PHOTOVOLTAIC ENERGY PRODUCTION FORECASTING USING LSTM AND CROSS-VALIDATION

Nairon Augusto Monari Gonçalves[1], Nicolas Fourmaux[2], Antonio Cesar Germano Martins[3]

[1]*ICTS – UNESP, Sorocaba, Sao Paulo, Brazil, nairon.goncalves@unesp.br*

[2]*CESI École d'ingénieurs – campus d'Arras, Arras, Pas-de-Calais, Hauts-de-France, France*

[3]*ICTS – UNESP, Sorocaba, São Paulo, Brazil*

ABSTRACT

The recent water shortage faced in Brazil and being the seventh largest in terms of population size according to the United Nations (UN), requires efficient and sustainable energy management strategies. Photovoltaic (PV) energy is a promising renewable source to meet this demand, but accurately forecasting its production remains a critical challenge. This study proposes the implementation of Long Short-Term Memory (LSTM) neural networks, a variant of recurrent neural networks (RNNs), for reliable prediction of photovoltaic production. To ensure the robustness and reliability of the predictive model, this work will use the Cross-validation statistical method. This technique helps validate model performance by dividing the dataset into training and testing sets, thereby providing insights into its generation and predictive capabilities. This is crucial to help energy suppliers in their decision-making processes and improve the regulation of photovoltaic production.

RESUMO

A recente escassez hídrica enfrentada no Brasil, e sendo o sétimo maior em termos de tamanho populacional segundo a Organização das Nações Unidas (ONU), exige estratégias de gestão de energia eficientes e sustentáveis. A energia fotovoltaica (PV) é uma fonte renovável promissora para satisfazer esta procura, mas a previsão precisa da sua produção continua a ser um desafio crítico. Este estudo propõe a implementação de redes neurais *Long Short-Term Memory* (LSTM), uma variante das redes neurais recorrentes (RNNs), para previsão confiável da produção fotovoltaica. Para garantir a robustez e confiabilidade do modelo preditivo, este trabalho utilizará o método estatístico *Cross-validation*. Esta técnica ajuda a validar o desempenho do modelo, dividindo o conjunto de dados em conjuntos de treinamento e teste, fornecendo assim *insights* sobre sua geração e capacidades preditivas. Isto é crucial para ajudar os fornecedores de energia nos seus processos de tomada de decisão e melhorar a regulação da produção fotovoltaica.

Keywords: Cross-validation. Forecasting. LSTM. Photovoltaic power systems. Renewable energy sources

1. INTRODUCTION

The constant growth of energy consumption is a global challenge, especially in populous and industrially developing nations like Brazil that has a vast geographic territory

and considerable population making it one of the highest energy consumers globally, highlighting the urgent need for sustainable energy solutions to meet escalating demand.

Renewable energy, particularly photovoltaic (PV) energy, is a promising solution to Brazil's energy demands. PV systems convert solar energy into electricity, presenting an environmentally friendly alternative (GUTIÉRREZ et al., 2021). Accurately predicting PV production is crucial for effective energy management and integration into the grid due to solar radiation's intermittent (LUO et al., 2021).

Brazil has a significant potential for extensive electricity generation through photovoltaic systems (ZAMAN et al., 2021). The notable growth in solar energy adoption reflects the nation's response to water scarcity challenges, encouraging a shift towards other sustainable energy sources. Anticipating 2025, a reliable predictive model for daily energy production becomes pivotal for efficient regulation and structuring incentives for energy producers striving to meet predetermined targets (URBANETZ et al., 2021).

Emphasizing increased reliance on solar sources aligns with the 7th UN Sustainable Development Goal (SDG 7), aiming for universal access to reliable, sustainable, modern, and affordable energy. The rising adoption of photovoltaic systems in recent years, especially in smart structures, underlines the need for accurate prediction of photovoltaic generation for seamless integration into existing energy systems (DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS, UN, 2023).

