



# **Detecção de anomalias em geração de energia solar utilizando LSTM Autoencoder**

**Trabalho de Conclusão de Curso**

**Engenharia da Computação**

**AILSON RAMON COSTA SILVA**

**Orientador: Prof. Alexandre Magno Andrade Maciel**

**Universidade de Pernambuco  
Escola Politécnica de Pernambuco  
Graduação em Engenharia de Computação**

**AILSON RAMON COSTA SILVA**

**DETECÇÃO DE ANOMALIAS EM  
GERAÇÃO DE ENERGIA SOLAR  
UTILIZANDO LSTM AUTOENCODER**

**Monografia** apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

**Recife, maio de 2022.**

Ramon, Ailson Costa Silva

Detecção de anomalias em geração de energia solar utilizando LSTM  
Autoencoder/ Ailson Ramon Costa Silva. - Recife, 2022.  
xiv, 42 f.: il. ; 29 cm.

Trabalho de Conclusão de Curso (Graduação em Engenharia de  
Computação) Universidade de Pernambuco, Escola Politécnica de  
Pernambuco local, ano

Orientador (a): Profº. Drº. Alexandre Magno Andrade Maciel.

1. Série temporal. 2. Detecção de anomalia. 3. Redes neurais. I.  
Detecção de anomalias em geração de energia solar utilizando  
LSTM Autoencoder. II. Orientador Maciel, Alexandre. III.  
Universidade de Pernambuco.

## MONOGRAFIA DE FINAL DE CURSO

### Avaliação Final (para o presidente da banca)\*

No dia 27/05/2022, às 8:00, reuniu-se para deliberar sobre a defesa da monografia de conclusão de curso do(a) discente **AILSON RAMON COSTA SILVA**, orientado(a) pelo(a) professor(a) **ALEXANDRE MAGNO ANDRADE MACIEL**, sob título Detecção de anomalias em geração de energia solar utilizando LSTM Autoencoder, a banca composta pelos professores:

**CARMELO JOSE ALBANEZ BASTOS FILHO (PRESIDENTE)**

**ALEXANDRE MAGNO ANDRADE MACIEL (ORIENTADOR)**

Após a apresentação da monografia e discussão entre os membros da Banca, a mesma foi considerada:

Aprovada       Aprovada com Restrições\*       Reprovada

e foi-lhe atribuída nota: **8** ( oito)

\*(Obrigatório o preenchimento do campo abaixo com comentários para o autor)

O(A) discente terá **7** dias para entrega da versão final da monografia a contar da data deste documento.



Documento assinado digitalmente  
Carmelo Jose Albanez Bastos Filho  
Data: 05/05/2022 08:19:08-0300  
Verifique em <https://verificador.itl.br>

---

AVALIADOR: Profº. Drº. Carmelo Jose Albanez Basto Filho



Documento assinado digitalmente  
ALEXANDRE MAGNO ANDRADE MACIEL  
Data: 05/05/2022 08:04:48-0300  
Verifique em <https://verificador.itl.br>

---

AVALIADOR: Profº. Drº. Alexandre Magno Andrade Maciel

\* Este documento deverá ser encadernado juntamente com a monografia em versão final.

## **AGRADECIMENTOS**

Agradeço primeiramente a Deus, pela minha vida, por me ajudar a vencer todos os obstáculos ao longo da graduação.

Aos meus pais e irmão, que me incentivaram e apoiaram nos momentos mais difíceis para realização deste trabalho.

Aos professores, pelos ensinamentos e avaliações que me permitiram apresentar um melhor desempenho na minha formação profissional.

E ao Instituto Edson Mororó Moura - ITEM, por disponibilizar os dados de energia solar utilizados neste trabalho.

## Autorização de publicação de PFC

Eu, **Ailson Ramon Costa Silva** autor(a) do projeto de final de curso intitulado: **Detecção de Anomalias em Geração de Energia Solar utilizando LSTM Autoencoder**; autorizo a publicação de seu conteúdo na internet nos portais da Escola Politécnica de Pernambuco e Universidade de Pernambuco.

O conteúdo do projeto de final de curso é de responsabilidade do autor.

Documento assinado digitalmente  
 AILSON RAMON COSTA SILVA  
Data: 27/05/2022 11:30:38-0300  
Verifique em <https://verificador.itl.br>

---

**Ailson Ramon Costa Silva**

Documento assinado digitalmente  
 ALEXANDRE MAGNO ANDRADE MACIEL  
Data: 27/05/2022 09:50:36-0300  
Verifique em <https://verificador.itl.br>

---

Orientador(a): **Alexandre Magno Andrade Maciel**

---

Coorientador(a):

---

Prof, de TCC: **Daniel Augusto Ribeiro Chaves**

---

Data: 27/5/2022

## RESUMO

O cenário mundial da sustentabilidade energética está testemunhando um alto crescimento de energia solar fotovoltaica(FV) sendo talvez a energia renovável mais promissora e preste a se tornar a maior fonte de eletricidade do mundo até 2050. Entretanto, diversos fatores anômalos influenciam negativamente sobre a eficiência dos painéis FV. Com isso, sistemas de detecção de falhas auxiliam na operação e manutenção de eficiência energética. Diante deste cenário, esta pesquisa destinou-se a propor um modelo de detecção de anomalias não-supervisionada utilizando LSTM e Autoencoder baseado no erro de reconstrução da previsão e o valor real medido. Os resultados revelaram que o modelo utilizando todos os dados com janela temporal de 32 para previsão de 1 hora a frente obteve o melhor resultado para encontrar falhas de operação no período.

Palavras-chaves: Detecção de anomalias; Geração fotovoltaica; Redes Neurais; LSTM; Autoencoder.

## **ABSTRACT**

The world scenario of energy sustainability is witnessing a high growth of solar photovoltaic (PV) being perhaps the most promising renewable energy and about to become the largest source of electricity in the world by 2050. However, several anomalous factors negatively influence efficiency of PV panels. With this, fault detection systems help in the operation and maintenance of energy efficiency. Given this scenario, this research aimed to propose an unsupervised anomaly detection model using LSTM and Autoencoder based on the forecast reconstruction error and the actual measured value. The results revealed that the model using all data with a time window of 32 for 1 hour ahead forecast obtained the best result to find operation failures in the period.

Keywords: Anomaly detection; Photovoltaic generation; Neural networks; LSTM; Autoencoder.

## LISTA DE FIGURAS

FIGURA 1 — Exemplo de um modelo de autoencoder genérico.....	14
FIGURA 2 — Formato dos dados de entrada para o modelo do tipo LSTM.....	16
FIGURA 3 — Passos para detecção de anomalia.....	21
FIGURA 4 — Distribuição de frequência da base de dados por completo.....	23
FIGURA 5 — Distribuição de frequência da base de dados sem os dados noturnos... .....	24
FIGURA 6 — Média horária da geração de energia solar ao longo do dia.....	24
FIGURA 7 — Média horária da geração de energia solar sem horário noturno.....	25
FIGURA 8 — Distribuição da quantidade de dados faltantes ao longo do dia.....	26
FIGURA 9 — Modelo de perda do processo de treinamento e validação.....	29
FIGURA 10 — Distribuição dos erros de treinamento.....	30
FIGURA 11 — Definição do limite de <i>threshold</i> .....	31
FIGURA 12 — <i>Threshold</i> para modelo sem dados noturno.....	34
FIGURA 13 — Comparação entre previsões e dados reais e detecção de anomalias.....	35
FIGURA 14 — Código em python do LSTM autoencoder.....	40

## LISTA DE TABELAS

TABELA 1 — Estatística básica para o conjunto de dados em média horárias.....	22
TABELA 2 — Ajustes nos hiperparâmetros do modelo.....	29
TABELA 3 — Avaliação da janela temporal.....	32
TABELA 4 — Resumo dos modelos testados .....	33

## LISTA DE ABREVIATURAS E SIGLAS

ABNT	<i>AntiBody NETwork</i>
LSTM	<i>Long Short Term Memory</i>
MLP	<i>Multilayer Perceptron</i>
PSOM	<i>Non Parametric Self-Organizing Map</i>

