UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Identificação de anomalias ofensoras à geração de usinas solares fotovoltaicas

Guilherme Théo Bredemann da Silva

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m	Silva, Guilherme Théo Bredemann Identificação de anomalias ofensoras à geração de usinas solares fotovoltaicas / Guilherme Théo Bredemann da Silva ; orientador Jean Roberto Ponciano. – São Carlos, 2024. 79 p. : il. (algumas color.) ; 30 cm. Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024. 1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Ponciano, Jean Roberto, orient. II. Título.

Guilherme Théo Bredemann da Silva

Identificação de anomalias ofensoras à geração de usinas solares fotovoltaicas

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial e Big Data

Orientador: Prof. Dr. Jean Roberto Ponciano

Versão original

São Carlos

2024

Este trabalho é dedicado aos meus pais que nunca mediram esforços e dedicaram seu tempo de vida para que eu pudesse ter uma boa educação e instrução para abrir portas e chegar onde eu quiser.

AGRADECIMENTOS

Com este trabalho, desejo expressar minha profunda gratidão a todos que, de alguma forma, contribuíram para essa realização.

Em primeiro lugar aos meus pais, que me capacitaram para que pudesse escrever este trabalho.

Em segundo lugar, a minha companheira Laís, que me apoiou e suportou nos momentos difícieis, incluindo não me deixando jogar tudo para o ar e ir vender arte na praia. Sem ela não existiria esse trabalho.

Um enorme agradecimento ao meu orientador Jean que sempre esteve disponível e presente, me auxiliando e orientado de forma exímia por todo o percurso deste projeto.

E por fim, o MBA em Ciências de Dados oferecido pelo Centro de Ciências Matemáticas Aplicadas à Indústria (CeMEAI) e pelo Instituto de Ciências Matemáticas e de Computação (ICMC/USP) proporcionou um vasto conhecimento em Inteligência Artificial, toda a estrutura foi EAD e o suporte dos professores, tutores e equipe técnica/administrativa foi essencial para o melhor desenvolvimento dos alunos.

"Acredito que no final do século, o uso da palavra e opinião "máquina pensante" não será mais motivo de contradição, e que se poderá falar de máquinas pensando sem que alguém se oponha a isso."

Alan Turing

RESUMO

Silva, G. T S. **Identificação de anomalias ofensoras à geração de usinas solares fotovoltaicas**. 2024. 79p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

A operação e manutenção de usinas solares destinadas à geração distribuída em larga escala apresentam desafios significativos no que se refere à manutenção da rentabilidade dos projetos, exigindo o desenvolvimento de soluções confiáveis e de menor custo para garantir a eficiência da geração de energia. Neste contexto, o presente trabalho propõe e testa diferentes modelos com variados custos computacionais e abordagens metodológicas para a identificação de anomalias que não são reportadas diretamente pelos inversores, como acúmulo de sujeira, trincas em painéis, células queimadas, entre outros problemas que impactam a geração de energia. Dentre os métodos empregados, destacam-se a detecção de *drifts* e o uso de redes neurais recorrentes do tipo *Long Short-Term Memory* (LSTM). Os dados utilizados no desenvolvimento deste estudo foram coletados a partir de equipamentos com tecnologia de Internet das Coisas (IoT), abrangendo informações reais de usinas solares, obtidas de inversores, estações solarimétricas e medidores de concessionárias, com periodicidade de 15 minutos e histórico iniciado em 2020. As variáveis monitoradas incluem geração de energia, irradiação solar e temperatura dos módulos, permitindo avaliar as condições mínimas necessárias para a aplicação de técnicas de aprendizado de máquina. Além disso, buscou-se evitar a necessidade de telemetrias mais sofisticadas que aumentariam os custos do projeto. Entre os resultados, o modelo de detecção de drift utilizando o método Page-Hinkley apresentou uma assertividade superior a 50% em comparação com os problemas identificados pela equipe de Operação e Manutenção. Devido ao seu baixo custo computacional, este modelo se mostrou adequado para ser embarcado em dispositivos de telemetria, realizando a detecção de drifts diretamente na borda. Já o modelo LSTM demonstrou bom desempenho na previsão da geração de energia, combinando variáveis de irradiação e temperatura, configurando-se como uma solução mais robusta para o monitoramento do desempenho da usina. Esse modelo possibilita a comparação entre os valores preditos e os valores reais de geração, além de ser compatível com a aplicação de detectores de drift na saída, permitindo a previsão de quedas consistentes na geração.

Palavras-chave: *Drift. Page-Hinkley.* LSTM. Geração. Temperatura do módulo. Irradiação. Inversor. Estação solarimétrica.

ABSTRACT

Silva, G. T S. **Identification of Anomalies Affecting the Power Generation of Photovoltaic Solar Plants**. 2024. 79p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

The operation and maintenance of solar power plants dedicated to large-scale distributed generation present significant challenges in maintaining project profitability, requiring the development of reliable and cost-effective solutions to ensure generation efficiency. In this context, the present work proposes and tests different models with varying computational costs and methodological approaches to identify anomalies not directly reported by inverters, such as dirt accumulation, cracked panels, burned cells, among other issues that impact energy generation. Among the methods employed, the drift detection and Long Short-Term Memory (LSTM) recurrent neural networks are highlighted. The data used in this study were collected from Internet of Things (IoT) devices, including real data from solar plants, obtained from inverters, solarimetric stations, and utility meters, with 15-minute periodicity and historical data starting from 2020. The monitored variables include energy generation, solar irradiation, and module temperature, allowing for the assessment of the minimum conditions necessary for applying machine learning techniques. Additionally, the need for more sophisticated telemetry systems, which would increase project costs, was avoided. Among the results, the drift detection model using the Page-Hinkley method showed an accuracy rate above 50% compared to issues identified by the Operation and Maintenance team. Due to its low computational cost, this model proved suitable for being embedded in telemetry devices, performing drift detection at the edge. On the other hand, the LSTM model demonstrated good performance in forecasting energy generation when combined with irradiation and temperature variables, making it a more robust solution for monitoring plant performance. This model allows for the comparison between predicted and actual generation values, and it is also compatible with the application of drift detectors at the output, enabling the prediction of consistent drops in generation.

Keywords: LSTM, Generation, Module Temperature, Irradiation, Inverter, Solarimetric Station.

LISTA DE FIGURAS

Figura 1 – UFV Rio das Flores - Solarian.[1] 25

Figura 2 – Processo de transição de um elétron da banda de valência para a banda de condução. Extraído de (MARQUES *et al.*, 1997). 29

Figura 3 – Desenho de uma estação solarimétrica à esquerda, extraído de (SCHUBERT, 2019) e estação solarimétrica real, extraído de (HUKSEFLUX, 2024). 31

Figura 4 – À esquerda, curva *I-V* do painel solar série CS3W extraído de (CANADIANSOLAR, 2020) e curva *I-V* com indicação de SPPM por irradiação extraído de (SEYEDMAHMOUDIAN *et al.*, 2016). 34

Figura 5 – Componentes da radiação solar. Adaptado de (PINHO *et al.*, 2008). . . 35

Figura 6 – Curva de *derating* de potência por temperatura ambiente. Extraído de (ELECTRONICS, 2021). 37

Figura 7 – Ilustração de tipos de velocidade de desvio de dados. Extraído de (LU *et al.*, 2018). 39

Figura 8 – Ilustração de um neurônio computacional. Extraído de (HAYKIN, 2001). 42

Figura 9 – Ilustração de um bloco LSTM. Adaptado de (YAN, 2016). 43

Figura 10 – Média de geração em kW explicada pela variável *pm_dc_w* por horas do dia. 54

Figura 11 – Gráficos de calor de inversores mostrando o percentual de dados válidos ao longo dos anos de 2022 a 2024 pela variável *pm_dc_w*. 55

Figura 12 – Gráficos de calor de estação solarimétrica mostrando o percentual de dados válidos ao longo dos anos de 2022 a 2024 pelas variáveis *global_irradiation_1*, *global_irradiation_3* e *temperatura_modulo_1*. 56

Figura 13 – Histograma das variáveis *global_irradiation_1*, *temperatura_modulo_1* e *pm_dc_w*. 57

Figura 14 – Dados históricos das variáveis *global_irradiation_1*, *temperatura_modulo_1* e *pm_dc_w*. 58

Figura 15 – Decomposição sazonal das variáveis *pm_dc_w*, *global_irradiation_1* e *temperatura_modulo_1*. 59

Figura 16 – Gráficos de *drifts* identificados nas variáveis *pm_dc_w* (GER), *temperatura_modulo_1* (TMP) e *global_irradiation_1* (IRR) representados por tracejados vermelhos verticais junto com ocorrências no transformador destacadas como faixas horizontais. 67

Figura 17 – Gráficos de *drifts* identificados nas variáveis *pm_dc_w* (GER), *temperatura_modulo_1* (TMP) e *global_irradiation_1* (IRR) representados por tracejados vermelhos verticais junto com ocorrências em inversores destacadas como faixas horizontais. 67
Figura 18 – Gráficos de *drifts* identificados nas variáveis *pm_dc_w* (GER), *temperatura_modulo_1* (TMP) e *global_irradiation_1* (IRR) representados por tracejados vermelhos verticais junto com ocorrências em *trackers* destacadas como faixas horizontais. 68
Figura 19 – Gráficos de *drifts* identificados nas variáveis *pm_dc_w* (GER), *temperatura_modulo_1* (TMP) e *global_irradiation_1* (IRR) representados por tracejados vermelhos verticais junto com ocorrências na rede da concessionária destacadas como faixas horizontais. 68
Figura 20 – Gráficos das métricas de erro da rede neural LSTM ao longo das épocas de aprendizado. 72

LISTA DE TABELAS

Tabela 1 – Variáveis de análise de desempenho de usinas fotovoltaicas. 36
Tabela 2 – Variáveis do conjunto de dados. 50
Tabela 3 – Colunas do banco de dados de estudo. 51
Tabela 4 – Quantidade de amostras por variável de interesse. 54
Tabela 5 – Ocorrências e Impacto na Geração 64
Tabela 6 – Avaliação de *drifts*. 65
Tabela 7 – Estrutura da Rede Neural LSTM. 71
Tabela 8 – Resultado das métricas de desempenho da Rede Neural LSTM. 71

LISTA DE ABREVIATURAS E SIGLAS

ACO	Otimização por Colônia de Formigas (em inglês, *Ant colony optimization*)
ANEEL	Agência Nacional de Energia Eletrica
ANN	Redes Neurais Artificiais (em inglês, *Artificial Neural Networks*)
BOS	Equilíbrio do Sistema (em inglês, *Balance of System*)
CA	Corrente Alternada
CC	Corrente Contínua
CNPJ	Cadastro Nacional de Pessoa Jurídica
CPF	Cadastro de Pessoa Física
DHI	Irradiação Difusa Horizontal (em inglês, *Difuse Horizontal Irradiation*)
DNI	Irradiação Direta Normal (em inglês, *Direct Normal Irradiation*)
EPE	Empresa de Pesquisa Energética
ETL	Extração, Transformação e Carregamento (do inglês, *Extract, Transform and Load*)
GHI	Irradiação Global Horizontal (em inglês, *Global Horizontal Irradiation*)
GW	Gigawatt
IEC	Comissão Eletrotécnica Internacional (do inglês, *International Electrotechnical Commission*)
IFAE	Incentivo às Fontes Alternativas de Energia Elétrica
IHM	Interface Homem-Máquina
IoT	Internet das Coisas (em inglês, *Internet of Things*)
KNN	N-Vizinhos Próximos (do inglês, Key Nearest Neightbors)
kW	Kilowatt
kWh	Kilowatt hora
LGPD	Lei Geral de Proteção de Dados

LSTM	Memória de Curto Longo Prazo (do inglês, Long Short-Term Memory)
MAE	Erro Absoluto Médio (do inglês, Mean Absolute Error)
MLP	Perceptron de multicamadas (em inglês, *Multilayer Perceptron*)
MSE	Erro Quadrático Médio (do inglês, Mean Squared Error)
MPPT	O mesmo que SPPM
MW	Megawatt
PL	Projeto de Lei
PLC	Controlador Lógico Programável (em inglês, *Programmable Logic Controller*)
PV	Fotovoltaico (em inglês, *Photovoltaic*)
PRONASOLAR	Política Nacional de Energia Solar Fotovoltaica
PSO	Otimização por Enxame de Partículas (em inglês, *Particle swarm optimization*)
REH	Resolução Homologatória
REN	Resolução Normativa
RMSE	Raiz do Erro Quadrático Médio (do inglês, Root Mean Squared Error)
SCADA	Sistema de Supervisão e Aquisição de Dados
SGBD	Sistema Gerenciador de Banco de Dados
SIN	Sistema Interligado Nacional
SPPM	Seguidor do Ponto de Máxima Potência (em inglês, *Maximum Power Point Tracking*)
SQL	Linguagem de Consulta Estruturada (do inglês, Structured Query Language)
UFV	Usina Solar Fotovoltaica
O&M	Operação e Manutenção
USP	Universidade de São Paulo
USPSC	Campus USP de São Carlos

LISTA DE SÍMBOLOS

γ	Letra grega Gama
m^2	Metro quadrado, unidade de medida de área
%	Percentual
°C	Grau Celsius, unidade de medida de temperatura
θz	Ângulo zenital
V	Volt, unidade de medida elétrica
W	Watt, unidade de medida elétrica
A	Ampère, unidade de medida elétrica
f	Hertz, unidade de medida de frequência
P	Potência, unidade de medida elétrica
Ω	Letra grega Ômega, representa a resistência elétrica, unidade de medida elétrica

SUMÁRIO

1 INTRODUÇÃO 25
1.1 Contextualização, Motivação e Lacunas 25
1.2 Problema 26
1.3 Objetivos 27
1.3.1 Objetivo Geral 27
1.3.2 Objetivos Específicos 27
1.4 Organização do Texto 27

2 FUNDAMENTAÇÃO TEÓRICA 29
2.1 Energia Solar Fotovoltaica 29
2.1.1 Células e Painéis Solares 29
2.1.2 Instrumentação de Plantas Fotovoltaicas 30
2.1.2.1 Inversor 30
2.1.2.2 Estação Solarimétrica 30
2.1.2.3 Seguidor Solar 31
2.1.2.4 Sistema Supervisório 31
2.1.3 Variáveis de Análise e Controle de Usinas Fotovoltaicas 32
2.1.4 Falhas em Sistemas Fotovoltaicos 35
2.1.5 Geração Solar Distribuída no Brasil 37
2.2 Detecção de Desvios em Fluxo de Dados 38
2.2.1 Método Page-Hinkley 40
2.3 Classificadores 41
2.3.1 Redes Neurais Artificiais 41
2.3.2 LSTM 42

3 REVISÃO BIBLIOGRÁFICA 45
3.1 Contribuições 47

4 METODOLOGIA 49
4.1 Considerações Iniciais 49
4.1.1 Máquina Utilizada 49
4.2 Coleta de Dados 49
4.3 Criação de Base de Dados Dedicada e Carga de Dados 50
4.4 Análise Exploratória 51
4.5 Filtragem dos dados 53
4.6 Disponibilidade e características dos dados 54

4.7 Bibliotecas 59
4.7.1 *Drift* 59
4.7.2 Predição 59
4.7.3 Treinamento dos Detectores de Desvios 60
4.7.4 Treinamento do Preditor 60

5 AVALIAÇÃO EXPERIMENTAL 61
5.1 Detector de *Drift* 61
5.1.1 Resultados 64
5.1.2 Considerações Finais 68
5.2 LSTM 69
5.2.1 Resultados 71
5.2.2 Considerações Finais 72

6 CONCLUSÕES 73
6.1 Trabalhos Futuros 74

Referências 77

1 INTRODUÇÃO

1.1 Contextualização, Motivação e Lacunas

Nos últimos anos, o Brasil desenvolveu incentivos à energia solar, visando a segurança energética, desenvolvimento sustentável e mudanças climáticas (UCZAI, 2012). Dentre todos os incentivos, é importante mencionar o Incentivo às Fontes Alternativas de Energia Elétrica (IFAE), que oferece subsídios para a instalação de sistemas de geração de energia solar e a instituição da Política Nacional de Energia Solar Fotovoltaica (PRONASOLAR) em 2018, estabelecendo de linhas de crédito facilitadas para a energia solar, isenções fiscais para a compra de equipamentos de geração de energia solar; e medidas regulatórias para facilitar o acesso à rede elétrica para os geradores de energia solar.

Em janeiro de 2022, foi sancionado o Projeto de Lei 5829/19, tornando-se o novo Marco Legal Solar no Brasil, pela Lei 14.300/22. Essa lei trouxe melhorias que facilitam a abertura de usinas solares, facilitam a criação de usinas e ainda permitem o abatimento de créditos de energia, bem como outras vantagens.

Essas vantagens fomentaram novos modelos de negócios em energia, como a Geração Distribuída, que é a geração elétrica realizada junto ou dentro da área de concessão da concessionária de energia, independente da potência, tecnologia e fonte de energia. Em 2023, a capacidade instalada da geração distribuída solar cresceu quase um Gigawatt (GW) nos dois primeiros meses do ano, atingindo 18 GW (SOCIAL, 2023).

As usinas de geração distribuída são, em sua maioria, pequenas, atingindo até 5MW e com pouca infraestrutura por dispenderem de pouca manutenção, como exemplo visto na Figura 1, construídas em todo o território nacional, inclusive em locais remotos.

Figura 1 – UFV Rio das Flores - Solarian.[1]

Tais características trazem desafios às equipes de operação e manutenção, responsáveis por manter as usinas com sua geração idealmente igual à estimada em projeto. Sendo

[1] Extraído de LinkedIn: <http://tiny.cc/n4bpxz> Acessado em 09 de abril de 2024.

inviável a presença de técnicos 24 horas nas usinas, a alternativa é uso de monitoramento remoto com sistemas IoT, integrados a um sistema, para que técnicos identifiquem falhas e reportem aos responsáveis para que as medidas necessárias para solução da falha sejam iniciadas (MORAIS; PONTES, 2022).

Apesar do olhar técnico, a análise gráfica nem sempre é efetiva devido a erro humano, o que impacta negativamente a geração de energia, assim como modelos matemáticos como os apresentados por (CHINE *et al.*, 2013) e (MANSOURI *et al.*, 2018) não são passíveis de serem aplicados em pequenas granularidades de tempo, como a cada 15 minutos.

