UNIVERSIDADE FEDERAL DO OESTE DO PARÁ

PRÓ-REITORIA DE PESQUISA, PÓS-GRADUAÇÃO E INOVAÇÃO TECNOLÓGICA

PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS NATURAIS DA AMAZÔNIA

PAULO GUILHERME SILVA DOS SANTOS

PREVISÃO DE VARIÁVEIS AMBIENTAIS NA AMAZÔNIA COM USO DE REDES NEURAIS ARTIFICIAIS DO TIPO LONG SHORT-TERM MEMORY

SANTARÉM - PARÁ

2021

PAULO GUILHERME SILVA DOS SANTOS

PREVISÃO DE VARIÁVEIS AMBIENTAIS NA AMAZÔNIA COM USO DE REDES NEURAIS ARTIFICIAIS DO TIPO LONG SHORT-TERM MEMORY

Projeto de dissertação apresentado à Universidade Federal do Oeste do Pará – UFOPA, como requisito para obtenção do título de Mestre em Ciências Ambientais, junto ao Programa de Pós-Graduação *Stricto Sensu* em Recursos Naturais da Amazônia.
Área de concentração: Processos de Interação Biosfera-Atmosfera.
Orientador: Prof. Dr. Anderson Alvarenga de Moura Meneses

SANTARÉM - PARÁ

2021

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema Integrado de Bibliotecas – SIBI/UFOPA

S237p Santos, Paulo Guilherme Silva dos

Previsão de variáveis ambientais na Amazônia com uso de redes neurais artificiais do tipo Long Short-Term Memory. / Feli Paulo Guilherme Silva dos Santos. – Santarém, 2021.

46 p. : il.

Inclui bibliografias.

Orientador: Anderson Alvarenga de Moura Meneses

Dissertação (Mestrado) – Universidade Federal do Oeste do Pará, Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação Tecnológica, Programa de Pós-Graduação em Recursos Naturais da Amazônia.

1. Predição de temperatura. 2. Predição de irradiância solar. 3. Deep Learning. I. Meneses, Anderson Alvarenga de Moura, *orient*. II. Título.

CDD: 23 ed. 621.47

Bibliotecária - Documentalista: Renata Ferreira – CRB/2 1440

PAULO GUILHERME SILVA DOS SANTOS

PREVISÃO DE VARIÁVEIS AMBIENTAIS NA AMAZÔNIA COM USO DE REDES NEURAIS ARTIFICIAIS DO TIPO LONG SHORT-TERM MEMORY

Dissertação apresentada ao Programa de Pós-Graduação em Recursos Naturais da Amazônia para obtenção do título de Mestre em Ciências Ambientais; Universidade Federal do Oeste do Pará; Área de concentração: Processos de Interação Biosfera-Atmosfera.

Conceito: 8,5

Data de Aprovação: 25/08/2021

Prof. Dr. Anderson Alvarenga de Moura Meneses
Universidade Federal do Oeste do Pará (UFOPA)

Prof. Dr. José Mauro Sousa de Moura
Universidade Federal do Oeste do Pará (UFOPA)

Profa. Dra. Helaine Cristina Moraes Furtado
Universidade Federal do Oeste do Pará (UFOPA)

Prof. Dr. Roberto Schirru
Universidade Federal do Rio de Janeiro (UFRJ)

Aos meus pais e minha noiva pelo apoio, carinho e incentivo

AGRADECIMENTOS

Agradeço ao Senhor meu Deus e meu Pai, pelo Seu amor incondicional em nossas vidas, pela conclusão deste trabalho e por ter me proporcionado não tudo o que pedi, mas tudo o que necessitei.

Aos meus pais Francisco Haroldo Ferreira dos Santos e Veraldina Ribeiro da Silva pois graças ao apoio deles que consegui chegar até este ponto, me ajudando em cada dificuldade que passei.

À minha noiva Niza Catarina Vaz Colares que esteve ao meu lado em todos os momentos me dando suporte e ajudando no que fosse necessário para que pudesse escrever esta dissertação.

Ao professor Dr. Anderson Alvarenga de Moura Meneses por toda contribuição, paciência e conhecimento que me foi passado.

Ao professor Dr. Wilson Negrão Macêdo e ao Grupo de Estudos e Desenvolvimento de Alternativas Energéticas pela disponibilidade de dados para realização deste trabalho.

Aos amigos que adquiri na graduação pelos momentos de alegria, apoio e descontração.

Aos professores que contribuíram para o meu conhecimento professional e crescimento como pessoa.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 132805/2019-7.

"Quando você passar por momentos difíceis e se perguntar onde estará Deus, lembre-se que durante uma prova, o professor está em silêncio"

Aline Barros

RESUMO

Com a expansão da indústria de energia solar fotovoltaica veio também a busca pela melhoria da eficiência de seus sistemas. A previsão de variáveis ambientais como temperatura e irradiância solar ajudam na tomada de decisão do uso destes sistemas. Neste trabalho, foi utilizada uma metodologia para avaliação de Redes Neurais Artificiais (RNA) de memórias de curto-longo prazo (LSTM), para predição de séries temporais destas duas variáveis. Os dados utilizados foram obtidos em duas cidades localizadas na região amazônica, sendo dois conjuntos de dados de temperatura e um conjunto de dados de irradiância solar. Durante o processamento dos conjuntos de dados foram verificadas características fundamentais para predição de séries temporais como autocorrelação e estacionariedade, e a divisão em conjuntos de treino-teste e validação. As arquiteturas utilizadas possuem diferenças em seus números de camadas, para analisar a influência de sua complexidade em seus desempenhos. Como resultado, na validação dos dados a arquitetura de 3 camadas apresentou diferença estatisticamente significativa em relação à arquitetura de 7 camadas, para um mesmo número de épocas. Para o conjunto de dados de temperatura do LABIC as médias de RMSE das duas arquiteturas foram de 0,9393°F e 1,4531°F, para 3 e 7 camadas, respectivamente; para o conjunto de dados de temperatura do GEDAE a média de RMSE foi de 1,6499°F e 1,9767°F, para 3 e 7 camadas, respectivamente; e no conjunto de dados de irradiância solar obtivemos RMSE médio de 170,6649 W/m² e 204,7825 W/m², para 3 e 7 camadas, respectivamente. A metodologia utilizada permitiu a comparação entre as arquiteturas e poderá ser utilizada futuramente para avaliação de outros modelos de RNAs de previsão de séries temporais.

Palavras chave: Predição de temperatura. Predição de irradiância solar. Deep Learning. Long short-term Memory.

ABSTRACT

With the expansion of the photovoltaic solar energy industry came the search for improving the efficiency of its systems. The prediction of environmental variables such as temperature and solar irradiance help in making decisions about the use of these systems. In this work, a methodology for the evaluation of Artificial Neural Networks (ANN) of short-long-term memories (LSTM) was used to predict the time series of these two variables. The data used were obtained from two cities located in the Amazon region, two sets of temperature data and a set of solar irradiance data. During the processing of data sets, fundamental characteristics for time series prediction were verified, such as autocorrelation and stationarity, and the division into training-test and validation sets. The architectures used have differences in their number of layers, to analyze the influence of their complexity on their performance. As a result, in data validation, the 3-layer architecture presented a statistically significant difference compared to the 7-layer architecture, for the same number of times. For the LABIC temperature dataset, the RMSE averages of the two architectures were 0.9393°F and 1.4531°F, for 3 and 7 layers, respectively; for the GEDAE temperature dataset the mean RMSE was 1.6499°F and 1.9767°F, for 3 and 7 layers, respectively; and in the solar irradiance dataset we obtained an average RMSE of 170.6649 W/m² and 204.7825 W/m², for 3 and 7 layers, respectively. The methodology used allowed the comparison between the architectures and could be used in the future to evaluate other models of ANNs for forecasting time series.

Key words: Temperature forecasting. Solar irradiance forecasting. Deep Learning. Long short-term Memory.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

A radiação solar é a maior fonte contínua de energia disponível para os seres humanos e é o principal fator meteorológico de estudos ambientais, ecológicos e econômicos. Seu potencial energético que incide sobre a superfície terrestre chega a aproximadamente 10.000 vezes o consumo anual de energia do nosso planeta (Mohanty et al., 2017).

Além desta capacidade energética, tem-se o atual interesse na busca de geração de energia através de fontes renováveis como alternativa para a emissão de poluentes e, consequentemente, a degradação do meio ambiente (Tan et al., 2012). Neste cenário, a indústria de energia solar fotovoltaica (FV) tem obtido destaque no seu crescimento ao longo dos últimos anos (IEA, 2017). No entanto, alguns fatores meteorológicos como temperatura, vento, pressão e umidade são parâmetros que moldam a eficiência e capacidade de geração deste tipo de energia (Das et al., 2018).

Com a importância econômica e ambiental deste mercado, torna-se necessário aprimorar a eficiência de seus sistemas. A realização de previsão da irradiância solar auxilia na otimização de *microgrids* supridos por painéis FV (Husein e Chung, 2019), além disso, como a temperatura está fortemente correlacionada a irradiância solar, sua previsão tem importância equivalente (Gao et al., 2019), além de seu impacto na eficiência energética dos painéis FV (Gnoatto et al., 2008).

As previsões de variáveis ambientais podem ser realizadas através de modelos físicos, estatísticos e de aprendizado de máquina (Aggarwal e Saini, 2014), sendo esta última a mais utilizada nos dias atuais. Os modelos de previsão baseados em aprendizado de máquina consistem em Redes Neurais Artificiais (RNAs), que podem aprender a partir de uma série de dados para desenvolver um mapeamento não linear entre dados de entrada e saída (Qing e Niu, 2018). Lewis (2016) classifica as RNAs em estáticas e dinâmicas, onde as estáticas calculam a saída diretamente dos dados de entrada, nas dinâmicas, os dados de saída dependem da entrada atual e das entradas e saídas do estado oculto das redes anteriores, assim são conhecidas como redes neurais recorrentes (RNR) (Elman, 1990).

No que se diz respeito ao horizonte de previsões de potencial energético de painéis FV, Vaz et al. (2016) dividiu em dois grupos: curto prazo (short-term) onde o intervalo de previsão pode ser de um minuto a um dia, e longo prazo (long-term) com intervalos de dias a um mês. Previsões de curto prazo auxiliam na manutenção do sistema e no aproveitamento energético,

enquanto previsões de longo prazo auxiliam nas tomadas de decisões e estudos de viabilidade de implantações de painéis FV.

Neste trabalho foi utilizado um método de RNA conhecido como Memorias de Longo e Curto-Prazo (Long Short-Term Memory - LSTM), desenvolvido por Hochreiter e Schmidhuber (1997), onde o modelo é capaz de compreender as dependências de curto e longo prazo, ou seja, assim que a rede recebe uma informação nova, ele decide se descarta a informação ou guarda para que a mesma possa ser utilizada em um período mais longo. Assim, realizou-se a predição de curto prazo para duas variáveis ambientais, temperatura do ar e irradiância solar, por meio de modelos LSTM com diferentes números de camadas e na avaliação seus desempenhos por meio de métodos estatísticos. Além disto propõe-se a avaliação da capacidade de adequação do modelo LSTM para as estas duas variáveis, que consiste de duas séries de dados de temperatura para diferentes municípios da Amazônia e uma série de dados de irradiância solar.

Entre as principais contribuições deste trabalho tem-se: (i) Previsão de variáveis ambientais para região Amazônica que possui poucos estudos deste tipo, e ainda se destaca com uma região com elevado potencial para a energia solar e variabilidade devido as condições de nebulosidade (Sousa et al., 2020), (ii) análise do desempenho de modelos LSTM com diferentes números de camadas, (iii) Proposta de metodologia para análise de modelos de previsão de séries temporais.