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO.....</b>	<b>11</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA.....</b>	<b>13</b>
2.1	Série temporal.....	13
2.2	Detecção de anomalia.....	13
2.3	Autoencoder(AE).....	13
2.4	Long Short Term Memory (LSTM).....	14
2.5	Autoencoder para detecção de anomalias.....	18
<b>3</b>	<b>TRABALHOS RELACIONADOS.....</b>	<b>19</b>
<b>4</b>	<b>METODOLOGIA.....</b>	<b>20</b>
4.1	Aquisição de dados.....	22
4.2	Análise dos dados.....	22
4.3	Pré-processamento.....	25
4.1	Definição do modelo.....	27
<b>5</b>	<b>AVALIAÇÃO E RESULTADO.....</b>	<b>28</b>
5.1	Treinamento.....	28
5.2	Threshold.....	30
5.3	Avaliação.....	31
5.4	Análise das anomalias.....	35
<b>6</b>	<b>CONCLUSÃO E TRABALHOS FUTUROS.....</b>	<b>37</b>
	<b>REFERÊNCIAS.....</b>	<b>38</b>
	<b>APÊNDICE A - CÓDIGO FONTE DO TRABALHO.....</b>	<b>40</b>

## 1 - INTRODUÇÃO

A energia solar e suas tecnologias mostraram um rápido crescimento nos últimos anos, hoje retrata uma tecnologia madura para produção de energia a partir de fontes renováveis [1]. Por consequência os sistemas fotovoltaicos têm um grande potencial para reduzir a dependência atual de fontes intensas de carbono, além disso sua eficiência energética tem aumentado cerca de 5% para mais de 40% nos últimos 60 anos [2]. Entretanto, os níveis de eficiência energética atuais são bastante baixos em comparação com as fontes alternativas de energia. Os sistemas fotovoltaicos(FV) devem operar perfeitamente sem anomalias para maximizar a eficiência e permanecer uma fonte de energia alternativa viável. No entanto, podem ocorrer vários tipos de anomalias que impedem os FV de operar em capacidade plena. Sendo assim, é de suma importância monitorar a atividade dos sistemas FV para que essas anomalias possam ser detectadas e reparadas para garantir a máxima eficiência [3-4].

Detecção de anomalias (DA) em sistemas FV pode ser realizada utilizando vários métodos, desde estatística clássica, mineração de dados e aprendizado de máquina[5-6]. Contudo, identificar anomalias não é uma tarefa simples, sendo conduzido a descoberta de dados extremos, esses dados têm diferentes características com dados normais. Quando esses dados são rotulados podemos utilizar uma abordagem de aprendizagem supervisionada, onde os dados rotulados são necessários para processo de treinamento para criar um modelo de detecção de anomalias [7]. Entretanto, os dados de sensores geralmente são dados não rotulados, com isso, dificulta uma abordagem supervisionada em modelos de detecção de anomalias. Além disso, o problema de previsão de sequência é desafiador, principalmente porque os algoritmos de aprendizado de máquinas e redes neurais em particular são projetados para funcionar com entradas de comprimento fixo e muitos problemas de modelagem preditiva envolvendo sequência e requerem uma previsão que também é uma sequência, sendo chamados de problemas de previsão sequência-a-sequência[34].

Desta maneira, a predição de sequências impõe uma ordem nas observações que devem ser preservadas ao treinar o modelo e fazer previsões. Com isso, redes neurais recorrentes como a rede *Long Short-Term Memory* ou LSTM são projetadas especificamente para suportar sequências de dados de entrada. Além disso, elas são capazes de aprender a dinâmica complexa dentro da ordenação temporal das sequências de entrada, bem como usar uma memória interna para lembrar ou usar informações em longas sequências de entrada[35]. Ademais, o *Autoencoder* é um modelo de rede neural que busca aprender uma representação compacta de uma entrada, utilizando método de aprendizado não supervisionado. Neste contexto, este trabalho busca apresentar uma arquitetura de rede neural utilizando LSTM junto com *Autoencoder* para um aprendizado não supervisionado para identificar padrões de comportamento e identificar anomalias em dados de séries temporais de geração de potência ativa em sistemas fotovoltaicos. O restante do trabalho está organizado da seguinte forma: fundamentação teórica no capítulo 2, trabalhos relacionados estão descritos no capítulo 3, metodologia no capítulo 4 e conclusão dada no capítulo 5.

## 2 - FUNDAMENTAÇÃO TEÓRICA

Este capítulo está dividido nos temas necessários para elaboração da metodologia do trabalho, onde está organizado; definição de série temporal, anomalia, autoencoder e LSTM.

### 2.1 - Série temporal

Uma série temporal pode ser definida como uma sequência de observações de uma variável ao longo do tempo [11]. Assim, uma série temporal pode ser entendida como uma sequência ordenada de pontos que ocorrem em intervalos de tempo igualmente espaçados. Análise de séries temporais tem sido utilizada em várias aplicações tais como: estudos de utilidade pública, análise de sensores, previsões econômicas, previsões de venda, análise do mercado de ações, detecção de anomalias e assim por diante. Normalmente, uma série temporal é representada por uma sequência de um vetor de observações d-dimensional ordenado:

$$X_t = X_{t1}, X_{t2}, X_{t3}, \dots, X_{td}$$

### 2.2 - Detecção de anomalia

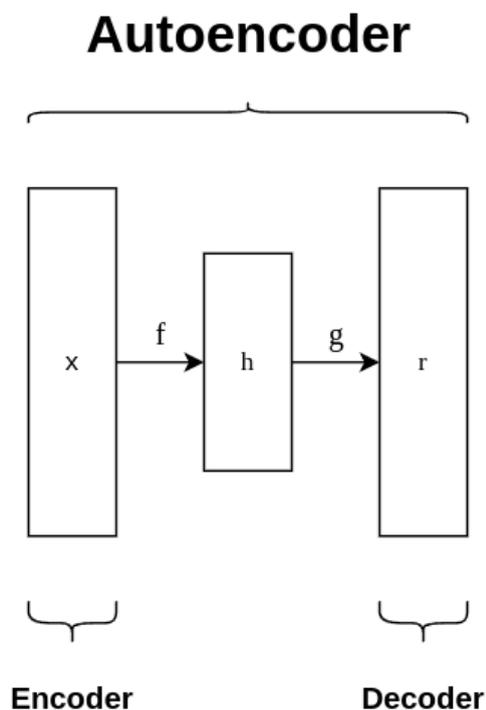
O conceito de Anomalia ou *outlier* de acordo com Hawkins[12] uma anomalia (outlier) é uma observação que se desvia muito em relação a outras observações que levam a suspeitas de serem geradas por algum mecanismo diferente. Outra definição em Chandola[5] aponta que as anomalias são padrões em dados que não estão em conformidade com uma noção bem definida de comportamento normal.

### 2.3 - Autoencoder(AE)

Um AutoEncoder tem sido tradicionalmente usado para redução de dimensionalidade e aprendizado de recursos[14]. A ideia fundamental do Autoencoder é a recuperação da informação que este é composto por uma camada oculta  $h$  que se refere à entrada e duas partes principais, ou seja, função codificadora  $h = f(x)$  e decodificador, também conhecido como reconstrução:  $r = g(h)$ . O principal objetivo deste conceito é que tanto o codificador quanto o

decodificador sejam treinados juntos, e a discrepância entre os dados originais e sua reconstrução possa ser minimizada. A estrutura geral do Autoencoder é representada na FIGURA 1.

FIGURA 1. Exemplo de um modelo de Autoencoder genérico.



Fonte: Compilação do autor.

O procedimento de treinamento dos Autoencoders não é supervisionado[27] e consiste em encontrar os parâmetros que fazem a reconstrução  $r$  o mais próximo possível da entrada original  $x$ , minimizando uma função de perda que mede a qualidade das reconstruções (por exemplo, erro quadrático médio).

#### 2.4 - Long Short Term Memory (LSTM)

Hoje em dia, modelos de aprendizado de máquina surgiram como o estado da arte para realizar previsões solares de uma para algumas horas de previsões[16]. Atualmente, muitos estudos relatam um desempenho superior exibido por modelos de aprendizado profundo (*Deep learning*) comparado com modelos de aprendizado de máquina (*Machine learning*) relacionado a problemas de classificação, regressão e previsão de séries temporais[17]. De acordo com observações de LenCun [18], modelos de redes neurais estavam se saindo melhor que modelos de aprendizado

de máquina aplicado em muitos domínios devido a sua capacidade superior de aprender padrões complexos extraindo dados brutos. LSTM é um modelo de aprendizado profundo projetado para lidar com dados em sequência. Sendo uma das vantagens do LSTM é que ele pode lidar bem com dados não lineares[15] e podem memorizar relacionamentos temporais longos em dados . Ao longo dos anos, modelos de LSTM têm mostrado alta eficiência em diversos domínios de aplicação como; modelos de linguagem[19-20], reconhecimento de voz [21], previsão de tempo[22], detecção de anomalias[23], etc.

