day. These adjustments ensure that both the detection and quantification of tracker faults reflect their full impact.

A seven-day period, characterized by high CSI and minimal curtailment, was selected from the complete analysis of the USPVP to evaluate and compare loss estimates. Fig. 9 presents the baseline string-related losses (Simão et al., 2024), the losses estimated with the method proposed in this work, and also the explicitly separated tracker and string related losses for this period, considering a seven-day recurrence and two string clusters. Although only seven days are shown in the figure, each loss estimate incorporates a seven-day recurrence analysis, meaning that alerts are validated based on their presence across the seven days preceding each sample. The proposed methodology yields significantly lower total estimated losses. This reduction can be attributed to the exclusion of alerts caused by isolated or transient deviations, which is particularly important for the efficiency of maintenance activities, as it reduces false positives that would trigger unnecessary field inspections, therefore avoiding waste of

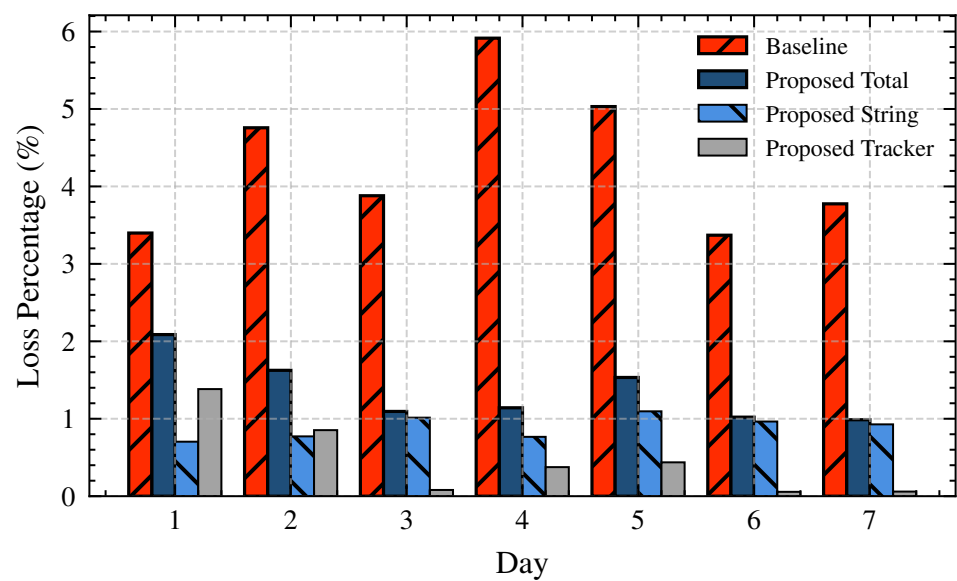


Figure 9 – Comparison of estimated energy losses using the baseline method (Simão et al., 2024) and the proposed method, which separates tracker-related and string-related losses.

financial and human resources.

6 FINAL REMARKS

This work set out to address one of the most persistent challenges in the monitoring of USPVPP: the reliable detection of missing strings under the constraints of limited instrumentation. By integrating tracker data and stringbox currents, supported by a regression-based model for power loss estimation and an alert recurrence criterion, the methodology demonstrated a significant advance over existing approaches. The proposed framework reduced false alarms by more than 70% and achieved an 83% accuracy rate in distinguishing between tracker-related and string-related failures. These outcomes highlight not only the robustness of the approach but also its compatibility with the operational realities of large solar plants, where fine-grained instrumentation is often not available.

From a scientific standpoint, this dissertation contributes to the literature by bridging two traditionally separate fault categories, tracker misalignments and missing strings, within a single analytical framework. While earlier works tended to analyze faults in isolation or required additional sensors, this study demonstrates that the combined use of SCADA data at different hierarchical levels can yield accurate, low-cost diagnostics. In addition, the choice of a polynomial regression model provided interpretability without sacrificing accuracy, strengthening the applicability of the approach in industrial contexts where transparent decision-making is valued. The introduction of a temporal recurrence mechanism further improved detection reliability, offering a means to filter out transient deviations that would otherwise burden operators with unnecessary maintenance actions.

In practical terms, the methodology aligns well with the pressing operational needs of USPVPP. By separating energy losses attributed to trackers from those caused by strings, plant operators gain a clearer picture of where to allocate maintenance resources. This capability reduces unnecessary field inspections, minimizes downtime, and ultimately improves the economic performance of the plant. Since the method relies exclusively on already available SCADA data, implementation is straightforward and does not incur additional capital or operational expenditures. These aspects are particularly relevant for plants in emerging markets, where cost constraints often limit the adoption of advanced diagnostic technologies.