Dessa forma, técnicas de inteligência artificial como apresentadas por (COSTA *et al.*, 2020), (CHINE *et al.*, 2013), (MEKKI; MELLIT; SALHI, 2016), (PAHWA *et al.*, 2020), (SPATARU *et al.*, 2015) e (SEYEDMAHMOUDIAN *et al.*, 2016) podem ser implementadas a fim de identificar falhas e classificá-las, além de quantificá-las e detectar a ocorrência. Ainda que os resultados desses trabalhos possam ser aplicados em qualquer tamanho de planta fotovoltaica com uma acurácia superior a 90% na detecção de falhas e, alguns de forma automática, não apresentam um meio de realizá-lo em tempo real à distância e sem a necessidade de gerar uma base com dados simulados para o conjunto de treinamento do algoritmo de aprendizado de máquina, bem como a análise de outras variáveis se não tensões e correntes, o que inviabiliza o emprego em grandes conjuntos de plantas fotovoltaicas sob condições distintas.

1.2 Problema

A concepção e modelo de negócio de usinas solares fotovoltaicas para a geração distribuída origina desafios complexos e árduos para as equipes de operação e manutenção (O&M) no que tange à manutenção da geração conforme contratos firmados. Um dos principais problemas é a identificação de anomalias que impactam negativamente a geração de energia.

Apesar das usinas solares fotovoltaicas em geração distribuída serem de menor porte (até 5MW conforme a REN 1000 (ANEEL, 2021)) em comparação com usinas de geração compartilhada, a quantidade de variáveis energéticas e climáticas é a mesma, onde ambas possuem os mesmos equipamentos elétricos que permitem o funcionamento da usina, porém em quantidades diferentes, sendo distintos apenas na quantidade de dados gerados.

Devido à modelagem financeira, especialmente devido ao retorno do investimento (ou do inglês, *payback*), a grande maioria das usinas solares voltadas à geração distribuída não dispõe de sistemas avançados de análises de variáveis em tempo real e em nuvem (GREENER, 2020), incindindo na necessidade de uma equipe robusta para análises individuais recorrentes e rotineiras, sendo um trabalho maçante e repetitivo, com propensão à falhas na identificação de anomalias e problemas.

Diante desses desafios, é crucial utilizar a tecnologia em prol da redução do trabalho redundante, dispendioso e sujeito a falhas, realizando a identificação de anomalias cruzando as variáveis monitoradas na usina, retornando ao analista apenas o alarme de cada anomalia identificada, permitindo ter uma análise cognitiva de maior valor agregado e possibilitando a trabalhar com uma quantidade maior de usinas em sua carteira. Tais ações permitem com que a empresa reduza os custos com mão-de-obra de O&M, aumente a rentabilidade dos projetos e mantenha um nível ótimo de performance.

1.3 Objetivos

Neste trabalho, o objetivo principal é o desenvolvimento de um modelo de aprendizado de máquina capaz de identificar falhas em sistemas fotovoltaicos em tempo real. Para atingimento esse objetivo, se faz necessária a aquisição de dados em tempo real de uma planta solar fotovoltaica em operação para que sirva de entrada a um modelo capaz de detectar mudanças nos padrões de dados de uma planta fotovoltaica.

Será explorada uma rede neural com as variávies disponíveis na base de dados de forma a avaliar a viabilidade de utilizá-la para a predição da geração com o objetivo de comparar com a geração atual e identificar possíveis anomalias na geração.

1.3.1 Objetivo Geral

Desenvolver um modelo de aprendizado de máquina capaz de identificar falhas de sistemas fotovoltaicos.

1.3.2 Objetivos Específicos

- Aquisição em tempo real de dados de equipamentos de uma planta solar fotovoltaica.
- Desenvolver um modelo de aprendizado de máquina capaz de detectar mudanças em padrões de dados de uma planta fotovoltaica.
- Desenvolver um modelo de aprendizado de máquina capaz de predizer a geração de energia.

1.4 Organização do Texto

O texto está organizado da seguinte forma. No capítulo 2 é apresentada a fundamentação teórica do trabalho, com uma breve descrição da detecção de anomalias e das redes neurais artificiais. A fundamentação é complementada com a apresentação de trabalhos relevantes no capítulo 3 para esse desenvolvimento.

No capítulo 4 é apresentado o desenvolvimento do projeto proposto, em divisões por temas e assuntos. As discussões e avalizações sobre o projeto são descritas no capítulo 5.

Por fim, no capítulo 6 são consolidadas as conclusões e apresentada uma proposta de trabalho futuro.

2 FUNDAMENTAÇÃO TEÓRICA

Este trabalho abrange duas áreas de conhecimentos, sendo elas energia solar e aprendizado de máquina. Dessa forma, serão apresentados os conceitos essenciais teóricos utilizados em sistemas fotovoltaicos, bem como suas características estruturais e as principais variáveis utilizadas para análises de desempenho de usinas solares.

2.1 Energia Solar Fotovoltaica

A energia solar fotovoltaica é uma fonte de energia limpa e renovável, cada vez mais utilizada no Brasil (ABSOLAR, 2024).

2.1.1 Células e Painéis Solares

A energia solar fotovoltaica é obtida a partir de materiais ou dispositivos que, quando expostos à luz transformam fótons em energia elétrica na forma de tensão e corrente. A descoberta foi feita por Alexandre-Edmond Becquerel em 1839. A partir de então, aliada com os estudos do campo da eletrônica, foram desenvolvidas células solares utilizando semicondutores. Essa descoberta propiciou o desenvolvimento de placas solares, compostas por células fotovoltaicas (BUBE, 1998).

A conversão da luz em energia elétrica é realizada por materiais semicondutores, como silício e que consiste na junção de duas regiões semicondutoras com concentrações distintas de elétrons (CHAAR; ZEIN *et al.*, 2011). Dessa forma, quando um fóton de energia absorvido é maior que a faixa proibida, são formados elétrons livres que transitam da banda de valência para a banda de condução (BUBE, 1998). A Figura 2 ilustra o processo.

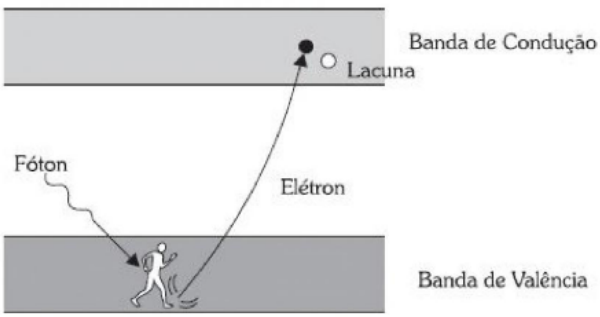

Figura 2 – Processo de transição de um elétron da banda de valência para a banda de condução. Extraído de (MARQUES *et al.*, 1997).

Os painéis solares são interligados em uma topologia tipo série e paralelo para que se obtenha um nível de tensão ótimo, formando uma *string*.

2.1.2 Instrumentação de Plantas Fotovoltaicas

O equilíbrio do sistema (em inglês, *balance of system* - BOS) compreende todos os componentes de um sistema fotovoltaico, exceto os painéis fotovoltaicos. A seguir serão tratados os principais equipamentos responsáveis pelo funcionamento e aumento de produtividade de uma planta fotovoltaica.

2.1.2.1 Inversor

Para que a energia gerada pelas células solares seja injetada na rede de energia, se faz necessária a conversão da energia gerada em corrente contínua (CC) para corrente alternada (CA) compatível com a rede elétrica, tanto em níveis de tensão quanto em níveis de frequência (MALLWITZ; ENGEL, 2010). Essa conversão é realizada por um equipamento denominado inversor e que é um componente crítico em um BOS.

Para cada *string*, ou conjunto de *strings*, há um inversor solar, equipamento responsável por transformar a energia contínua gerada pelas placas para energia alternada. Todos os inversores são conectados em paralelo e, o somatório de energia gerada convertida é então injetada na rede da concessionária de energia local (BOSMAN *et al.*, 2020).

Além da função de conversão de energia, os inversores são equipados com um controlador lógico programável (em inglês, *Programmable Logic Controller* - PLC) e dispõe de funções adicionais como controle de geração solar, medições de variáveis energéticas (tensão, corrente, frequência), bem como monitoramento e proteção de todo o sistema da usina fotovoltaica (MALLWITZ; ENGEL, 2010). Todas essas variáveis podem ser lidas por um equipamento externo e integradas em um sistema de controle.

2.1.2.2 Estação Solarimétrica

Uma estação solarimétrica, ilustrado pela Figura 3, é o conjunto de sensores e equipamentos para a medição de grandezas meteorológicas como: temperatura do ar, volume pluviométrico, direção e velocidade do vento, umidade, GHI (Irradiância Horizontal Global), DHI (Irradiância Horizontal Difusa), DNI (Irradiação Normal Direta), além de temperatura do painel solar e até sujidade do painel. A coleta desses dados é realizada por diferentes sensores, como piranômetro, barômetros, anemômetros, pluviômetros e termômetros, e integrados em um controlador (LIRA; SOARES; SANTOS, 2016).

A instalação de estações solarimétricas é obrigatória para projetos de usinas solares de grande porte que se destinam à venda de energia em leilões, exigido pela EPE (Empresa de Pesquisa Energética). Os dados devem ser coletados por um período de 12 meses, que antecedem a construção da usina para composição do estudo de potencial de geração da usina, com dados mais fidedignos para que o empreendimento possa ter uma garantia firme de entrega de energia (EPE, 2016).

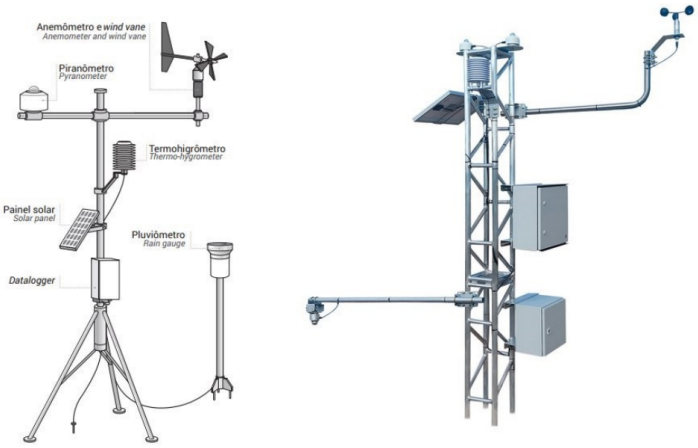

Figura 3 – Desenho de uma estação solarimétrica à esquerda, extraído de (SCHUBERT, 2019) e estação solarimétrica real, extraído de (HUKSEFLUX, 2024).

Apesar de ser um item obrigatório para usinas de grande porte, é possível observar sua presença em usinas de menor porte. Sua utilização é de grande relevância para estudos de potencial de geração e análise de desempenho da usina, uma vez que a geração tem correlação direta com variáveis meteorológicas.

2.1.2.3 Seguidor Solar

Um seguidor solar (do inglês, *tracker*) é um equipamento que realiza inclinação automática de placas solares visando a máxima incidência solar, acompanhando o caminho do sol diário e sazonal. Para que isso aconteça, são instalados sensores de iluminação nas placas solares para que os dados coletados sirvam de entrada para um cálculo em relação ao melhor ângulo para o painel durante cada momento do dia e ao longo do ano. São compostos por servomotores, controlador eletrônico que tem a função de definir o melhor ângulo de posicionamento e controlar os servomotores e pistões para estabilização do ângulo desejado (VALMONT, 2022).

Segundo estudos anteriores, sistemas com seguidores solares podem aumentar a produtividade de energia entre 20% a 50% dependendo da geografia em que a planta solar está localizada (QUESADA *et al.*, 2015).

2.1.2.4 Sistema Supervisório

Para que seja possível analisar o funcionamento de uma planta solar, se faz necessário um sistema supervisório ou Sistema de Supervisão e Aquisição de Dados (SCADA). Um sistema SCADA é um *software* que integra os equipamentos (em inglês, *hardwares*) instalados em campo, realizando a extração de dados gerados por cada equipamento, para que seja possível acompanhar, configurar, armazenar e disponibilizar informações para que a equipe de O&M de uma usina possa intervir manualmente ou automaticamente no processo, quando necessário (DANEELS; SALTER, 1999).

Esses sistemas normalmente oferecem painéis (em inglês, *dashboards*) ao responsável pela análise da usina com dados em tempo real coletados dos inversores, seguidores solares e demais equipamentos do BOS. Atualmente existem soluções em nuvem que realizam a integração dos equipamentos utilizado equipamentos com conceito de internet das coisas (em inglês, *Internet of Things* - IoT) (ENGECOMP, 2024), permitindo o acompanhamento do comportamento da usina à distância, dispensando a necessidade de um computador físico instalado na usina solar e uma pessoa dedicada presencialmente para garantir o correto funcionamento da usina.

2.1.3 Variáveis de Análise e Controle de Usinas Fotovoltaicas

A análise de desempenho e controle de usinas baseia-se em variáveis elétricas e meteorológicas, oriundas dos equipamentos instalados concentradas no sistema supervisório. Para o melhor entendimento dessa subseção e itens subsequentes, abaixo são definidas as principais variáveis elétricas e meteorológicas e suas respectivas unidades de medida:

- Tensão: É uma grandeza física escalar referente à diferença de potencial elétrico entre dois pontos, referenciada como "V" e sendo o *Volt* ("V") sua unidade de medida;
- Corrente elétrica: É uma grandeza física escalar referente ao deslocamento de cargas dentro de um condutor, referenciada como "I" e sendo o *Ampère* ("A") sua unidade de medida;
- Potência: É uma grandeza física escalar referente à quantidade de energia consumida ou cedida durante um tempo, referenciada como "P" e sendo o *Watt* ("W") sua unidade de medida;
- Frequência: É uma grandeza física escalar referente ao número de oscilações por segundo, referenciada como "f" e sendo o *Hertz* ("Hz") sua unidade de medida;
- Resistência elétrica: É uma grandeza física escalar referente à capacidade física de um material a se opor à passagem de corrente elétrica, referenciada como "R" e sendo o *Ohm* ("Ω") sua unidade de medida;
- Corrente Contínua: É a corrente sentido único mediante a presença de uma tensão, abreviada por "CC" na literatura brasileira e "V_{DC}" na literatura estrangeira;
- Corrente Alternada: É a corrente de sentido variado ao longo do tempo mediante a presença de uma tensão, abreviada por "CA" na literatura brasileira e "V_{AC}" na literatura estrangeira;
- "Radiação Solar": é um termo genérico e pode ser denominado como "irradiação solar" no contexto de energia por unidade de área como W/m^2 ou "irradiância solar" no contexto de fluxo de potência (PINHO; GALDINO, 2014).

As variáveis mais básicas e essenciais de análise de usinas fotovoltaicas são as de tensão, corrente e frequência mensuradas pelos inversores, com granularidade em *strings*. Essas variáveis permitem entender se a usina, ou *string*, está desligada pela ausência de tensão; ou com baixa ou nenhuma geração pela corrente; ou com problema de inversão de frequência pela frequência. As duas primeiras grandezas elétricas permitem o cálculo de outras variáveis baseando-se na lei de Ohm, como a potência instantânea sendo $P = V \times I$ e a geração horária da usina como descrita na Equação 2.1, considerando n a quantidade de amostras de dados coletados em um intervalo de 1 hora, permitindo a comparação com os valores estimados em projeto.

$$\text{Geração Horária} = \left(\sum_{i=1}^{n} P_i \times I_i \right) \times \frac{1}{n} \tag{2.1}$$

As variáveis elétricas são influenciadas por variáveis meteorológicas (esperadas e adversas como nuvens densas, sombreamento por árvores e construções e sujeira) como a corrente que é diretamente proporcional à irradiância solar e a tensão sendo diretamente proporcional à temperatura, ocasionando variações (PINHO; GALDINO, 2014) (SEYEDMAHMOUDIAN *et al.*, 2016). Esse comportamento é esperado e é documentado em manuais de placas solares, denominado como "curva *I-V*", como visto na Figura 4 que em condições de operação em temperaturas acima de 45°C e/ou altas irradiações, a geração do módulo fotovoltaico é degradada podendo chegar a zero quase de forma exponencial.

Inversores de frequência mais novos são dotados de um mecanismo de controle eletrônico denominado Seguidor do Ponto de Máxima Potência (SPPM, *em inglês Maximum Power Point Tracking* - MPPT). O SPPM tem a função de analisar as alterações na curva *I-V* e atua no inversor a fim de manter o gerador fotovoltaico operando com a tensão correspondente à de máxima potência, o que reduz perdas de energia nas células (PINHO; GALDINO, 2014). Os pontos de máxima potência tendem a ficar no limiar entre o início da perda de potência e a máxima potência nas condições em que o módulo está submetido, como visto na Figura 4 em que permanece em uma faixa de saída de tensão entre 17V a 18V. O SPPM pode ser feito através de dois métodos, sendo direto e indireto, onde o primeiro independe das características do gerador e utiliza dados do sensoriamento em tempo real, e o segundo realiza comparações das variáveis atuais com informações já conhecidas, como ensaios em laboratórios, armazenados em um banco de dados. Devido à sua característica de operação, é imprescindível que os controles de SPPM estejam funcionando corretamente, de forma a não ocasionar uma perda maior do que um equipamento que não dispõe dessa tecnologia.