Este trabalho está organizado da seguinte maneira. A seção dois apresenta o referencial teórico. A seção três apresenta os objetivos gerais e específicos, bem como perguntas e hipóteses do trabalho. A seção quatro apresenta os procedimentos metodológicos de aquisição de dados e composição dos modelos. A seção cinco apresenta os resultados obtidos na aplicação da rede LSTM com dados de temperatura e irradiância. A conclusão é apresentada na seção 6, e por fim, as referências utilizadas na seção 7.

2. OBJETIVOS

2.1. Objetivo Geral

Realizar a previsão de dados de temperatura e irradiância solar por meio de Redes Neurais Recorrentes do tipo LSTM.

2.2. Objetivos Específicos

- Aplicar diferentes arquiteturas LSTM com diferentes números de camadas e normalizações para efeito de comparação;
- Aplicar uma metodologia de comparação, analisando os resultados obtidos por meio de métricas e testes estatísticos e verificando se há diferença estatisticamente significativa entre as arquiteturas;
- Determinar e validar o modelo de previsão com melhor desempenho estatístico.

3. REFERENCIAL TEÓRICO

Esta seção está destinada a apresentação de uma fundamentação teórica relevante para o entendimento do problema. Onde serão tratados conceitos de radiação solar e temperatura, com suas devidas importâncias no sistema de geração de energia fotovoltaica. Serão apresentados ainda, as RNAs, RNRs e LSTM, bem como suas estruturas, equações e aplicação em diversas áreas de conhecimento.

3.1. Radiação Solar

A transferência de energia proveniente do interior da esfera solar é conhecida como radiação solar (Varejão, 2005). Assim, radiação solar é toda radiação eletromagnética com origem no Sol que chega na Terra (Querino et al., 2006; Querino et al., 2011). Ela possui comprimentos de onda que variam na faixa de 150 a 4.000 nm (Rosemberg, 1974; Slater, 1980). A radiação solar é a principal fonte primaria de energia, de onde derivam quase todas as outras formas de energia.

Além de ser responsável por aquecer a superfície terrestre, resultando em diversos fenômenos como a evaporação das águas dos rios e oceanos e, consequentemente, a formação de nuvens e precipitação, a radiação solar é também fonte de energia para o processo de fotossíntese, importante na reciclagem de CO_2 (Martins et al., 2014). Borges et al. (2010) afirmam que esta fonte de energia é a força motriz para os diversos processos de ordem física, química e biológica que ocorrem em nosso planeta.

Somente parte da radiação solar atinge a superfície terrestre, devido aos processos físicos que os raios solares sofrem ao atravessar a atmosfera. Cerca de 19% da radiação solar que chega no topo da atmosfera são absorvidos pelas nuvens, aerossóis e vapor d'água, e 30% são refletidas de volta por gases, nuvens e o solo, sobrando 51% que atinge o solo e é absorvido. A radiação que atinge a superfície é conhecida como radiação solar global é formada por duas componentes, denominadas radiação solar difusa e a radiação solar direta (Kleissl, 2013; Querino et al., 2011).

O estudo das três radiações (radiação solar global, radiação solar difusa e radiação solar direta) é importante porque torna possível obter os índices radiométricos, os quais indicam a transmissão das radiações na atmosfera em qualquer dia e local (Almeida et al., 2011). A

radiação solar é uma das principais fontes de energias renováveis, e apresenta-se como destaque na atualidade por ser uma fonte renovável, limpa, e gratuita o que faz com que seja considerada a maior fonte ininterrupta de energia disponível para os seres humanos (Al-Salaymeh, 2006). Assim, realização de medidas da radiação solar global é fundamental também para o dimensionamento de sistemas geradores de energia solar fotovoltaica e energia térmica.

A radiação solar direta é a porção que atinge diretamente a superfície, sofrendo apenas a refração em seu caminho devido à mudança de densidade entre as camadas atmosféricas (Querino et al., 2011). Se a atmosfera apresentar baixa nebulosidade, menor concentração de particulados provenientes da poluição, e baixa quantidade de partículas em suspensão, menor será a difusão sofrida pela radiação solar. Isso significa que maior será a proporção dos raios solares que atinge diretamente a superfície. Dessa forma, a radiação solar direta apresenta uma sensibilidade à profundidade óptica dos aerossóis, sendo afetada intensamente pelas nuvens que cobrem o sol (Kotti et al., 2014).

Já a radiação solar difusa inclui a radiação solar proveniente de todas as direções, sendo difundida na atmosfera devido à presença de partículas diversas (Inácio, 2009). Segundo Iqbal (1978) esta radiação, em determinados momentos, depende de algumas condições como altitude e latitude da região, da declinação e do ângulo de elevação do Sol, do índice de turbidez, do grau de concentração de vapor atmosférico e da presença de nuvens. A radiação solar global direta e a difusa serão influenciadas pelo albedo (taxa de reflexão) da superfície, uma vez que este irá determinar a quantidade de radiação de onda curta que ficará no sistema e a porção que será devolvida à atmosfera, caracterizando assim o Balanço de Ondas Curtas (Pavão, 2016).

Dentre os equipamentos mais comuns para medir a irradiância solar tem-se os piranômetros, utilizados para realizar a medição da radiação solar global e também nas estimativas de radiação difusa (Varejão, 2005). Existem piranômetros compostos por fotodiodo de células de silício, e também por termopares que funcionam através da formação de uma termopilha (Latimer, 1971).

3.2. Temperatura

O conceito de temperatura é o grau de agitação térmica das moléculas de um determinado corpo (Faucher e Physique, 1966; Hope, 1928; Macedo, 1981). Pelo fato de a temperatura ser

um dos principais fatores que permitem a subsistência de vida na Terra, ela já vem sendo estudada e seus dados obtidos desde o século passado. No entanto, têm-se discutido muito questões como mudanças climáticas nos dias atuais, em que a temperatura se tornou alvo de destaque nos estudos de monitoramento ambiental.

A temperatura influencia na eficiência de placas solares como afirmam Gnoatto et al. (2008), em que seus estudos constataram que o aumento da temperatura reduz o desempenho de produção de energia solar. No caso de Michels et al. (2010), que avaliaram a eficiência em diferentes níveis de potência, obtiveram valores de temperatura média de 36,85°C para irradiância de 500W/m^2 com eficiência média do painel fotovoltaico de 8,48%, e para irradiância de 1000W/m^2 onde a média de temperatura foi de 50,68°C, a média de eficiência do equipamento sofreu redução e atingiu 5,63%.

3.2.1. Temperatura do ar à superfície

Os estudos de temperatura do ar são importantes nas mais variadas áreas de conhecimento, como na meteorologia, oceanografia, climatologia e hidrologia (Cavalcanti et al. 2005). O termo *temperatura do ar à superfície* expressa, na meteorologia, a temperatura predominante em um ponto da atmosfera próximo a superfície, assim, é usualmente medida a uma altura que varia de 1,25 a 2,00m (Varejão, 2005). A oscilação desta variável ao longo do tempo não ocorre de maneira lenta, Middleton (1943) apresentou no seu estudo que mesmo em poucos minutos a temperatura à superfície pode variar até 2 °C.

3.3. Análise de séries temporais para predições

Os grupos de dados ordenados em um tempo cronológico são chamados de séries temporais (Parzen, 1961). Estas séries temporais possuem características que dizem respeito ao tipo de dado que ela apresenta, podendo ser uma série temporal com uma variável, conhecida como univariada, ou com múltiplas variáveis, chamada de multivariada. Suas características são fundamentais no processo de previsão das séries temporais, pois vão influenciar na capacidade dos modelos de previsão em prever seus dados futuros.

A primeira característica que devemos observar nas previsões de séries temporais é quanto a estacionariedade destas séries. Montgomery e Jennings (2015) definiram que uma série temporal é estritamente estacionária quando a mudança no tempo não altera suas propriedades, ou seja, se temos uma distribuição de dados $y_t, y_{t+1}, \ldots, y_{t+n}$ e o coeficiente de estacionariedade da série for igual a 0, assume-se a probabilidade da distribuição y_t ser a mesma para qualquer período. Em corroboração, Peña et al. (2001) afirma que estas séries estacionárias possuem variação constante em torno de uma média ao longo do período de dados. Já para séries não estacionárias, segundo o autor, as médias e comportamentos dos dados não rodeiam constantemente uma média ao longo das observações, o que traz uma aleatoriedade aos dados da série, o que resulta em uma maior dificuldade de realizar a sua previsão. Testes estatísticos como o teste aumentado de Dickey-Fuller são utilizados para verificar se uma série temporal é estacionária.

A segunda característica importante para previsão das séries temporais é sobre a autocorrelação de seus dados. Ao analisarmos uma série temporal temos que seus dados são igualmente espaçados por um período de tempo *k*, este intervalo de tempo é conhecido como *lag*. A autocorrelação dos dados indica se eles possuem dependências entre si, e até que nível (*lag*) chega esta dependência (Montgomery e Jennings, 2015), ou seja, se a observação y_t possui dependência com $y_{t-1}, y_{t-2}, \ldots, y_{t-n}$. Para realizar a estimativa da autocorrelação de uma série de dados, tem-se a função de autocorrelação, descrita pela Equação:

$$\rho_k = \frac{y_k}{y_0} \tag{1}$$

Onde, y_k é a covariância entre y_i e y_{i+k} para qualquer i.

3.4. Redes Neurais Recorrentes (RNR)

Redes Neurais Recorrentes são um tipo de RNA, desta forma elas tem o propósito de simular o funcionamento do cérebro para solucionar problemas. O esquema de uma rede neural artificial, como a *Multilayer Perceptron* (MLP), a qual apresenta apenas uma única direção, sendo sua sequência dada pelos dados de entrada (*input layer*), camadas ocultas (*hidden layers*), podendo ter uma ou mais, e as camadas de saída (*output layers*), a interação entre os dados de entrada e as camadas entre si se dá através dos pesos. Assim, estes tipos de RNAs não possuem boa resposta a problemas que envolvem sequência de dados (Castelão, 2018).

Lewis (2016) classifica as RNAs em estáticas (Figura 1a) e dinâmicas (Figura 1b), em que as estáticas calculam a saída diretamente dos dados de entrada, através das conexões diretas. Já as dinâmicas os dados de saída dependem da entrada atual e das entradas e saídas do estado oculto das redes anteriores, sendo conhecidas assim, como redes neurais recorrentes. Elas apresentam conexões recorrentes dos neurônios das camadas ocultas, em que as saídas destes neurônios são armazenadas por uma etapa de tempo e então alimentam de volta a camada de entrada (Haykin, 2005), isto alinhado ao treinamento da rede, assim, cada neurônio possui interação com os neurônios das camadas adjacentes.