Overall, this dissertation reinforces the importance of exploiting existing operational data to enhance fault detection in USPVPP. The results indicate that a careful combination of statistical modeling, clustering, and temporal validation can achieve accuracy levels comparable to more instrumented approaches, while maintaining compatibility with real-world monitoring infrastructures. By doing so, the study not only contributes to the academic understanding of PV fault detection but also provides a practical tool for the photovoltaic industry, supporting more efficient, reliable, and cost-

effective renewable energy generation.

6.1 FUTURE WORK

As clipping becomes more frequent with the increasing DC/AC ratio in photovoltaic systems, its impact can obscure underlying performance issues by artificially capping power output during periods of high irradiance. Therefore, incorporating clipping awareness into fault detection and performance analysis methodologies is a highly relevant direction for future work.

A major challenge faced during data preprocessing was the identification of inaccurate measurements, particularly from trackers. Future work could focus on developing automatic approaches to distinguish between real and imputed data. One possibility is the design of a methodology that validates tracker data using correlated measurements, such as current sensors, ensuring greater reliability of the input data.

Finally, the methodology presented in this work achieved promising results in the studied USPVPP. Extending and validating the proposed approach in other plants, with different configurations, geographic locations, and operating conditions, represents an important next step. Such generalization would strengthen the applicability of the method across diverse real-world scenarios.

REFERENCES

- AARSETH, Bjorn L.; SKOMEDAL, Asmund F.; OGAARD, Mari B.; MARSTEIN, Erik Stensrud. Detecting Permanently Activated Bypass Diodes in Utility-Scale PV Plant Monitoring Data. **IEEE Journal of Photovoltaics**, v. 12, n. 5, p. 1230–1236, 2022.
- ALHMOUD, Lina. Why Does the PV Solar Power Plant Operate Ineffectively? en. **Energies**, v. 16, n. 10, p. 4074, jan. 2023. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute. ISSN 1996-1073.
- ALI KHAN, Muhammad Yasir; LIU, Haoming; YANG, Zhihao; YUAN, Xiaoling. A Comprehensive Review on Grid Connected Photovoltaic Inverters, Their Modulation Techniques, and Control Strategies. en. **Energies**, v. 13, n. 16, p. 4185, jan. 2020. Number: 16 Publisher: Multidisciplinary Digital Publishing Institute. ISSN 1996-1073.
- ANDERSON, Kevin; DOWNS, Christopher; ANEJA, Saurabh; MULLER, Matthew. A Method for Estimating Time-Series PV Production Loss From Solar Tracking Failures. **IEEE Journal of Photovoltaics**, v. 12, n. 1, p. 119–126, jan. 2022. ISSN 2156-3403.
- AWAN, Kamran S.; ELGENDY, Mohammed A.; HUO, Dehong; ADHIKARI, Kabita; HASANIEN, Hany M. Impact of the Photovoltaic Array Configuration on its Performance under Partial Shading Conditions. In: 2022 57th International Universities Power Engineering Conference (UPEC). [S.l.: s.n.], 2022. p. 1–6.
- AWEDAT, Khalfalla; COMERT, Gurcan; AYAD, Mustafa; MREBIT, Abdulmajid. Advanced fault detection in photovoltaic panels using enhanced U-Net architectures. **Machine Learning with Applications**, v. 20, p. 100636, jun. 2025. ISSN 2666-8270.
- CHANG, Maoyi; CHEN, Kun-Hong; CHEN, Yu-Sheng; HSU, Chung-Chian; CHU, Chia-Chi. Developments of AI-Assisted Fault Detection and Failure Mode Diagnosis for Operation and Maintenance of Photovoltaic Power Stations in Taiwan. **IEEE Transactions on Industry Applications**, v. 60, n. 4, p. 5269–5281, 2024.
- GAALOUL, Yasmine; BEL HADJ BRAHIM KECHICHE, Olfa; OUDIRA, Houcine; CHOUDER, Aissa; HAMOUDA, Mahmoud; SILVESTRE, Santiago; KICHOU, Sofiane. Faults Detection and Diagnosis of a Large-Scale PV System by Analyzing Power Losses and Electric Indicators Computed Using Random Forest and KNN-Based Prediction Models. en. **Energies**, v. 18, n. 10, p. 2482, jan. 2025. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute. ISSN 1996-1073.
- HOLMGREN, William F.; HANSEN, Clifford W.; MIKOFSKI, Mark A. pvlib python: a python package for modeling solar energy systems. en. **Journal of Open Source Software**, v. 3, n. 29, p. 884, set. 2018. ISSN 2475-9066.

IEA. **Renewables 2024**. [S.l.: s.n.], 2024.

<https://www.iea.org/reports/renewables-2024>. IEA, Paris. Licence: CC BY 4.0.