No âmbito das variáveis meteorológicas, as variáveis coletadas restringem-se às disponibilizadas pela estação solarimétrica. De acordo com o regulamento da EPE (EPE,

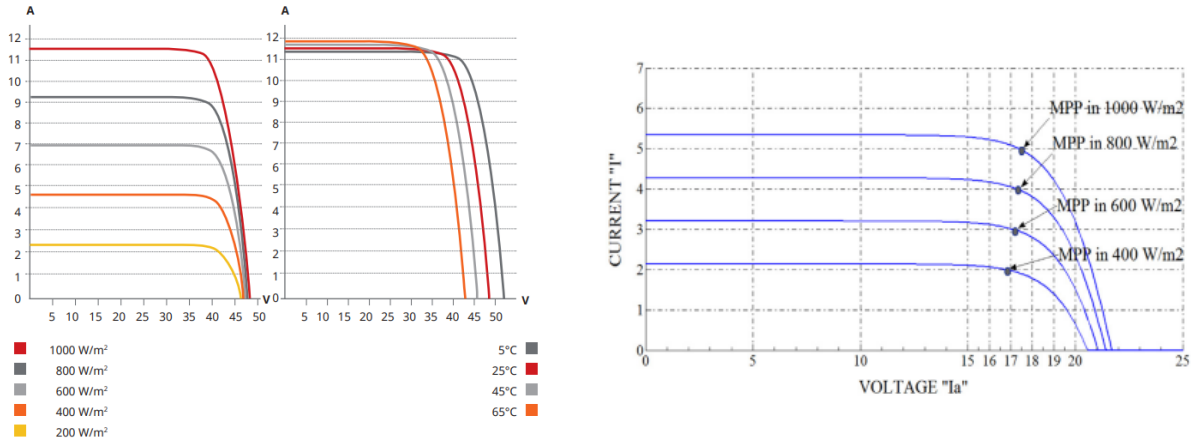

Figura 4 – À esquerda, curva *I-V* do painel solar série CS3W extraído de (CANADIAN-SOLAR, 2020) e curva *I-V* com indicação de SPPM por irradiação extraído de (SEYEDMAHMOUDIAN *et al.*, 2016).

2016), a estação solarimétrica deve possuir piranômetro para a medição da irradiação global horizontal, anemômetro para a medição da velocidade do vento, higrômetro para a medição da umidade relativa do ar e termômetro para a medição da temperatura do ar, indicados na Figura 3.

No contexto de energia, as variáveis meteorológicas mais comuns de pertencerem à análise de desempenho de uma usina fotovoltaica são: irradiação global horizontal, irradiação normal direta e irradiação difusa horizontal. A Figura 5 ilustra componentes citadas, além da componente de espelhamento que reflete parte da irradiação solar para o espaço ocasionada pela atmosfera terrestre e corpos refratores (gelo, neve, areia branca, como exemplo) e o albedo que é a componente refletida e que é absorvida por outro corpo.

A irradiação normal direta (em inglês *Direct Normal Irradiation* - DNI), marcada como “1” na Figura 5, é a irradiação que incide diretamente sobre uma superfície plana a 90° em relação ao feixe e é excluída a luz dispersa vinda do céu. É a radiação solar que passa diretamente através da atmosfera até a superfície terrestre (PARAíBA, 2023).

A irradiação difusa horizontal (em inglês *Difuse Horizontal Irradiation* - DHI), marcada como “2” na Figura 5, é a irradiação solar que atinge a superfície após sofrer espelhamento pela atmosfera terrestre por moléculas de ar, vapor d’água, aerosols e nuvens da radiação solar direta assim como refletida pelo entorno, como topografia, cobertura do solo e edificações (PARAíBA, 2023).

A irradiação global horizontal (em inglês *Global Horizontal Irradiation* - GHI) é a relação da irradiância normal direta e difusa do hemisfério incidente em uma superfície horizontal. Quando o sol está a exatos 90° do plano horizontal, é produzido um feixe circular, mas à medida que se move para baixo no céu, o feixe se converte em uma elipse – fenômeno análogo ao alongamento de sombras em finais de tarde (PARAíBA, 2023). A

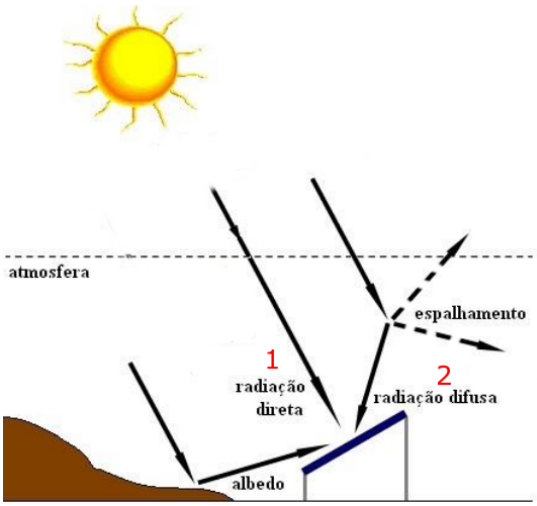

Figura 5 – Componentes da radiação solar. Adaptado de (PINHO *et al.*, 2008).

relação entre GHI, DHI e DNI é expressada na equação (2.2) abaixo, onde θz é o ângulo zenital, correspondente ao ângulo entre a posição do sol e a placa solar. As medições locais do piranômetro GHI permitem a comparação da energia solar disponível entre os locais e entre os conjuntos de dados e a validação das estimativas de satélite e modelo para o local específico.

$$\text{GHI} = \text{DNI} \times \cos(\theta_z) + \text{DHI} \tag{2.2}$$

Com a combinação das variáveis elétricas e meteorológicas é realizada a análise de desempenho e comportamento da usina e controle da usina. As análises podem ser em tempo real com foco na geração positiva da usina e através de simulações entre o estimado em projeto contra o real em um período de tempo, utilizando *softwares* de projeto como o *PVSyst*[1].

Demais equipamentos como relés de proteção, transformadores, equipamentos de proteção dentre outros, também disponibilizam extração de dados, porém não estão fortemente relacionados com o desempenho de usina e não são foco desse trabalho.

A Tabela 2.1.3 resume as variáveis supramencionadas.

2.1.4 Falhas em Sistemas Fotovoltaicos

A correta classificação de uma falha em uma usina solar, através da análise de variáveis disponíveis, é pré-requisito para a correta correção do problema. Por ser uma tecnologia consolidada e com padronização de construção, os tipos de falhas são conhecidos e registrados na literatura. No trabalho de (MADETI; SINGH, 2017) são elencados os tipos de falhas divididos em duas categorias, sendo falhas em energia contínua (CC) e falhas

[1] Software PVSyst: <https://www.pvsyst.com/> Acessado em 01 de abril de 2024.

Tabela 1 – Variáveis de análise de desempenho de usinas fotovoltaicas.

Variável	Unidade de Medida	Equipamento	Sensor
Tensão	V	Inversor de Frequência	
Corrente	A	Inversor de Frequência	
Potência	W	Inversor de Frequência	
Frequência	Hz	Inversor de Frequência	
SPPM	W	Inversor de Frequência	
Temperatura do Ar	ºC	Estação Solarimétrica	Termômetro
GHI	W/m²	Estação Solarimétrica	Piranômetro
DNI	W/m²	Estação Solarimétrica	Piranômetro
DHI	W/m	Estação Solarimétrica	Piranômetro
Velocidade do Vento	m/s	Estação Solarimétrica	Anemômetro
Umidade Relativa do Ar	%	Estação Solarimétrica	Higrômetro

em corrente alternada (CA). Em corrente alternada são falhas decorrentes das saídas dos inversores e da rede elétrica da concessionária local, com poucas variáveis de análise. Em corrente contínua são falhas decorrentes de SPPMs e placas solares em si, estendendo para falhas por descasamento, curto-circuito, circuito aberto, aterramento, diodo de passagem (em inglês, *bypass*), assimetria, arco e ponte.

Dentre as falhas, a que mais impacta a geração da usina é a falha de circuito aberto, ocasionada por problemas como má conexão e/ou conexão frouxa entre placas solares e caixas de junção, placas e/ou células quebradas, conectores danificados e/ou desgastados. Esse tipo de falha é a interrupção do fluxo de corrente, onde a depender de onde esteja tal interrupção e da topologia da usina, pode causar a interrupção de todo o sistema. A detecção é relativamente fácil, uma vez que a potência caia a zero no ponto de interrupção, sendo identificada pelo inversor de frequência, que por sua vez indica em qual das *strings* o problema acontece pela medição de tensão e corrente individualizada. Os mesmos problemas causadores de circuito aberto somados a defeitos de fabricação podem causar falhas de curto-circuito que é um ponto de baixa impedância no sistema, ocasionando a queima do equipamento. Como extensão, falhas de aterramento e arco também podem ser causadas pelos mesmos problemas causadores de circuito aberto e curto-circuito e podem ser preliminares às falhas de circuito aberto e curto-circuito

A falha por descasamento ocorre quando células ou módulos solares estão com grandes diferenças elétricas em comparação aos demais que compõem o sistema. (MADETI; SINGH, 2017) dividem a falha nas categorias permanente e temporária. A categoria permanente é quando ocorre algum dano físico no material, como quebra de vidro de proteção, descoloração e demais danos. A categoria temporária, que é a categoria a ser trabalhada nesse estudo, é causada por eventos como sombreamento por nuvem, construções próximas e árvores por exemplo, bem como materiais acumulados sobre a placa como

poeira e neve, que reduzem a irradiação incidente nas células solares.

Ainda como ofensor à geração da usina, há uma proteção presente em inversores de frequência (e demais equipamentos eletrônicos como celulares e ar condicionado) chamada *derating* (sem tradução). O *derating* é a operação abaixo da capacidade máxima a fim de prolongar a vida útil do equipamento e, no caso de inversores de frequência, ocorre em temperaturas muito elevadas, onde o PLC opera com redução controlada de potência e/ou corrente ou desliga momentaneamente para evitar danos aos circuitos eletrônicos. Dessa forma, mesmo que a irradiação esteja ótima, os módulos limpos e sem nenhum problema de conexão física, a geração é menor do que a esperada. Pela Figura 6 é possível observar o efeito de *derating* sob diferentes tensões de operação de um inversor de exemplo, sendo mais sensível em tensões mais altas, registrando perdas a partir de 25°C, e em tensões menores um pouco acima de 45°C, reduzindo a capacidade total praticamente linearmente.

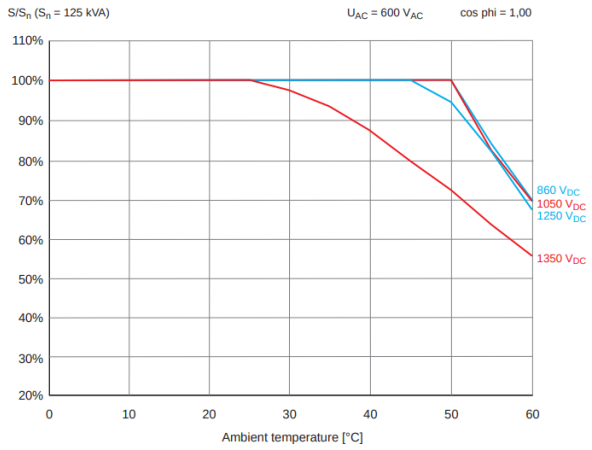

Figura 6 – Curva de *derating* de potência por temperatura ambiente. Extraído de (ELECTRONICS, 2021).

2.1.5 Geração Solar Distribuída no Brasil

A geração distribuída no Brasil é uma modalidade de geração de energia elétrica, onde a geração ocorre junto ou próxima à carga (em um telhado ou dentro da área de concessão da concessionária de energia em que a unidade está alocada, como exemplos). Essa modalidade foi instituída pelo Artigo 14 do Decreto-Lei n.º 5.163 de 2004 (BRASIL, 2004). Como limitantes, em caso de geração hidrelétrica a capacidade deve ser inferior a 30MW e para geração termelétrica, com ou sem cogeração, a eficiência energética deve ser superior a 75%.

Em 2012 foi criada a Resolução Normativa (REN) n.º 482 (BRASIL, 2012), estabelecendo condições regulatórias para a inserção da geração distribuída na matriz energética brasileira, visando a ampliação da matriz energética nacional. Essa resolução trouxe as definições de Microgeração distribuída com potência até 75 kW e Minigeração distribuída com potência superior a 75 kW e inferior a 5 MW, ambas sendo sistemas de geração de

energia renovável ou cogeração qualificada conectados à rede (BRASIL, 2012). Em 2015 a REN 482 foi aprimorada pela REN 687, trazendo o direito à utilização dos créditos por excedente de energia injetada na rede, definições de procedimentos burocráticos para a conexão à rede do Sistema Interligado Nacional (SIN), a forma de autoconsumo remoto e a geração compartilhada, na qual um grupo de unidades consumidoras é responsável por uma única unidade de geração (BRASIL, 2015).

Em janeiro de 2022 foi decretada a Lei 14300, conhecida como o marco legal da geração distribuída, com o objetivo de garantir segurança jurídica ao mercado, impedindo que mudanças abruptas na regulação. Essa Lei também proveu mudanças técnicas na geração distribuída, reduzindo a potência máxima de 5MW para 3MW, realocação de créditos excedentes para outras unidades consumidoras do mesmo titular na mesma área de concessão e unificação de titularidade (BRASIL, 2021) como exemplos de itens com maior relevância no mercado.

2.2 Detecção de Desvios em Fluxo de Dados

A detecção de falhas considerando a natureza dos dados a serem trabalhados neste projeto pode ser feita com a utilização de algoritmos detectores de anomalias e desvio. Porém será utilizado o método por desvio (em inglês, *drift*) uma vez que anomalias são numerosas e recorrentes e nem sempre indicam uma falha propriamente dita, como por exemplo a queda na geração devido a passagem de uma nuvem densa sobre a planta solar e a detecção para alarme desse tipo de erro é dispensável, além de que os inversores de frequência possuem identificadores de desvios elétricos como subtensão e sobretensão como exemplo, disponíveis no sistema SCADA. A utilização de detecção de desvios tem como objetivo identificar mudanças mais sutis e duradouras, que possam indicar um novo padrão daquele conjunto de dados que não é esperado em projeto.

Um fluxo de dados (em inglês, *data stream*) é uma sequência de dados com um vínculo de data e hora (do inglês, *timestamp*) visando a avalização de tais dados na linha do tempo. No aspecto desse projeto, variáveis monitoradas como tensão e corrente variam com o tempo, tornando o fluxo de dados com distribuição não estacionária. Dessa forma a ordem de observação dos dados deve ser seguida em ordem cronológica com os *timestamps*. Em uma aplicação passível de falhas, como uma planta solar, a mudança na distribuição dos dados precisa ser detectada (AGRAHARI; SINGH, 2022) (WANG; ABRAHAM, 2015). Diante desse desafio foram desenvolvidos algoritmos de aprendizado de máquina específicos para *data stream*.

As alterações dos dados podem ser divididas em desvio de conceito real e virtual (em inglês, *Real concept drift* e *Virtual concept drift*, respectivamente). Em termos de velocidade de alteração, os desvios podem ser abrupto (em inglês, *Sudden drift*), gradual(em inglês, *Gradual drift*), incremental (em inglês, *Incremental drift*) e recorrente (em inglês,

Recurrent drift). Desvios abruptos são caracterizados pela mudança instantânea no padrão normal dos dados e perdurando ao longo do tempo em comparação aos dados anteriores; desvios graduais são caracterizados pela mudança abrupta no padrão normal dos dados, porém oscilando entre o padrão antigo e novo padrão por um certo período de tempo; desvios incrementais são caracterizados pelo incremento ou decremento dos valores dos dados ao longo do tempo, de forma suave até que haja a estabilização do novo padrão; e a mudança recorrente é quando um comportamento antigo é observado por um período em comparação ao comportamento atual(AGRAHARI; SINGH, 2022) (LU *et al.*, 2018). A Figura 7 ilustra tais comportamentos.

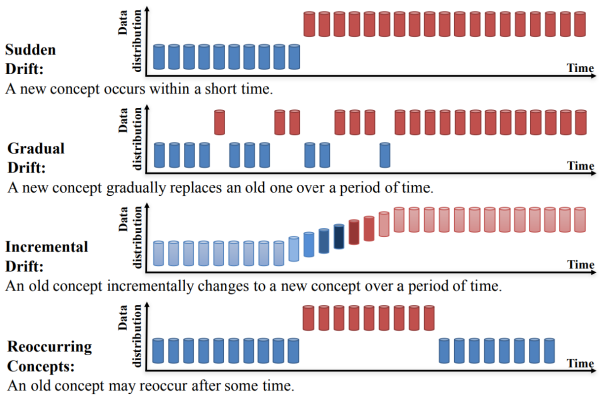

Figura 7 – Ilustração de tipos de velocidade de desvio de dados. Extraído de (LU *et al.*, 2018).

Devido à característica de alteração de padrão na linha do tempo, se faz necessário um aprendizado adaptativo. Modelos de aprendizado em tempo real desconsideram aprendizados antigos e se adaptam aos novos padrões de dados, ou em outras palavras, aos desvios desses dados ao longo do tempo. Para essa tarefa há dois métodos principais, sendo o método de análise sequencial e o método em janela, ambos com o intuito de sinalizar quando há a detecção de padrão dos dados. O método de análise sequencial compara os novos lotes de dados com o padrão atual, armazenando toda a massa de dados em memória para que seja possível captar mudanças graduais e incrementais. Já o método janela define datas de início e fim de dois lotes de dados para que sejam comparados, armazenados em memória, onde o primeiro é utilizado como base e o segundo como comparador (AGRAHARI; SINGH, 2022).

Com a identificação automática e não supervisionada de desvios de padrão de dados é possível a combinação com algoritmos mais avançados de aprendizado de máquina para que possam classificá-los quanto ao tipo, usando como exemplo falhas em sistemas fotovoltaicos.

2.2.1 Método Page-Hinkley

O método Page-Hinkley para detecção de mudança de conceito (do inglês, *concept drift*) baseia-se no cálculo dos valores observados e suas respectivas médias até o momento atual. Diferente de outros métodos que podem sinalizar zonas de alerta, o Page-Hinkley apenas identifica a ocorrência de mudanças de maneira direta. Esse detector utiliza o gráfico de controle de soma cumulativa (do inglês, *Cumulative Sum Control Chart* - CUSUM) para detectar alterações. Além disso, a implementação suporta o teste bilateral de Page-Hinkley, permitindo a detecção de variações tanto crescentes quanto decrescentes na média dos valores de entrada (JRAD *et al.*, 2017)

A aplicação do teste de Page-Hinkley pressupõe o conhecimento da média do conjunto de dados antes da ocorrência de uma mudança significativa, ou então, que esta média é estimada de forma recursiva a partir dos primeiros dados disponíveis. Define-se, então, o valor absoluto mínimo δ da amplitude da mudança que se deseja detectar, e o procedimento consiste na realização de dois testes simultâneos para monitorar possíveis alterações (JRAD *et al.*, 2017)

Seja $X_1, X_2, \ldots, X_n$ uma sequência de observações de um processo ao longo do tempo e a definição do valor inicial das variáveis $S_0 = 0$ e $n_0 = 0$.