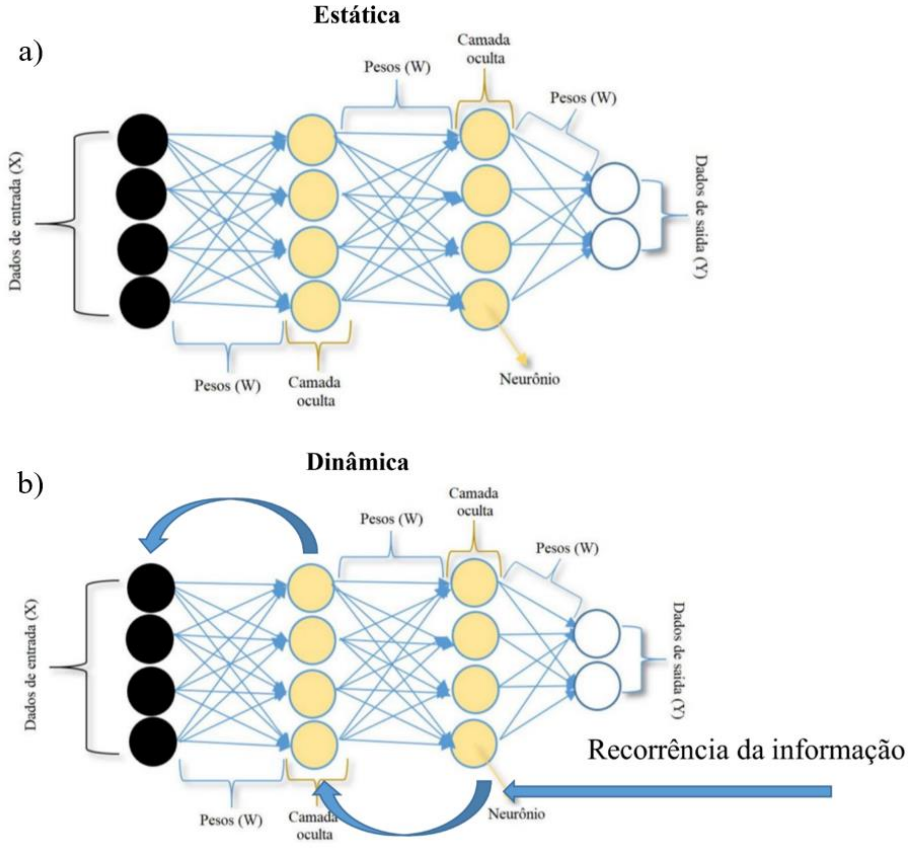
As Equações 2 e 3 descrevem o conceito de RNR elaborado por Elman (1990):

$$h^t = \sigma(W_h X + W_r h^{t-1}) \quad (2)$$

$$y = \sigma(W_y h^t) \quad (3)$$

Onde h indica o estado oculto, W são pesos, X entrada, y é a saída e σ representa uma função sigmoide.

Figura 1. a) Esquema de RNA estática; e b) Esquema de RNA dinâmica propostas por Lewis 2016.



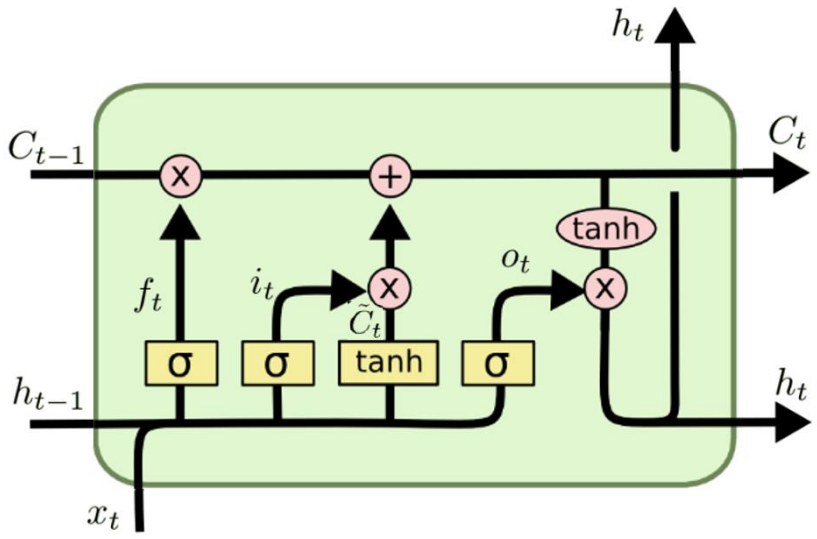
Fonte: Autor.

3.5. Redes Neurais Recorrentes do tipo LSTM

Em 1997, Hochreiter e Schmidhuber propuseram um modelo de Rede Neural Recorrente que solucionasse alguns problemas encontrados com a retropropagação das RNR, como erros nos valores das camadas de saída que tendem a ter uma explosão ou a desaparecer, isto se deve ao fato de que a evolução temporal do erro retro propagado depende de maneira exponencial dos tamanhos dos pesos (Hochreiter, 1991).

Para solucionar o problema de retropropagação dos erros, a Rede LSTM é composta por unidades de memórias longas de curto prazo. Desta forma, ela possui *portões* (*gates*) que são compostos por camadas sigmoide e uma função de multiplicação, onde serão decididos se a informação irá passar ou não por toda a estrutura (Hochreiter e Schmidhuber, 1997). Assim, a LSTM tem um estado de célula (*Cell State*) (Figura 2), onde as informações distribuídas são ponderadas pelas entradas em cada tempo.

Figura 2. Estado de célula e portões de uma rede LSTM



Fonte: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

O primeiro *portão* da rede é o *portão do esquecimento* (*forget* gate) que é o responsável por esquecer informações do estado interno, então ele irá descartar conteúdos que não sejam relevantes para o modelo proposto. É composto por uma função sigmoide, ou seja, obtém

valores de 0 ou 1, onde 1 mantém o estado interno intacto e 0 limpa o estado. O *forget gate* é representado pela equação a seguir:

$$f_t = \sigma(W_f[h^{<t-1>}, x^{<t>}] + b_f) \quad (4)$$

onde σ é a função sigmoide, W_f é uma matriz de parâmetros que recebe $h^{<t-1>}$ (estado anterior) e $x^{<t>}$ (entrada no tempo *t*), e b_f é um valor de correção.

Posteriormente a informação passará pelo *portão de entrada* (*input gate*) que define o que será inserido, proveniente da entrada, no estado celular, através de uma função sigmoide. Daí, vem uma segunda camada com função *tanh* (tangente hiperbólica) a qual criará um novo candidato ao estado celular, $\tilde{C}^{<t>}$. Na função *tanh* os valores serão convertidos em valores entre -1 e 1, esta função é fundamental para evitar o erro, que ocorre usualmente nas RNRs, em previsões de longo prazo, onde os valores não tenderão a explosão ou desaparecimento exponencial. Desta forma, tem-se as equações:

$$u_t = \sigma(W_u[h^{<t-1>}, x^{<t>}] + b_u) \quad (5)$$

$$\tilde{C}^{<t>} = \tanh(W_c[h^{<t-1>}, x^{<t>}] + b_c) \quad (6)$$

Onde W_u e W_c constituem matrizes de parâmetros em que entram $h^{<t-1>}$ e $x^{<t>}$, e b_u e b_c são valores de correção.

Daí o antigo $\tilde{C}^{<t-1>}$é atualizado com um novo $\tilde{C}^{<t>}$, multiplicando o estado antigo por f_t e depois adicionando $i_t * \tilde{C}^{<t>}$. Assim, o *input gate* virá com uma soma para acrescentar a informação:

$$C^{<t>} = f_t * C^{<t-1>} + i_t\tilde{C}^{<t>} \quad (7)$$

O *portão de saída* (*output gate*) irá definir qual informação nova será passada para o próximo tempo e para a saída da rede. Primeiro irá ser determinado qual parte do estado da célula será enviado para a saída por meio de uma função sigmoide, e então, usa-se uma função *tanh* para gerar as saídas que serão multiplicadas a saída da função sigmoide, como descreve as equações a seguir:

$$o^{<t>} = \sigma(W_o[h^{<t-1>}, C^{<t>}] + b_o) \quad (8)$$

$$h^{<t>} = o_t * \tanh(C^{<t>}) \quad (9)$$

Os trabalhos que envolvem a aplicação de LSTM envolvem análise de consumo de energia (Junior, 2019), previsões no mercado de ações (Castelão, 2018), vazão de rio (VassalI, 2018), nível de lençol freático em áreas de agricultura (Zhang et al. 2018), fluxo de turismo (Li e Cao, 2017). Geller e Meneses (2021) modelaram o software EnergySaver (ver Silva et al., 2021) com Unified Modelling Language (UML), o software tem a funcionalidade de usar redes LSTM para realizar a previsão de consumo de energia em um sistema de Internet das Coisas (IoT). Desta forma, o sistema realiza o monitoramento e previsão do consumo energético. As redes utilizadas no presente trabalho poderão futuramente formar um módulo para variáveis ambientais a ser acoplado ao software de análise de consumo de energia.

3.6. Uso de LSTM nas séries temporais de temperatura e irradiância solar

Para os estudos de temperatura, Karevan e Suykens (2018) realizaram a predição em cinco cidades: Bruxelas, Antuérpia, Liège, Amsterdã e Eindhoven, e obtiveram baixos valores de *MAE* e *Mean Squared Error* (MSE) em seu estudo para previsão do tempo e concluíram que a utilização deste método melhora os desempenhos das previsões.

Xu et al. (2019) utilizaram as redes LSTM para predições de temperatura do ar em interiores de edifícios, onde propôs ainda uma LSTM modificada do original, a qual obteve leve melhora em termos de predição da previsão direcional e o acompanhamento da tendência das variações, com *Root Mean Square Error* (RMSE) de 0,364 e 0,358 para LSTM e LSTM modificada respectivamente, na predição de um passo à frente, e 0,526 e 0,520 para LSTM e LSTM modificada, respectivamente, na predição de múltiplos passos à frente.

Zhang et al. (2018) propuseram um modelo hibrido de LSTM com *Ensemble Empirical Mode Composition* (EEMD) com objetivo de reduzir a dificuldade de modelagem e melhorar a precisão da previsão, no seu trabalho, comparou o modelo hibrido com outros cinco modelos, com destaque para o hibrido que apresentou melhor desempenho no seu trabalho.

Já para previsão de radiação solar, Kara (2019), interessado na geração de energia solar, analisou quatro métricas no seu trabalho, sendo elas MAE, RMSE, *Mean Absolut Percentage*

Error e R^2, e concluiu que o método LSTM de precisão é mais eficiente do que outros métodos de aprendizagem de máquina.

Qing e Niu (2018) compararam a rede neural LSTM com a RNR simples, onde para dados de 2 anos o modelo LSTM foi 18,34% superior em desempenho e para dados históricos de 10 anos o erro estudado (RMSE) reduziu 42,9%.

4. METODOLOGIA

Neste tópico, será abordado a metodologia utilizada, onde serão apresentados os conjuntos de dados, pré-processamento dos dados, modelos e arquiteturas utilizados para realizar a predição das séries temporais, como critérios comparativos para a validação da Rede Neural LSTM. Na análise estatística dos resultados, os métodos comumente utilizados para este tipo de trabalho (*MAE, MAPE e RMSE*) serão abordados, além de testes estatísticos para comparar as diferentes arquiteturas utilizadas. O procedimento metodológico pode ser observado no fluxograma a seguir:

Figura 3. Fluxograma do esquema metodológico.

Fonte: Autor.

4.1. Caso de estudo e Conjunto de dados

4.1.1. Conjunto de dados 1 (*TempLabic*): Temperatura do ar adquirida no Laboratório de Inteligência Computacional da UFOPA

O conjunto de dados *TempLabic* consiste em medições de temperatura realizadas através da estação meteorológica profissional ITWH – 1080. É um equipamento com transmissão *wireless* que realiza medições de velocidade e direção do vento, precipitação, temperatura e umidade

(interna e externa), pressão barométrica, ponto de orvalho e sensação térmica. Embora o software da fabricante seja do sistema operacional Windows, o equipamento está acoplado a um *Raspberry Pi 3 Model B* com sistema operacional Linux. Na Figura 4, tem-se o esquema da aquisição dos dados.