IFTIKHAR, Hamid; SARQUIS, Eduardo; BRANCO, P. J. Costa. Why Can Simple Operation and Maintenance (O&M) Practices in Large-Scale Grid-Connected PV Power Plants Play a Key Role in Improving Its Energy Output? en. **Energies**, v. 14, n. 13, p. 3798, jan. 2021. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute. ISSN 1996-1073.

KLYVE, Øyvind Sommer; GRAB, Robin; OLKKONEN, Ville; MARSTEIN, Erik Stensrud. Influence of high-resolution data on accurate curtailment loss estimation and optimal design of hybrid PV–wind power plants. **Applied Energy**, v. 372, p. 123784, out. 2024. ISSN 0306-2619.

KOLANTLA, Dharani; MIKKILI, Suresh; PENDEM, Suneel Raju; DESAI, Aditi Atul. Critical review on various inverter topologies for PV system architectures. **IET Renewable Power Generation**, v. 14, n. 17, p. 3418–3438, 2020. eprint: <https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-rpg.2020.0317>.

LE, Trang T; FU, Weixuan; MOORE, Jason H. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. **Bioinformatics**, Oxford University Press, v. 36, n. 1, p. 250–256, 2020.

LI, Xiaoxia; LI, Wei; YANG, Qiang; YAN, Wenjun; ZOMAYA, Albert Y. Edge-Computing-Enabled Unmanned Module Defect Detection and Diagnosis System for Large-Scale Photovoltaic Plants. **IEEE Internet of Things Journal**, v. 7, n. 10, p. 9651–9663, 2020.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: PROCEEDINGS of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press. [S.l.: s.n.], 1967.

MARTY, Christoph; PHILIPONA, Rolf. The clear-sky index to separate clear-sky from cloudy-sky situations in climate research. **Geophysical Research Letters**, v. 27, n. 17, p. 2649–2652, 2000. eprint: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2000GL011743>.

MORENO, Pablo; LAGUNA, Gerard; CIPRIANO, Jordi; LUNA, Alvaro. Detection of Abnormal Operation of PV Inverters Based on Regressive Prediction Models With Recursive Least Squares Training. **IEEE Journal of Emerging and Selected Topics in Industrial Electronics**, v. 5, n. 1, p. 138–148, 2024.

MORENO-GARCIA, Isabel M.; PALACIOS-GARCIA, Emilio J.; PALLARES-LOPEZ, Victor; SANTIAGO, Isabel; GONZALEZ-REDONDO, Miguel J.; VARO-MARTINEZ, Marta; REAL-CALVO, Rafael J. Real-Time Monitoring System for a

Utility-Scale Photovoltaic Power Plant. en. **Sensors**, v. 16, n. 6, p. 770, jun. 2016. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute. ISSN 1424-8220.

OLIVEIRA, Aline Kirsten Vidal de; BRACHT, Matheus Körbes; AGHAEI, Mohammadreza; RÜTHER, Ricardo. Automatic fault detection of utility-scale photovoltaic solar generators applying aerial infrared thermography and orthomosaicking. **Solar Energy**, v. 252, p. 272–283, mar. 2023. ISSN 0038-092X.

OPERADOR NACIONAL DO SISTEMA ELÉTRICO (ONS). **Plano da Operação Elétrica de Médio Prazo - PAR/PEL 2024**. [S.l.: s.n.], 2024.

<https://www.ons.org.br/paginas/energia-no-futuro/suprimento-eletrico/parpel2024/sumario-executivo>. Online; accessed: 2025-06-05.

QORBANI, Mahya; FREGOSI, Daniel; WIDRICK, Devin; PAYNABAR, Kamran. Advanced Signal Decomposition Analysis and Anomaly Detection in Photovoltaic Systems. **IEEE Journal of Photovoltaics**, v. 15, n. 1, p. 155–163, 2025.

RAMIREZ, David F.; PUJARA, Deep; TEPEDELENLIOGLU, Cihan; SRINIVASAN, Devarajan; SPANIAS, Andreas. Infrared Computer Vision for Utility-Scale Photovoltaic Array Inspection. In: 2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA). [S.l.: s.n.], 2024. p. 1–4.

RAO, Sunil; MUNIRAJU, Gowtham; TEPEDELENLIOGLU, Cihan; SRINIVASAN, Devarajan; TAMIZHMANI, Govindasamy; SPANIAS, Andreas. Dropout and Pruned Neural Networks for Fault Classification in Photovoltaic Arrays. **IEEE Access**, v. 9, p. 120034–120042, 2021.

RAPTOR MAPS. **2025 Raptor Maps Global Solar Report**. [S.l.], 2025. 2025 Edition. Disponível em: <https://raptormaps.com/resources/global-solar-report-2025>.