Accurate estimation of photovoltaic energy production poses a challenge due to the stochastic nature of these systems, relying on uncertain and dynamic climate variables (NESPOLI et al., 2019). Inaccuracies in estimation can jeopardize grid stability, efficiency, and security. Leveraging LSTM neural networks, known for their ability to model time-series data and capture long-term dependencies, is a promising approach to predict complex energy generation patterns, especially in the context of photovoltaic production (ZAMAN et al., 2021).

This study proposes the application of LSTM neural networks to predict photovoltaic production precisely and accurately, aiming to aid energy suppliers in making supported decisions about energy distribution and allocation. The integration of the Cross-validation statistical method is used to validate and ensure the robustness of the predictive model. Specifically, in the context of varying seasons, the Cross-validation statistical model plays a

critical role when training an LSTM neural network for predicting photovoltaic energy production.

The Cross-validation statistical model implies partitioning the dataset into subsets for training and validation, enabling the LSTM neural network to learn and adapt to diverse patterns within the data. Given the cyclical and periodic nature of seasonal variations, this approach proves particularly valuable, allowing the model to effectively predict photovoltaic energy production across different seasons (ABEDINIA et al., 2021).

2. METHODOLOGY

2.1. Data Acquisition

The data for the present study was obtained from Solar Centre Desert Knowledge Australia—a research institution dedicated to solar energy (DKASC, 2023). This center has openly provided comprehensive data regarding its photovoltaic systems for over a decade.

The selected dataset originated from a monocrystalline silicon photovoltaic panel, manufactured by Trina Solar, boasting a capacity of 10.5kWp. This panel was situated at the Alice Springs complex in Australia. The dataset encompassed data from January 01, 2019, to December 31, 2021. There are no missing data points in the dataset, and the data was gathered at 5-minute intervals. Subsequently, data from September 13, 2023, was used, to validate the network.

2.2. Software Used

The LSTM neural network was developed, and the data was partitioned into training and testing sets using Colab (GOOGLE COLAB, 2023), a Python development platform hosted by Google, featuring Python language version 3.7.12. Within this platform, several essential libraries were employed, including Numpy, MatPlotLib, Pandas, Tensor Flow, Keras, Sklearn, and PyLab, to facilitate the model creation and analysis process.

2.3. LSTM Architecture

The RNN are a class of neural networks that have recurrent connections, allowing previous information to be maintained and influence future decisions. This type of architecture is widely used in sequence processing tasks, such as speech recognition, machine translation and text generation (GAO et al., 2019).

LSTM is a special type of RNN, first proposed by Hochreiter and Schmidhuber in 1997 (LI et al., 2020), it can handle complex temporal sequences and capture long-term dependencies in the data. Unlike traditional regression models, which have difficulty capturing long-term temporal relationships, LSTM is specifically designed to overcome this challenge (SHARMA et al., 2022). LSTM has a forget gate, an input gate, an update gate, and an output gate, contained in the same cell and generally defined by the following functions:

$$i_t = \sigma(W_i h_{t-1} + W_i h_t) \quad (1)$$

$$f_t = \sigma(W_f h_{t-1} + W_f h_t) \quad (2)$$

$$o_t = \sigma(W_o h_{t-1} + W_o h_t) \quad (3)$$

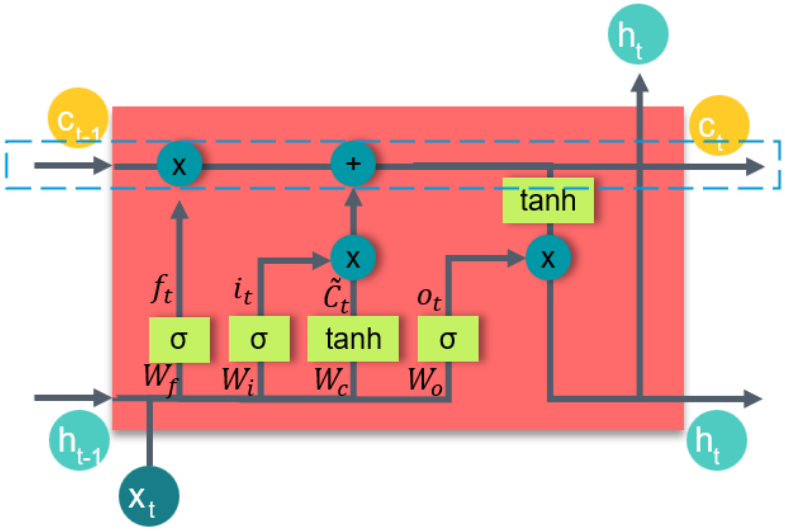
$$C = \tanh\,(W_c h_{t-1} + W_c h_t) \quad (4)$$