Figura 4. Esquema de aquisição de dados de temperatura do ar.

Fonte: Autor.

Assim, a estação instalada envia os dados a cada 5 minutos para o seu receptor que está conectado a um *Raspberry Pi 3 Model B* que, através do software *WeeWX* (software livre escrito em Python), disponibilizado no site http://www.weewx.com/, armazena os dados em um banco de dados, além disto o *WeeWX* cria uma página HTML para visualização dos dados obtidos da estação conectada (Figura 5). Embora os dados de temperatura obtidos pela estação estejam em graus Celsius, no armazenamento dos dados o software os converte para graus Fahrenheit.

Figura 5. Interface HTML criada pelo software WeeWX para visualização de dados.

Fonte: Autor.

4.1.2. Conjunto de dados 2 (*TempGedae*) e Conjunto de dados 3 (*IrradGedae*): Dados de Temperatura do ar e irradiância solar obtidos pelo Grupo de Estudos e Desenvolvimento de Energias Alternativas da UFPA

O segundo conjunto de dados de temperatura, foi obtido no Grupo de Estudos e Desenvolvimento de Energias Alternativas – GEDAE (Sousa et al., 2020), na Universidade Federal do Pará, em Belém - Pará (1º27'S, 48º29'O), com 10.248 registros no período de dezembro de 2015 a novembro de 2016, entre 05h e 19h, convertidos em graus *Fahrenheit*. Os dados de irradiância solar também foram coletados pelo GEDAE, no mesmo período descrito anteriormente em W/m², no mesmo intervalo de tempo. Este conjunto de dados foi utilizado para previsão no trabalho de Sousa et al. (2020).

4.2. Pré-Processamento e preparação dos dados

Os dados foram pré-processados com a utilização de códigos em *Python* através da biblioteca *Pandas*, que fornece recursos para análise e manipulação de dados, onde linhas com dados nulos ou ausentes foram excluídos, além da conversão de temperatura dos dados do GEDAE de graus $^{\circ}C$ para $^{\circ}F$.

Os dados foram divididos em dados de treino e teste, 80/20 (Dangeti, 2017) onde os primeiros 80% dos dados são utilizados para o treino do modelo e os 20% restantes para validação do modelo treinado. Estes dados, que estarão dispostos em uma lista serão convertidos para o tipo *array* (vetor) - para que os dados possam ser processados, transformá-los em vetor torna-se mais adequado. O conjunto de dados de validação consiste de um conjunto de dados que não participou da etapa de treinamento e teste dos modelos, assim o esquema de divisão é descrito na Figura 6. As avaliações de resultados provenientes apenas de treinamento podem implicar em desempenhos tendenciosos. Assim, o conjunto de dados de validação é aplicado para verificar a imparcialidade dos modelos.

Figura 6. Divisão dos conjuntos de dados de treino e teste, e validação.

Fonte: Autor.

Como os modelos de *machine learning* e LSTM são sensíveis as escalas das entradas (Bouktif et al., 2018) e que as redes neurais artificiais trabalham melhor com valores entre 0 e 1, pois acelera o processo e evita grandes erros da rede. Neste trabalho houve a normalização – conversão dos valores originais para um intervalo definido - de 0,1 a 0,9 e de 0 a 1, dos dados através da função *MinMaxScaler* importada da biblioteca *sklearn.preprocessing*. Outras bibliotecas serão utilizadas nas arquiteturas das redes, como *keras*, *numpy*, *matplotlib.pyplot* e *time* (a fim de verificar os tempos de execuções), as funções utilizadas foram *Sequential* de *keras.models*, e *Dense*, *Dropout* e *LSTM* de *sklearn.preprocessing*.

4.3. Análise exploratória das séries temporais

Em estudos de previsão de séries temporais considera-se que os dados são dependentes de seus valores passados, assim, a análise exploratória das séries temporais auxilia na identificação de dependências, padrões e tendências na série.

A função de autocorrelação (ACF) auxilia na análise exploratória das séries temporais, pois indica o grau de dependência dos dados destas séries (Box e Jenkins, 1970; Chatfield, 2000),

onde esta função realiza medidas da correlação de Pearson entre os dados da série que para valores de 0 indicam uma não dependência dos dados e valores de 1 uma dependência linear perfeita (Zhou, 2012). Com base em resultados preliminares no estudo da ACF (função de autocorrelação) foi definido empiricamente o *Sliding Window* (Paoli et al., 2010) - parâmetro que utiliza um determinado intervalo para realizar a previsão do dado seguinte - de 90 para todas as arquiteturas.

Outro passo importante sobre uma série temporal em estudo, é a identificação da estacionariedade da série. A estacionariedade de uma série indica que seus parâmetros estatísticos e suas propriedades estruturais como média, variância e autocorrelação tendem a não variar com o tempo (Chatfield, 2000). Para realizar essa análise utilizou-se o teste aumentado de Dickey-Fuller (ADF) que para p-valor < 0,05, rejeita-se a hipótese nula da existência de uma raiz unitária na série, com evidência estatística de que as séries temporais analisadas são estacionárias.

4.4. Descrição das Arquitetura dos modelos LSTM

Para avaliar o desempenho dos modelos LSTM com diferentes camadas foram definidas três arquiteturas para cada série temporal. As características principais de treinamento das redes neurais artificias deste estudo consistem na utilização de 100 épocas – quantidade de vezes que o conjunto completo de dados é passado durante o treinamento. Em testes preliminares foi utilizado o *batch_size* - tamanho do lote, ou seja, número de exemplos de treinamento utilizados em uma interação - de 128, que não apresentou diferenças estatisticamente significativas para o valor padrão de 32, que diminui significativamente o tempo de execução da rede. Foi definido o número de 100 neurônios para cada camada e função de ativação *linear* - comumente utilizada em problemas de regressão, é uma função que não altera as saídas dos neurônios. Para o estudo de camadas, para cada conjunto de dados foram utilizadas 3, 5 e 7 camadas (Tabela 1).

Tabela 1. Arquiteturas de Redes Neurais Artificias LSTM utilizadas no presente trabalho

Arquitetura	N˚ de neurônios	N˚ de camadas	Sliding Window
LSTM-3L	100	3	90
LSTM-5L	100	5	90
LSTM-7L	100	7	90

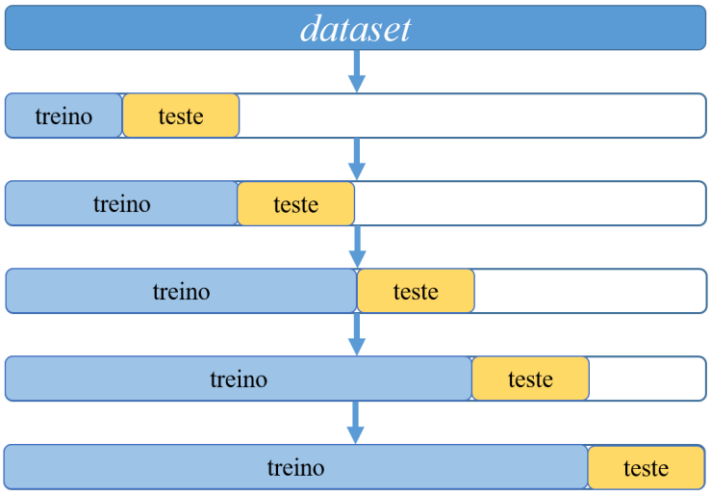
Fonte: Autor.

As execuções de treinos e testes deste trabalho foram realizadas via browser no site colab.research.google.com da Google que conta em seu setup computacional as GPUs Nvidia K80 com 24GB de VRAM um poderoso hardware para processamento de modelos de aprendizado de máquina. Nesta plataforma as entradas e saídas dos códigos ficam salvos na nuvem do próprio site.

4.5. Treinamento e Teste das Séries Temporais com *Cross-Validation*

A etapa de validação cruzada nas previsões de séries temporais é importante para garantir a robustez dos modelos (Dangeti, 2017). Neste trabalho, foi utilizado o *TimeSeriesSplit*, que é uma função do sklearn a qual divide o conjunto de dados em *k folds* e realiza treino e teste *k* vezes, em que em cada etapa incorpora os dados utilizados no treino e teste anterior (Figura 7). Trabalhos de validação cruzada comumente utilizam $k = 5$ ou $k = 10$, sendo que para este estudo foi definido $k = 10$.

Figura 7: Cross-Validation com *TimeSeriesSplit*, onde *k* = 5



Fonte: Autor.

4.6. Avaliação dos Modelos Treinados

Com objetivo de avaliar o desempenho estatístico das arquiteturas foram avaliados quatro métricas que são comumente utilizadas para trabalhos de previsões de séries temporais, sendo elas *mean absolute error (MAE), mean absolute percentage error (MAPE)* e *root mean square error (RMSE)* descritas por:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{10}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - x_i}{y_i} \right| * 100 \tag{11}$$

$$RMSE = \sqrt{[\frac{1}{n} \sum_{i=1}^{n} (|x_i - y_i|)^2]} \tag{12}$$

Onde y_i é o valor previsto pelas redes e x_i o valor observado.

4.6.1. Testes estatísticos

Em ordem de determinar a melhor arquitetura, e assumindo que os resultados não possuem distribuição normal, para os resultados foram aplicados os testes de Kruskal-Wallis (Dmitrienko et al., 2007), que é um teste não paramétrico utilizado para comparar três ou mais grupos de dados, sua hipótese nula é de que não há diferença estatisticamente significativa entre os grupos, enquanto a hipótese alternativa indica que pelo menos um grupo possui diferença estatisticamente significativa.

Utilizou-se o teste de Friedman (1937) (veja também Garcia et al., 2010) para análise dos resultados, uma vez que este teste realiza a comparação dos resultados dentro de cada conjunto por meio de um método de rankings, desta forma este teste coloca os resultados em uma classificação. Sua hipótese nula indica que os algoritmos possuem comportamentos similares. Se rejeitada sua hipótese nula, pelo menos um algoritmo se difere dos demais, sendo necessário um teste post hoc.