Define-se a média cumulativa até o instante n como:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

A soma de custo S_n é dada por:

$$S_n = \sum_{i=1}^{n} (X_i - \bar{X}_n)$$

O objetivo é determinar se houve uma mudança significativa no processo. Para isso, utiliza-se um limite de controle definido como h. Quando S_n ultrapassa h, considera-se que ocorreu uma mudança:

$$S_n > h \implies \text{Mudança Detectada}$$

A soma de custo pode ser atualizada de forma recursiva, onde para cada nova observação de X_n, há:

$$S_n = S_{n-1} + (X_n - \bar{X}_n)$$

O limite de controle h pode ser definido com base em um desvio padrão esperado σ do processo:

$$h = k \cdot \sigma$$

2.3 Classificadores

Classificadores são uma parte do aprendizado de máquina, e são funções que separam dados com o objetivo rotular saídas destas funções dentro de um contexto, sendo análogos à regressão quando as variáveis de análise são discretas. Tais funções possuem diferentes abordagens e complexidades de implementação, tanto quanto em custo quanto eficácia. O treinamento de classificadores consiste na inserção de dados rotulados (modo supervisionado) ou não (modo não supervisionado) para observação de padrões rotulações. Uma vez treinado, o classificador pode determinar qual rótulo deve ser atribuído àquela informação ou dado. Nessa mesma linha de raciocínio, uma vez que o modelo contenha uma acurácia desejável, existe a capacidade de rotular dados jamais vistos (PEREIRA; MITCHELL; BOTVINICK, 2009).

Dentre todos os classificadores existentes, pode-se elencar como exemplo a regressão logística para classificação binária, Ávores de Decisão para multiclasses, K-Vizinhos Mais Próximos (em inglês, *K-Nearest Neighbours* - KNN) utilizando a distância entre pontos como método de classificação e Redes Neurais Artificiais (em inglês, *Artificial Neural Networks* - ANN) (MAHESH, 2020). Todos os classificadores são baseados em modelos matemáticos com larga utilização de álgebra linear.

2.3.1 Redes Neurais Artificiais

Uma rede neural é uma forma computacional com intuito de simular o aprendizado humano, utilizando como exemplo um neurônio biológico, para gerar uma resposta à uma entrada ou um conjunto de entradas. O aprendizado ocorre com a alteração de pesos sinápticos de forma a atingir ao máximo a certeza da resposta de um problema (HAYKIN, 2001).

Um neurônio de uma rede neural é responsável por processar informação. Ele é formado por três componentes sendo os pesos sinápticos, junção aditiva e função de ativação, ilustrado pela Figura 8. Os pesos sinápticos definem o peso no cálculo de cada sinal de entrada, podendo ser negativo ou positivo; a junção aditiva realiza a soma das entradas ponderando pelos seus respectivos pesos e, a função de ativação calcula um valor finito e em intervalos normalizados que define se o neurônio deve ser ativado ou não com base em sua entrada ponderada. Além dessas funções, neurônios podem contar um viés (do inglês, *bias*) com o efeito de aumentar ou diminuir a junção aditiva, sendo um número real e constante, permitindo o aprendizado mesmo quando todas as entradas são zero.

Por ser um classificador, uma rede neural artificial pode ser treinada com método supervisionado, não supervisionado e por reforço (do inglês, *reinforced learning*), sendo este último o aprendizado por objetivo ou maximização do acerto através de várias rodadas de aprendizado, utilizando as saídas como novas entradas com ganho (recompensa) de

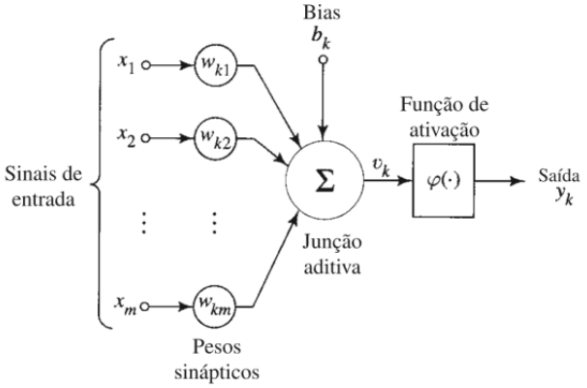

Figura 8 – Ilustração de um neurônio computacional. Extraído de (HAYKIN, 2001).

modo a alterar os pesos de cada entrada (MAHESH, 2020). Dentre essas características de aprendizado há diferentes tipos de redes neurais artificiais para diferentes aplicações, seja para textos, imagens, sons, vídeos. São redes neurais artificiais: memória de curto longo prazo (do inglês, *Long Short-Term Memory* - LSTM), redes neurais convolucionais (em inglês, *Convolutional Neural Networks* - CNN), Perceptron de multicamadas (em inglês, *Multilayer Perceptron* -MLP) (HAYKIN, 2001). Estudos correlatos à este utilizam em sua maioria redes neurais convolucionais como forma de classificação de erros em usinas solares fotovoltaicas.

2.3.2 LSTM

A rede de Memória de Curto Longo Prazo (do inglês, *Long Short-Term Memory* - LSTM) foi proposta por Sepp Hochreiter e Jürgen Schmidhuber em 1997, com o objetivo de resolver o problema do desaparecimento dos gradientes (informação usada para atualizar os parâmetros da rede) que ocorre em redes neurais tradicionais quando submetido ao aprendizado com dependências de longo prazo, como tarefas de tradução de linguagem, reconhecimento de fala e previsão de séries temporais. A LSTM possui controles não lineares e dependentes de dados na célula da RNN, que podem ser treinados para garantir que o gradiente da função objetivo em relação ao sinal de estado não desapareça ou que sejam esquecidas de maneira mais eficaz. Os dados a serem utilizados em um treinamento de uma LSTM devem ser padronizados, de modo que todos os elementos da entrada para a rede tenham média 0 e desvio padrão 1 (SHERSTINSKY, 2020).

Uma LSTM é composta por células de memória e três portas principais responsáveis pelo controle do fluxo de informações, conforme ilustrado na Figura 9. A porta de entrada (do inglês, *Input Gate*) tem a função de decidir quais valores de entrada devem ser atualizados na célula de memória, aplicando uma função de ativação (geralmente a sigmoide) de forma a determinar quais informações são relevantes; A porta de esquecimento (do inglês, *Forget Gate*) tem a função de determinar quais informações da célula de memória devem ser descartadas, também aplicando uma função de ativação de forma a determinar

quais partes da memória anterior devem ser mantidas ou removidas; por fim, a porta de saída (do inglês, *Output Gate*) tem a função de decidir quais informações da célula de memória devem ser enviadas para o próximo estado oculto e para a saída da LSTM, também aplicando uma função de ativação de forma a determinar quais partes da célula de memória devem influenciar a saída (SHERSTINSKY, 2020).

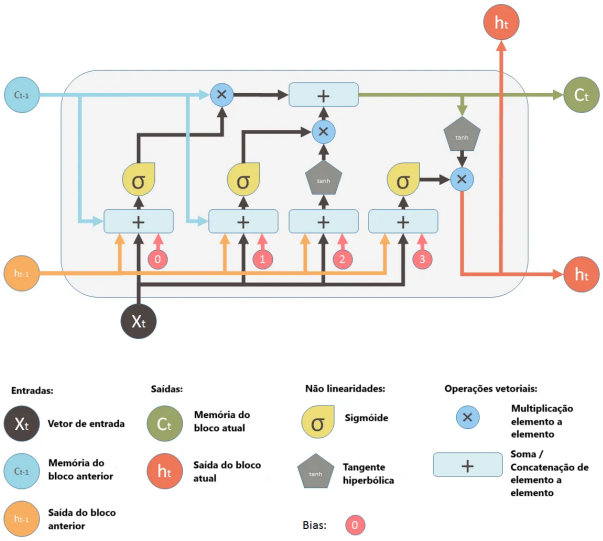

Figura 9 – Ilustração de um bloco LSTM. Adaptado de (YAN, 2016).

3 REVISÃO BIBLIOGRÁFICA

Este capítulo se dedica a explorar a literatura sobre monitoramento de sistemas fotovoltaicos e métodos de detecção de falhas, com complemento da respectiva classificação, com foco em técnicas de inteligência artificial. A análise comparativa com pesquisas existentes busca evidenciar as contribuições originais desta investigação.

Para esta revisão, foram consultadas bases de dados acadêmicas e científicas, como IEEE Xplore, ScienceDirect e Google Scholar, utilizando termos de busca como "*solar PV faults detection*", "*solar PV failure analysis*" e "*solar PV maintenance*". Foram selecionados estudos publicados entre os anos de 2013 e 2024, com foco em métodos de identificação de *drifts* e previsão de geração de energia, salvo conceitos básicos de elétrica e computação.

Um desenvolvimento relevante no contexto desse trabalho foi publicado por Costa (COSTA *et al.*, 2020), no qual o pesquisador aplicou quatro classificadores para detecção de falhas de sistemas fotovoltaicos, sendo eles k-Vizinhos mais próximos, Árvore de Decisão, Máquina de Vetores Suporte e Rede Neural Artificial. Para o treinamento do classificador, foi desenvolvido um sistema fotovoltaico teórico e aplicados ruídos equivalentes à falhas de Curto-circuito, degradação, circuito aberto e sombreamento para a geração da base de treino. Após o teste dos classificadores com dados coletados de uma usina fotovoltaica real, a maior acurácia obtida foi através de Rede Neural Artificial com 92,08% das novas amostras de teste e k-Vizinhos mais próximos obtendo a menor acurácia com 76,48% das novas amostras de teste. O sistema proposto conta com análise em tempo real utilizando um programa de computador com uma IHM (interface homem-máquina), sem necessidade de desconexão da planta da rede da concessionária ou de parte da carga para realização da classificação de falhas, com dados coletados via sistema de monitoramento desenvolvido pelo autor.

No trabalho de (CHINE *et al.*, 2013), é proposto a identificação em duas etapas utilizando uma rede neural artificial tipo Perceptron com análise de curva VxI do sistema, alcançando uma acurácia de 90,3%. A primeira etapa identifica anomalias na produção de energia comparando a potência real da planta com a potência simulada. Já a segunda etapa classifica as falhas em quatro categorias sendo circuito aberto, degradação, curto-circuito e sombreamento. Devido à não utilização de monitoramento embarcado, a detecção de falhas é feita com um computador conectado próximo à planta, inviabilizando a análise remota da planta. Nessa mesma linha, (CHEN *et al.*, 2019) propõem uma alternativa utilizando Rede neural convolucional e ResNet obtendo acurácias de 92,4% e 99,9% respectivamente.

Em um outro trabalho com a proposta de criar um modelo para detectar falhas em painéis solares individuais causadas por sombreamento parcial foi feito por (MEKKI;

MELLIT; SALHI, 2016). O modelo utiliza redes neurais para analisar dados de irradiância, temperatura, tensão e corrente do painel em tempo real. A partir da estimativa da tensão e corrente utilizando as variáveis de irradiância e temperatura do módulo e realizando a comparação da tensão e corrente reais, o sistema identifica automaticamente falhas por sombras no painel. Embora a abordagem seja específica para falhas por sombreamento e necessite de treinamento da rede neural com dados históricos, apresenta vantagens como a detecção automática em tempo real e a não necessidade de desconexão do sistema. A acurácia do modelo depende da qualidade dos dados, mas a proposta abre caminho para soluções mais eficientes de monitoramento e manutenção de painéis solares.

De forma a comparar e avaliar o desempenho de diferentes técnicas de aprendizado de máquina na classificação automática de falhas em painéis fotovoltaicos levou ao desenvolvimento do trabalho de (PAHWA *et al.*, 2020). As falhas elencadas para teste foram sombreamento parcial, diodo de *bypass*, ponte, temperatura, sombreamento completo e curto circuito. O estudo conta com bases estáticas e não há forma de identificação de falhas em tempo real. Foram avaliados os algoritmos Árvore de Decisão, XGBoost, Random Forest e Redes Neurais. A maior acurácia foi obtida com o algorítimo de Rede Neural acima de 99,5%, contra 95,91% de acurácia utilizando árvore de decisão, sendo a mais imprecisa dentre os modelos avaliados.

Análises utilizando a curva *I-V* foram realizadas por (SPATARU *et al.*, 2015) utilizando uma planta solar pequena por eles construída para a realização dos estudos. O diagnóstico é feito com três classificadores do tipo *fuzzy* identificando falhas de sombreamento parcial, aumento de perda por resistência em série e degradação induzida por potencial. As falhas foram induzidas manualmente para geração dos dados a serem comparados com a base de treinamento, sendo coletados utilizando um computador conectado próximo à planta. A acurácia dos classificadores *fuzzy* variou para cada tipo de falha, sendo 90% para sombreamento parcial, 97% para resistência em série e 80% para degradação induzida por potencial. Apesar das acurácias serem menores em comparação a algoritmos mais sofisticados de aprendizado de máquina, a implementação de classificadores *fuzzy* é relativamente fácil.

Dentre as técnicas exploradas, um estudo intenso sobre técnicas de análise de variáveis SPPM para identificação de sombreamento parcial utilizando métodos de inteligência artificial foi realizado por (SEYEDMAHMOUDIAN *et al.*, 2016) . No estudo foram comparados métodos como rede neural artificial, controle lógico por *fuzzy*, otimizações por enxame de partículas (do inglês *Particle swarm optimization* - PSO) e colônia de formigas (do inglês *Ant colony optimization* - ACO), algorítimo genético, evolução diferencial e outros métodos emergentes. O estudo comparou estes métodos em dez qualificadores, dentre eles velocidade de convergência, eficiência e ajuste periódico. Por ser um estudo com viés de estado da arte, não foram elencadas as acurácias médias de cada método,

porém pela análise dos qualificadores, os algoritmos com a combinação mais interessante são PSO e ACO.

3.1 Contribuições

Ainda que alguns dos trabalhos acima apresentados contenham uma análise automática de falhas, com algoritmos de aprendizado de máquina com acurácias superiores a 90%, ainda dependem de um computador conectado próximo à planta, impedindo a análise em tempo real à distância. Ainda como limitação, as variáveis de análise mais comuns, como a curva *I-V*, dependem da desconexão da usina para a classificação de erros, bem como valer-se de dados simulados para a verificação da acurácia dos algoritmos propostos. Em aplicações práticas e em escala, considerando empresas com grande número de usinas solares construídas em todo o território brasileiro, já em operação e sem a possibilidade de simulação de falhas e/ou desconexão do sistema para a geração de dados de treino e testes, o presente trabalho propõe um método de identificação e classificação de falhas em sistemas fotovoltaicos contendo:

- Identificação de possíveis falhas utilizando análise de dados em *streaming*.
- Uso de dados reais tanto para treinamento quanto para testes, valendo-se de variáveis de erros sinalizados por inversores como forma de classificação e composição da base de treinamento.

4 METODOLOGIA

4.1 Considerações Iniciais

Neste capítulo, são descritas as etapas para o desenvolvimento do modelo capaz identificar quedas na geração de uma usina solar fotovoltaica que não sejam originadas por problemas técnicos já identificados por inversores de frequência, assim como reduções por condições climáticas.

Os estágios foram descritos nos tópicos a seguir, onde cada fase do projeto é dependente da etapa anterior, de forma que o trabalho segue uma metodologia linear.

Para atingimento do objetivo se faz necessária a extração das informações e realização de tratamentos nos dados, como a verificação de dados faltantes, alinhamento de marca temporal (do inglês, *timestamp*) e possíveis inconsistências. Com o conjunto de dados estatisticamente satisfatório, os modelos preditivos serão treinados, aplicados e validados para a obtenção do modelo mais adequado para esta problemática.

Por fim, a comparação de desempenho entre os modelos candidatos deve ser feita utilizando múltiplas métricas para evitar qualquer tipo de viés. Além disso, o modelo mais adequado deve se encaixar à rotina do negócio para que seja, de fato, útil no contexto da equipe comercial.

4.1.1 Máquina Utilizada

- Processador: AMD Rayzen 7 5800X
- Memória RAM: 32Gb, DDR4, 3200MHz
- Armazenamento: 1Tb, NVMe, Leitura 7000MB/s e Gravação 6000MB/s
- Placa de vídeo: NVIDIA Geforce RTX 3050 Ventus 2X, 8Gb GDDR6
- Sistema Operacional: Ubuntu 22.04.4 LTS, kernel 6.5.0-35-generic

4.2 Coleta de Dados

Os dados utilizados foram coletados da mesma base de dados da empresa Engecomp, pertencentes a um de seus clientes detentor de usinas solares fotovoltaicas. Em consonância com as regras da Lei Geral de Proteção de Dados (LGPD), foram escritos e assinados termos de disponibilização e confidencialidade dos dados para fins acadêmicos entre a Engecomp e seu cliente e Engecomp e autor do projeto. De forma a manter a confidencialidade dos dados, as menções as usinas serão por números inteiros.

As informações foram extraídas utilizando linguagem Python para melhor controle de carga, uma vez que foram extraídos direto da base de dados de produção da empresa e, devido a essa condição, os códigos dessa atividade não serão disponibilizados por questões de segurança.

Para a extração foram consideradas as usinas monitoradas do cliente da Engecomp que contém uma estação solarimétrica instalada e com dados coletados. Para cada usina foram extraídos dados de cada inversor de frequência e sua respectiva estação solarimétrica. Foi extraído todo o histórico de dados correspondentes às condições descritas.

Como o objetivo do projeto é lidar com dados em *stream*, foram necessários dois passos para a coleta dos dados. O primeiro foi a extração dos dados históricos para a carga em uma base de dados dedicada a esse projeto, tanto para ser usada para a análise exploratória e elencar eventuais tratamentos a serem feitos, quanto para o treinamento dos modelos propostos. A segunda foi a construção de um método de extração, transformação e carregamento (do inglês, *Extract, Transform and Load* - ETL) para os dados em tempo real.

A Tabela 2 exibe os atributos disponíveis no conjunto de dados selecionados, bem como suas respectivas frequências de coleta. Dados definidos com n são repetidos pela quantidade existente em cada inversor de cada usina.

Tabela 2 – Variáveis do conjunto de dados.