$$c_t = (i_t * C) + (f_t * c_{t-1}) \quad (5)$$

$$h_t = o_t * \tanh\,(c_t) \quad (6)$$

where the input function is fed by the input vectors x_t generating an output h_t, through modification by the network of the parameters W, U and σ, where W is the weight matrices of the equations, U is the number of neurons, and σ is the activation function as Figure 1.

Figure 1. LSTM Architecture



The construction of the neural network hinged on the Keras Layers library, employing a sequential model comprising an LSTM layer and a "dense layer" for result output. This design facilitated training exclusively through the specification of parameters outlined in Table 1. The crucial parameter was determined through meticulous testing during the training phase.

Table 1. Information for training LSTM

Neurons	Keras Default
Activation Function	Linear
Error Measure	Mean Squared Error (MSE)
Optimizer	Adam
Batch Size	64
Time Steps	4
Epochs	100
Early Stopping	10

2.4. Cross-validation

It is essential to understand that the temporal nature of data in time series prevents the application of a random training and testing split, as is often done in other types of data.

In time series, the order of data is crucial to understanding temporal relationships. Shuffling the data would destroy this order, resulting in a representation of the data that does not reflect the actual temporal structure. LSTM models, which are designed to capture temporal dependencies, need this order intact to learn and generalize properly.

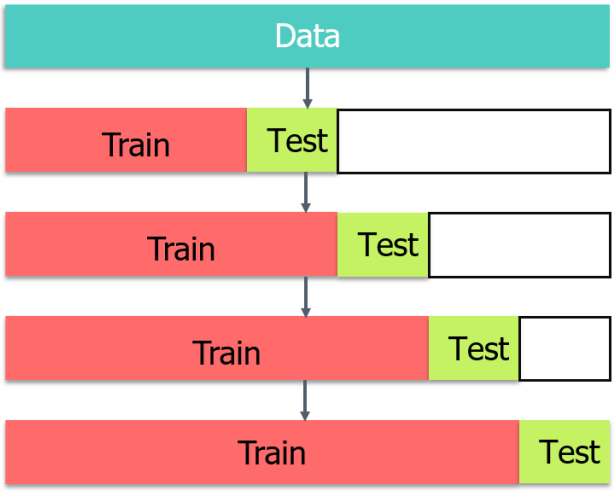
A way to overcome that is to split the data in large groups used for training and testing. By using this approach in the photovoltaic energy prediction, patterns associated to seasons of the year can be missed.

Another approach is the use of cross-validation techniques that maintains the temporal order of the data (ABEDINIA et al., 2021), ensuring, reflecting its effectiveness in predicting future events based on information available to date, without violating the fundamental premise of data temporality.

K-Fold is a specific cross-validation method that divides the dataset into k sequential subsets (or "folds", "chunks"). In each iteration, one subset is used as the test set and the previous subsets are used as the training set.

Starting with the first iteration, the LSTM model is trained with the data from the training subsets, which are arranged in temporal order as shown in Figure 2. Then, the model is evaluated with data from the test subset, which represents the next temporal sequence in the timeline. For the next training step, the test subset is incorporated into the training set. This process is repeated for each iteration, advancing sequentially throughout the time series (KONSTANTINOU et al., 2021).