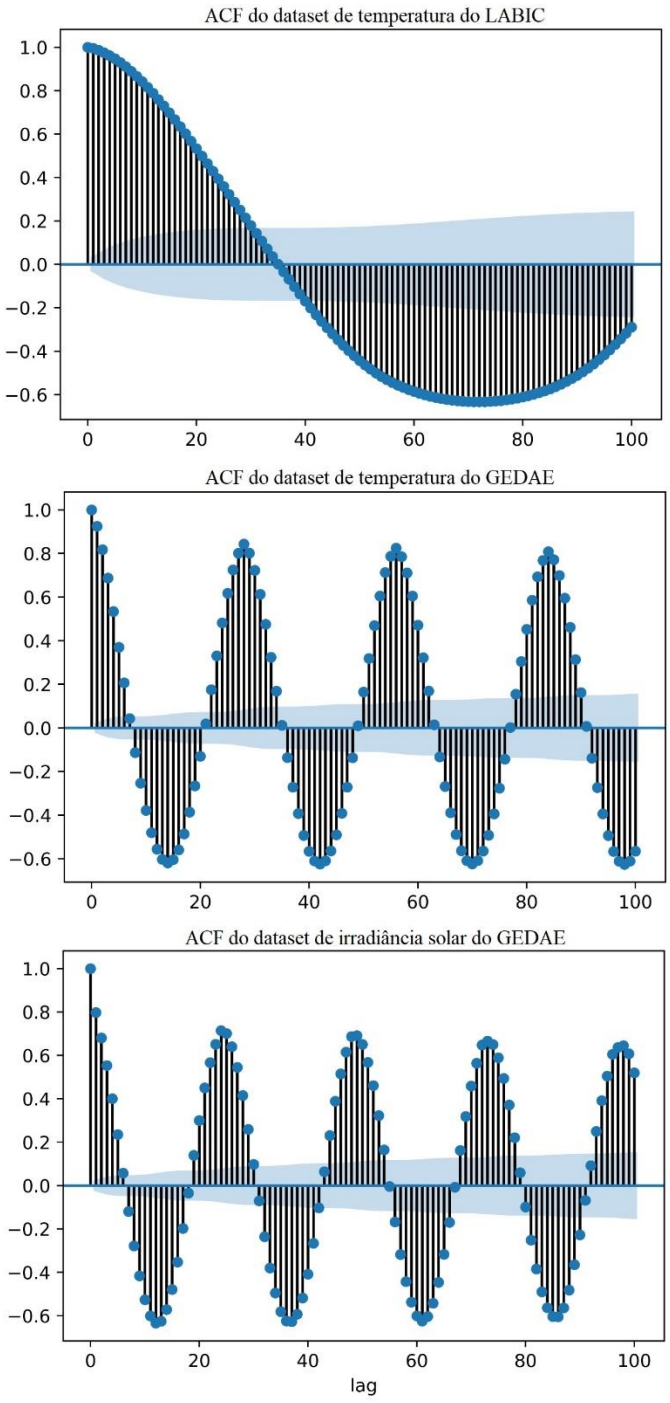
O teste post hoc de Nemenyi (Barrow et al., 2013) foi utilizado para determinar quais grupos de algoritmos possuem diferenças estatisticamente significativas. Neste teste, se a diferença no ranking médio entre dois algoritmos for superior a uma diferença crítica o teste indicará qual deles obteve melhor performance com base em seus rankings com nível de confiança α.

5. RESULTADOS

5.1. Análise Exploratória dos Conjuntos de dados

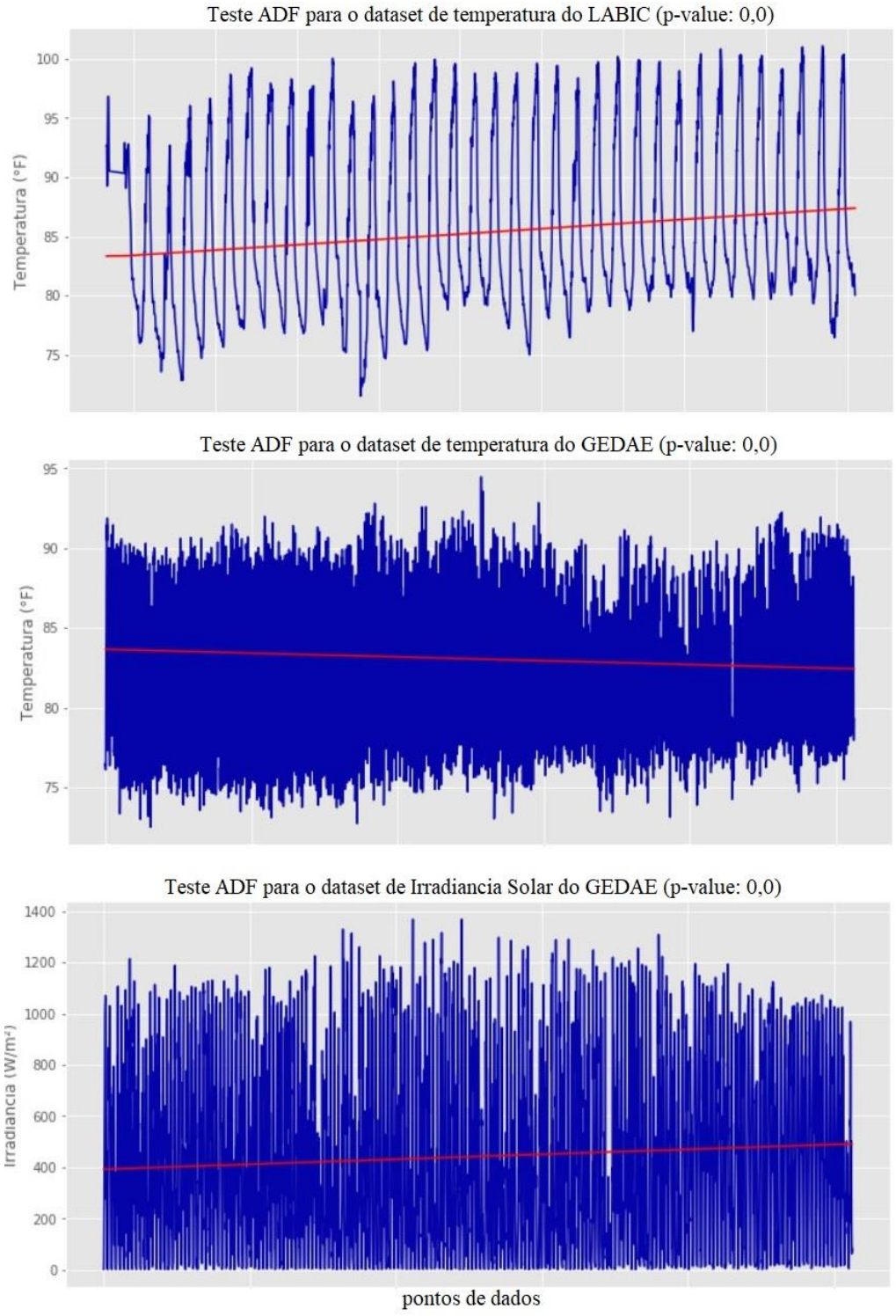
A Figura 8 apresenta o resultado dos testes da ACF para os 3 conjuntos de dados deste trabalho, em que se pode observar a dependência dos dados ao longo do tempo, onde os dados que estão fora da zona azul clara possuem correlações. Os conjuntos de Temperatura e Irradiância Solar do GEDAE apresentam ciclos sazonais de dependências positivas e negativas mais curtos, enquanto o conjunto de dados de Temperatura do LABIC apresenta um ciclo de dependência positivas e negativas mais longos em seus dados. O teste ADF (Figura 9) apresentou p-valor < 0,05, para os 3 conjuntos de dados, em que se assume que as séries temporais analisadas são estacionárias.

Figura 8. Plot dos testes de ACF para os 3 conjuntos de dados utilizados neste estudo.



Fonte: Autor.

Figura 9. Plot dos testes aumentado de Dickey-Fuller para os 3 conjuntos de dados utilizados neste estudo.



Fonte: Autor.

A Tabela 2 apresenta a correlação entre a temperatura do ar e a irradiância solar para os conjuntos de dados do GEDAE, onde observamos que há uma correlação positiva entre as variáveis.

Tabela 2. Correlação entre as variáveis ambientais dos conjuntos de dados do GEDAE.

	Irradiância solar	Temperatura
Temperatura	0,5782	1
Irradiância solar	1	0,5782

Fonte: Autor.

5.2. Conjunto de Treino e Teste

5.2.1. Normalizações 0,1 a 0,9 e 0 a 1

As Tabelas 3 e 4 exibem as médias das métricas obtidas pelos 10 modelos salvos pela função *TimeSeriesSplit* de cada uma das três arquiteturas para os três conjuntos de dados, para normalizações de 0,1 a 0,9 (grupo n1) e de 0 a 1 (grupo n2), respectivamente. Na normalização do grupo n1, dentre todas as médias das arquiteturas o modelo de 3 Camadas obteve menores valores de métricas no conjunto de dados de temperatura do LABIC, com 0,0230, 5,0472 e 0,0298 para MAE, MAPE e RMSE, respectivamente. Já na comparação entre arquiteturas por conjunto de dados os menores valores foram obtidos para as arquiteturas com 3 camadas. O mesmo comportamento se deu nos resultados dos dados do grupo n2. Entretanto, os resultados desta normalização obtiveram maiores valores nas métricas de erros, comparados com os da Tabela 3.

Tabela 3. Média das métricas do grupo n1, obtidas na etapa de treino-teste.

Conjunto de dados	Arquiteturas	MAE	MAPE	RMSE	Tempo (s)
LABIC Temperatura	LSTM-3L	0,0230	5,0472	0,0298	1029,32
	LSTM-5L	0,0296	6,4962	0,0369	1809,13
	LSTM-7L	0,0346	6,9965	0,0030	2693,58
GEDAE Temperatura	LSTM-3L	0,0411	10,7519	0,0600	1997,35
	LSTM-5L	0,0417	10,9129	0,0614	3479,71
	LSTM-7L	0,0453	11,6625	0,0635	5089,51
GEDAE Irradiância Solar	LSTM-3L	0,0831	28,6955	0,1124	1786,55
	LSTM-5L	0,0851	30,4020	0,1152	3063,71
	LSTM-7L	0,0883	30,0491	0,1192	4650,80

Fonte: Autor.

Tabela 4. Média de MAE e RMSE para os resultados do grupo n2, obtidas na etapa de treino e teste.

Conjunto de dados	Arquiteturas	MAE	RMSE
LABIC Temperatura	LSTM-3L	0,0265	0,0363
	LSTM-5L	0,1136	0,1539
	LSTM-7L	0,1033	0,1443
GEDAE Temperatura	LSTM-3L	0,0459	0,0685
	LSTM-5L	0,0928	0,1299
	LSTM-7L	0,1017	0,1380
GEDAE Irradiância Solar	LSTM-3L	0,1002	0,1397
	LSTM-5L	0,1433	0,1809
	LSTM-7L	0,1510	0,1869

Fonte: Autor.

A Tabela 5 apresenta os resultados do teste de *Wilcoxon* para as métricas de MAE e RMSE entre os resultados obtidos nas Tabelas 3 e 4, onde podemos observar que há uma diferença estatisticamente significativa entre os resultados obtidos para diferentes intervalos de normalização. Sendo assim, como o grupo n1 obteve menores valores nas métricas de erros e apresentou diferença estatisticamente significativa para o grupo n2, este foi utilizado para os resultados adiante.

Tabela 5. Resultados do teste de Wilcoxon para as duas normalizações diferentes.

