Solar Energy Forecasting Using Neural Networks: Insights from Real Systems in the Amazon Region

Maykon S. Carvalho[1], Fábio L. Bezerra[1] and Licinius D. S. Alcantara[1]

[1] Cyberspatial Institute/Federal Rural University of the Amazon, Belém (Brazil)

Abstract

This work presents a neural network-based methodology to predict the daily energy output of photovoltaic (PV) systems by using meteorological and geographical information. The study relies on a dataset of 1,444 real operational records collected from two different clients in the state of Pará, Brazil, a region with high solar potential and significant seasonal variability. The input variables include solar radiation, ambient temperature, wind speed, latitude, longitude, and day of the year, which are used to train a multilayer perceptron (MLP) with two hidden layers containing 100 and 50 neurons, respectively. The model was trained and validated separately for each PV system, allowing for a comparison of performance under different local conditions. Results show that the neural network achieved satisfactory predictive accuracy for Client 1 ($R^2 = 0.68$), successfully capturing seasonal trends and the influence of environmental parameters. However, for Client 2, the performance was lower ($R^2 = 0.49$), likely due to system anomalies such as inverter malfunctions or operational irregularities, which introduced unexpected deviations in the dataset. These findings highlight both the potential and the limitations of data-driven models when dealing with real-world distributed generation systems. Despite discrepancies, the proposed approach demonstrates that ANNs can provide valuable insights for short-term solar energy forecasting in tropical regions, supporting the integration of PV systems into the grid.

Keywords: photovoltaic systems, energy prediction, neural networks, machine learning, solar forecasting

1. Introduction

The north region of Brazil presents a great strategic location for the installation of photovoltaic (PV) energy systems due to its proximity to the Equator. As a result, the demand for sustainable energy generation methods increases, accelerating the pace of production and installation of photovoltaic systems in the north region (Lima et al., 2020). In 2014, the state of Pará had only two photovoltaic systems, whereas by June 2019, it already had 821 photovoltaic systems, with a growth rate of 233% per year for photovoltaic systems and 216% for installed capacity (Lima et al., 2020).

The accurate forecasting of energy generation in photovoltaic (PV) systems is essential for improving the integration of solar power into the energy grid. The variability of solar radiation due to atmospheric conditions, combined with the nonlinear behavior of PV systems, poses significant challenges for traditional prediction models. Linear statistical models, while computationally inexpensive, often fail to capture the complex, time-varying patterns inherent in solar energy production (Mohandes et al., 1998).

Artificial Neural Networks (ANNs) offer a powerful alternative due to their ability to model complex, nonlinear relationships directly from empirical data (Mellit and Kalogirou, 2008; Yona et al., 2008). In particular, multilayer perceptron (MLP) networks have been successfully applied to predict solar radiation and PV output with improved accuracy compared to conventional approaches. Furthermore, hybrid models combining ANNs with optimization algorithms such as genetic algorithms or particle swarm optimization have shown even greater performance improvements in forecasting tasks (Voyant et al., 2017).

In recent years, the growing installation of PV systems in tropical and equatorial regions has intensified the need for more robust forecasting tools. These regions face specific challenges due to rapid cloud dynamics, high aerosol concentrations, and frequent seasonal variations that significantly affect solar irradiance. In the

10.18086/swc.2025.05.19 Available at http://proceedings.ises.org

Amazon region, solar irradiance is highly sensitive to cloud cover and rainy periods (Campos and Alcantara, 2016), reinforcing the need for models that incorporate meteorological variability when estimating or forecasting PV generation. Accurate forecasting is particularly relevant in the Amazon basin, where PV systems are increasingly deployed to support both urban grids and isolated communities with limited storage capacity. Recent studies highlight the benefits of applying deep learning techniques—such as deep neural networks (DNNs), long short-term memory (LSTM) models, and convolutional architectures—for enhancing predictive accuracy under such highly variable conditions (Souza et al., 2024).

Complementary research emphasizes the role of multivariate and seasonally tuned models. For example, Moreira et al. (2023) demonstrated that artificial neural networks incorporating diverse meteorological inputs can better capture seasonal variability in PV generation, outperforming simpler approaches. Likewise, Marques et al. (2023) explored solar energy forecasting in the Amazon Basin using data-driven models trained with satellite-derived irradiance and meteorological data, revealing both the opportunities and the difficulties of modeling in regions affected by strong environmental variability and land-use change.