Figure 2. Cross-validation K-Fold method



To evaluate the performance of the Cross Validation approach, a comparison was made with results obtained by dividing the data into larger groups.

2.5. Error measurement

Evaluation of the post-training network was carried out by analyzing the Root Mean Square Error (RMSE), calculated by their respective functions from the Sklearn Metrics library. The RMSE is a widely used metric to evaluate the performance of regression or prediction models. It is a variation of Mean Squared Error (MSE), however it is more intuitive, as it returns a measure of the average error in original units. It is especially useful for aa more direct idea of the magnitude of forecast errors. The RMSE formula is given in (7):

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (7)$$

where:

n = total number of data points in the test set.

y_i = actual value of data point i.

$\hat{y}_i$ = value predicted by the model for data point i.

Σ = indicates the sum of terms for all data points in the test set.

$\sqrt{}$ = represents the square root operation.

The RMSE error has some important characteristics:

- **Original units:** it returns an error measure in original units of the data, which makes it more interpretable. This is useful because one can directly understand the average size of prediction errors in relation to data units.
- **Sensitivity to outliers:** it is also sensitive to outliers in the data, because larger errors are amplified by the square root operation.
- **Interpretation:** a lower RMSE value indicates a better fit of the model to the real data, while a higher value indicates a less accurate fit. Therefore, the objective is to minimize the RMSE.

The RMSE is often used in forecasting and regression problems such as time series forecasting, demand forecasting, financial modeling. It provides a useful way to quantify the accuracy of a model's predictions in terms of average error in original units.

3. RESULTS AND DISCUSSIONS

Two LSTM-based neural network models using Keras (a part of TensorFlow) were created. A model was trained with data splitting between 80% for training and 20% for testing, and a model was trained through K-Fold Cross Validation method. The models were trained using the specifications from Table 1.

To test the data, the training dataset containing historical data of energy production from January 1, 2019, to December 31, 2021, was read. The data was processed by converting Active Power to Energy (kWh) and stored it in a NumPy array. A MinMaxScaler was defined and fit to the trained data. This scaler was used to normalize the data. Then, a test dataset containing data for September 12, 2023 was read and processed similarly to the training data.

Finally, the pre-trained LSTM model was loaded, the input data was reshaped to match the model's input shape, which is expected to be in the form of sequences with the specified time step window, and a prediction was made on the test data using the loaded LSTM models. After, the scaled predictions and test data were inverse transformed to their original scales, and the evaluation metric was calculated to assess the model's performance in predicting energy production. Table 2 shows the RMSE evaluation metric from both models.

Table 2. Comparison of training results

Error	Training Through Data Splitting	Training Through Cross-validation
RMSE	0.0038	0.0023

The difference in RMSE results between the two trained models provides valuable insights into the performance and generalization of photovoltaic energy prediction models. The model trained with Cross Validation achieved a lower RMSE compared to the model trained on data splitting. This suggests that the Cross Validation model is more accurate in its predictions.

The data splitting method can be more susceptible to issues such as overfitting, where the model fits the training data very well but does not generalize well to new data.

Cross Validation helps mitigate this issue by providing a more robust assessment of model performance in different scenarios.

The lower RMSE result in the Cross Validation model indicates that it is more reliable in making accurate predictions in real-world situations. However, it is important to note that the Cross Validation model can be more time-consuming and computationally intensive, as it involves several training and testing iterations.

In Table 3, the result of the RMSE error from other papers in the literature is presented, to compare with the results obtained in the two methods tested in this work.