As equações para o LSTM podem ser vista da seguinte forma:

$$f_t = \sigma_g(W_f \times x_t + U_f \times h_{t-1} + b_f) \quad (6)$$

$$i_t = \sigma_g(W_i \times x_t + U_i \times h_{t-1} + b_i) \quad (7)$$

$$o_t = \sigma_g(W_o \times x_t + U_o \times h_{t-1} + b_o) \quad (8)$$

$$c'_t = \sigma_g(W_c \times x_t + U_v \times h_{t-1} + b_c) \quad (9)$$

$$c_t = f_t \cdot c_{t-1} + i_{t-1} \cdot c'_t \quad (10)$$

$$h_t = o_t \cdot \sigma_g(c_t) \quad (11)$$

Onde,

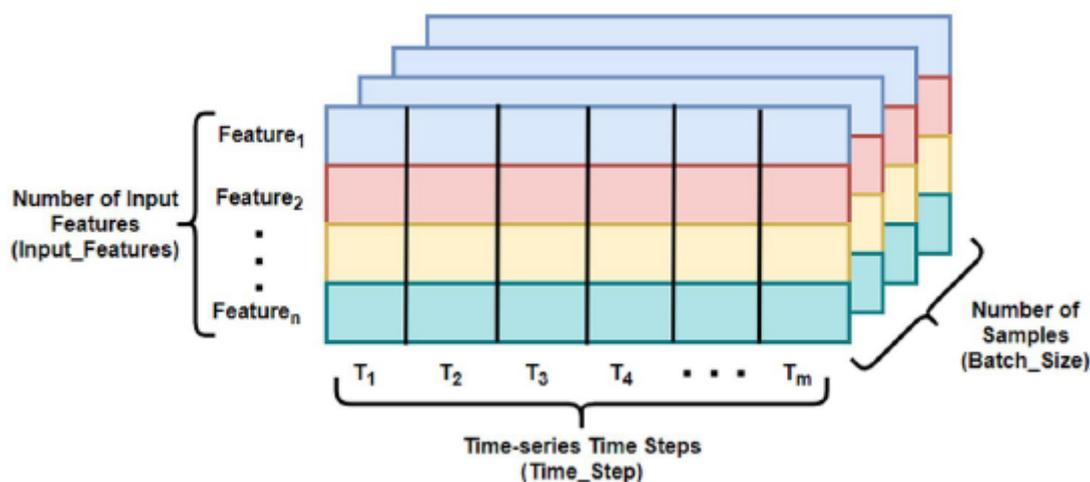
- $f_t$  é o *Forget gate*, ou seja canal do esquecimento, dependendo da entrada atual  $x_t$  e a saída da camada oculta anterior  $h_{t-1}$ , baseado em uma camada sigmóide, o canal do esquecimento produz 0 ou 1. Se 1, as informações são retidas na memória, caso contrário é descartada.
- O  $i_t$  é o canal de entrada, no qual ajuda a decidir se a novas informações a serem adicionadas ao estado atual da célula com base em novos valores candidatos fornecidos por  $c'_t$ .
- $c_t$  é o estado da célula, esse novo estado depende do estado da célula anterior  $c_{t-1}$  e  $c_{t-1} \cdot f_t$  é a fração do estado da célula antiga que será descartada com a ajuda do *Forget gate*, enquanto novas informações serão adicionadas através de  $c'_t \cdot i_t$ . A soma dessas duas atualizações simultâneas é o estado atual da célula.

- $o_t$  é o canal de saída do estado de longo prazo, sendo determinado por uma função de ativação  $\tanh$ .
- $h_t$  é o estado oculto, por fim o resultado é multiplicado com o estado da célula através da  $\tanh$  para calcular o valor do estado oculto atual.

E  $W_f$ ,  $W_i$ ,  $W_c$  e  $W_o$  são as matrizes de pesos e  $b_f$ ,  $b_i$ ,  $b_c$  e  $b_o$  é o viés dado para cada canal.  $\sigma_g$  é indicado para a função de ativação  $\tanh$ . O elemento  $\cdot$  é a multiplicação e  $+$  implica na adição dos elementos.

Para LSTM, a preparação dos dados é tratada de forma diferente dos modelos tradicionais de aprendizado de máquina. O dado precisa estar formatado em um *array* de 3 dimensões, onde 1 representa o tamanho do *batch*, 2 número de *time-steps* (Janela temporal) e 3 o número de características de entradas. Exemplificado melhor na FIGURA 2 a seguir:

FIGURA 2. Formato dos dados de entrada para o modelo do tipo LSTM



Fonte: Malakar, S., Goswami, S., Ganguli, B [24]

Ao utilizar o modelo LSTM em uma série temporal, devemos atentar em duas apresentações para o problema:

- 1 Supervisionada: onde os passos de tempo anteriores são considerados independentes uns dos outros e são tratados como características separadas[24].
- 2 Não-supervisionado: no qual a ordenação e os passos de tempo são dados importantes que influenciam no comportamento do modelo.

No artigo [25] e [26], Gensler e Yagli trataram os recursos de entrada como independentes do tempo, desenvolvendo um modelo de previsão. Como o modelo utilizado, leva em consideração a apresentação 2, no modo não-supervisionado, precisaremos realizar alguns ajustes na entrada da série temporal para o modelo LSTM.

Suponha que tenhamos uma série temporal de comprimento  $n$  dado por:

$$X_1, X_2, X_3, \dots, X_n. \quad (1)$$

Para aplicação de dados de entrada para um modelo *LSMT*, a série é convertida na seguinte representação:

$$1 \quad \text{Passo} \quad \{[X_1, X_2, X_3][X_4]\} \quad (2)$$

$$2 \quad \text{Passo} \quad \{[X_2, X_3, X_4][X_5]\} \quad (3)$$

$$3 \quad \text{Passo} \quad \{[X_{n-3}, X_{n-2}, X_{n-1}][X_n]\} \quad (4)$$

Assumindo que a janela temporal tenha tamanho 3, essas três primeiras amostras se tornam o meu vetor de entrada e a quarta amostra é a minha saída, demonstrado no passo 1. No passo 2 segue a mesma abordagem e até chegar na  $n$ ésima posição da série temporal a ser previsto, no passo 3. Ou seja, as amostras são separadas por vírgulas e o conjunto de entrada  $\{[X_1, X_2, X_3]\}$  e a saída prevista  $\{[X_4]\}$  estão separadas entre chaves [24].

Em uma configuração não-supervisionada de LSTM precisamos nos atentar em 3 parâmetros:

a) Número de passos no tempo; define o intervalo de tempos passados a ser utilizado no modelo.

b) Números de recursos de entrada; informa quantas variáveis de entrada iremos utilizar na previsão.

c) Tamanho de lote: é um hiperparâmetro que define o número de amostras a serem trabalhadas antes de atualizar os parâmetros do modelo interno. Geralmente, os modelos são mais eficientes ao encontrar o gradiente em lotes menores, em vez de utilizar toda a base de dados.

## 2.5 - Autoencoder para detecção de anomalias

A principal abordagem para detecção de anomalias baseadas em Autoencoder é focar no seu comportamento “normal” ao invés de modelar o que é anômalo[27].

O Autoencoder é treinado para reconstruir dados com padrão normal, minimizando uma função de perda que mede a qualidade das reconstruções. Logo após seu treinamento, o modelo é capaz de reconstruir os dados da série temporal padrão normal, enquanto a falha na reconstrução de dados anômalos, uma vez que nunca o viu durante o treinamento, é feita a detecção usando as métricas de reconstrução definido um *threshold* para anomalias. Ou seja, caso o erro da reconstrução seja maior que o limite estipulado, ele informa que essas observações são anômalas.

---

### 3 - TRABALHOS RELACIONADOS

Um trabalho recente sobre detecção de anomalias em painéis FV utiliza redes ABNET, PSOM e MLP foi elaborado por Fonseca [8], neste trabalho ele utiliza imagens termográficas para reconhecimento de padrões de superaquecimento dos painéis fotovoltaicos. Por meio de validação cruzada a taxa de acertos foi de 87,5% para ABNET e 96% para PSOM e MLP. Todavia, essa abordagem precisa de um instrumental bem específico para colher os dados termográficos dos painéis FV inviabilizando a utilização em sistemas mais simples sem esses recursos disponíveis.

Além disso, no trabalho Platon [4] desenvolve um algoritmo para DA em sistemas FV que primeiro modela a produção de energia CA utilizando irradiância solar e dados de temperatura do painel FV. Em seguida, o algoritmo executa a DA com base na comparação temporal entre a produção de energia CA observada e modelada e assim, atinge taxas de detecção superiores a 90%. É de suma importância notar, no entanto, que tais algoritmos são consideravelmente mais complexos e requerem dados de ambientes para operar. Todavia, os algoritmos desenvolvidos neste estudo têm a vantagem de serem mais amplamente aplicáveis [9].