These advances suggest that incorporating richer input features, accounting for site-specific anomalies, and adapting model architectures to local seasonal effects are key to improving solar forecasting in equatorial regions. Building on these insights, this work applies a feedforward ANN to predict daily energy output from PV installations in Pará, Brazil, using environmental and geographical data. By assessing model performance across two different clients, the study provides evidence of the effectiveness and limitations of ANN-based forecasting, as well as the impact of system-specific anomalies on predictive outcomes.

2. Methodology

2.1. Data Source

The dataset used in this study was constructed from real operational data collected from two grid-connected photovoltaic (PV) systems installed in the state of Pará, Brazil. This region, located in the Amazon basin, is characterized by high solar irradiance potential but also by significant climatic variability, such as frequent cloud cover, high humidity, and intense seasonal rainfall. These conditions make the region an interesting case for studying PV energy forecasting, as they introduce additional challenges compared to more stable climates.

The database contains a total of 1,444 daily records, covering different periods of operation for the two selected clients. Each record corresponds to a single day and aggregates both environmental and technical variables. The environmental inputs include global solar radiation (kWh m^{-2} day^{-1}), ambient air temperature (°C), and wind velocity (m s^{-1}), which are fundamental to characterize the weather patterns that influence PV module performance. In addition, geographical coordinates (latitude and longitude) were included as static features, ensuring that the neural network can associate system location with seasonal effects such as solar path and day length variation.

The outputs correspond to the measured daily energy generation of the PV systems (kWh day^{-1}). By using measured values rather than simulated outputs, the dataset reflects the actual operational conditions of PV installations, including possible effects of shading, dust accumulation, system degradation, or equipment malfunctions. This is particularly important because models trained on purely theoretical or simulated datasets tend to overlook the operational irregularities that are common in real-world systems.

Two distinct PV installations were used as case studies. Client 1, located in Belém, operates a residential system with 4.52 kWp of installed capacity. The system comprises 8 Jinko modules rated at 565 W each, connected to a 6 kW Growatt inverter. On average, this system produces around 500 kWh per month, and data collection covered 869 days of operation. Client 2, situated in Salinópolis, has a larger residential/commercial system with 9.90 kWp of installed capacity. It consists of 18 Jinko modules rated at 550 W each, also paired with a 6 kW Growatt inverter. This system has an average monthly generation of 1,150 kWh, with data covering 575 days.

The inclusion of two clients with different system sizes and locations provides a useful basis for comparison.

On one hand, Client 1's dataset offers a longer time series, enabling the model to better learn seasonal and interannual patterns. On the other hand, Client 2's system provides insights into the challenges of forecasting in systems where operational anomalies—such as inverter failures or abrupt drops in output—may occur more frequently. Together, these two cases allow us to test the robustness of the forecasting model under heterogeneous conditions.

Finally, all collected data were organized into a CSV file format, which facilitated preprocessing and integration into the machine learning pipeline. Before training the neural network, the dataset was cleaned and standardized, ensuring consistency across variables and eliminating potential biases caused by differences in measurement units. This careful preparation of the dataset was essential to guarantee the reliability of the forecasting model and to provide a solid foundation for evaluating the performance of the proposed approach.

2.2. Neural Network Theory

Artificial Neural Networks (ANNs) are computational models inspired by the functioning of the human brain, designed to identify patterns and approximate complex nonlinear relationships directly from data (Haykin, 2009). They are particularly useful in forecasting problems, where the target variable depends on multiple interrelated inputs and traditional statistical models may fail to capture nonlinear interactions. In the context of photovoltaic (PV) systems, ANNs are attractive because solar energy generation depends simultaneously on meteorological conditions, geographic location, and system-specific characteristics, all of which interact in nonlinear ways.

A neural network is typically structured into three main types of layers: the input layer, one or more hidden layers, and the output layer. Each layer consists of nodes, or artificial "neurons," that receive numerical inputs, apply a weighted sum, and pass the result through an activation function. This function introduces nonlinearity, enabling the network to approximate complex mappings that would not be possible with purely linear models. In this study, the Rectified Linear Unit (ReLU) activation function was employed, which is widely used in regression and classification tasks due to its computational efficiency and ability to mitigate the vanishing gradient problem (Glorot et al., 2011).

The architecture chosen for this work was a Multilayer Perceptron (MLP), one of the most common ANN configurations, as shown in Figure 1. The network implemented here has two hidden layers: the first with 100 neurons and the second with 50 neurons. This depth allows the model to learn hierarchical representations of the input data, progressively transforming raw meteorological and geographical variables into higher-level abstractions that are more closely related to PV energy output. The final output layer consists of a single neuron, responsible for producing the daily energy generation prediction (kWh/day).