Atributo	**Descrição**	**Unidade**	**Frequência**
ts	*Timestamp* do dado	datetime	-
usine	Nome da usina	string	-
sensor_type	Tipo do sensor	string	-
sensor_id	ID do sensor	string	-
uom	Tipo de medição	string	-
global_irradiation_(n)	GHI e DNI	float	5 minutos
temperatura_modulo_(n)	Temperatura de referência dos módulos	float	10 minutos
pm_fault_code	Código de falha do inversor	int	Acontecimento
pm_op_mode	Modo de operação do inversor	int	Acontecimento
pm_dc_w	Potência de saída do inversor	float	15 minutos
pm_i(n)	Corrente da string	float	15 minutos
pm_mppt_(n)_i	Corrente MPPT da string	float	15 minutos
pm_mppt_(n)_v	Tensão MPPT da string	float	15 mintuos

4.3 Criação de Base de Dados Dedicada e Carga de Dados

A fim de não utilizar a base de dados de produção da empresa Engecomp, os dados foram inseridos em uma base dedicada ao projeto, instalada localmente na máquina utilizada no desenvolvimento deste projeto. Foi escolhido o sistema gerenciador de banco

de dados (SGBD) PostgreSQL com a extensão *Timescale DB*[1]. A escolha desse banco de dados em específico foi por ser modelado para trabalhar com séries de dados temporais (do inglês, *timeseries*) e com linguagem SQL, além de possuir ferramentas para agregações temporais realizadas pelo próprio gerenciador.

Para a carga dos dados, foi criada uma única tabela com as características conforme elencadas na Tabela 3. A chave primária definida é composta pelas colunas *ts*, *sensor_id* e *uom*, garantindo valores únicos para cada variável de cada sensor para cada coleta em um único *timestamp*. Para essa tabela, foi criada uma hiper tabela (do inglês, *hypertable*) e com índices (do inglês, *index*) nas colunas *sensor_id*, *uom* e *ts* para a redução do tempo de consulta.

Tabela 3 – Colunas do banco de dados de estudo.

Coluna	Descrição	Tipo
ts	*Timestamp* do dado	timestampz
client	Nome do cliente Engecomp	text
usine	Nome da usina	text
sensor_type	Tipo do sensor	text
sensor_id	ID do sensor	text
uom	Variável medida	double precision

A extração resultou em uma tabela única com dimensão de 111.530.467 linhas e 6 colunas, ocupando 21Gb de armazenamento.

4.4 Análise Exploratória

A base de dados analisada inclui trinta e sete usinas solares fotovoltaicas. Cada usina possui entre um e n inversores, cada um com sete variáveis correspondentes às que possuem frequência de coleta definida e descritas na Tabela 2. Dessa forma, é essencial identificar quais usinas têm maior probabilidade de fornecer dados adequados para esta análise. Para abordar este problema, foram formuladas as seguintes perguntas:

- Qual usina possui o inversor e a estação solarimétrica com o dado mais antigo?
- Qual usina possui a menor diferença entre o dado mais antigo do inversor e a sua respectiva estação solarimétrica?
- Qual usina possui a maior quantidade de inversores?
- Qual usina possui o inversor com menor quantidade de ausência de dados?

[1] TimescaleDB: <https://www.timescale.com/> Acessado em 03 de junho de 2024.

A primeira pergunta visa identificar a usina com o registro de dados mais antigo, aumentando as chances de detectar reduções na geração devido ao desgaste das placas ou outros problemas não capturados pelos dispositivos de monitoramento.

A segunda pergunta investiga se as variáveis dos inversores e as estações solarimétricas compartilham o mesmo período de coleta de dados. Discrepâncias podem levar a identificação de "*drifts*" que na realidade representam eventos temporários e não problemas substanciais.

A terceira pergunta serve como critério adicional para avaliar a usina com mais inversores, proporcionando uma margem de segurança em caso de ausência de dados em determinados equipamentos para o desenvolvimento deste projeto.

A quarta pergunta identifica o equipamento com maior potencial de uso para estudo, dispensando análise visual e quantificando a qualidade máxima dos dados a serem trabalhados, além de evidenciar eventual necessidade de tratamentos durante o processo de uso desses dados.

Um algoritmo desenvolvido em Python foi empregado para analisar e responder às perguntas relacionadas às usinas solares fotovoltaicas. Por padrão, este algoritmo classifica e retorna as três usinas que mais atendem a cada questão, embora esse número possa ser ajustado conforme necessário para uma análise mais ampla via parâmetro de entrada ou específica. As usinas são classificadas do maior ao menor grau de atendimento às questões propostas, oferecendo flexibilidade para ajustar os critérios de classificação conforme os objetivos da análise. Um sistema de pesos é aplicado para refinar ainda mais a classificação, descrito pela Equação 4.1. O peso de cada usina é determinado pela sua posição na lista de retorno de cada questão, com pesos decrescentes à medida que a posição avança. Esse sistema de ponderação é configurável, permitindo ajustes para atender a diferentes critérios analíticos.

$$P_i = \begin{cases} 1 & \text{se } i = 0, \\ 0.5^i & \text{se } i > 0. \end{cases} \tag{4.1}$$

Onde:

- i é o índice da usina no vetor, começando de 0 até $n - 1$.
- P_i é o peso atribuído à usina no índice i.

A usina mais indicada a ser explorada é a usina 2 com 2,25 pontos, contendo 13 inversores. Os registros de dados desta usina datam de 18 de agosto de 2022, marcando o início do período de coleta completa. Portanto, as análises subsequentes serão baseadas exclusivamente nesses dados.

4.5 Filtragem dos dados

De forma a reduzir a quantidade de variáveis a serem analisadas e, consequentemente, a complexidade do algoritmo a ser desenvolvido, foram consideradas as variáveis utilizadas no cálculo de performance de usinas solares fotovoltaicos, conforme a norma número 61724 da Comissão Eletrotécnica Internacional (do inglês, *International Electrotechnical Commission* - IEC) (SENSORS, 2024). Um dos métodos é a performance corrigida pela temperatura $\mathrm{PR}'_{\mathrm{STC}}$, que leva em consideração a temperatura para normalizar a variação sazonal, como as estações do ano, definida pela Equação 4.2. O método indica a utilização do DNI como variável de irradiação para usinas que contém *trackers* uma vez que esses equipamentos maximizam a incidência de irradiação nas placas solares, e que é o caso da usina a ser estudada. A utilização dessas variáveis permite, além da detecção de *drifts*, um eventual cálculo de performance que também pode ser uma variável a ser monitorada. Com isso, as variáveis a serem trabalhadas no decorrer do projeto serão: DNI (global_-irradiation_1), temperatura de referência dos módulos (temperatura_modulo_1) e geração instantânea somada de todos os inversores da usina (pm_dc_w).

$$PR'_{\mathrm{STC}} = \frac{\frac{E_{\mathrm{out}}}{C_k P_0}}{\frac{H_i}{G_{\mathrm{i,ref}}}} \tag{4.2}$$

com

$$C_k = 1 + \gamma \times (T_{\mathrm{mod},k} - T_{\mathrm{referência}})$$

- E_{out}: [kWh] Energia gerada pelo sistema fotovoltaico (AC), ou seja, após o inversor (pm_dc_w);
- P_0: [kW] Valor fixo de projeto da potência total de saída em DC de todos os módulos fotovoltaicos instalados nas condições padrão de teste (1 000 W/m² de irradiância e 25 °C de temperatura da célula fotovoltaica);
- H_i: [kWh/m²] Irradiância no plano de inclinação (global_irradiation_1);
- γ: Coeficiente de temperatura do módulo fotovoltaico;
- $T_{\mathrm{mod},k}$: Temperatura do módulo fotovoltaico em um determinado instante k (temperatura_modulo_1);
- $T_{\mathrm{referência}}$: Temperatura de referência.

Como uma usina solar depende da irradiação solar para gerar energia e o equipamento de telemetria coleta dados de forma ininterrupta, é imprescindível determinar os horários em que a usina está gerando energia. Essa análise permitirá a filtragem dos dados, auxiliando na velocidade do processamento e exclusão de períodos indesejados para análise.

Dessa forma, a Figura 10 permite observar que os horários de geração plena da usina concentram-se a partir das 8 horas da manhã até 16 horas da tarde. Assim, as análises subsequentes e o desenvolvimento do projeto utilizarão essa faixa horária.

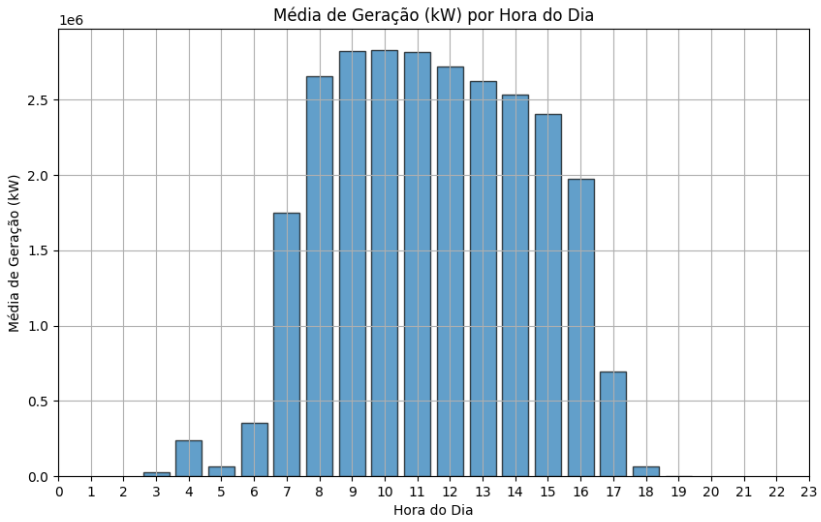

Figura 10 – Média de geração em kW explicada pela variável *pm_dc_w* por horas do dia.

4.6 Disponibilidade e características dos dados

Com a usina determinada, variáveis de análise definidas e os intervalos horários definidos, a base histórica de análise foi construída, com a quantidade de amostras para cada variável a ser analisada exposta na Tabela 4.

Tabela 4 – Quantidade de amostras por variável de interesse.

Variável	**Amostras**
global_irradiation_1	44.826
temperatura_modulo_1	25.313
pm_dc_w	15.034

Para determinar quais inversores apresentam a melhor qualidade de dados, considerando a consistência nos períodos de análise, realizou-se uma análise visual e quantitativa. O foco é determinar o percentual de dados coletados conforme a frequência de cada variável definida na Tabela 2. Períodos de ausência de dados serão considerados com valores nulos e ofensores ao percentual de coleta.

A variável *pm_dc_w*, que representa o produto da tensão pela corrente, foi utilizada devido à sua relevância e ao fato de que os dados estão disponíveis consistentemente no mesmo período amostral. Esta análise é ilustrada na Figura 11. É possível observar pelos

números em cada célula da matriz que os inversores apresentaram entre 90% e 97% de dados não nulos por mês. O não atingimento de percentuais maiores ocorre por diversas razões como manutenção preventiva e queima de equipamento de telemetria como exemplos. Nota-se considerável falta de dados em outubro de 2022 e janeiro de 2023, momentos em que necessitam de atenção no treinamento dos algoritmos de aprendizado de máquina. Devido a tais ausências, a base de dados considerará dados a partir de 01 de fevereiro de 2023.

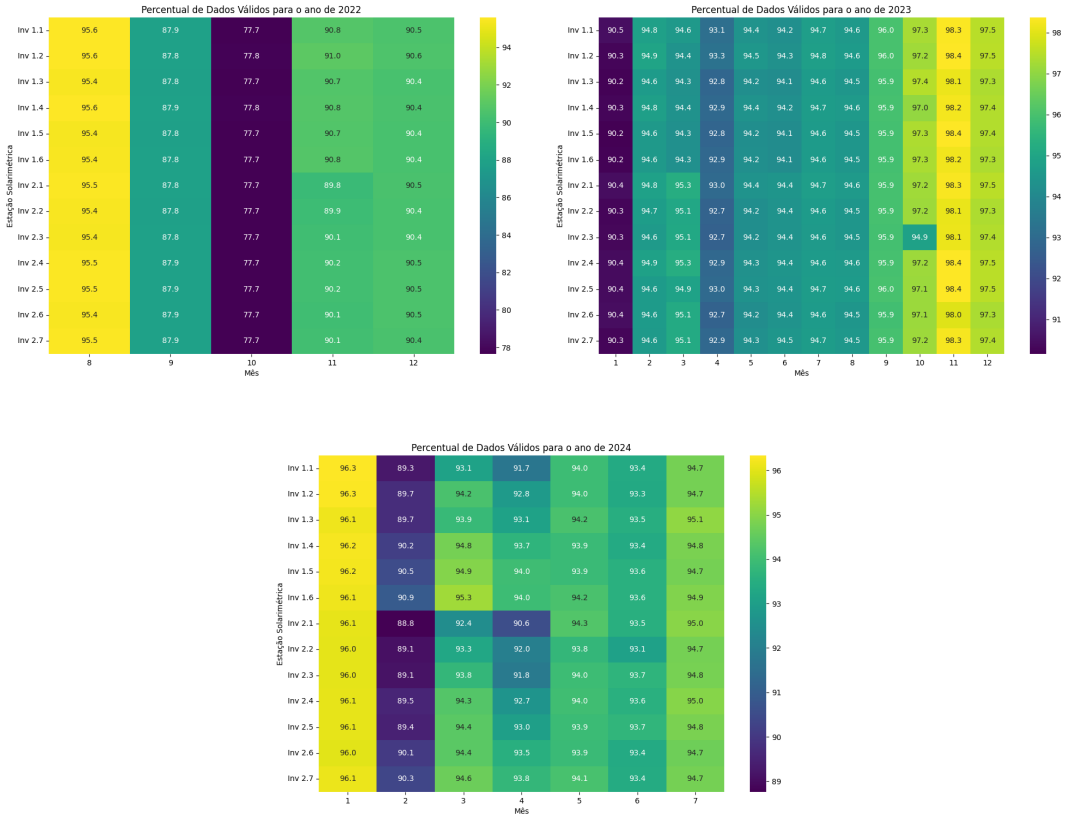

Figura 11 – Gráficos de calor de inversores mostrando o percentual de dados válidos ao longo dos anos de 2022 a 2024 pela variável *pm_dc_w*.

A mesma análise foi realizada para a estação solarimétrica da usina, utilizando as variáveis *global_irradiation_1*, sendo o DNI, *global_irradiation_3* sendo o GHI e *temperatura_modulo_1* sendo a temperatura de referência dos módulos fotovoltaicos para avaliação da qualidade dos dados. Por serem sensores físicos distintos é importante analisar se ambos possuem as mesmas características quanto à disponibilidade de dados. É possível observar através da Figura 12 que os dados possuem mais ausências no início das operações em 2022 e com disponibilidade de dados não nulos acima de 90% para ambas as variáveis.

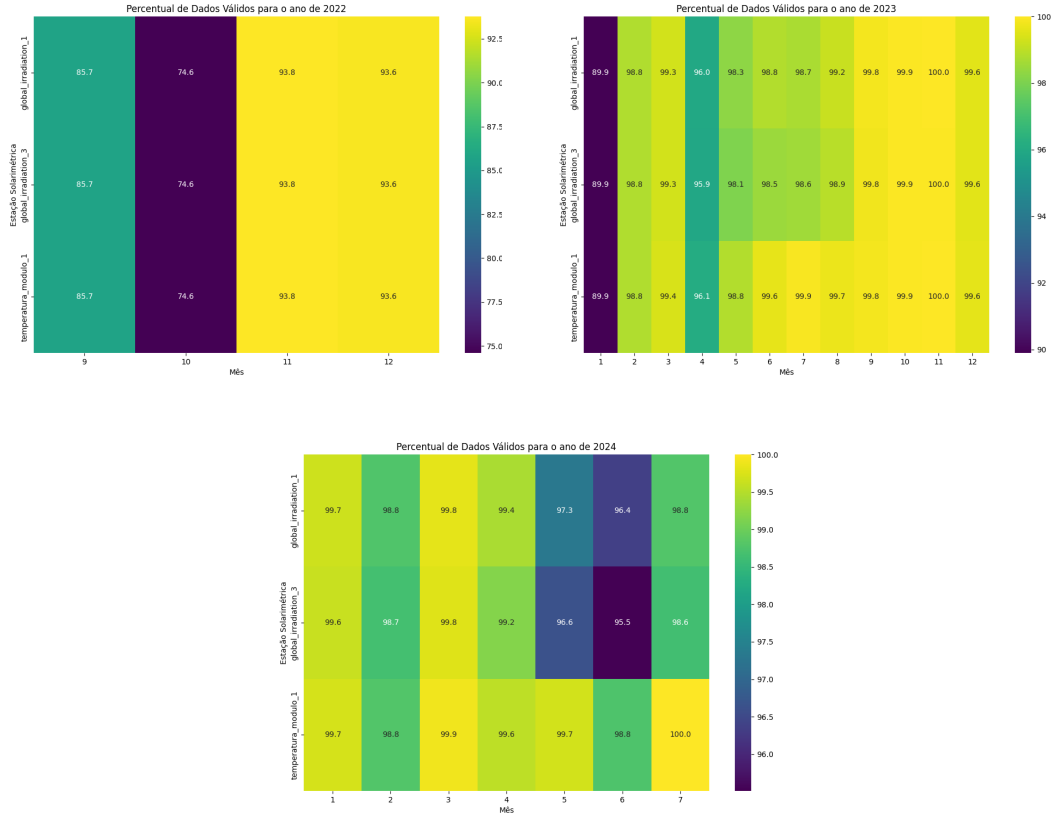

Figura 12 – Gráficos de calor de estação solarimétrica mostrando o percentual de dados válidos ao longo dos anos de 2022 a 2024 pelas variáveis *global_irradiation_1*, *global_irradiation_3* e *temperatura_modulo_1*.

De forma a compreender a distribuição dos dados, foram elaborados os histogramas presentes na Figura 13. A variável *global_irradiation_1* possui valores de 0 a cerca de 1.200 W/m², com pico de frequência próximo a 800 W/m², com histograma assimétrico com distribuição negativa e curtose leptocúrtica; A variável *temperatura_modulo_1* possui valores de 20 °C a 70 °C, com pico de frequência em 50 °C, com histograma simétrico com ligeira tendência para distribuição negativa e curtose mesocúrtica; A variável *pm_dc_w* possui valores de 0 a cerca de 250.000 W, com uma moda em torno do limite superior, com histograma assimétrico com distribuição positiva e curtose platicúrtica.