Os trabalhos sobre detecção de anomalias em dados de séries temporais têm aumentado significativamente nos últimos anos, em particular utilizando redes neurais profundas. Neste caso o modelo Seq2Seq e Autoencoder foram aplicados com sucesso em tarefas DA de serie temporal. Com esse *framework* Malhotra[10] propuseram um modelo baseado em previsão utilizando *Long Short Term Memory (LSTM)* e usou a distribuição dos erros de previsão para calcular pontuações de anomalias.

Esses trabalhos relacionados auxiliaram na pesquisa ao utilizar uma abordagem de previsão, sem muito custo ou aparato tecnológico, como imagens termográficas para utilizar na detecção de anomalias em dados de geração de energia solar e sanar alguns problemas de arquitetura de LSTM.

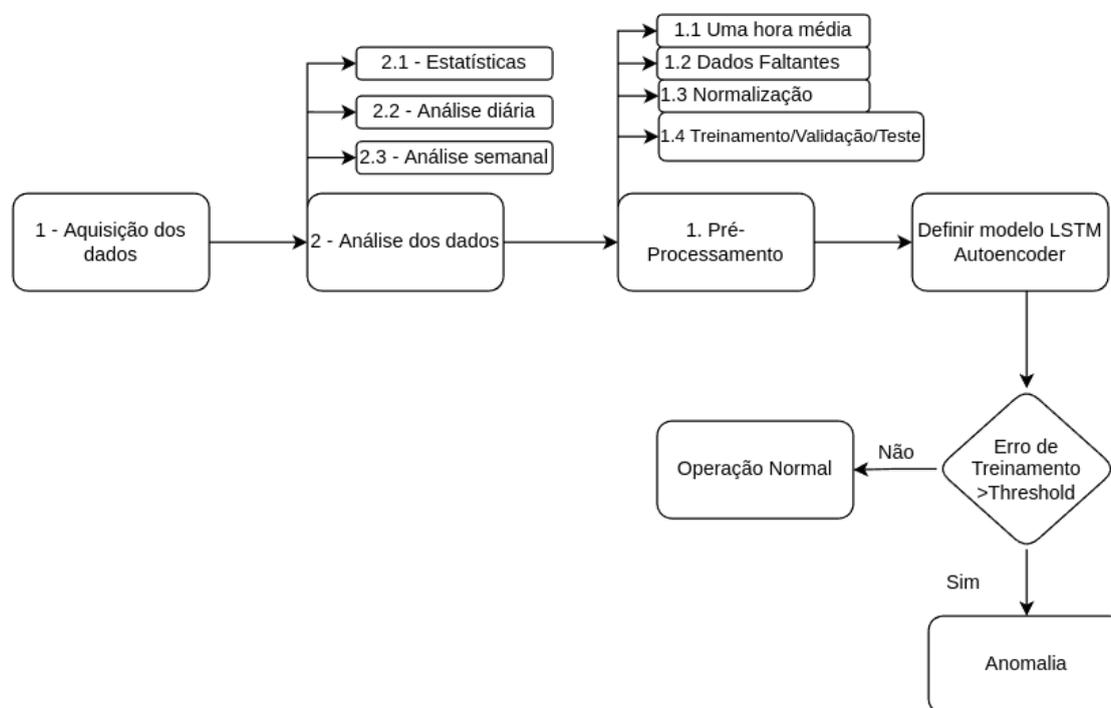
#### 4 - METODOLOGIA

A rede LSTM tem um problema relacionado à arquitetura, pois ainda não existe um acordo geral entre os modelos a serem escolhidos, por exemplo: como preservar a ordem temporal dos dados, a necessidade de remoção dos dados noturnos[24], visto que eles representam um volume muito alto no conjunto dos dados. Além disso, problemas relacionados ao tamanho de lote, um dos hiperparâmetros mais importantes para sintonizar o modelo de aprendizagem, variando entre lotes menores e maiores para achar uma melhor convergência, qual seria o melhor horizonte de previsão etc. Por causa disso, este trabalho considerou-se duas hipóteses para sanar problemas de arquitetura da rede LSTM juntamente com Autoencoder na previsão e detecção de anomalia em geração de energia solar, as hipóteses estão listadas abaixo:

- Hipótese 1: De que forma o tamanho da janela temporal escolhida influencia na performance do modelo para previsão de 1 hora a frente, 1 hora, 1 dia, 1 semana. Além disso, se o formato da janela temporal estiver em representação binária, exemplo, 2, 4, 6, 16, 32 o modelo terá melhor desempenho?
- Hipótese 2: Qual é o desempenho das previsões de anomalias caso os dados das horas noturnas sejam removidos?

Portanto, neste trabalho realizamos a implementação de um modelo de detecção de anomalias utilizando *Long Short Term Memory Autoencoder*. Com a representação da metodologia utilizada na FIGURA 3, a seguir.

FIGURA 3. Passos para detecção de anomalias



Fonte: Compilação do autor.

A seção está separada na seguinte forma a seguir; (1) Aquisição dos dados para o modelo, (2) Análise de dados para investigar os comportamento diário e semanal, (3) aplicação do pré-processamento para normalização e transformação dos dados para o modelo LSTM, (4) os hiperparâmetros utilizados na criação do modelo e (5) os resultados obtidos.

#### 4.1 - Aquisição de dados

Os dados históricos foram coletados de uma microrrede, localizada na cidade de Belo Jardim – PE. A microrrede apresenta uma unidade fotovoltaica (UFV) com 820 painéis de 380 W, totalizando 311,6 kWp divididos em quarenta e uma fileiras. Além da UFV, a microrrede possui um sistema de armazenamento de energia de 250 kW e 560 kWh com tecnologia de armazenamento em baterias do tipo chumbo carbono.

No período de coleta dos dados, entre os meses de janeiro a junho do ano de 2021, o Sistema de Armazenamento de Energia em Baterias (SAEB) e a unidade fotovoltaica (UFV) eram acoplados no barramento de corrente alternada, caracterizando a instalação como uma microrrede CA. Nessa configuração, as 41 fileiras fotovoltaicas se conectam a cinco inversores fotovoltaicos, dois deles com potência nominal de 36 kW e três deles com potência nominal de 60 kW. De forma geral, o registro do intervalo de medição das amostras é, em média, a cada 1 minuto e 10 segundos.

#### 4.2 - Análise dos dados

Esta pesquisa utiliza os dados de Potência Ativa gerada pelos painéis fotovoltaicos no período de 2021-01-01 até 2021-07-06. As amostras de 1 minuto e 10 segundos foram calculadas em média ao longo de 1 hora, no trabalho de Platon[4] foi demonstrado que a variabilidade e previsão do modelo tem aumento significativo. Os modelos utilizados neste trabalho também observou-se esse comportamento.

Além disso, foram extraídos algumas estatísticas da base de dados para os dados completos e sem os horários noturnos.

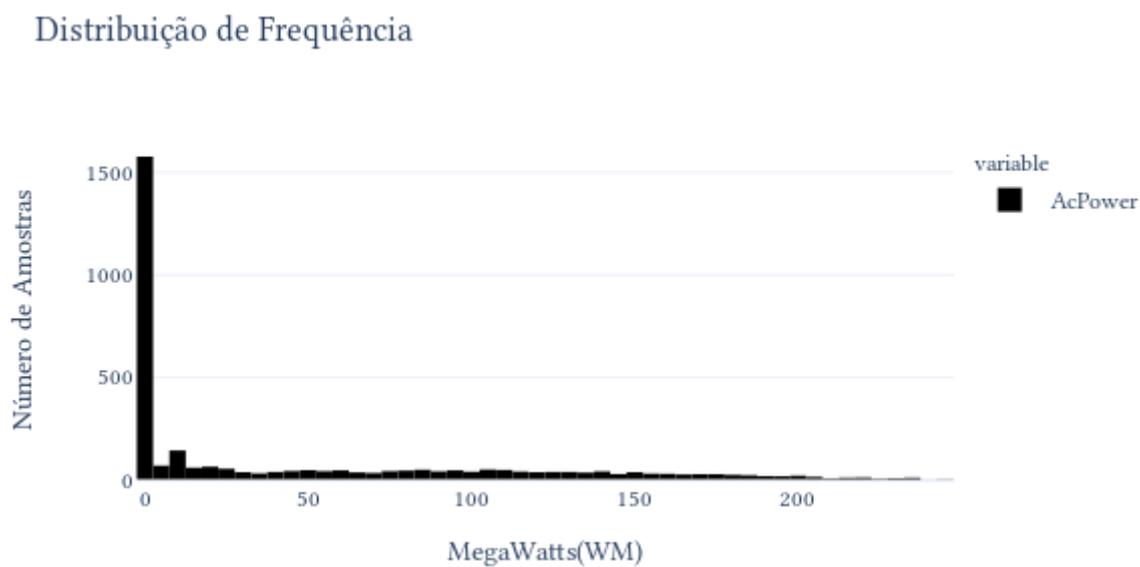
Tabela 1 - Estatística básica para o conjunto de dados em médias horárias

Potência Ativa(MW)	amostras	Média	Mínimo	25%	50%	75%	Máximo
Completo	4355	41,928	0	0	2,004	79,28	243,478
Sem noite	1873	95,941	15,055	53,334	91,852	132,619	243,478

Fonte: Compilação do autor.