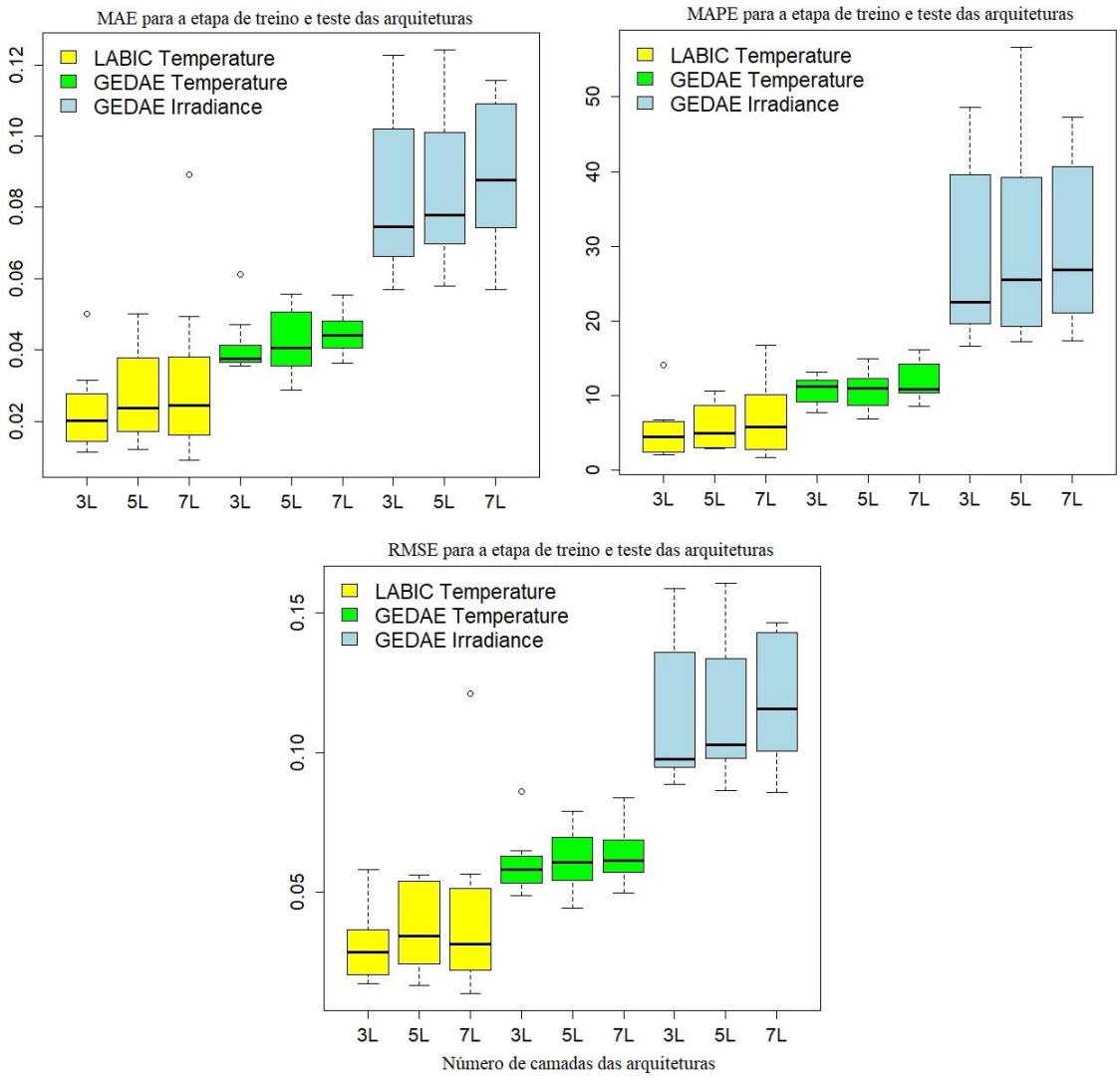
	MAE	RMSE
p-valor do teste de *Wilcoxon*	0,0078	0,0056

Fonte: Autor.

Os tempos de execução de cada arquitetura (Tabela 3) levam em consideração o total de execuções da validação cruzada pela *TimeSeriesSplit*. Com a menor quantidade de dados de entrada as execuções do conjunto de dados LABIC Temperature foram mais rápidas. Na comparação por arquiteturas entre conjuntos de dados, as de 3 camadas levaram menor tempo de execução. Neste caso, as arquiteturas de 7 camadas além de levarem mais tempo de processamento ainda obtiveram maiores médias em suas métricas de erro.

A figura 10 apresenta os boxplots das métricas obtidas por cada modelo de cada arquitetura para os três conjuntos de dados. Através da sua observação em conjunto com a Tabela 3 pode-se observar que as arquiteturas tiveram maior dificuldade na previsão da irradiância solar. Na avaliação entre camadas por conjunto de dados, os modelos de LSTM-3L obtiveram menores valores nas métricas, seguidas LSTM-5L e LSTM-7L, respectivamente.

Figura 10. Boxplots para os resultados do grupo n1 de MAE, MAPE e RMSE dos modelos na etapa de treino-teste



Fonte: Autor.

Como podemos observar os resultados obtidos pelas arquiteturas foram próximos quando comparado suas métricas de erro, tendo sua maior diferença no tempo de processamento. Para verificar se há diferença estatísticamente siginificativa entre as arquiteturas foi utilizado o teste

de Kruskal-Wallis, levando em consideração os 10 modelos gerados para cada arquitetura. A Tabela 6 apresenta os resultados do teste comparando as métricas de MAE, MAPE e RMSE, onde podemos observar que a hipótese nula do teste foi aceita a um nível de confiança de 95%, significando que não há diferença estatisticamente significativa entre as arquiteturas.

Tabela 6. Resultados do teste de Kruskal-Wallis comparando as 3 arquiteturas por métricas do grupo n1.

Conjuntos de Dados	MAE (p-value)	MAPE (p-value)	RMSE (p-value)
LABIC Temperature	0,5066	0,4173	0,4787
GEDAE Temperature	0,2938	0,7806	0,6796
GEDAE Irradiância Solar	0,7506	0,8599	0,6429

Fonte: Autor.

5.3. Conjunto de Validação

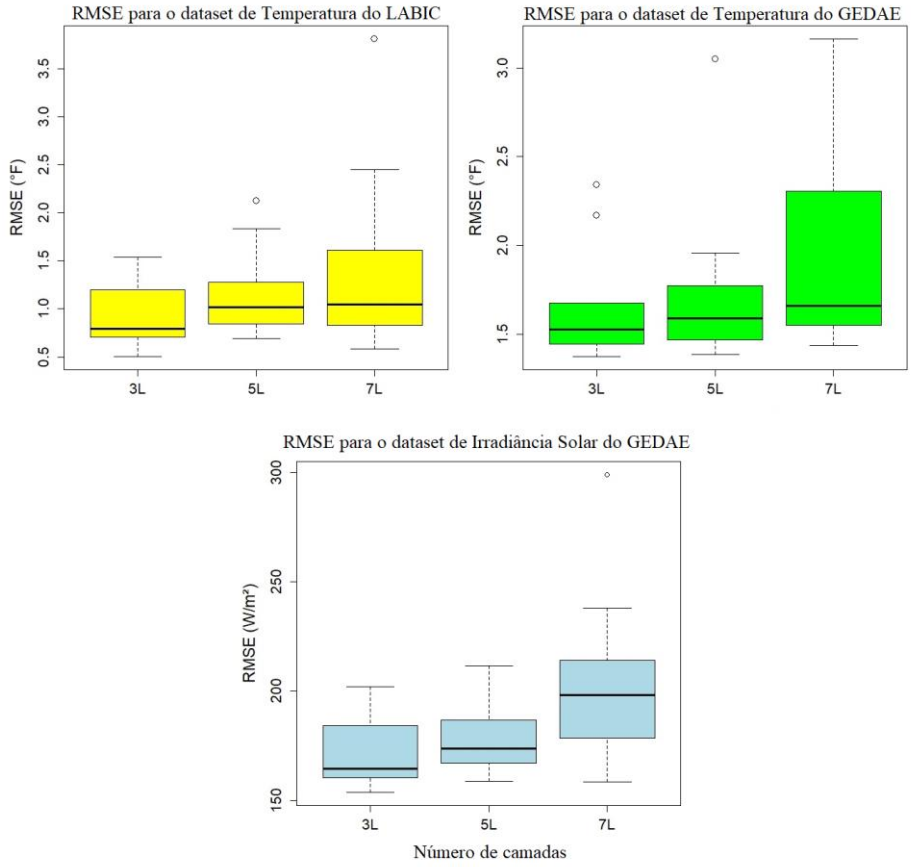
Esta etapa consiste no resultado da aplicação dos 20% de dados para validação dos modelos. A Tabela 7 apresenta os parâmetros estatísticos obtidos por meio dos resultados dos modelos, gerados na etapa de treino e teste, submetidos ao conjunto de dados de validação, sendo: média, desvio padrão, máximo, mínimo e mediana. Estes resultados são provenientes dos valores desnormalizados, por meio da função *inverse_transform* da biblioteca do *sklearn*, dos dados reais e dados previstos para a métrica RMSE, uma vez que esta métrica possui um bom indicativo de o quanto os dados previstos pelos modelos estão próximos dos dados reais. Ao observar a Tabela 4, temos que a arquitetura *LSTM-3L* obteve menores valores em todos os parâmetros estatísticos, em comparação a *LSTM-5L* e *LSTM-7L*. A figura 11 apresenta os boxplots dos resultados obtidos, em que se observa que a diferença entre as arquiteturas não parece tão evidente, entretanto, a arquitetura *LSTM-7L* apresentou maiores amplitudes.

Tabela 7. Parâmetros estatísticos dos modelos gerados pela *TimeSeriesSplit* aplicados no conjunto de dados de validação.

Conjunto de dados	*Model*	LSTM-3L	LSTM-5L	LSTM-7L
LABIC Temperatura	Média	0,9393	1,1530	1,4531
	Desv. Padrão	0,3489	0,4475	0,9299
	Máx	1,5417	2,1256	3,8108
	Mín	0,5028	0,6953	0,5871
	Mediana	0,7964	1,0200	1,0491
GEDAE Temperatura	Média	1,6499	1,7563	1,9767
	Desv. Padrão	0,3169	0,4629	0,5925
	Máx	2,3410	3,0494	3,1642
	Mín	1,3739	1,3849	1,4375
	Mediana	1,5279	1,5914	1,6615
GEDAE Irradiância Solar	Média	170,6649	178,3429	204,7825
	Desv. Padrão	14,7964	15,9513	38,7843
	Máx	202,1510	211,4770	299,1990
	Mín	153,5882	158,7389	158,4912
	Mediana	164,3430	173,5689	198,1844

Fonte: Autor.

Figura 11. Boxplots dos resultados de RMSE obtidos para o conjunto de validação.



Fonte: Autor.

Com objetivo de determinar se houve diferença estatisticamente significativa entre as arquiteturas utilizou-se o teste de Friedman, que leva em consideração as arquiteturas para todos os três conjuntos de dados. Com $\alpha = 5\%$, o resultado do teste foi de p-valor = 0,049, assim, existem evidências de que os resultados das arquiteturas possuem diferença estatísticamente significativa em pelo menos uma arquitetura.

Para conhecer quais pares de arquiteturas divergiram foi aplicado o teste *post hoc* de comparação múltipla de Nemenyi. A Tabela 8 apresenta o resultado do teste, onde temos que há diferença significativa na entre as arquiteturas de 3 camadas e 7 camadas com p-valor = 0,038. Já para as comparações de 3 e 5 camadas e 5 e 7 camadas não houve evidências de diferença estatísticamente significativa, ambas com p-valor = 0,438.

Tabela 8. Resultados do teste *post hoc* de Nemenyi.

	LSTM-3L	LSTM-5L
LSTM-5L	0,438	-
LSTM-7L	0,038	0,438

Fonte: Autor.

A Tabela 9 apresenta os resultados normalizados da métrica RMSE dos 10 modelos gerados para cada arquitetura, com um total de 90 modelos.