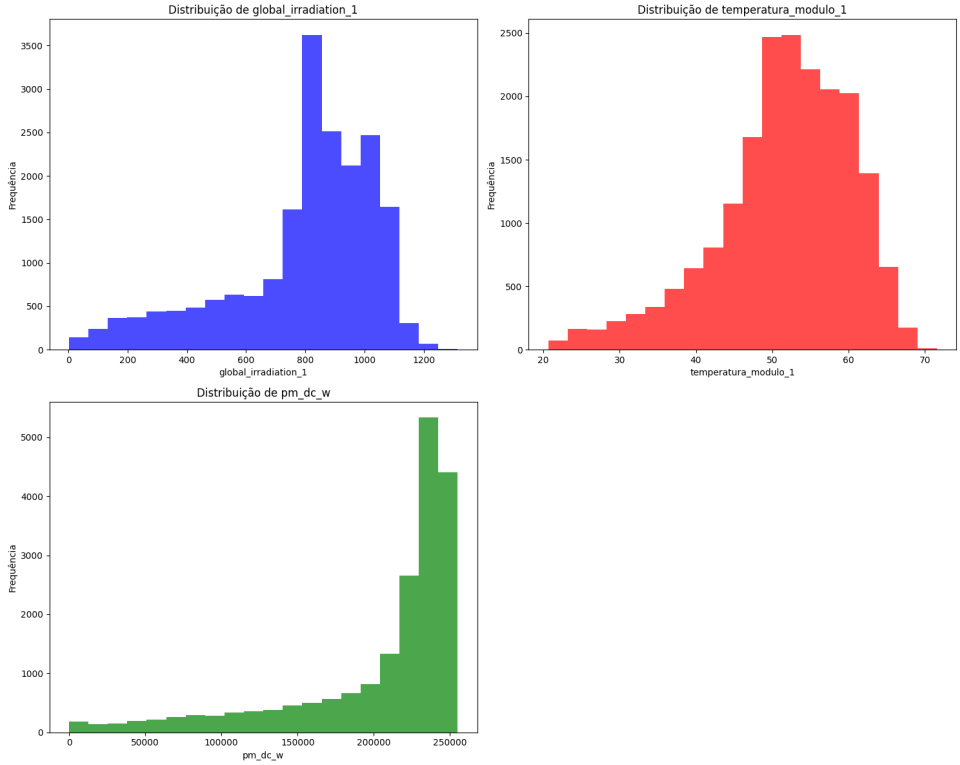

Figura 13 – Histograma das variáveis *global_irradiation_1*, *temperatura_modulo_1* e *pm_dc_w*.

Para visualizar os dados reais ao longo do tempo foi feita uma agregação diária pela média de cada variável de forma a melhorar a visualização dos dados e deixá-los menos condensados. Pela Figura 14 é possível observar que há uma sazonalização na irradiação e temperatura, principalmente entre o final do outono e final do inverno (maio a setembro), o que pode indicar uma alteração de conceito para o detector de *drift*.

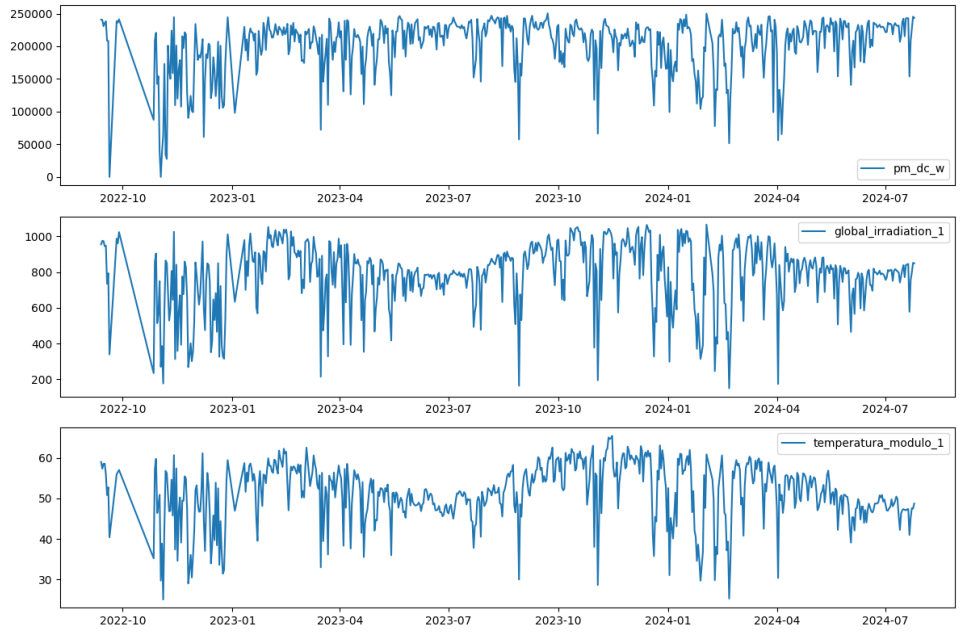

Figura 14 – Dados históricos das variáveis *global_irradiation_1*, *temperatura_modulo_1* e *pm_dc_w*.

Para comprovar a presença de sazonalização dos dados, foi feita uma decomposição sazonal para cada variável, com modelo tipo aditivo e com período trimestral, de forma a abranger cada estação do ano e ilustrada na Figura 15. É possível observar que a irradiação (*global_irradiation_1*) e temperatura (*temperatura_modulo_1*) são fortemente correlacionadas, enquanto a geração (*pm_dc_w*) comportamento contraintuitivo ante as demais, com valores mais altos em períodos de valores mais baixos de irradiação e temperatura. Esse comportamento pode ser explicado pela perda de eficiência das placas fotovoltaicas por temperatura, bem como condições de proteção eletrônica contra sobrecarga dos inversores como *derating*, discutido na Seção 2.1.4.

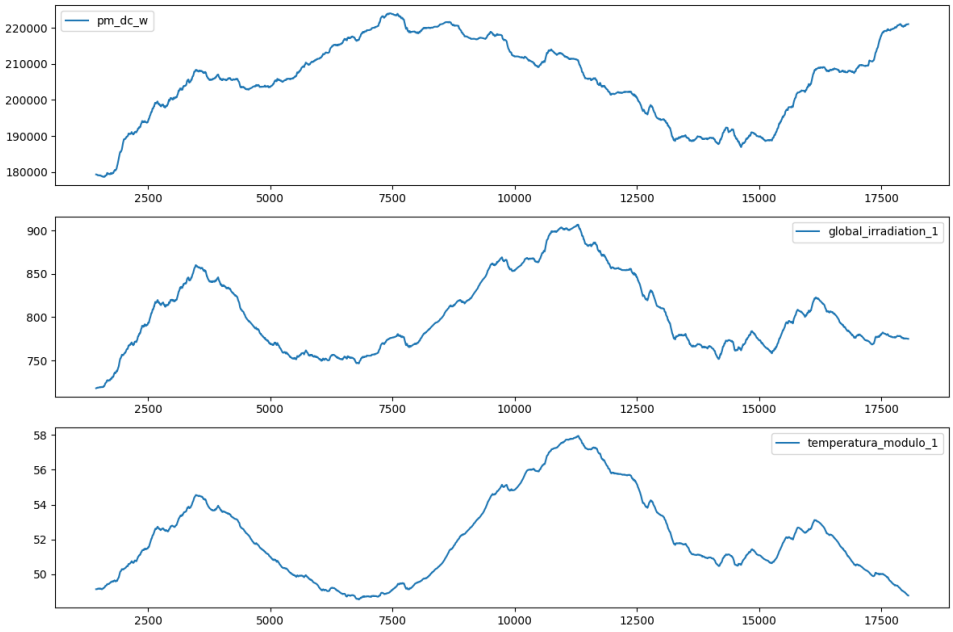

Figura 15 – Decomposição sazonal das variáveis *pm_dc_w*, *global_irradiation_1* e *temperatura_modulo_1*.

4.7 Bibliotecas

Para o atingimento dos objetivos, foram estudados métodos e bibliotecas que possuíam as ferramentas necessárias, usos comprovados, desenvolvimento ativo e facilidade de uso. Essas características visam simplificar as atividades de aprendizado de máquina e reduzir a complexidade do algoritmo em desenvolvimento.

4.7.1 *Drift*

Para o desenvolvimento da identificação de desvios será explorado o método *Page-Hinkley*. .

A biblioteca escolhida foi a *River*[2]. Essa biblioteca tem como foco o desenvolvimento online de modelos de aprendizado de máquina e contempla os principais métodos de identificações de *drifts*, com utilização de dicionários de forma a facilitar o entendimento de seus algoritmos (RIVER, 2024).

4.7.2 Predição

Como uma forma alternativa de análise de desempenho da usina, será explorada a rede neural LSTM para predição da geração. As bibliotecas utilizadas serão *Keras* e *Sklearn*.

[1] River: <https://riverml.xyz/latest/> Acessado em 06 de junho de 2024.

Essas bibliotecas são amplamente utilizadas em projetos de aprendizado de máquina, com consolidação no mercado e considerável comunidade.

4.7.3 Treinamento dos Detectores de Desvios

Os modelos serão treinados utilizando os dados históricos de forma a conter o comportamento normal das variáveis a serem monitoradas. O objetivo é que estes modelos sejam capazes de detectar desvios significativos que possam indicar desvios, como falhas ou degradação de desempenho dos equipamentos persistentes ao serem atualizados com dados em tempo real. As variáveis utilizadas nessa etapa serão: *global_irradiation_1*, *pm_dc_w* e *temperatura_modulo_1*.

4.7.4 Treinamento do Preditor

Nessa etapa serão utilizadas as variáveis *global_irradiation_1*, *pm_fault_code*, *pm_dc_w* e hora obtida pela variável *ts*. OS dados serão enriquecidos com novas variáveis temporatis derivadas da variável *ts*, como dia da semana, hora e quarto de hora como forma de auxiliar o algoritmo a identificar padrões que se repetem. Como forma de validação, será utilizada a técnica *KFold* e avaliação das métricas como MAE, MSE e RMSE.

5 AVALIAÇÃO EXPERIMENTAL

A seguir serão apresentados os desenvolvimentos e resultados dos detectores de *drift* e predição utilizando LSTM com a base de dados gerada conforme as análises do capítulo anterior.

Ambiente de desenvolvimento:

- IDE: PyCharm 2024 Professional
- Python: v3.11.9
- Ambiente virtual: Poetry v1.8.3
- River: v0.21.1
- Keras: v3.4.1
- Sklearn: v1.5.0

5.1 Detector de *Drift*

Para o desenvolvimento do detector de *drift* foi utilizado o método Page-Hinkley da biblioteca River.

A escolha do método Page-Hinkley se deu principalmente pelo baixo requisito de recursos computacionais, podendo ser embarcado em microcontroladores de equipamentos de telemetria na borda, se tornando uma alternativa interessante para alarmes em tempo real. Essa alternativa extingue a necessidade de acréscimo de poder computacional de servidores e/ou máquinas para este fim.

O objetivo do detector é identificar *drifts* do tipo gradual e incremental, visando as perdas de geração por motivos físicos e não mensuráveis por equipamentos instalados em campo, como sujidade das placas, placas com vidro trincado ou quebrado e situações similares. Dessa forma a configuração do detector deve ser menos sensível a mudanças abruptas. Para isso, uma forma de configurar o detector Page-Hinkley é utilizando a média e desvio padrão da base de dados para definição dos parâmetros de *delta* e *threshold* (JRAD *et al.*, 2017).

Um detector Page-Hinkley foi instanciado para cada variável a ser analisada. A configuração dos parâmetros de cada detector seguiu a mesma métrica de forma a padronizar o formato de configuração dos parâmetros e possibilitar a escalabilidade do algoritmo para demais variáveis seguindo os seguintes critérios:

- ***min_instances***: Valor fixo 32, sendo 4 dados por hora e 8 horas de operação, totalizando 1 dia útil de medição;
- ***delta***: Média da variável de análise multiplicado por um fator atenuante como sensibilidade;
- ***threshold***: Soma da média e desvio padrão dos dados históricos da variável de análise;
- ***alpha***: Valor fixo 0,9999 (padrão do método implementado pela biblioteca *River*);
- ***mode***: *"up"* para *temperatura_modulo_1* e *"down"* para *pm_dc_w* e *global_irradiation_1*. A escolha do modo *"up"* para *temperatura_modulo_1* e *"down"* para *global_irradiation_1* deve-se à influência dessas variáveis na queda de geração da usina, confome Equação 4.2.

Como os dados analisados possuem tempos de coleta distintos, além de possuir eventual não coleta de dados de um dos equipamentos, fez-se necessário o alinhamento dos *timestamp* sem perder a característica dos dados. Foi utilizado um *buffer* para cada variável, sendo o *deque* optado para tal tarefa. O tamanho de cada *buffer* foi definido de acordo com a janela de coleta dos dados de maior intervalo, sendo o *pm_dc_w* com coleta a cada 15 minutos. Para cada dado novo de geração, são esperados 3 dados de irradiação e 1 dado de temperatura, vide frequências de coletas descritas na Tabela 2. Essas características definem o tamanho máximo do deque para cada variável.

Para cada dado novo de geração é verificado nos demais *buffers* se há dados, na quantidade de dados esperados para cada variável, em uma janela entre o *timestamp* do dado de geração e 15 minutos a menos, como sendo $t_{gen} - 14:59 \text{ minutos} \leq t_{buffer} \leq t_{gen}$. Em caso positivo, para um *buffer* com tamanho maior que 1, é calculada a média dos dados que nele estão contidos e os dados são inseridos em seus respectivos detectores. Isso garante que cada entrada de cada detector esteja com o mesmo intervalo de tempo.

Para avaliar o modelo, as detecções de *drifts* são armazenadas em listas específicas para cada variável analisada, de forma a permitir comparações e quantificações subsequentes. No entanto, essa abordagem não é viável em um ambiente de produção, pois o armazenamento contínuo das detecções ao longo do tempo pode levar ao consumo excessivo de memória.

Dessa forma, os *timestamps* das detecções de cada variável são comparados com base nas detecções de *drifts* da geração, considerando um intervalo de até 1 hora anterior, a fim de identificar detecções concorrentes. Esse processo visa correlacionar um *drift* de geração com *drifts* simultâneos de irradiação e/ou temperatura, validando a hipótese de que as oscilações nessas variáveis meteorológicas são responsáveis pelas variações na geração de energia. Essa é a lógica do *ensemble* dos detectores.

O objetivo principal dessa análise é duplo: primeiramente, verificar a acurácia do modelo na detecção de desvios na geração que são explicados por flutuações de irradiação e temperatura; e, em segundo lugar, identificar os *drifts* não concorrentes, que são de maior interesse para o projeto, pois podem indicar problemas na geração de energia não atribuíveis a variáveis meteorológicas.

O algoritmo 1 exemplifica a rotina criada para a detecção de *drifts*.

Algorithm 1 Configuração e Avaliação do Detector de *Drifts* com Page-Hinkley

Require: Dados de geração, irradiação, temperatura
Require: Parâmetros: *min_instances*, *delta*, *threshold*, *alpha*, *mode*
Inicializar um *deque* para cada variável com tamanho baseado no maior intervalo de coleta (15 minutos)
Instanciar um detector Page-Hinkley para cada variável (*geração*, *irradiação*, *temperatura*) com os parâmetros configurados
for cada novo dado de geração t_{gen} **do**
    Verificar nos demais *buffers* se há dados no intervalo $[t_{gen} - 15 \text{ minutos}, t_{gen}]$
    if dados presentes nos *buffers* **then**
        for cada *buffer* com tamanho maior que 1 **do**
            Calcular a média dos dados no *buffer*
        end for
        Inserir os dados nos respectivos detectores Page-Hinkley
    end if
end for
Armazenar detecções de *drifts* em listas específicas para cada variável
for cada detecção de *drift* na geração t_{drift_gen} **do**
    Comparar *timestamps* com detecções de *drifts* em irradiação e temperatura em um intervalo $[t_{drift\_gen} - 1 \text{ hora}, t_{drift\_gen}]$
    if detecções concorrentes **then**
        Correlacionar *drifts* de geração com *drifts* de irradiação e/ou temperatura
    else
        Marcar *drift* de geração como não concorrente (potencial problema físico)
    end if
end for
Resultado: Validar a acurácia do modelo e identificar *drifts* não concorrentes

Para identificar a configuração ideal de sensibilidade dos detectores, diversos experimentos foram realizados, abrangendo todas as variáveis analisadas. A sensibilidade, representada pelo parâmetro *delta* do detector, foi ajustada variando-se a multiplicação da média por diferentes percentuais: 10%, 5% e 1%. Além disso, a agregação dos dados pela média foi realizada em intervalos de tempo maiores visando a redução da oscilação dos dados, incluindo 30 minutos, 45 minutos, e intervalos de uma hora, conforme ilustrado na Equação 2.1. Também foram considerados intervalos de 2 horas e 4 horas para a análise.

Para validar os *drifts* identificados, foi necessário solicitar ao cliente da Engecomp um levantamento dos problemas registrados pela equipe de O&M no período de fevereiro

de 2022 a junho de 2024. Vale destacar que os dados de erro começaram a ser coletados pelo cliente a partir de agosto de 2023. Os problemas registrados estão apresentados na Tabela 5. Observa-se que ocorreram problemas com os *trackers*, os quais persistiram por aproximadamente dois meses, além de problemas recorrentes na rede da concessionária entre os meses de abril e junho de 2024. As ações de lavagem dos painéis fotovoltaicos foram coletadas para verificar se houve um impacto positivo na geração após a execução do serviço.