Observou-se com a distribuição de frequência um alto volume de dados baixos próximos a 0 (zero) na base de dados, visto na FIGURA 4:

FIGURA 4. Distribuição de frequência da base de dados por completo

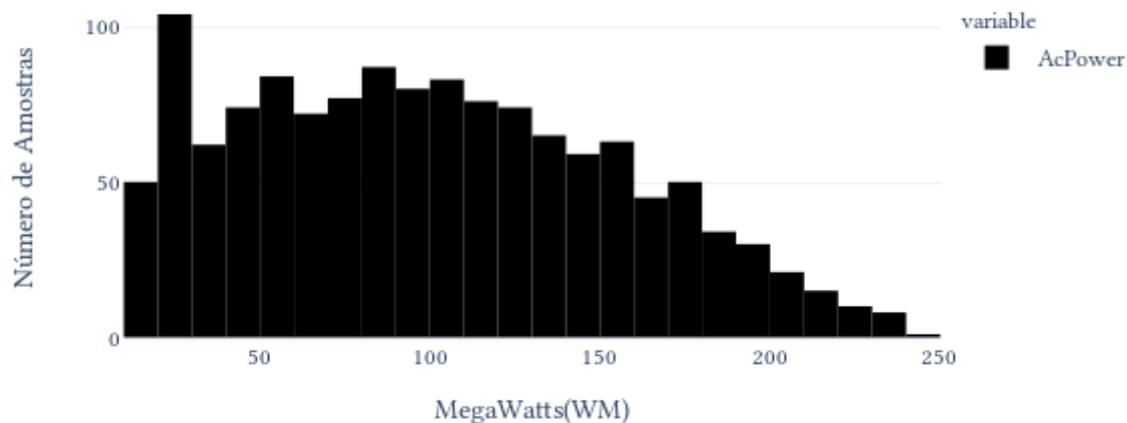


Fonte: Compilação do autor.

Com a remoção dos dados noturno[28] conseguiu-se ter uma visão menos poluída da distribuição dos dados, sendo observado na FIGURA 5.

FIGURA 5. Distribuição de frequência da base de dados sem os dados noturnos.

### Distribuição de frequência sem dados noturno

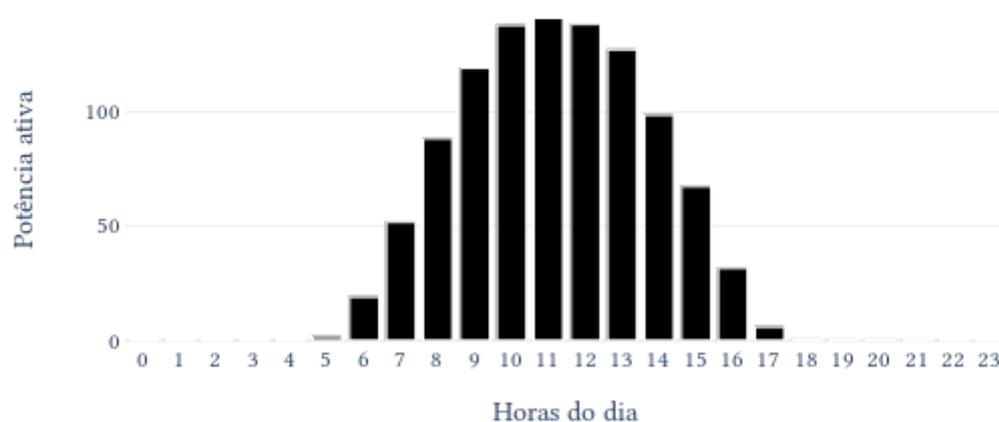


Fonte: Compilação do autor.

Realizou-se a mesma abordagem para as médias dos valores da base de dados para obtermos uma visão de seu comportamento ao longo de 1 dia com os dados completos e sem os dados noturno, visto respectivamente nas FIGURA 6 e 7.

FIGURA 6. Média horária da geração de energia solar ao longo do dia

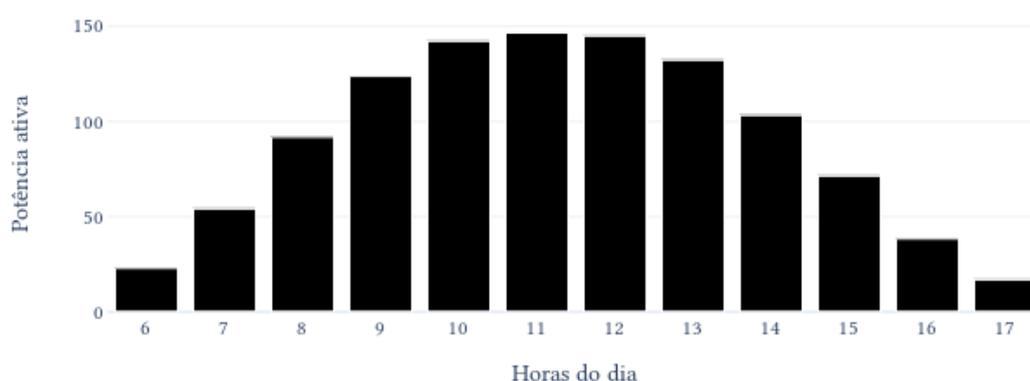
### Média horária de geração de energia solar ao longo do dia



Fonte: Compilação do autor.

FIGURA 7. Média horária da geração de energia solar sem horário noturno

Média horária de geração de energia solar ao longo do dia sem dados noturnos



Fonte: Compilação do autor.

A partir desta análise conseguiu-se observar que os horários de maior geração de potência ativa estão no intervalo de 10-12 horas. E também que o horário noturno acumula maior volume de amostras com dados próximos de zero.

#### 4.3 - Pré-processamento

O próximo passo é o pré-processamento do dado, neste processo realizou-se a remoção dos dados faltantes, no primeiro momento foi realizado uma visualização da distribuição dos dados faltantes e observado que a probabilidade em quase todos os dados dados faltantes são similares a inferência *missing completely at random* (MCAR) [29], reparado na FIGURA 8.

FIGURA 8. Distribuição da quantidade de dados faltantes ao longo do dia.

## Dados faltantes por hora



Fonte: Compilação do autor.

Com a análise feita, realizou a remoção dos dados faltantes e aplicou a técnica de normalização para os dados, isso visando uma melhor performance na rede utilizada[30]. A função de normalização utilizada foi a *Standard Scaler* que modifica a média para 0 e o desvio padrão para 1, calculando com a seguinte equação:

$$Z = (x - u) / \sigma \quad (5)$$

Onde  $u$  é a média das amostras de treinamento e  $\sigma$  é o desvio padrão das amostras. Isso é feito para que os dados possam ser implementados corretamente no modelo usado. Além disso, os dados constam 4355 amostras e foram divididos em 3035(70%) treinamento usado para criar o modelo, 671 (15%) validação onde podemos obter uma estimativa inicial da habilidade do modelo e ajuste dos parâmetros e por fim (649) 14,9% teste usado para avaliar o desempenho do modelo[32].

#### 4.1 - Definição do modelo

Neste trabalho o LSTM representa o conceito principal do modelo proposto. Isso acontece, porque o LSTM mostra-se a capacidade para superar dependências de longo prazo mais facilmente que uma simples arquitetura interativa[13].

Na arquitetura de rede do Autoencoder LSTM, o primeiro par de camadas da rede neural é criar a representação dos dados de entrada compactados, codificador. Em seguida, usamos uma camada de vetor de repetição para distribuir o vetor representacional comprimido ao longo das etapas de tempo do decodificador. A camada de saída final do decodificador nos fornece os dados de entrada reconstruídos.

Em seguida, instanciou-se o modelo e compilou-se usando o otimizador Adam e o erro absoluto médio para calcular a função de perda, sendo expressada abaixo:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - y'_i| \quad (12)$$

Onde,  $n$  é o número de observações,  $y_i$  é o valor real e  $y'_i$  é a previsão do modelo.