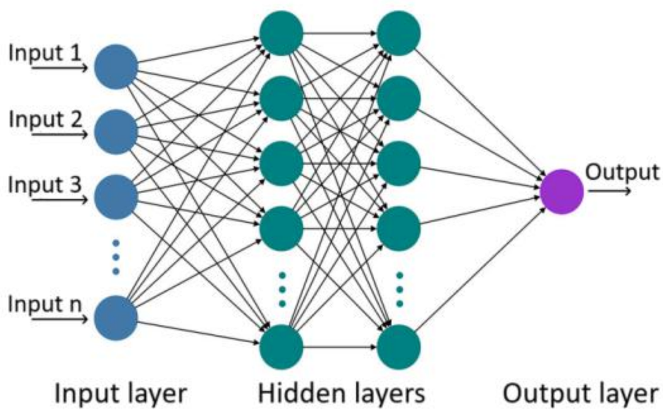

Figure 1: Diagram of the multi-layer perceptron for regression (Sahraei et al., 2021).

Training the neural network involves adjusting the weights associated with each connection between neurons. This is achieved through an iterative process called backpropagation, in which the network's predictions are compared with the actual measured outputs, and the error is propagated backward to update the weights. To perform this optimization, the Adam algorithm (Kingma and Ba, 2015) was used. Adam is a stochastic optimization method that combines the advantages of momentum and adaptive learning rates, leading to faster and more stable convergence compared to classical gradient descent.

The chosen neural network architecture offers a balance between complexity and interpretability. While deep networks with many layers and neurons may offer marginal improvements in accuracy, they require larger datasets and computational resources. The MLP structure adopted here is simple enough to be trained with a modest dataset of 869 for Client 1 and 575 records for Client 2, yet powerful enough to capture nonlinearities and seasonal trends in PV energy generation, making it suitable for practical applications in the Amazon region.

2.3. Training and Validation

To evaluate the predictive capability of the proposed neural network model, the dataset was divided into two subsets: training and testing. The training set, comprising 80% of the records, was used to fit the model parameters, while the remaining 20% served as an independent testing set to assess the model's generalization ability. This split ensures that the network does not simply memorize the training data, but rather learns patterns that can be applied to unseen cases. Such an approach is essential for forecasting tasks, where the ultimate goal is to predict future outputs from new data.

Before training, the input variables (solar radiation, ambient temperature, wind velocity, latitude, longitude, and day of the year) were preprocessed using the *StandardScaler* from the *scikit-learn* library. This procedure standardizes each feature by centering it around zero and scaling it to unit variance. Standardization prevents variables with larger magnitudes from dominating the learning process, improves numerical stability, and often accelerates model convergence. By ensuring that all features contribute equally, the neural network can more effectively capture the relationships between environmental inputs and PV energy generation.

The training process relied on the backpropagation algorithm, in which the prediction error is calculated by comparing the network's output with the measured daily energy generation. This error is then propagated backward through the network layers, updating the weights to minimize future errors. The Adam optimizer (Kingma and Ba, 2015) was employed to perform these updates. Adam is particularly well-suited for problems with noisy or sparse data because it adaptively adjusts learning rates for each parameter, combining the benefits of stochastic gradient descent with momentum-based approaches.

The training was performed for a maximum of 1,000 iterations (epochs). However, to prevent overfitting — a situation where the network memorizes the training data and performs poorly on new data — model performance was monitored throughout the process. If the error on the validation portion of the dataset did not improve after a number of epochs, early stopping criteria could be applied to halt training and preserve the best-performing model. This approach helps balance learning capacity with generalization ability.

For model evaluation, both quantitative and qualitative metrics were employed. Quantitatively, two key indicators were used: the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). RMSE provides a direct measure of the average magnitude of prediction errors in kWh/day, allowing interpretation in physical units relevant to PV system operation. Meanwhile, R^2 evaluates the proportion of variance in the observed data that is explained by the model, serving as a normalized measure of fit quality. Together, these metrics provide a robust assessment of accuracy and explanatory power.

In addition to numerical evaluation, visual analyses were conducted by comparing predicted and measured outputs across the study period. Time-series plots helped verify whether the network successfully reproduced seasonal variations, such as the expected drop in solar generation during the rainy season. Scatter plots of predicted versus measured outputs were also employed to assess how closely the model predictions aligned with actual system performance, identifying any systematic biases or anomalies.

Overall, the training and validation strategy adopted in this study ensures that the neural network is not only able to achieve a good fit on historical data but also capable of generalizing its predictive ability to new, unseen records. This careful combination of preprocessing, optimization, and evaluation strengthens the reliability of the proposed framework and provides a solid foundation for its application in real PV system management.