Tabela 5 – Ocorrências e Impacto na Geração

Data Inicial	**Data Final**	**Tipo de Problema**	**Impacto na Geração**
2023-08-02 00:00:00	2023-08-05 23:59:00	Lavagem	Não
2024-04-18 00:00:00	2024-04-21 23:59:00	Lavagem	Não
2024-07-24 00:00:00	2024-07-27 23:59:00	Lavagem	Não
2023-08-15 08:30:00	2023-08-15 11:00:00	Concessionaria	Sim
2024-02-01 06:00:00	2024-02-01 06:46:00	Concessionaria	Sim
2024-04-05 06:00:00	2024-04-05 12:15:00	Concessionaria	Sim
2024-04-08 12:45:00	2024-04-08 13:36:00	Concessionaria	Sim
2024-04-10 14:15:00	2024-04-12 15:15:00	Concessionaria	Sim
2024-05-13 06:00:00	2024-05-13 07:44:00	Concessionaria	Sim
2024-06-10 06:00:00	2024-06-10 07:30:00	Concessionaria	Sim
2024-06-23 06:00:00	2024-06-23 07:20:00	Concessionaria	Sim
2024-07-26 09:30:00	2024-07-26 13:36:00	Concessionaria	Sim
2023-10-02 14:30:00	2023-10-03 07:32:00	Inversor	Sim
2023-10-17 06:00:00	2023-10-17 10:15:00	Inversor	Sim
2023-10-17 07:30:00	2023-10-18 12:00:00	Inversor	Sim
2024-03-06 07:59:00	2024-03-06 11:00:00	Inversor	Sim
2024-03-25 09:01:00	2024-03-25 09:01:00	Inversor	Sim
2024-04-03 12:45:00	2024-04-11 15:32:00	Inversor	Sim
2024-06-05 07:00:00	2024-06-05 07:00:00	Inversor	Sim
2023-12-28 13:45:00	2024-01-16 15:02:00	Tracker	Sim
2024-01-31 15:00:00	2024-01-02 07:23:00	Tracker	Sim
2024-01-30 16:00:00	2024-01-31 06:30:00	Transformador	Sim

5.1.1 Resultados

Os resultados das variações de sensibilidade do detector, juntamente com os dados agregados, estão apresentados na Tabela 6, destacando-se a melhor configuração em negrito. As colunas denominadas com "Percentual" corresponde à quantidade de *drifts* ante ao total de amostras da base de dados utilizada, sendo o total de 16.254 intervalos de 15 minutos. A configuração que demonstrou a maior eficácia na identificação de *drifts* na geração, explicados por quedas na irradiação e/ou aumentos na temperatura, foi obtida utilizando um *delta* de 5% da média dos dados, com os dados mantidos na sua frequência de coleta

padrão, resultando em 50,59% de *drifts* concorrentes. Adicionalmente, a quantidade de *drifts* detectados representou menos de 1,1% de toda a base de dados analisada para cada variável.

Tabela 6 – Avaliação de *drifts*.

Sensibilidade (delta)	**Frequência**	**Percentual Geração**	**Percentual Irradiação**	**Percentual Temperatura**	**Percentual Concorrentes**
10%	15m	0.79%	0.89%	0.06%	43.75%
5%	**15m**	**1.05%**	**1.09%**	**0.22%**	**50.59%**
1%	15m	1.37%	1.26%	0.65%	31.53%
10%	30m	0.65%	0.83%	0.06%	35.29%
5%	30m	0.76%	0.97%	0.14%	35.0%
1%	30m	1.12%	1.11%	0.56%	42.05%
10%	45m	0.71%	0.71%	0.08%	37.84%
5%	45m	0.85%	0.87%	0.17%	45.45%
1%	45m	1.08%	1.04%	0.52%	37.5%
10%	1h	0.73%	0.82%	0.10%	43.33%
5%	1h	0.90%	0.90%	0.19%	29.73%
1%	1h	1.38%	1.16%	0.65%	19.3%
10%	2h	0.73%	0.94%	0.10%	0.0%
5%	2h	0.94%	1.26%	0.10%	11.11%
1%	2h	1.26%	1.36%	0.63%	16.67%
10%	4h	0.73%	0.94%	0.10%	0.0%
5%	4h	0.94%	1.26%	0.10%	11.11%
1%	4h	1.26%	1.36%	0.63%	16.67%
10%	8h	0.91%	0.68%	0.23%	0.0%
5%	8h	0.91%	1.13%	0.23%	25.0%
1%	8h	1.36%	1.36%	0.68%	16.67%

Para a visualização dos *drifts* na geração de energia, foram removidas as detecções que coincidiram com variações simultâneas de irradiação e temperatura, isto é, aquelas concorrentes descritas na Tabela 6, mantendo-se todas as detecções para as demais variáveis. Com isso, restaram apenas as detecções de *drifts* na geração que não puderam ser explicadas por interferências meteorológicas, sugerindo a possibilidade de problemas na usina. Para validar essas detecções e verificar se, de alguma forma, elas precederam problemas identificados pela equipe de campo, os problemas listados na Tabela 5 foram destacados nas Figuras 16, 19, 17 e 18 como faixas translúcidas horizontais com cores distintas para cada problema (lavagem em verde, concessionaria em preto, inversor em marrom, *tracker* em laranja e transformador em roxo). Para todos os problemas detectados pela equipe de campo houve uma detecção de *drift*, representada em linhas vermelhas tracejadas na vertical, nos dados de geração capturada no intervalo de 60 minutos que precede o problema, mesmo para problemas que ocorreram em um curto período de tempo.

Utilizando os gráficos da Figura 16, é possível verificar que mesmo para o problema de transformador, tendo apenas 1 ocorrência verificada, foi detectado um *drift* no dia 30/01/2024 às 12:00 e outro no dia 31/01/2024 às 15:30 (marcados como "A" na Figura 16), sendo este último o registro de uma geração zero, característica de um problema de desligamento total da usina e que confronta a informação de solução do problema enviado pela equipe de O&M.

Para os problemas de inversor não foram detectados *drifts* em 3 ocorrências sendo as das datas 17/10/2023 e 25/03/2024 (círculos verde e amarelo na Figura 17), mas com detecção com 1 dia de atraso para a data 02/10/2023 (círculo preto na Figura 17) e detecções para as demais datas dentro do intervalo de tempo do problema registrado pela equipe de O&M. A não detecção no intervalo de acontecimento do problema pode ser explicada pelo motivo do problema estar em apenas 1 dos 13 inversores, tendo baixo impacto na geração total da usina.

Para os problemas relacionados aos *trackers*, foi identificado um *drift* com um dia de atraso em relação à ocorrência registrada em 28/12/2023 (marcado como "B" na Figura 18). Durante o período de maior duração registrado, observam-se alterações na geração de energia que não correspondem às variações na irradiação, o que pode indicar a realização de testes durante a manutenção dos equipamentos. Além disso, foi detectado um aumento súbito e persistente na geração em momentos que antecederam o *drift* identificado em 17/01/2024, ocorrido um dia após a conclusão da manutenção dos equipamentos. Em relação ao problema reportado em 31/01/2024 (marcado como "C" na Figura 18), também foi constatado um incidente envolvendo o transformador, que resultou em geração nula. Isso sugere que o *drift* identificado nesse momento deve-se ao problema com o transformador, e não ao *tracker*.

Para os problemas relacionados à concessionária, foram detectados *drifts* apenas para os registros de 05/04/2024 e 10/06/2024 (marcados como "D" e "E" na Figura 19, respectivamente). A ausência de detecção nas demais ocorrências pode ser semelhante à não detecção dos problemas relacionados ao inversor, apesar de representarem um problema, devido ao seu baixo impacto na geração total da usina. Para uma análise mais detalhada, seriam necessários mais detalhes sobre a natureza específica do problema, informações que a equipe de O&M não possuía no momento.

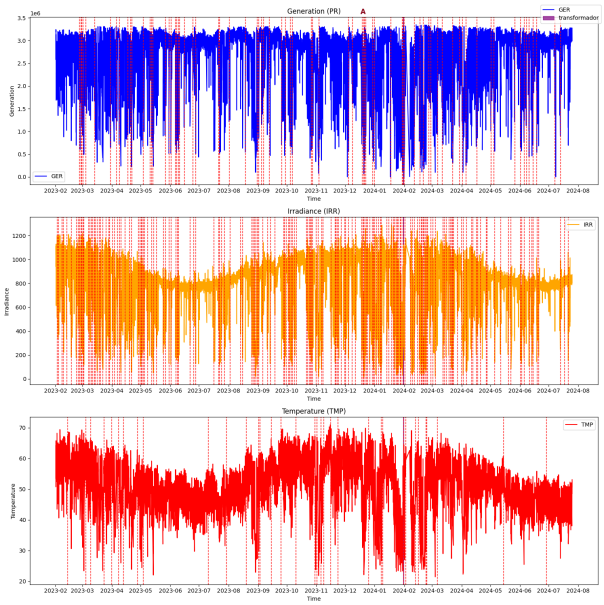

Figura 16 – Gráficos de *drifts* identificados nas variáveis *pm_dc_w* (GER), *temperatura_modulo_1* (TMP) e *global_irradiation_1* (IRR) representados por tracejados vermelhos verticais junto com ocorrências no transformador destacadas como faixas horizontais.

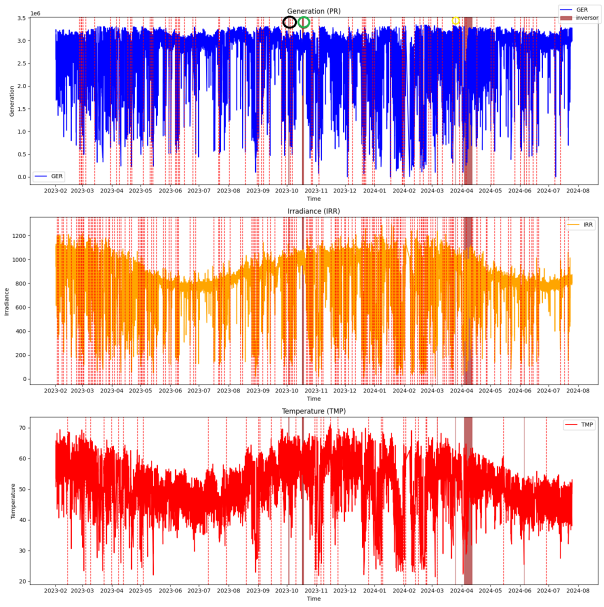

Figura 17 – Gráficos de *drifts* identificados nas variáveis *pm_dc_w* (GER), *temperatura_modulo_1* (TMP) e *global_irradiation_1* (IRR) representados por tracejados vermelhos verticais junto com ocorrências em inversores destacadas como faixas horizontais.

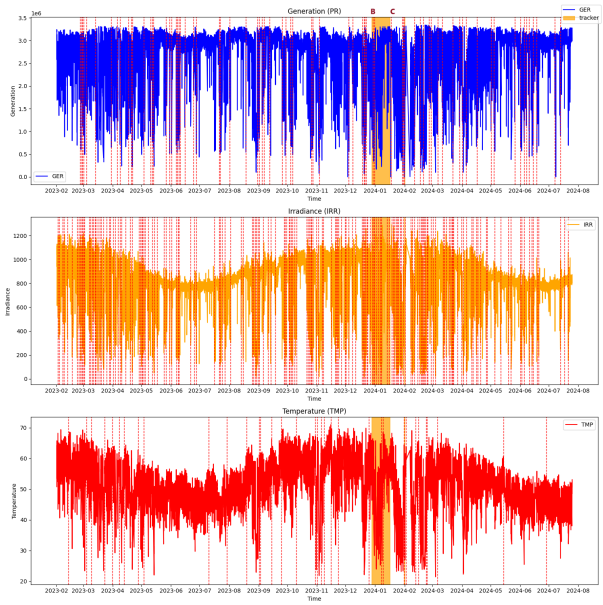

Figura 18 – Gráficos de *drifts* identificados nas variáveis *pm_dc_w* (GER), *temperatura_modulo_1* (TMP) e *global_irradiation_1* (IRR) representados por tracejados vermelhos verticais junto com ocorrências em *trackers* destacadas como faixas horizontais.

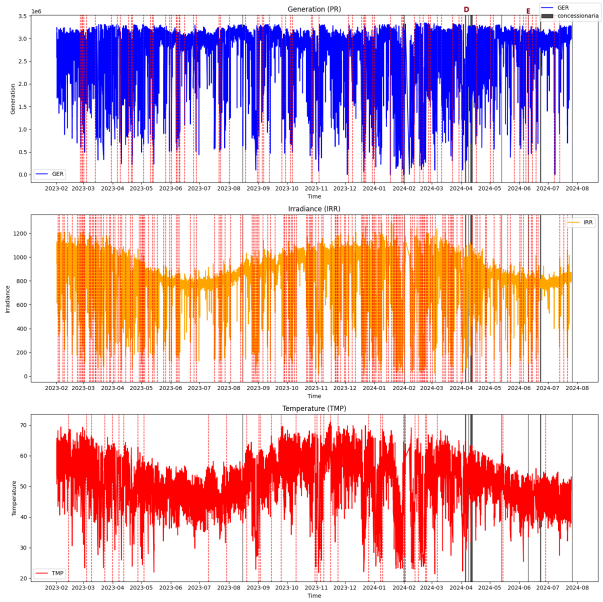

Figura 19 – Gráficos de *drifts* identificados nas variáveis *pm_dc_w* (GER), *temperatura_modulo_1* (TMP) e *global_irradiation_1* (IRR) representados por tracejados vermelhos verticais junto com ocorrências na rede da concessionária destacadas como faixas horizontais.

5.1.2 Considerações Finais

Os detectores foram de implementação simples e com razoável explicabilidade nos *drifts* de geração pelas variáveis de temperatura e irradiação. A principal dificuldade é

validar o modelo para cada usina, invalidando uma implementação sem as informações das ocorrências na usina pela equipe de O&M, não invalidando porém dificultando a escalabilidade do projeto em curto prazo.

Outra dificuldade encontrada foi capturar mudanças sutis nos dados devido às grandes oscilações nos dados.

5.2 LSTM

A rede neural LSTM foi escolhida para a tarefa de previsão de geração pela ampla utilização em séries temporais e por ser mais eficiente na predição com depedências de longo prazo, visto que a geração de energia possui uma tendência senoidal pelas estações do ano.

O objetivo da rede neural é prever pelo menos 7 dias de geração futura usando como base os dados históricos de geração, irradiação e temperatura da usina. Isso permitirá avaliar a precisão do modelo ao comparar a geração predita com a real, o que é essencial para detectar discrepâncias significativas ou desvios na performance ao longo do tempo. Esse processo de previsão é fundamental para antecipar possíveis problemas e melhorar a manutenção preventiva, garantindo a eficiência contínua da usina.

O conjunto de dados foi montado utilizando o algoritmo da etapa de deteção de *drift*, sendo os mesmos dados de entrada de cada detector. Dessa forma obteve-se um conjunto com células com os *timestamps* alinhados para todas as variáveis de interesse.

Para o treinamento do modelo, os dados de geração, temperatura e irradiação foram normalizados utilizando a técnica *"MinMax"* de forma a igualar as escalas porém mantendo as características de oscilações de cada variável, além de contribuir no treinamento da rede neural prevenindo saturação da função de ativação e acelerando o processo de treino. Após a normalização, foram inseridos mais 3 colunas, sendo *"month"*, *"day_of_week"* e *"hour"* sendo dados numéricos de tipo inteiro. Essa inserção visa incluir no modelo demais variáveis que auxiliam a identificação de padrões e repetições da variável alvo *"pm_dc_w"*. Por fim, o conjunto de dados foi dividido em 80% para treino e 20% para teste.

O modelo consiste em duas camadas LSTM, cada uma seguida de uma camada de *dropout* como forma de evitar *overfitting* e uma camada densa de saída com um único neurônio. O modelo é compilado utilizando o otimizador *Adam*, com métricas de validação MSE, RMSE e MAE. O tamanho da janela foi definido por 224, sendo o total de 7 dias de operação da usina com 8 horas de coleta de dados a cada 15 minutos.

Como validação, foi implementada a técnica *KFold* com 5 divisões e com embaralhamento dos dados, além de filtrar os dados de treinamento e teste até o dia 30/06/2024 e utilizar os dados de 01/07/2024 em diante para validar as predições do modelo.

A fim de identificar a melhor combinação de hiperparâmetros, foi utilizada a técnica

de *"GridSearchCV"*, alterando os parâmetros *neurons, dropout_rate, learning_rate, epochs, batch_size*. De forma a interromper o treinamento em caso de estabilização da métrica *loss*, foi utilizada a técnica de *EarlyStopping* com uma paciência de 10 épocas.

Assim que os melhores parâmetros foram encontrados pelo *"GridSearchCV"*, o modelo foi novamente treinado e salvo em formato *.pkl* para que pudesse ser reutilizado.

O algoritmo 2 exemplifica a rotina criada para o treinamento da rede neural LSTM.

Algorithm 2 Algoritmo de Treinamento do Modelo LSTM para Previsão da Geração de Energia

Require: Dados de entrada X e y, Tamanho da Janela $window\_size$, Parâmetros do Modelo $param\_grid$
Ensure: Modelo LSTM treinado e avaliação de desempenho
Preprocessamento:
Carregar os dados e extrair características temporais (*mês*, *dia da semana*, *hora*)
Normalizar as características usando *MinMaxScaler*
Criar sequências multivariadas de X e y com o tamanho da janela $window\_size$
Divisão do Conjunto de Dados:
Dividir os dados em conjuntos de treinamento, validação e teste usando *train_test_split*
Dividir os dados em k **folds** para validação cruzada com $KFold$
Construção e Treinamento do Modelo:
Definir a estrutura do modelo LSTM com camadas *LSTM* e *Dropout*
Definir o otimizador e a função de perda
Inicializar o *GridSearchCV* com o modelo LSTM e os parâmetros $param\_grid$
Treinar o modelo com *GridSearchCV*, usando o MSE negativo como métrica de avaliação
Seleção dos Melhores Hiperparâmetros:
Identificar os melhores hiperparâmetros a partir dos resultados do *GridSearchCV*
Re-treinar o modelo final utilizando todo o conjunto de treinamento com os melhores hiperparâmetros
Avaliação do Modelo:
Avaliar o desempenho do modelo no conjunto de teste calculando as métricas MSE, MAE, e RMSE
Previsão para 7 Dias Futuros:
Usar a última sequência do conjunto de teste para prever a geração de energia para os próximos 7 dias
Inverter a normalização dos dados previstos para obter os valores reais
Visualização dos Resultados:
Plotar as previsões futuras em comparação com os dados reais

Os melhores hiperparâmetros encontrados pelo *GridSearchCV* estão descritos abaixo e as caractísticas da rede estão descritas na Tabela 7.

- *batch_size*: 128
- *dropout_rate*: 0.2
- *epochs*: 100

- *learning_rate*: 0.001
- *neurons*: 50

Tabela 7 – Estrutura da Rede Neural LSTM.