## 5 - AVALIAÇÃO E RESULTADO

Neste capítulo apresenta-se os resultados obtidos para elucidar as hipóteses da pesquisa. Está dividido em (1) treinamento e avaliação do modelo, (2) definição do *threshold* e distribuição do erro e (3) visualização das anomalias.

### 5.1- Treinamento

Antes de computar os modelos preditivos, os dados são separados em conjuntos de treinamento, validação e teste. O conjunto de treinamento é usado para desenvolver o modelo e definir o *threshold*, quanto o conjunto de validação é utilizado para ajustar os parâmetros do classificador, juntamente com o conjunto de teste para avaliar o desempenho do modelo[31].

Realizamos diferentes experimentos para obter os melhores valores de hiperparâmetros para a iniciação do modelo. Durante esses experimentos, alteramos o valor da taxa de aprendizado, função de ativação entre Relu, sigmoid e tanh, tamanho de lote e épocas para fornecerem uma maior taxa de precisão.

Nesta pesquisa o melhor resultado obtido para horizonte temporal de previsão de 1 hora utilizou-se tamanho de lote com 32, taxa de aprendizado para 0.0001, 50 épocas, junto a função Adam de otimizador e a camada de ativação com a função *tanh*. Utilizamos 4 camadas LSTM com tamanho de 32 e 16 cada. Em seguida, utilizou-se a camada distribuída no tempo e a repetição de vetores para filtrar os valores compactados para a camada de decodificação. Aplicou-se a função de perda utilizando o erro médio absoluto (MAE) para calcular o valor de perda do modelo, visto na TABELA 2.

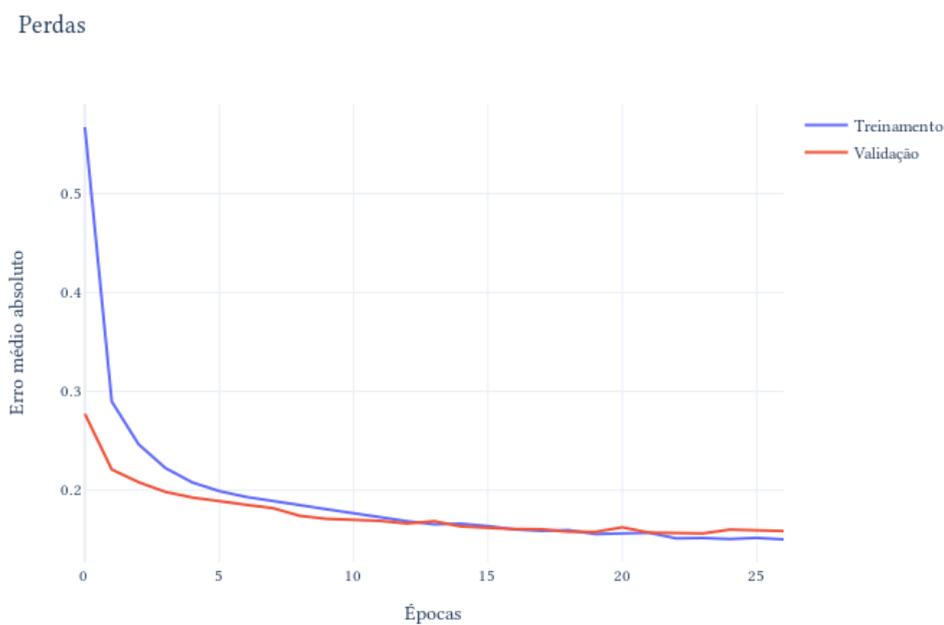
TABELA 2. Ajustes nos hiperparâmetros do modelo

Parâmetros	Melhores Valores
LSTM	4
Tamanho da camada	32, 16
Otimizador	Adam
Ativação	tanh
Taxa de aprendizagem	0,001
Épocas	50
Tamanho do lote	32
Perda	MAE

Fonte: Compilação do autor.

Baseado nos parâmetros de ajuste para estimativa do modelo proposto, neste trabalho obteve-se uma boa curva de treinamento e validação no modelo de perda após 25 épocas conforme mostrado na FIGURA 9 .

FIGURA 9. Modelo de perda do processo de treinamento e validação

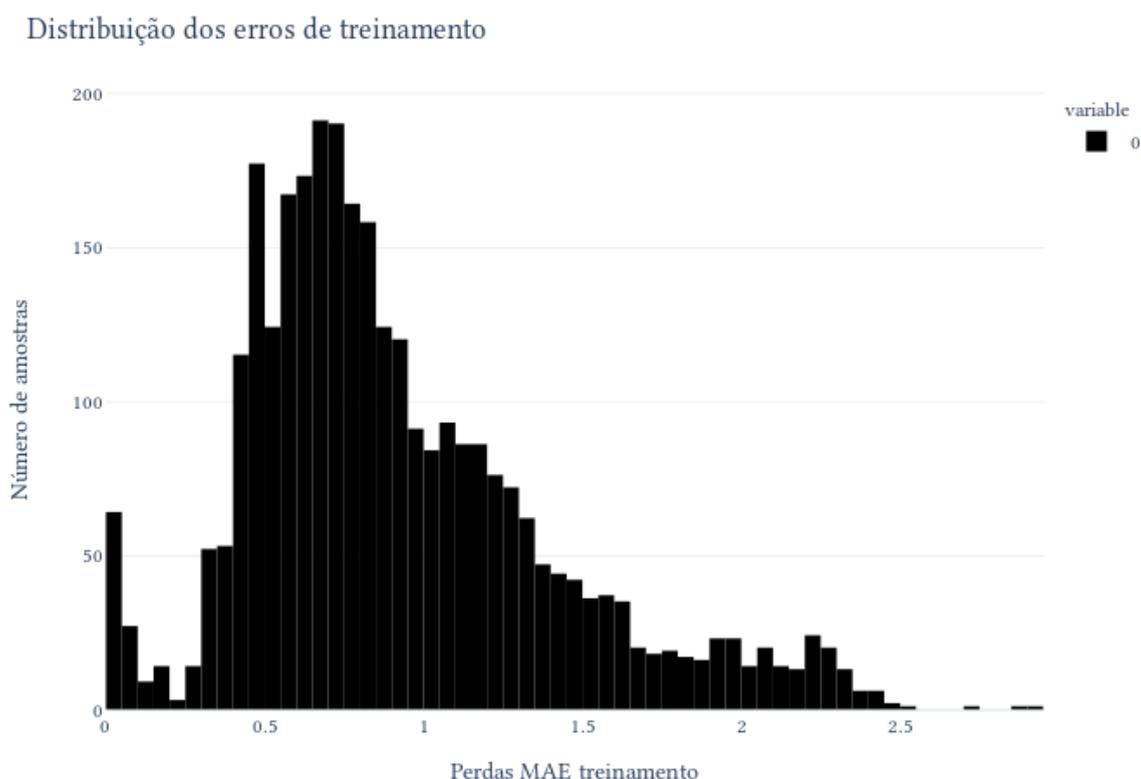


Fonte: Compilação do autor.

## 5.2 - Threshold

Ao traçar a distribuição da perda calculada no conjunto de treinamento, pode-se determinar um valor limite adequado para identificar uma anomalia. O gráfico da distribuição dos erros absolutos médios pode ser observado na FIGURA 10.

FIGURA 10. Distribuição dos erros de treinamento

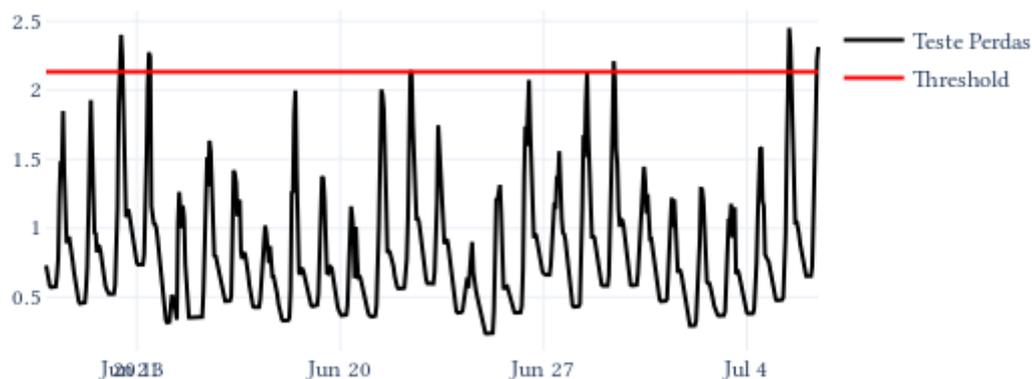


Fonte: Compilação do autor.