Although the proposed model performs same-day estimation, it still provides valid forecasting capabilities due to the periodic nature of solar resource availability. In tropical climates such as the Amazon region, seasonal irradiance patterns repeat annually with limited interannual variability (Campos and Alcantara, 2016). Consequently, once the model has learned the mapping between environmental inputs and energy generation, it can be applied to forecast PV output in future seasons under similar climatic conditions. This climatology-based generalization is an accepted approach in seasonal energy forecasting literature, where recurrent annual irradiation patterns are used to enable predictive modeling even without explicit temporal autoregression (Voyant et al., 2017; Antonanzas et al., 2016).

3. Results and discussion

The inclusion of multiple environmental and geographical variables in the neural network model significantly improved predictive performance compared to using solar radiation alone. By integrating temperature, wind velocity, and geographical coordinates, the model was better able to capture the nonlinear effects of weather and seasonal patterns on PV output. This aligns with findings in recent literature, where multivariate approaches consistently outperform univariate ones in renewable energy forecasting (Moreira et al., 2023; Marques et al., 2023).

Figures 2 and 3 present the predicted and measured energy outputs across the days of the year for Client 1 and Client 2, respectively. For Client 1, the predictions closely follow the seasonal trends of the measured data. Peaks during the dry season and reductions during the rainy season were well captured by the model, suggesting that the network effectively learned the relationship between meteorological inputs and PV generation. The relatively low dispersion between predicted and actual values indicates that, for this client, the system operated under stable conditions, allowing the model to generalize effectively.

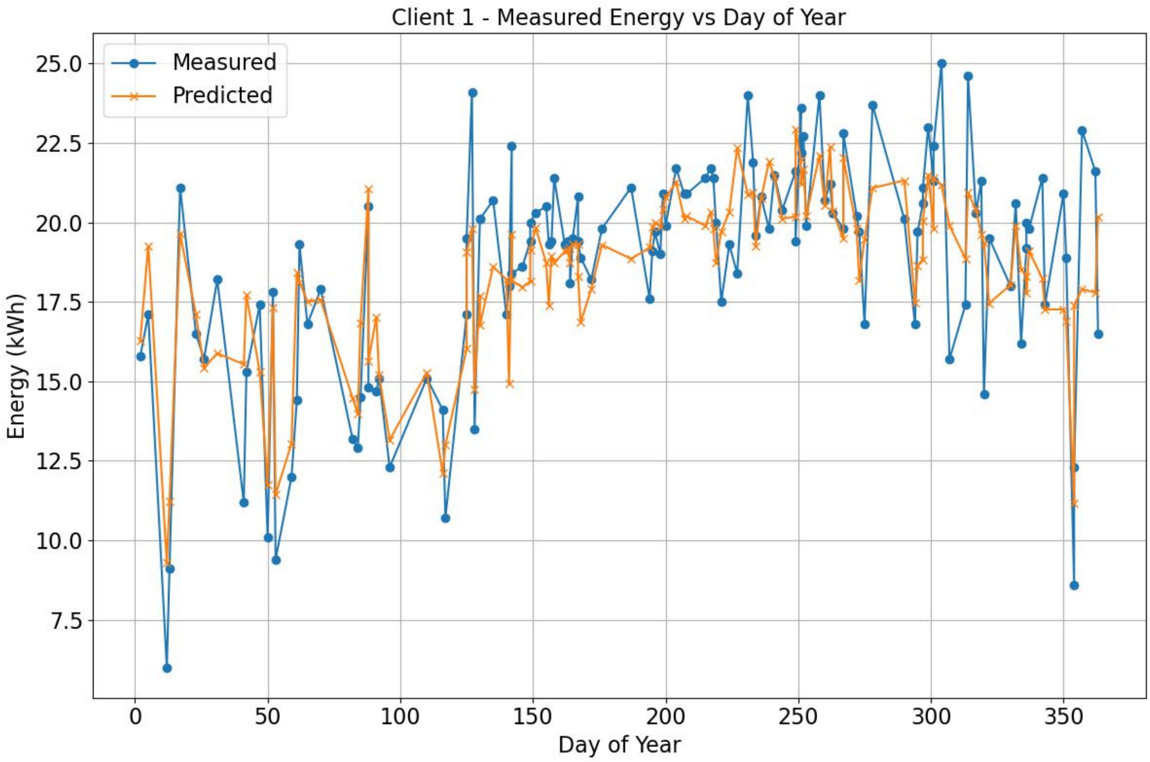

Figure 2: Predicted and measured energy over the days of the year for Client 1.