Table 3. Result of other papers in the literature

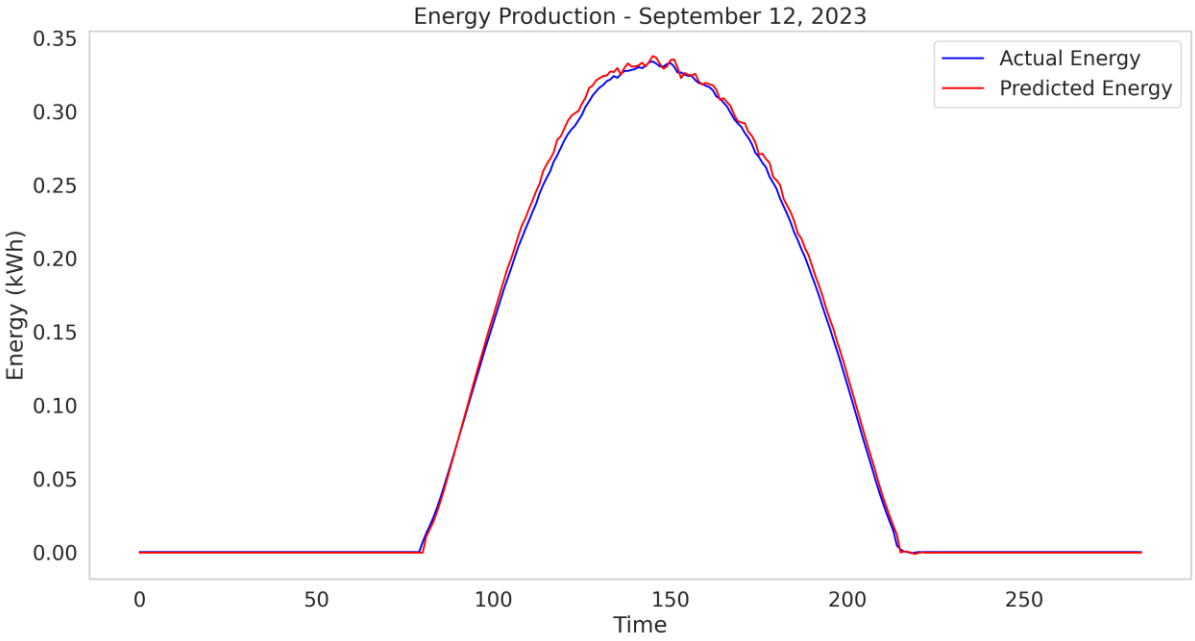
Study	Forecasting Error in RMSE
GENSLER et al., 2016	0.0713
LEE et al., 2018	0.0987 – 0.2520
LEE, D., KIM, K., 2019	0.563 – 0.874
WANG et al., 2019	0.621

Compared with the literature, the results obtained in this work are significantly better than most values mentioned in the literature. The RMSE values are close to one of the lowest values mentioned (0.0713) and are much better than the higher ranges and other RMSE values mentioned.

This suggests that the model trained in this work is highly competent in predicting photovoltaic energy production, outperforming many models mentioned in the scientific literature. However, it is worth highlighting that results may vary depending on the data, preprocessing method and algorithm used, and it is important to maintain continuous validation and evaluation to ensure the robustness of the model in different scenarios.

Figure 3 shows a plot with the actual energy values (in blue) and the predicted energy values (in red) for September 12, 2023, from the model trained with data split. This visually shows the model's performance in predicting energy production on that specific day.

Figure 3. Actual vs Forecast in data splitting method



4. CONCLUSION

In summary, based on the presented results, using Cross-validation for training the LSTM model has resulted in an improved prediction performance compared to the data splitting method. The lower error metrics obtained through cross-validation indicate enhanced robustness and generalization capabilities of the model, making it a preferable approach for training LSTM models in the context of predicting photovoltaic energy production in time series data.

REFERÊNCIAS BIBLIOGRÁFICAS

ABEDINIA, O., BAGHERI, M. (2021), **Execution of synthetic Bayesian model average for solar energy forecasting**, *IET Renewable Power Generation*.

CAO, Y., LIU, G., LUO, D., BAVIRISETTI, D. P., XIAO, G. (2023), **Multi-timescale photovoltaic power forecasting using an improved Stacking ensemble algorithm based LSTM-Informer model**, *Energy 8*, https://doi.org/10.1016/j.energy.2023.128669.

DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS, UN (2023), **The 2030 Agenda for Sustainable Development, Goal 7**, Available in: https://sdgs.un.org/goals/goal7. Accessed Oct 04, 2023.

DKASC (2023), **Desert Knowledge Australia Solar Centre**, Available in: https://dkasolarcentre.com.au.

GAO, M., LI, J., HONG, F., LONG, D. (2019), **Day-ahead power forecasting in a large-scale photovoltaic plant based on weather classification using LSTM**, *Energy, 187*, https://doi.org/10.1016/j.energy.2019.07.168.

GENSLER, A., HENZE, J., SICK, B., RAABE, N. (2016), **Deep Learning for solar power forecasting—An approach using AutoEncoder and LSTM Neural Networks**, *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, https://doi.org/10.1109/SMC.2016.7844673.

GOOGLE COLAB (2023) **COLABORATORY**, Available in: https://colab.research.google.com.

GUTIÉRREZ, L., PATIÑO, J., DUQUE-GRISALES, E. (2021), **A comparison of the performance of supervised learning algorithms for solar power prediction**, *Energies, 14(15)*.

LEE, D., KIM, K. (2019), **Recurrent Neural Network-Based Hourly Prediction of Photovoltaic Power Output Using Meteorological Information**, *Energies 2019, 12(2), 215*, https://doi.org/10.3390/en12020215.

LEE, W., KIM, K., PARK, J., KIM, J., KIM, Y. (2018), **Forecasting Solar Power Using Long-Short Term Memory and Convolutional Neural Networks**, *IEEE Access (Volume: 6)*, https://doi.org/10.1109/ACCESS.2018.2883330.

LI, P., ZHOU, K., LU, X., YANG, S. (2020), **A hybrid deep learning model for short-term PV power forecasting**, *Applied Energy, 259*, https://doi.org/10.1016/j.apenergy.2019.114216.

LUO, X., ZHANG, D., & ZHU, X. (2021), **Deep learning-based forecasting of photovoltaic power generation by incorporating domain knowledge**, *Energy, 225*, https://doi.org/10.1016/j.energy.2021.120240.

KONSTANTINOU, M., PERATIKOU, S, CHARALAMBIDES, A. G. (2021), **Solar Photovoltaic Forecasting of Power Output Using LSTM Networks**, *208 MDPI Journals Awarded Impact Factor, Networks. Atmosphere 2021, 12, 124*, https://doi.org/10.3390/atmos12010124.

NESPOLI, A. et al. (2019), **Day-ahead photovoltaic forecasting: A comparison of the most effective techniques**, *Energies, Multidisciplinary Digital Publishing Institute, v. 12, n. 9, p. 1621*. Accessed Oct 04, 2023.

SHARMA, J., SONI, S., PALIWAL, P., SABOOR, S., CHAURASIYA, P. K., SHARIFPUR, M., KHALILPOOR, N., Afzal, A. (2022), **A novel long term solar photovoltaic power forecasting approach using LSTM with Nadam optimizer: A case study of India**, *Energy Science and Engineering, 10(8), 2909–2929*, https://doi.org/10.1002/ese3.1178.

URBANETZ, I. V. et al. (2019), **Current Panorama and 2025 Scenario of Photovoltaic Solar Energy in Brazil**, *Braz. arch. biol. technol., Curitiba, v. 62, n. spe, e19190011*, Available in:

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-89132019000200210&lng=en&nrm=iso, Accessed Oct 04, 2023.

WANG, K., QI, X., LIU, H. (2019), **Photovoltaic power forecasting based LSTM-Convolutional Network**, *Energy 2018, 189, 116225*, https://doi.org/10.1016/j.energy.2019.116225.

ZAMAN, M., SAHA, S., EINI, R., ABDELWAHED, S. (2021), **A Deep Learning Model for Forecasting Photovoltaic Energy with Uncertainties**, *IEEE Green Energy and Smart Systems Conference*, https://doi.org/10.1109/IGESSC53124.2021.9618681.