Com base na distribuição de perdas acima, seguimos a conhecida técnica em levar em conta a distância entre os dados observados e a média dos valores previstos[33]. De acordo com a nossa análise, observamos que a média dos erros mais 2,5 vezes o desvio padrão seria o limite definido para separar dados normais e anormal. Como pode-se observar na FIGURA 11.

FIGURA 11. Definição do limite de *threshold*

## Perdas em dados de testes e definição de Threshold: 2.13352



Fonte: Compilação do autor.

### 5.3 - Avaliação

Nesta pesquisa para responder às questões da hipótese 1, no qual realizou-se os testes para janela temporal de 1 hora, 1 dia e 1 semana a frente, junto com os ajustes nos parâmetros e suas respectivas representações binárias. Os resultados dos experimentos obtidos mostraram que a representação binária mais próxima de 1 dia ou 24 horas, para 32 horas, se mostrou com a menor taxa de erro para o horizonte temporal de previsão de 1 hora, conforme é visto na TABELA 3, abaixo.

TABELA 3 - Avaliação da janela temporal

Passo de tempo	MAE
<b>32</b>	<b>0,12818</b>
24	0,13884
168	0,14068
256	0,14481
2	0,18484
1	0,60952

Fonte: Compilação do autor.

Como o resultado do modelo é um comportamento estocástico, a TABELA 3 é o resultado da média de 10 treinamentos para cada passo de tempo. Resultando na representação binária de 1 dia, ou seja, 32 passos de tempo conquistando o menor erro médio absoluto.

Além disso, para avaliar a segunda hipótese, realizou-se a remoção dos dados noturno[28] e comparamos o erro absoluto médio com o modelo contendo os dados completos. Observou-se que o modelo obteve melhor desempenho ao utilizar os dados completos, em relação aos dados sem os horários noturno, atrelado ao tamanho da janela de 32 horas atrás conseguimos . Sumarizado na TABELA 4.

TABELA 4 - Resumo dos modelos testados

Janela temporal (horas)	Estatística dos erros dos modelos			Anomalias Detectadas		
	Média	Desvio Padrão	<i>Threshold</i>	Validação	Teste	MAE
1	0,9880	0,7113	0,3182	2	0	0,6095
2	0,2876	0,3620	0,1192	3	4	0,1848
24	0,9495	0,4816	0,2153	2	9	0,1388
<b>32</b>	<b>0,9197</b>	<b>0,4917</b>	<b>0,2149</b>	<b>2</b>	<b>11</b>	<b>0,1281</b>
168	0,1197	0,4603	0,2158	1	9	0,1406
256	0,1193	0,4500	0,2128	7	7	0,1448
1-N	0,8349	0,5516	0,2213	0	0	0,7441
2-N	0,6144	0,3318	0,1444	0	0	0,4255
24-N	0,9914	0,3940	0,1976	1	1	0,3683
32-N	0,9856	0,3210	0,1788	0	0	0,3797
168-N	0,1055	0,2818	0,1760	0	3	0,3618
256-N	0,11067	0,3087	0,1878	0	0	0,2607

Fonte: Compilação do autor.

Percebemos que quanto maior é o erro médio absoluto, menor é a capacidade do modelo em reconhecer as anomalias, pois o limite de *threshold* definido não é alcançado pelos modelos de maior erro.

Ademais, identificou-se que na medida que o erro absoluto médio aumenta no modelo, menores são as chances da abordagem de escolha do *threshold* identificar anomalias, sendo visto na FIGURA 12

FIGURA 12. *Threshold* para modelo sem dados noturno.

Perdas em dados de teste sem dados noturnos e definição de Threshold: 1.8465

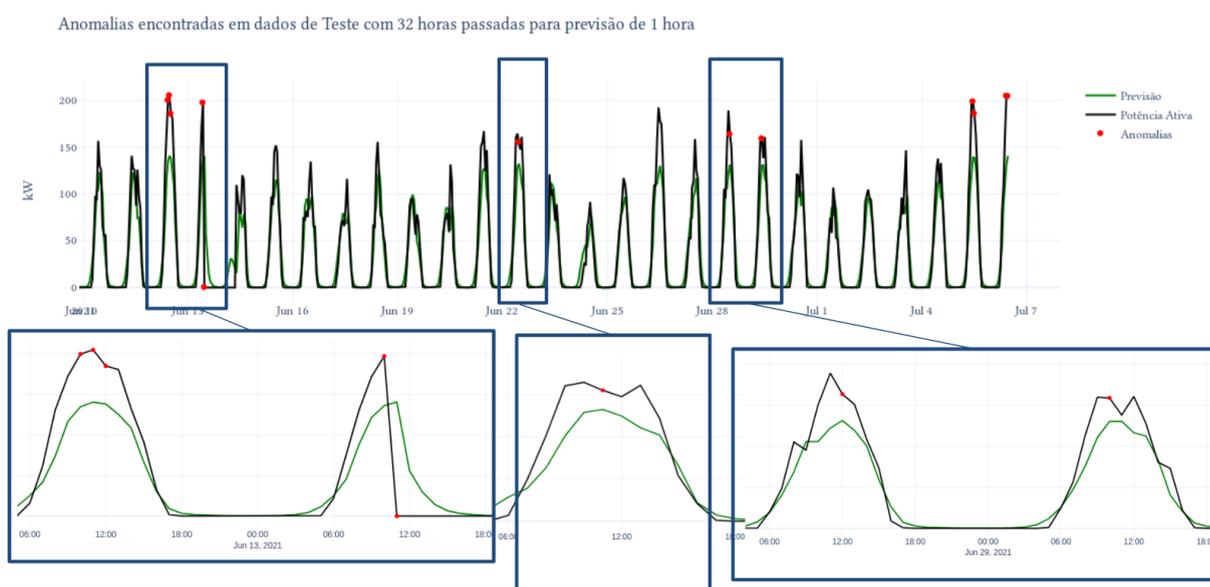


Fonte: Compilação do autor.

#### 5.4 - Análise das anomalias

Os sistemas fotovoltaicos podem ter muitos tipos de anomalias[6], causados por efeitos de fatores internos e externos, as anomalias encontradas nos dados de teste podem ser vistas na FIGURA 13.

FIGURA 13. Comparação entre previsões e dados reais e detecção de anomalias.



Fonte: Compilação do autor.

O número total de anomalias encontradas no sinal foram 11, sendo distribuída no período de 08-06-2021 até 06-07-2021. Na primeira imagem destacada, pode-se notar que o modelo identificou anomalia de contexto, visto que no horário de pico houve uma queda na geração de potência ativa, no dia seguinte, o modelo foi capaz de identificar o momento de desligamento da geração de potência. Na segunda imagem destacada, observamos novamente uma anomalia de contexto, com quedas na produção em horário de picos. Novamente, na terceira imagem destacada observados mais falhas de contextos

com quedas de produção de potência ativa nos horários de 11:00 do dia 29/06/21.

## **6 - CONCLUSÃO E TRABALHOS FUTUROS**

Neste trabalho realizou-se a análise no desenvolvimento de um modelo para detecção de anomalias de potência ativa na geração de energia fotovoltaica. Levantou-se duas hipóteses de arquitetura no seu desenvolvimento para verificar qual seria o melhor modelo não-supervisionado do LSTM Autoencoder. Identificou-se que para o horizonte temporal de previsão de 1 hora a frente, o modelo tem seu melhor desempenho para uma janela temporal de 32 horas passadas, ou seja, 1 dia em seu binário mais próximo, em conjunto com todos os seus dados, sem a remoção dos horários noturno, no qual não há geração de energia.

Em desenvolvimento e pesquisas futuras, o modelo pode ser testado em dados supervisionados para verificar a quantidade de falsos-positivos gerados.

Além disso, utiliza-se em modelos mais complexos com múltiplas variáveis tais como, tensão, corrente, irradiação solar e temperaturas dos painéis. Afim de se obter um grau maior de confiabilidade.