REN, Bo; WANG, Qianggang; ZHOU, Niancheng; LU, Jiangang; MEKHILEF, Saad. Fault diagnosis of photovoltaic systems based on global I-V features and similarity learning. **International Journal of Electrical Power & Energy Systems**, v. 171, p. 110987, out. 2025. ISSN 0142-0615.

ROCHA, Daniel; ALVES, João; LOPES, Vitor; TEIXEIRA, Jennifer P.; FERNANDES, Paulo A.; COSTA, Mauro; MORAIS, Modesto; SALOMÉ, Pedro M. P. Multidefect Detection Tool for Large-Scale PV Plants: Segmentation and Classification. **IEEE Journal of Photovoltaics**, v. 13, n. 2, p. 291–295, 2023.

RODRÍGUEZ-GALLEGOS, Carlos D.; GANDHI, Oktoviano; PANDA, S. K.; REINDL, Thomas. On the PV Tracker Performance: Tracking the Sun Versus Tracking the Best Orientation. **IEEE Journal of Photovoltaics**, v. 10, n. 5, p. 1474–1480, 2020.

SHEPPARD, Scott; COOK, Tim; FREGOSI, Daniel; PERULLO, Christopher; BOLEN, Michael. Field Experience Detecting PV Underperformance in Real Time

Using Existing Instrumentation. In: 2022 IEEE 49th Photovoltaics Specialists Conference (PVSC). [S.l.: s.n.], 2022. p. 0307–0313.

SILVA, João Lucas de Souza; PAULA, Marcelo Vinícius de; BARROS, Juliana de Souza Granja; BARROS, Tércio André Dos Santos. Anomaly Detection Workflow Using Random Forest Regressor in Large-Scale Photovoltaic Power Plants. **IEEE Access**, v. 13, p. 54168–54176, 2025.

SIMÃO, Tiago Edmir; VALLE, Bruno Castro; ROSA, Yago Castro; VARELA, Fernando Santos; HOELLER, Arliones; NETO, Mario de Noronha; FRIES, Carlos Ernani; SOUZA, Richard Demo. A Method for the Estimation of Missing Strings in Very-Large-Scale Photovoltaic Power Plants. **IEEE Journal of Photovoltaics**, v. 14, n. 5, p. 839–847, set. 2024. ISSN 2156-3403.

SKOMEDAL, Asmund F.; OGAARD, Mari B.; HAUG, Halvard; MARSTEIN, Erik Stensrud. Robust and Fast Detection of Small Power Losses in Large-Scale PV Systems. **IEEE Journal of Photovoltaics**, v. 11, n. 3, p. 819–826, maio 2021. ISSN 2156-3403.

TUKEY, J.W. **Exploratory Data Analysis**. [S.l.]: Addison-Wesley Publishing Company, 1977. (Addison-Wesley series in behavioral science, v. 2). ISBN 9780201076165.

WANG, Dejun; DUAN, Zhenqing; WANG, Wenbin; CHU, Jingchun; CUI, Qingru; ZHU, Runze; CUI, Yahui; ZHANG, You; YOU, Zedong. Self-supervised pre-training in photovoltaic systems via supervisory control and data acquisition data. en. **IET Cyber-Physical Systems: Theory & Applications**, v. 8, n. 4, 2023. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1049/cps2.12056>. ISSN 2398-3396.

WANG, Yifan; XU, Jianrong; XU, Fei; DENG, Long; JIN, Chensheng; BAO, Wenzhong; XIE, Yufeng. Contactless Diagnosis Method and Unsupervised Learning for Panel-Level Photovoltaic Plant Operation and Maintenance. **IEEE Access**, v. 12, p. 174808–174817, 2024.

ZHAO, Yingying; LI, Dongsheng; LU, Tun; LV, Qin; GU, Ning; SHANG, Li. Collaborative Fault Detection for Large-Scale Photovoltaic Systems. **IEEE Transactions on Sustainable Energy**, v. 11, n. 4, p. 2745–2754, out. 2020. ISSN 1949-3037.

ZHAO, Yingying; LIU, Qi; LI, Dongsheng; KANG, Dahai; LV, Qin; SHANG, Li. Hierarchical Anomaly Detection and Multimodal Classification in Large-Scale Photovoltaic Systems. **IEEE Transactions on Sustainable Energy**, v. 10, n. 3, p. 1351–1361, jul. 2019. ISSN 1949-3037.

ZIDANE, Tekai Eddine Khalil; AZIZ, Ali Saleh; ZAHRAOUI, Younes; KOTB, Hossam; ABORAS, Kareem M.; KITMO; JEMBER, Yosef Berhan. Grid-Connected Solar PV Power Plants Optimization: A Review. **IEEE Access**, v. 11, p. 79588–79608, 2023.