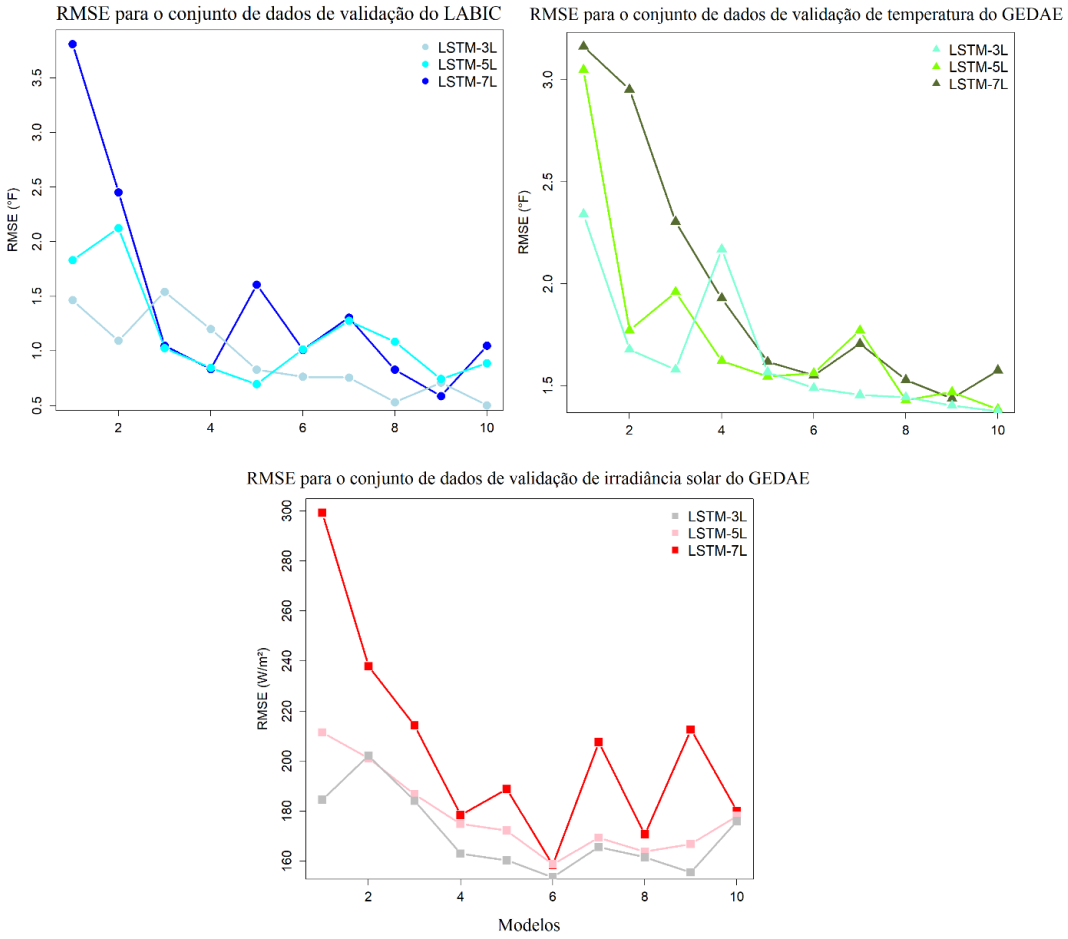
Tabela 9. Resultados normalizados de RMSE dos modelos para a etapa de validação.

Conjunto de dados	Modelo	*LSTM-3L*	*LSTM-5L*	*LSTM-7L*
LABIC Temperatura	1	0,0480	0,0592	0,1241
	2	0,0361	0,0693	0,0800
	3	0,0500	0,0332	0,0346
	4	0,0387	0,0283	0,0265
	5	0,0265	0,0224	0,0520
	6	0,0245	0,0332	0,0332
	7	0,0245	0,0412	0,0424
	8	0,0173	0,0346	0,0265
	9	0,0224	0,0245	0,0200
	10	0,0173	0,0283	0,0346
GEDAE Temperatura	1	0,1039	0,1356	0,1407
	2	0,0748	0,0787	0,1311
	3	0,0700	0,0872	0,1025
	4	0,0964	0,0721	0,0860
	5	0,0693	0,0686	0,0721
	6	0,0663	0,0693	0,0686
	7	0,0648	0,0787	0,0755
	8	0,0640	0,0632	0,0678
	9	0,0624	0,0656	0,0640
	10	0,0608	0,0616	0,0700
GEDAE Irradiancia solar	1	0,1158	0,1327	0,1876
	2	0,1269	0,1261	0,1493
	3	0,1153	0,1170	0,1345
	4	0,1025	0,1095	0,1118
	5	0,1005	0,1082	0,1183
	6	0,0964	0,0995	0,0995
	7	0,1039	0,1063	0,1304
	8	0,1015	0,1030	0,1072
	9	0,0975	0,1049	0,1334
	10	0,1105	0,1118	0,1127

Fonte: Autor.

Nos conjuntos de dados de temperatura dentre os modelos que obtiveram menor valor foram os da última etapa do *TimeSeriesSplit*, na *LSTM-3L*, são modelos que mais tiveram ajustes dos pesos e também maior quantidade de dados no treino, com 0,0173 no conjunto de dados LABIC e 0,0608 no conjunto de dados GEDAE. Já para a irradiância o modelo 6 apresentou menores valores dentre as 3 arquiteturas utilizadas com 0,0964, também na *LSTM-3L*. A Figura 12 apresenta os gráficos de linhas com os valores de RMSE dos conjuntos de dados de temperatura LABIC e GEDAE e irradiancia solar do GEDAE, onde observamos que os modelos menos treinados possuem uma tendencia a ter um desempenho pior, não necessariamento indica que um modelo mais treinado possua melhor desempenho. Em média *LSTM-3L* obeteve menores valores da métrica de erro.

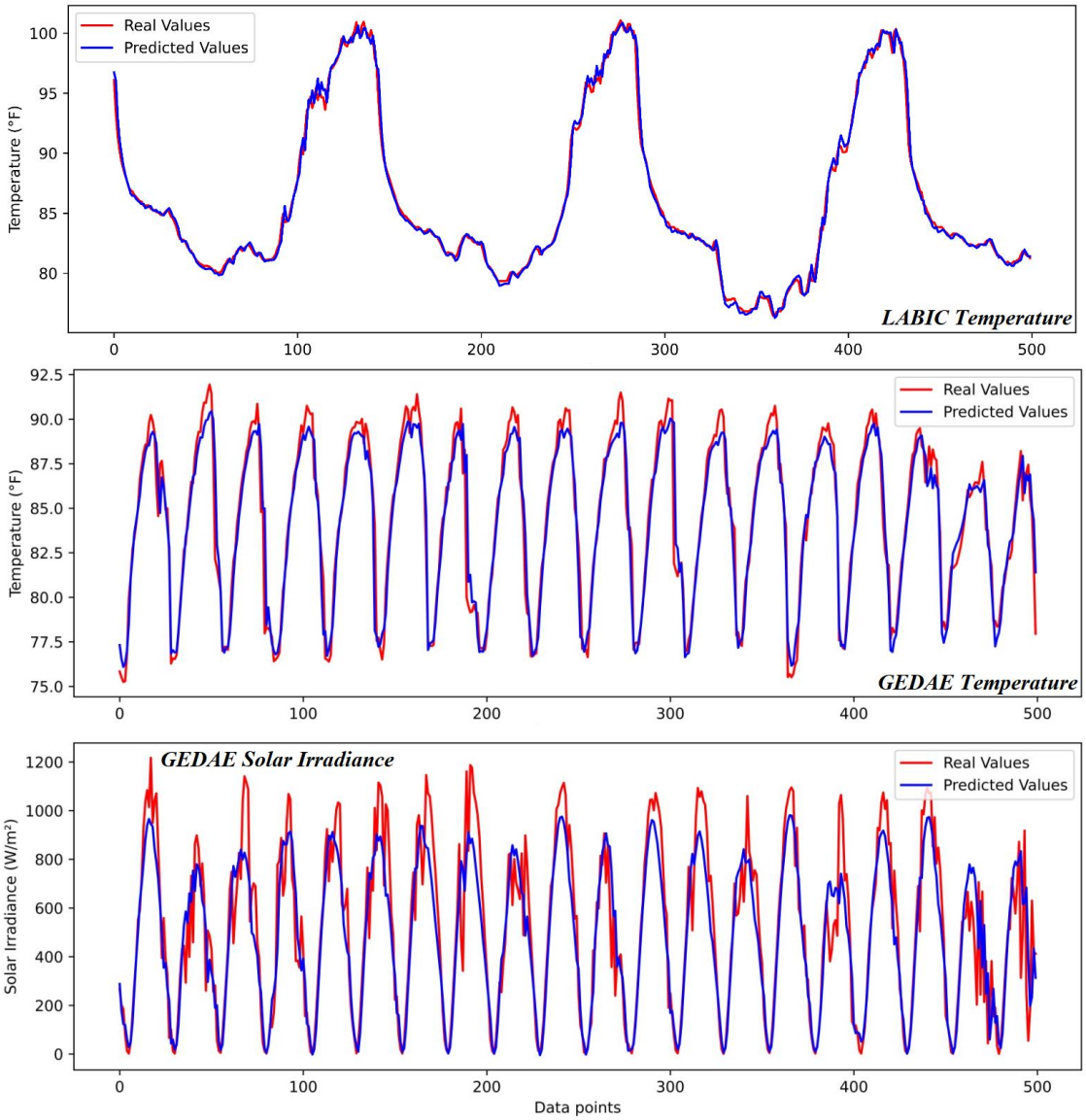
Figura 12. Gráficos de linhas dos modelos obtidos na etapa de treino-teste aplicados na validação com dados desnormalizados na métrica RMSE.



Fonte: Autor.

A Figura 13 apresenta os últimos 500 pontos dos dados do conjunto de dados de validação, para melhorar a visualização dos dados, em comparação com os modelos que obtiveram menor valor na métrica RMSE citados anteriormente. Na comparação com o conjunto de dados do LABIC, houve praticamente uma sobreposição total com os resultados da previsão sobre a temperatura real, os valores previstos acompanharam os valores de máximos e mínimos da série temporal. No conjunto de dados de temperatura GEDAE os dados previstos seguiram bem a variação dos dados reais, porém o modelo não acompanhou bem os valores máximos e mínimos. Já para os dados de irradiância solar os valores de previsão não seguiram as máximas alcançadas pelos valores reais, dentre os 3 conjuntos de dados este foi o que menos sobrepôs os valores reais, entretanto, o modelo acompanhou bem as variações da irradiância solar.

Figura 13. Plot dos dados de validação e dados reais para os conjuntos de dados de Temperatura do LABIC e GEDAE, e o conjunto de dados de Irradiância Solar do GEDAE.



Fonte: Autor.

Nos gráficos do conjunto de dados GEDAE, os pontos de dados do resultado não foram sobrepostos com os dados atuais como no conjunto de dados LABIC. Considera-se que os dados do LABIC consistem em pouco mais de um mês de dados, em um período conhecido como "verão amazônico" onde as temperaturas são mais elevadas, com aumento do número de incêndios e escassez de chuvas, enquanto o conjunto de dados do GEDAE mostra toda a variação sazonal durante o período de um ano. Nota-se também que os dados LABIC possuem maior dependência de seus dados quando comparados aos dados GEDAE, o que pode estar

relacionado ao seu melhor desempenho, o que não compromete a qualidade das previsões no conjunto de dados GEDAE.

6. CONCLUSÕES

Neste trabalho, diferentes arquiteturas de RNA LSTM foram utilizadas para a previsão e comparação dos resultados de temperatura e irradiância solar na região amazônica. Os dados ambientais foram divididos em dados de treino-teste e validação.

Considerando que os grupos de normalizações n1 e n2 apresentaram diferença estatisticamente significativa e ainda que, o grupo n1 obteve menores valores de métricas de erro, temos que a normalização entre 0,1 e 0,9 obteve melhor performance para este trabalho, em concordância com os resultados obtidos por Ak et al. (2015).