## REFERÊNCIAS

- [1] Natural Resources Canada. About Renewable Energy (2014) [Online]. Available: <http://www.nrcan.gc.ca/energy/renewable-electricity/7295#solar>
- [2] Dimroth, F.; Grave, M.; Beutel, P.; Fiedeler, U.; Karcher, C.; Tibbits, T.N.D.; Oliva, E.; Siefer, G.; Schachtner, M.; Wekkeli, A.; et al. Waferbonded four-junction GaInP/GaAs//GaInAsP/GaInAs concentrator solar cells with 44.7% efficiency. *Prog. Photovolt.* 2014, 458, 277–282. [[CrossRef](#)]
- [3] Chouder, A.; Silvestre, S. Automatic supervision and fault detection of PV systems based on power losses analysis. *Energy Convers. Manag.* 2010, 15, 1929–1937.
- [4] Platon, R.; Martel, J.; Woodruff, N.; Chau, T.Y.: Online fault detection in PV systems. *IEEE Trans. Sustain. Energy* 2015, 6, 1200–1207. [[CrossRef](#)]
- [5] Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* 2009, 41, 15. [[CrossRef](#)]
- [6] Pillai D. S.; Blaabjerg F.; Rajasekar N.: "A Comparative Evaluation of Advanced Fault Detection Approaches for PV Systems" in *IEEE Journal of Photovoltaics*, vol. 9, no. 2, pp. 513-527, March 2019, doi: 10.1109/JPHOTOV.2019.2892189. [[CrossRef](#)]
- [7] M. Said Elsayed, N. A. Le-Khac, S. Dev, and A. D. Jurcut, "Network Anomaly Detection Using LSTM Based Autoencoder" Q2SWinet 2020 - Proc. 16th ACM Symp. QoS Secur. Wirel. Mob. Networks, no. November, pp. 37–45, 2020, doi: 10.1145/3416013.3426457 [[CrossRef](#)]
- [8] Fonseca Alves, Ricardo Henrique & Deus Júnior, Getúlio & Marra, Enes & Lemos, Rodrigo. (2021). *Automatic fault classification in photovoltaic modules using Convolutional Neural Networks*. *Renewable Energy*. 10.1016/j.renene.2021.07.070.
- [9] Branco P, Gonçalves F, Costa AC. *Tailored Algorithms for Anomaly Detection in Photovoltaic Systems*. *Energies*. 2020; 13(1):225. <https://doi.org/10.3390/en13010225>
- [10] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long Short Term Memory Networks for Anomaly Detection in Time Series" in *Proceedings of the 23rd European Symposium on Artificial Neural Networks*, 2015.
- [11] WOOLDRIGE, J. M. *Introductory Econometrics: a Modern Approach*. [S.l.: s.n.], 2019.
- [12] HAWKINS, D. *Identification of outliers*. Chapman and Hall, 1980.
- [13] LeCun Y, Bengio Y, Hinton G.: "Deep learning," *Nat.* 2015 5217553, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [14] Hinton, G.E., Salakhutdinov, R.: *Reducing the dimensionality of data with neural networks*. *Science* . 313, 504–507, 2006.
- [15] Laptev N, Yosinski J, Li LE, Smyl S.: *Time-series extreme event forecasting with neural networks at uber*. *Int Conf Mach Learn* 34:1–5, 2017.
- [16] Voyant C, Notton G, Kalogirou S, Nivet ML, Paoli C, Motte F, Fouilloy.: *A Machine learning methods for solar radiation forecasting: a review*. *Renew Energy* 105:569–582, 2017. <https://doi.org/10.1016/j.renene.2016.12.095>
- [17] Alzahrani A, Shamsi P, Dagli C, Ferdowsi M.: *Solar irradiance forecasting using deep neural networks*. *Proc Comput Sci* 114:304–313, 2017. <https://doi.org/10.1016/j.procs.2017.09.045>
- [18] LeCun Y, Bengio Y, Hinton G.: *Deep learning*. *Nature*. 521(7553):436–444, 2015. <https://doi.org/10.1038/nature14539>
- [19] Chiu JP, Nichols E.: *Named entity recognition with bidirectional lstm-cnns*. *Trans Assoc Comput Linguist* 4:357–370, 2016. [https://doi.org/10.1162/tacl\\_a\\_00104](https://doi.org/10.1162/tacl_a_00104)
- [20] Sundermeyer M, Schlüter R, Ney H.: *Lstm neural networks for language modeling, in thirteenth annual conference of the international speech communication association*, 2012.
- [21] Soltau H, Liao H, Sak H.: *Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition*, 2016. arXiv:161009975

- 
- [22] Zhao Z, Chen W, Wu X, Chen PC, Liu J.: *Lstm network: a deep learning approach for short-term traffic forecast*. *IET Intell Transp Syst.* 11(2):68–75, 2017. <https://doi.org/10.1049/iet-its.2016.0208>
- [23] Ibrahim M, Alsheikh A, Awaysheh M F, Alshehri M D.: *Machine learning Schemes for Anomaly Detection in Solar Power Plants*. *Energies*, 15, 1082, 2022. <https://doi.org/10.3390/en15031082>
- [24] Malakar, S., Goswami, S., Ganguli, B.: *Designing a long short-term network for short-term forecasting of global horizontal irradiance*. *SN Appl. Sci.* 3, 477,2021. <https://doi.org/10.1007/s42452-021-04421-x>
- [25] Gensler A, Henze J, Sick B, Raabe N.: *Deep learning for solar power forecasting an approach using autoencoder and lstm neural networks*. In: 2016 IEEE international conference on systems, man, and cybernetics (SMC), IEEE, pp 002858–002865, 2016. <https://doi.org/10.1109/SMC.2016.7844673>
- [26] Yagli GM, Yang D, Srinivasan D.: *Automatic hourly solar forecasting using machine learning models*. *Renew Sustain Energy Rev* 105:487–498, 2019. <https://doi.org/10.1016/j.rser.2019.02.006>
- [27] Pereira J, Silveira M.: "Unsupervised Anomaly Detection in Energy Time Series Data Using Variational Recurrent Autoencoders with Attention". 17th IEEE International Conference on Machine Learning and Applications(ICMLA),pp. 1275-1282,2018. doi: 10.1109/ICMLA.2018.00207
- [28] Iqbal M.: *An introduction to solar radiation*. Elsevier, Amsterdam, 2012.
- [29] Rubin B, D.: "Inference and Missing Data." *Biometrika* 63 (3): 581–90, 1976.
- [30] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *IEEE Trans. Nucl. Sci.*, vol. 44, no. 3 PART 3, pp. 1464–1468, 1997, doi: 10.1109/23.589532
- [31] Kunh M, Johnson K.: "Applied Predictive Modeling", 1, New York, Springer, 2013. <https://doi.org/10.1007/978-1-4614-6849-3>
- [32] Ripley, B.: *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press, 1996. doi:10.1017/CBO9780511812651
- [33] Gopali S, Abri F, Namini S , Namin S , Akbar.: *A Comparative Study of Detecting Anomalies in Time Series Data Using LSTM and TCN Models*. Texas Tech University, 2021. <https://arxiv.org/pdf/2112.09293.pdf> Acesso em: 16 de maio de 2022.
- [34] Sutskever I, Vinyals O, Le V. Q.: Sequence to Sequence Learning with Neural Networks. arXiv, 2014. <https://doi.org/10.48550/arxiv.1409.3215>
- [35] Hochreiter S, Schmidhuber J.: Long Short-term Memory. *Neural computation*, . 1735-80, 1997. doi:10.1162/neco.1997.9.8.1735.

---

## APÊNDICE A - CÓDIGO FONTE DO TRABALHO

Neste apêndices estamos disponibilizando o script responsável pelo desenvolvimento deste trabalho. O scripts completos podem ser consultados pelo repositório no *Github*: <https://github.com/ailsonramon/AnomalyDetectionSolarPower>

### A.1 Implementação do Algoritmo Autoencoder LSTM

A FIGURA 14 define o algoritmo LSTM autoencoder, em python, utilizado neste trabalho.

FIGURA 14. Código em python do LSTM autoencoder

#### ▼ Creater an LSTM Autoencoder Network

```
[15] 1 def build_autoencoder_model(timeserie):
2     act_func = 'tanh'
3     n_feature = int(timeserie.shape[2])
4     n_time_steps = int(timeserie.shape[1])
5     steps_hidden_layer = int(n_time_steps / 2)
6     model = Sequential()
7     #Encoder layer
8     model.add(LSTM(n_time_steps, activation=act_func, input_shape=(n_time_steps, n_feature), return_sequences=True))
9     #Encoder hidden layers
10    model.add(LSTM(steps_hidden_layer, activation=act_func, return_sequences=False))
11    #Encoder output -> Decoder input
12    model.add(RepeatVector(n_time_steps))
13    #Decoder hidden layer
14    model.add(LSTM(steps_hidden_layer, activation=act_func, return_sequences=True))
15    #Decoder layer
16    model.add(LSTM(n_time_steps, activation=act_func, return_sequences=True))
17    #Decoder prediction
18    model.add(TimeDistributed(Dense(n_feature)))
19    model.summary()
20    return model
```

Fonte:Compilação do Autor.