Camada (tipo)	**Formato de Saída**	**Parâmetros #**
lstm_2 (LSTM)	(None, 56, 50)	11,400
dropout_2 (Dropout)	(None, 56, 50)	0
lstm_3 (LSTM)	(None, 50)	20,200
dropout_3 (Dropout)	(None, 50)	0
dense_1 (Dense)	(None, 1)	51

5.2.1 Resultados

Os resultados das métricas de avaliação da rede neural LSTM estão descritas na Tabela 8 e seus respectivos gráficos ao longo das épocas ilustrados na Figura 20. Ao todo foram utilizados 31.651 (123.64KB) de parâmetros treináveis e zero parâmetros não treináveis. Ainda que seja uma rede neural relativamente simples, os resultados obtidos foram satisfatórios ante à natureza de cada variável, com erros de até 12%. O resultado negativo do *KFold* indica subestimação nas previsões, porém apresenta um valor próximo de zero. As métricas MAE, MSE e RMSE apresentaram bons resultados, em especial ao MSE por penalizar erros grandes de forma mais severa, apresentou um valor de 0.01562.

Tabela 8 – Resultado das métricas de desempenho da Rede Neural LSTM.

Métrica	**Resultado**
KFold	0.01550
MAE	0.08697
MSE	0.01562
RMSE	0.12452
Loss	0.01634

Ao analisar as curvas das métricas da Figura 20, nota-se uma queda acentuada dos erros a partir da 20ª época, apresentando uma estabilidade e repetição de padrões a partir da 80ª época, demonstrando que a aprendizagem do modelo convergiu rapidamente.

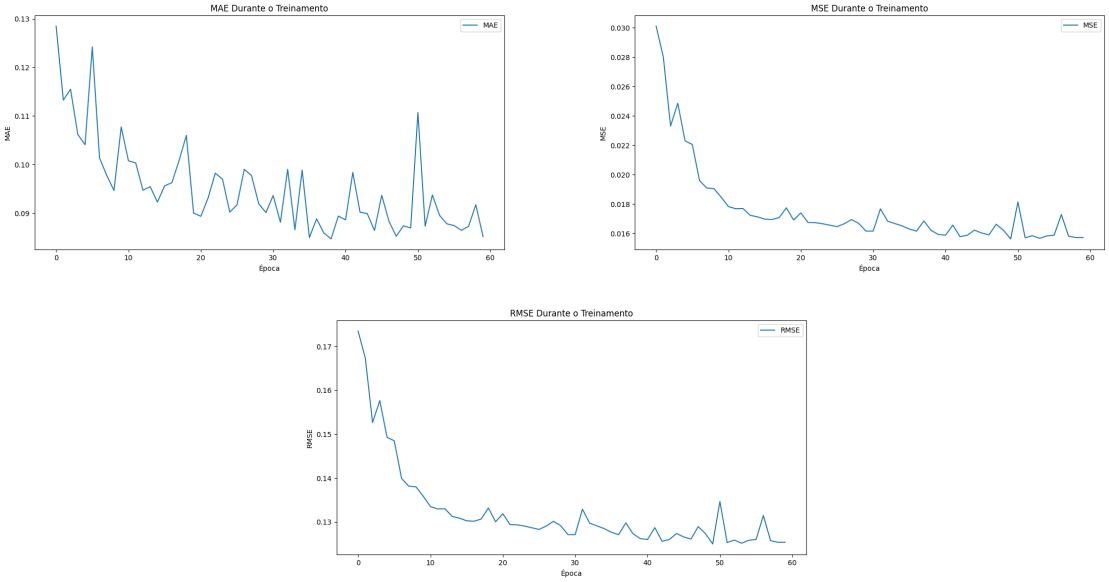

Figura 20 – Gráficos das métricas de erro da rede neural LSTM ao longo das épocas de aprendizado.

5.2.2 Considerações Finais

Ainda que as métricas apresentaram bom desempenho e se mostrando uma alternativa robusta para o monitoramento do desempenho da usina, há espaço para melhorias na rede neural, mas com o ônus de exigir mais capacidade computacional e dificultar a escalabilidade para um ambiente com diversas usinas em diferentes locais do território brasileiro.

6 CONCLUSÕES

O principal problema de uma usina de geração distribuída é o custo de Operação e Manutenção. Com a aplicação de tecnologias de aprendizado de máquina eficientes e computacionalmente menos custosas os resultados financeiros ao longo prazo serão maiores, além de possuir informações em tempo real de prováveis problemas, mesmo em áreas remotas, dispensando a necessidade de acompanhamento humano assíduo.

Nesse contexto, a análise e previsão de séries temporais de dados de telemetria é um importante recurso na manutenção preventiva e preditiva da usina fotovoltaica com impacto financeiro.

Neste trabalho foram apresentados modelos com diferentes custos computacionais e diferentes abordagens para a identificação de desvios na geração de energia de uma usina fotovoltaica, visando identificar problemas que não são reportados pelos inversores como sujidade, painéis trincados, células queimadas, dentre outros tipos de problemas que afetam a geração.

Durante a análise exploratória foram encontrados algumas dificuldades como a equalização dos *timestamps* para os modelos de aprendizado de máquina, além da possibilide de ausência de dados em qualquer momento. Outro ponto de dificuldade foram as caudas alongadas vistas nos histogramas dificultando a modelagem dos dados e se fazendo necessária a aplicação de filtros nas séries temporais para minimizar a presença de *outliers*.

A utilização de detectores de *drifts*, em especial o modelo Page-Hinkley, se mostrou como uma alternativa simples e passível de ser embarcada em microcontroladores, realizando as análises na borda, com uma explicabilidade de desvios de geração de mais de 50% pelas variações de irradiação e temperatura. Ainda que esse número não seja o ideal, para aplicações onde não há investimentos em equipamentos com mais tecnologia, especialmente inversores, pode ser uma boa alternativa.

Uma desvantagem dos algoritmos de detecção de *drifts* é a dificuldade em capturar mudanças sutis nos dados, como por exemplo capturar o desgaste das placas solares ao longo dos anos. Essa dificuldade foi reportada em outros trabalhos como em (MAHDI *et al.*, 2024), onde o autor explicita a escassez de estudos de detecção de *drifts* incrementais e a necessidade de tratar mudanças abruptas de forma específica sem que interfira no modelo. Ainda, o autor comenta a necessidade de haver um *buffer* para armazenar comportamentos anteriores e reutilizar o conhecimento aprendido em observações anteriores, sendo análogo ao funcionamento de uma LSTM.

A rede neural desenvolvida com 2 camadas LSTM com 50 neurônios cada apresentou um bom resultado na previsão da geração de energia combinado com as variáveis de

irradiação e temperatura, se mostrando uma alternativa mais robusta para o monitoramento do desempenho da usina, podendo ser utilizada para comparar os valores preditos com os valores reais e ainda utilizar detectores de *drifts* na saída de forma a prever quedas consistentes na geração. Uma outra vantagem dessa rede neural é que a possibilidade do treinamento incremental após um período desejado, mantendo a rede confiável para as predições.

As desvantagens de se utilizar esse tipo de abordagem é a exigência computacional e a complexidade do algoritmo para implementação em larga escala. Como usinas possuem características construtivas diferentes e podem estar em condições climáticas diferentes, seria necessário o treinamento para cada caso específico.

Por fim, ambas as abordagens são promissoras para a avaliação de desempenho de uma usina fotovoltaica utilizando apenas 3 variáveis de interesse, disponíveis na grande maioria de usinas fotovoltaicas, tanto em geração distribuída quanto em outros fins. O principal desafio dessas abordagens é a validação em campo, uma vez que nem todas as usinas possuem um controle detalhado dos ocorridos para que seja possível avaliar a acurácia das detecções e previsões.

6.1 Trabalhos Futuros

Como trabalhos futuros, pode-se detectar *drifts* na performance da usina ajustada pela temperatura, conforme a Equação 4.2 e dentro de intervalos específicos como, por exemplo, acima de 85% para avaliar a deteção de *drifts* mais sutis ao longo do tempo, podendo indicar problemas emergentes como degradação dos módulos acima do esperado, problemas no cabeamento, além de garantir uma melhor eficiência na manutenção da usina.

Ainda que os inversores dispõem de alarmes e registro em memória de ocorrências nos sensores de energia, os problemas alheios a esses sistemas podem ser registrados e assim criado um classificador para determinar a natureza do *drift* encontrado, tornando a avaliação mais robusta e automática.

Para a rede neural LSTM, utilizando máquinas com maior capacidade computacional que a utilizada no desenvolvimento deste projeto, pode-se variar o tamanho de janela para períodos mais longos como trimestres, bem como aumentar o número de unidades LSTM e realizar mais validações cruzadas para aprimorar o modelo e garantir que esteja generalizando bem os dados.

Uma vez que um desvio é detectado, é interessante classificá-lo a fim de entender a qual erro pertence e se esse desvio é conhecido dentre os parâmetros de detecção do inversor. Desvios desconhecidos são de maior interesse para o projeto e podem indicar degradação do painel, alta sujidade ou demais condições físicas que diminuem a geração prevista

da usina. Para o desenvolvimento da classificação de erros pode ser explorado o método de N-Vizinhos Próximos (do inglês, *Key Nearest Neightbors* - KNN). A biblioteca *River* também possui algoritmos de classificação, dentre eles o método KNN. Essa biblioteca pode ser utilizada para as tarefas de classificação, mantendo todo o ecossistema de aprendizado de máquina dependendo de uma única biblioteca, visando a facilidade de implementação e eximindo pós tratamentos de dados para intercâmbio entre as tarefas de detecção de desvios e classificação. Casos em que o erro não esteja catalogado serão de grande interesse para o projeto.

REFERÊNCIAS

ABSOLAR. **Panorama da solar fotovoltaica no Brasil e no mundo**. 2024. Disponível em: <https://www.absolar.org.br/mercado/infografico/>.

AGRAHARI, S.; SINGH, A. K. Concept drift detection in data stream mining : A literature review. **Journal of King Saud University - Computer and Information Sciences**, v. 34, n. 10, Part B, p. 9523–9540, 2022. ISSN 1319-1578.

ANEEL. ResoluÇÃo normativa aneel nº 1.000, de 7 de dezembro de 2021. **Diário Oficial da República Federativa do Brasil**, 2021.

BOSMAN, L. B. *et al.* Pv system predictive maintenance: Challenges, current approaches, and opportunities. **Energies**, v. 13, n. 6, 2020. ISSN 1996-1073.

BRASIL. Decreto nº 5.163 de 30 de julho de 2004. **Diário Oficial [da] República Federativa do Brasil**, 2004.

BRASIL. ResoluÇÃo normativa nº 482, de 17 de abril de 2012. **Diário Oficial [da] República Federativa do Brasil**, 2012.

BRASIL. ResoluÇÃo normativa nº 687, de 24 de novembro de 2015. 2015.

BRASIL. Lei nº 14.300, de 6 de janeiro de 2022. **Diário Oficial [da] República Federativa do Brasil**, 2021.

BUBE, R. H. **Photovoltaic materials**. [*S.l.: s.n.*]: World Scientific, 1998. v. 1.

CANADIANSOLAR. **Canadian Solar Datasheet HiKu CS3W**. [*S.l.*], 2020.

CHAAR, L. E.; ZEIN, N. E. *et al.* Review of photovoltaic technologies. **Renewable and sustainable energy reviews**, Elsevier, v. 15, n. 5, p. 2165–2175, 2011.

CHEN, Z. *et al.* Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions. **Energy Conversion and Management**, v. 198, p. 111793, 2019. ISSN 0196-8904.

CHINE, W. *et al.* Fault detection method for grid-connected photovoltaic plants. **Renewable Energy Volume 66**, 2013.

COSTA, C. H. d. *et al.* **Classificação de falhas em plantas fotovoltaicas usando aprendizado de máquina**. 2020. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2020.

DANEELS, A.; SALTER, W. What is scada? 1999.

ELECTRONICS, D. **M125HV Operation and Installation Manual**. [*S.l.*], 2021.

ENGECOMP. **Sistemas para Geração Distribuída**. 2024. Disponível em: <https://engecomp.com.br/gera%C3%A7%C3%A3o-distribu%C3%ADda-1>.

EPE, E. de P. E. **Instruções para Solicitação de Cadastramento e Habilitação Técnica com vistas à participação nos Leilões de Energia Elétrica**. [*S.l.*], 2016.

GREENER. **Modelos de Negócio em Geração Distribuída**. 2020. Disponível em: <https://sebrae.com.br/Sebrae/Portal%20Sebrae/UFs/PI/Anexos/greener.pdf>.

HAYKIN, S. **Redes neurais: princípios e prática**. [*S.l.: s.n.*]: Bookman Editora, 2001.

HUKSEFLUX. **Estação Solarimétrica GD**. [*S.l.*], 2024.

JRAD, N. *et al.* A page-hinkley based method for hfos detection in epileptic depth-eeg. *In*: **2017 25th European Signal Processing Conference (EUSIPCO)**. [*S.l.: s.n.*], 2017. p. 1295–1299.

LIRA, A. L. d. O.; SOARES, B. de L.; SANTOS, S. de A. Estação solarimétrica de referência: Instalação, operação e manutenção. *In*: **Congresso Brasileiro de Energia Solar-CBENS**. [*S.l.: s.n.*], 2016. p. 1–8.

LU, J. *et al.* Learning under concept drift: A review. **IEEE transactions on knowledge and data engineering**, IEEE, v. 31, n. 12, p. 2346–2363, 2018.

MADETI, S. R.; SINGH, S. A comprehensive study on different types of faults and detection techniques for solar photovoltaic system. **Solar Energy**, v. 158, p. 161–185, 2017. ISSN 0038-092X.

MAHDI, O. A. *et al.* Roadmap of concept drift adaptation in data stream mining, years later. **IEEE Access**, IEEE, 2024.

MAHESH, B. Machine learning algorithms-a review. **International Journal of Science and Research (IJSR).[Internet]**, v. 9, n. 1, p. 381–386, 2020.

MALLWITZ, R.; ENGEL, B. Solar power inverters. *In*: **2010 6th International Conference on Integrated Power Electronics Systems**. [*S.l.: s.n.*], 2010. p. 1–7.

MANSOURI, M. *et al.* Wavelet optimized ewma for fault detection and application to photovoltaic systems. **Solar Energy**, Elsevier, v. 167, p. 125–136, 2018.

MARQUES, Â. E. B. *et al.* **Dispositivos Semicondutores Diodos e transistores**. [*S.l.: s.n.*]: Saraiva Educação SA, 1997.

MEKKI, H.; MELLIT, A.; SALHI, H. Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules. **Simulation Modelling Practice and Theory**, v. 67, p. 1–13, 2016. ISSN 1569-190X.

MORAIS, C. V. d. S.; PONTES, I. C. d. C. **Boas práticas de operação de manutenção em usinas fotovoltaicas para uma maior eficiência e confiabilidade**. 2022. 104 f. Monografia (Graduação) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2022.

PAHWA, K. *et al.* Performance evaluation of machine learning techniques for fault detection and classification in pv array systems. *In*: **2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)**. [*S.l.: s.n.*], 2020. p. 791–796.

PARAíBA, G. da. **Atlas Solarimétrico da Paraíba**. 2023. Disponível em: <https://atlassolar.pb.gov.br/atlas-pt/metodologia-pt.html>.

PEREIRA, F.; MITCHELL, T.; BOTVINICK, M. Machine learning classifiers and fmri: A tutorial overview. **NeuroImage**, v. 45, n. 1, Supplement 1, p. S199–S209, 2009. ISSN 1053-8119. Mathematics in Brain Imaging.

PINHO, J. T. *et al.* Sistemas híbridos. **Brasília: Ministério de Minas e Energia**, 2008.

PINHO, J. T.; GALDINO, M. A. **Manual de Engenharia para Sistemas Fotovoltaicos**. [*S.l.*], 2014.

QUESADA, G. *et al.* Tracking strategy for photovoltaic solar systems in high latitudes. **Energy Conversion and Management**, v. 103, p. 147–156, 2015. ISSN 0196-8904.

RIVER. **Basic concepts**. 2024. Disponível em: https://riverml.xyz/latest/introduction/basic-concepts/https://atlassolar.pb.gov.br/atlas-pt/metodologia-pt.html.

SCHUBERT, C. **Atlas Eólico e Solar: Ceará**. Fortaleza: ADECE, FIEC, SEBRAE, 2019. 188 p. ISBN 978-85-67342-05-4.

SENSORS, H. T. **How to calculate PV performance ratio and performance index**. [*S.l.*], 2024.

SEYEDMAHMOUDIAN, M. *et al.* State of the art artificial intelligence-based mppt techniques for mitigating partial shading effects on pv systems – a review. **Renewable and Sustainable Energy Reviews**, v. 64, p. 435–455, 2016. ISSN 1364-0321.

SHERSTINSKY, A. Fundamentals of recurrent neural network rnn and long short-term memory lstm network. **Physica D: Nonlinear Phenomena**, v. 404, p. 132306, 2020. ISSN 0167-2789. Available at: <https://www.sciencedirect.com/science/article/pii/S0167278919305974>.

SOCIAL, A. de C. **Capacidade instalada de geração distribuída solar cresce e atinge 18 GW**. 2023. Disponível em: <https://www.gov.br/mme/pt-br/assuntos/noticias/capacidade-instalada-de-geracao-solar-cresce-e-atinge-18-gw>.

SPATARU, S. *et al.* Diagnostic method for photovoltaic systems based on light i–v measurements. **Solar Energy**, v. 119, p. 29–44, 2015. ISSN 0038-092X.

UCZAI, D. P. **Energias Renováveis riqueza sustentável ao alcance da sociedade**. [*S.l.: s.n.*]: Biblioteca Digital da Câmara dos Deputados, 2012.

VALMONT. **CONVERT-1P SINGLE-AXIS SOLAR TRACKER | 1-IN-PORTRAIT**. [*S.l.*], 2022.

WANG, H.; ABRAHAM, Z. Concept drift detection for streaming data. *In*: **2015 International Joint Conference on Neural Networks (IJCNN)**. [*S.l.: s.n.*], 2015. p. 1–9.

Waqar Akram, M. *et al.* Failures of photovoltaic modules and their detection: A review. **Applied Energy**, v. 313, p. 118822, 2022. ISSN 0306-2619.

YAN, S. **Understanding LSTM and its diagrams**. 2016. Disponível em: <https://blog.mlreview.com/understanding-lstm-and-its-diagrams-37e2f46f1714>.