Para o mesmo número de épocas, a arquitetura LSTM-3L obteve menores médias nos erros das métricas utilizadas. Na avaliação do conjunto de dados de validação ao aplicarmos testes estatísticos, encontramos uma diferença estatisticamente significativa entre as arquiteturas LSTM-3L e LSTM-7L, onde obtivemos, para o conjunto de dados de temperatura LABIC, valores mais baixos de RMSE médio: 0,9393 e 1,4531 para LSTM-3L e LSTM-7L, respectivamente. Desta forma temos que a arquitetura LSTM-3L possui melhor desempenho tanto na previsão quanto em processamento devido ao reduzido tempo de execução.

Com os resultados obtidos podemos determinar que o método LSTM possui boa resposta na previsão das variáveis ambientais utilizadas neste estudo para a região amazônica.

Para trabalhos futuros, a possibilidade de aplicação da metodologia utilizada neste trabalho para avaliar outros modelos de previsão, ou até mesmo modelos híbridos da LSTM como o CNN-LSTM e MRC-LSTM, além da incorporação destas arquiteturas a um sistema com IoT, como o EnergySaver para monitoramento, análise e previsão de séries temporais em tempo real.

REFERÊNCIAS BIBLIOGRÁFICAS

AL-SALAYMEH, A. **Model for the prediction of global daily solar radiation on horizontal surfaces for Amman city**. Emirates Journal of Engineering Research, v.11, 2006.

ALMEIDA, T. do N. S.; ESCOBEDO, J. F.; OLIVEIRA, A. P. de; SOARES, J. Análise climática anual das radiações global, direta e difusa de botucatu/sp. **XVII Congresso Brasileiro de Agrometeorologia**, 2011.

AGGARWAL, S.K.; SAINI, L.M. **Solar energy prediction using linear and non-linear regularization models: A study on AMS (American Meteorological Society) 2013–14 Solar Energy Prediction Contest**. Energy. Elsevier, vol. 78(C), pages 247-256. 2014.

BORGES, V. P.; OLIVEIRA, A. S. DE; COELHO FILHO, M. C., SILVA, T. S. M. DA; PAMPONET, R. M. **Avaliação de modelos de estimativa da radiação solar incidente em Cruz das Almas, Bahia**. Revista Brasileira de Engenharia Agrícola e Ambiental, v.14, p.74–80, 2010.

BOUKTIF, S., FIAZ, A., OUNI, A., SERHANIM M. A. **Optimal Deep Learning LSTM Model for Electric Load Forecasting using Feature Selection and Genetic Algorithm: Comparison with Machine Learning Approaches**. Energies. 2018, 11, 1636; doi:10.3390/en11071636.

BOX, G., JENKINS, G. (1970) **Time Series Analysis: Forecasting and Control**. Holden-Day, San Francisco.

CASTELÃO, R. **Utilização de Redes Neurais para Previsões no Mercado de Ações**. Projeto final de graduação. Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP. 2018.

CAVALCANTI, E. P.; SILVA, V. de P. R.; SOUSA, F. de A. S. de. **Programa computacional para a estimativa da temperatura do ar para a região Nordeste do Brasil**. Revista Brasileira de Engenharia Agrícola e Ambiental, 2006.

DANGETI, P. **Statistics for Machine Learning: Techniques for Exploring Supervised, Unsupervised, and Reinforcement Learning Models with Python and R**. Packt Publishing: Birmingham, UK, 2017.

DAS, U. K., TEY, K. S., SEYEDMAHMOUDIAN, M., MEKHILEF, S., IDRIS, M. Y I., DEVENTER, W. V., HORAN, B., STOJCEVSKI, A. **Forecasting of photovoltaic power generation and model optimization: A review. Renewable and Sustainable Energy Reviews**, Volume 81, Pages 912-928, ISSN 1364-0321, 2018.

DMITRIENKO, A., CHRISTY C., RALPH D. **Pharmaceutical Statistics Using SAS® : A Practical Guide**. Cary, NC: SAS Institute Inc. 2007.

DUNN, O. J. **Multiple comparisons using rank sums**. Technometrics 6: 241–252.

EICKER, U., DEMIR, E., GÜRLICH, D. **Strategies for cost efficient refurbishment and solar energy integration in European Case Study buildings**. Energy Build. 102, 237-249. 2015.

ELMAN, J. **Finding Structure in Time**. Cognitive Science. 1990.

ERKAL, C., CECEN, A. **Empirical Fokker-Planck-based test of stationarity for time series**. American Physical Society - PHYSICAL REVIEW E 89, 062907. 2014. DOI: 10.1103/PhysRevE.89.062907

FAUCHER, R.; PHYSIQUE, **Hatier**. Paris, 1966.

FRIEDMAN, M. (1937) **The use ofranks to avoid the assumptions of normality implicit in the analysis of variance**. J. Amer. Statist. Assoc., 32, 675-701.

GNOATTO, E.; DALLACORT, R.; RICIERI, R. P.; SILVA, S. L.; FERRUZI, Y. **Eficiência de um conjunto fotovoltaico em condições reais de trabalho na região de Cascavel**. Acta Sci. Technol. Maringá, v. 30, n. 2, p. 215-219, 2008.

HAYKIN, S. Neural networks: **A comprehensive Foundation**. Segunda edição. Pearson Education. 2005.

Hochreiter, J. **Untersuchungen zu dynamischen neuronalen Netzen**. Diploma thesis, Institut fur Informatik, Lehrstuhl Prof. Brauer, Technische Universit at Munchen. 1991.

HOCHREITER, S; SCHMIDHUBER, J. **Long Short-Term Memory**. Neural Computation, 1997.

HOPPE, E.; **Histoire de la Physique**, Payot: Paris, 1928.

INÁCIO, T. **Potencial solar das radiações global, difusa e direta em Botucatu**. 2009. 84p. Dissertação (Agronomia). Faculdade de Ciências Agronômicas da UNESP – Campus de Botucatu. Botucatu – SP. 2009.

IQBAL, M. 1978. **Hourly vs daily method of computing isolation on inclined surfaces**. Solar Energy, v. 21, p. 485-489, 1978.

JUNIOR, J. F. da S.; **Estratégias para Analisar e Estimar a Eficiência do Consumo de Energia em Centros de Dados**. Programa de Pós-Graduação em Ciências da Computação da Universidade Federal de Pernambuco. 2019.

KARA, A. **Global Solar Irradiance Time Series Prediction Using Long Short-Term Memory Network**. Fen Bilimleri Dergisi. GU J Sci, Part C. 2019.

KAREVAN, Z.; SUYKENS, J. A. K. **Spatio-temporal Stacked LSTM for Temperature Prediction in Weather Forecasting**. arXiv:1811.06341v1 [cs.LG]. 2018.

KLEISSL, J. **Solar Energy Forecasting and Resource Assessment**. Elsevier Inc. 2013.

KOTTI, M.C.; ARGIRIOU, A.A.; KAZANTZIDIS, A. **Estimation of direct normal irradiance from measured global and corrected diffuse horizontal irradiance**. Energy, v. 70, n. 1, p. 382-392, 2014.

LATIMER, J. R. **Radiation Measurement. National Researrch Council of Canada**. International Field Year for the Great Lakes, Toronto, 1971.

LEWIS, N. D. **Deep Time Series Forecasting With Python: An Intuitive Introduction to Deep Learning for Applied Time Series Modeling**. N.D. Lewis. 2016.

LI, Y.; CAO, H. **Prediction for Tourism Flow based on LSTM Neural Network**. Procedia Computer Science, 2018.

MACEDO, H. C. M.; **Físico-Química I**, Guanabara Dois: Rio de Janeiro, 1981

MARTINS, P. A. da S. **Verificação da turbidez atmosférica em Humaitá-AM**. Revista EDUCAmazônia - Educação Sociedade e Meio Ambiente, Humaitá - Ano 7, Vol XII, Número 1, Jan-Jun, 2014.

MICHELS, R. N.; GNOATTO, E.; SANTOS, J. A. A.; KAVANAGH, E.; HALMEMAN, M. C. **A influência da temperatura na eficiência de painéis fotovoltaicos em diferentes níveis de incidência da radiância solar**. Revista Agrogeoambiental. 2010.

MIDDLETON, W. E. K. **Meteorological Instruments**. University of Toronto, Toronto, 1943.

MONTGOMERY, D. C., JENNINGS, C.L. **Introduction to Time Series Analysis and Forecasting**. Wiley. 2015.

PARZEN, E. **An Approach to Time Series Analysis**. The Annals of Mathematical Statistics Vol. 32, No. 4. 1961.

PAVÃO, V. M. **Efeitos da correção atmosférica em imagens landsat 8 e diferentes modelos de radiação solar global na estimativa do saldo de radiação superficial**. Universidade Federal de Mato Grosso, 2016.

PEÑA, D., TIAO, G. C., TSAY, R. R. **A Course in Time Series Analysis**. Wiley. 2001.

QING, X.; NIU, Y. **Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM**. Energy, 2018.

QUERINO, C. A. S.; MOURA, M.A.L.; QUERINO, J.K.A.S.; VON RADOW, C.; MARQUES FILHO, A.O. **Estudo da radiação solar global e do índice de transmissividade (kt), externo e interno, em uma floresta de mangue em Alagoas – Brasil**. Revista Brasileira de Meteorologia, v.26, n.2, p. 204 – 294, 2011.

QUERINO, C. A. S.; MOURA, M.A.L.; LYRA, R.F.F.; MARIANO, G.L.; **Avaliação e comparação de radiação solar global e albedo com ângulo zenital na região Amazônica**. Revista Brasileira de Meteorologia, v. 21, n.3a, p. 42 – 49, 2006.

ROSENBERG, N. J**. Microclimate: the biological environmet**. New York: J. Wiley & Sons, 1974.

SLATER, P. N. **Remote sensing, optics and optical systems**. Massachussets: Addison-Wesley, 1980.

SOUSA, A. F. G., FURTADO, H. C. M., MACÊDO, W. N., MENESES, A. A. de M. **Analysis of Artificial Neural Network point forecasting models and Prediction Intervals for solar irradiance estimation**. American Journal of Engineering and Applied Sciences. 2020, 13 (3): 347.357 DOI: 10.3844/ajeassp.2020.347.357

VASSALI, L. C.; **Aplicação de Redes Neurais LSTM para previsão de curto prazo de vazão do Rio Paraíba do Sul**. Projeto final de graduação. Faculdade de Engenharia, Universidade Federal de Juiz De Fora. 2018.

VAREJÃO, M. **METEOROLOGIA E CLIMATOLOGIA** – Recife, 2005.

XU, C.; CHEN, H.; WANG, J.; GUO, Y.; YUAN, Y**. Improving prediction performance for indoor temperature in public buildings based on a novel deep learning method**. Building and Environment. Volume 148, 2019.

ZHANG, J.; ZHU, Y.; ZHANG, X.; Y, M.; **Developing a Long Short-Term Memory (LSTM) based Model for Predicting Water Table Depth in Agricultural Areas**. Journal of Hydrology. 2018

ZHANG, X.; ZHANG, Q.; ZHANG, G.; NIE, Z.; GUI, Z.; QUE, H. **A Novel Hybrid Data-Driven Model for Daily Land Surface Temperature Forecasting Using Long Short-Term Memory Neural Network Based on Ensemble Empirical Mode Decomposition**. *Int.* J. Environ. Res. Public Health 2018.

ZHOU, Z. **Measuring nonlinear dependence in time-series, a distance correlation approach**. Journal of Time Series Analysis - J TIME SER ANAL. 2012. 33. 10.1111/j.1467-9892.2011.00780.x.