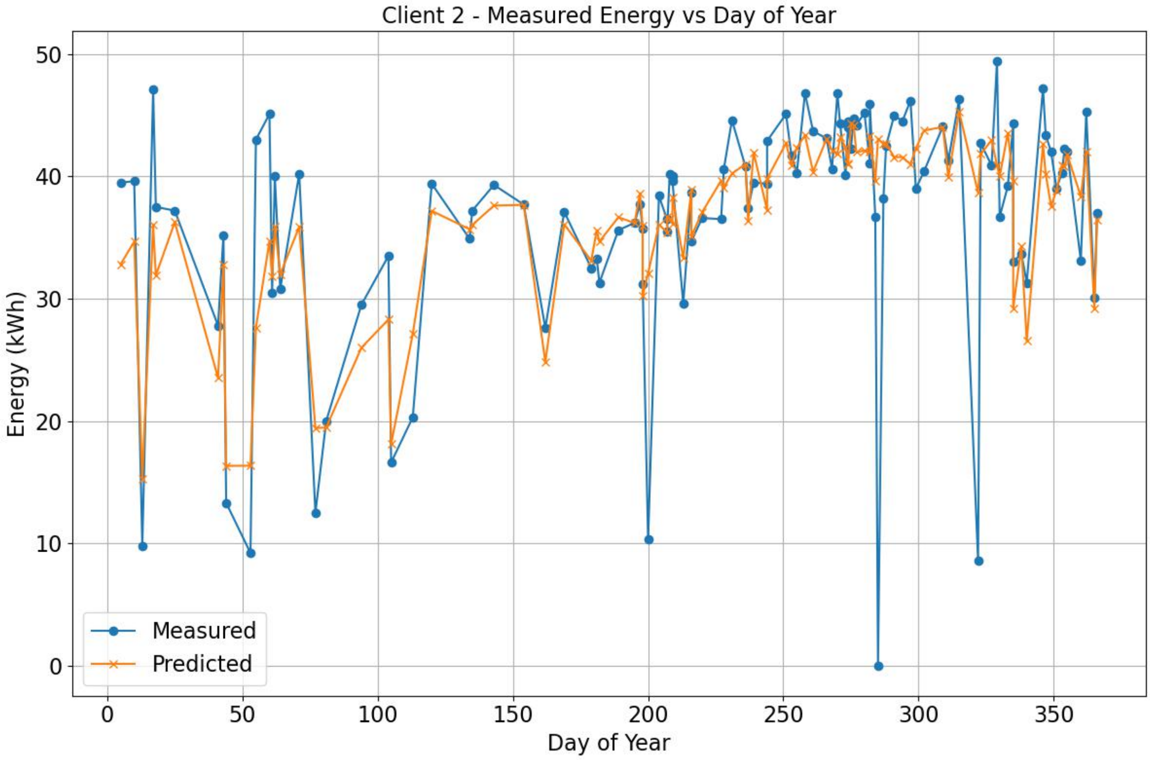

Figure 3: Predicted and measured energy over the days of the year for Client 2.

For Client 2, however, discrepancies were more evident. While the model was still able to capture general seasonal variations, certain days showed abrupt deviations between predicted and measured energy. These anomalies were reflected in the reduced correlation value (R^2 = 0.49) and higher error (RMSE = 6.63) compared to Client 1 (R^2 = 0.68, RMSE = 1.97), as shown in Table 1. Upon closer inspection, these outliers cannot be fully explained by environmental variables alone, suggesting the influence of system-specific anomalies, such as inverter malfunctions, temporary shading events, or operational irregularities. Such events introduce noise into the dataset, making it more difficult for the ANN to generalize accurately.

Table 1 Error and correlation parameters for each client.

Client	RMSE	R^2
1	1.97	0.68
2	6.63	0.49

Figure 4, which plots predicted versus measured outputs, reinforces this interpretation. For Client 1, points cluster closely around the 1:1 line, reflecting a strong agreement between the model and actual system performance. For Client 2, the dispersion is more pronounced, with several points falling significantly below the ideal line, highlighting periods where the measured output was abnormally low. These results suggest that the forecasting model is sensitive to operational anomalies—while it can represent environmental effects, it struggles when unexpected system failures occur.

Figure 5 presents the correlation matrix of the input features used in the photovoltaic energy prediction model. Solar radiation shows a moderate positive correlation with energy generation, as expected given its direct influence on PV output. Temperature also exhibits a weak to moderate positive correlation, consistent with its indirect impact through module operating efficiency. Geographic variables (latitude and longitude) display negligible correlation with the target variable since they are static attributes and therefore contribute indirectly by influencing seasonal patterns already captured by the day-of-year feature. These results justify the selection of the chosen input parameters and confirm the absence of strong multicollinearity among variables.

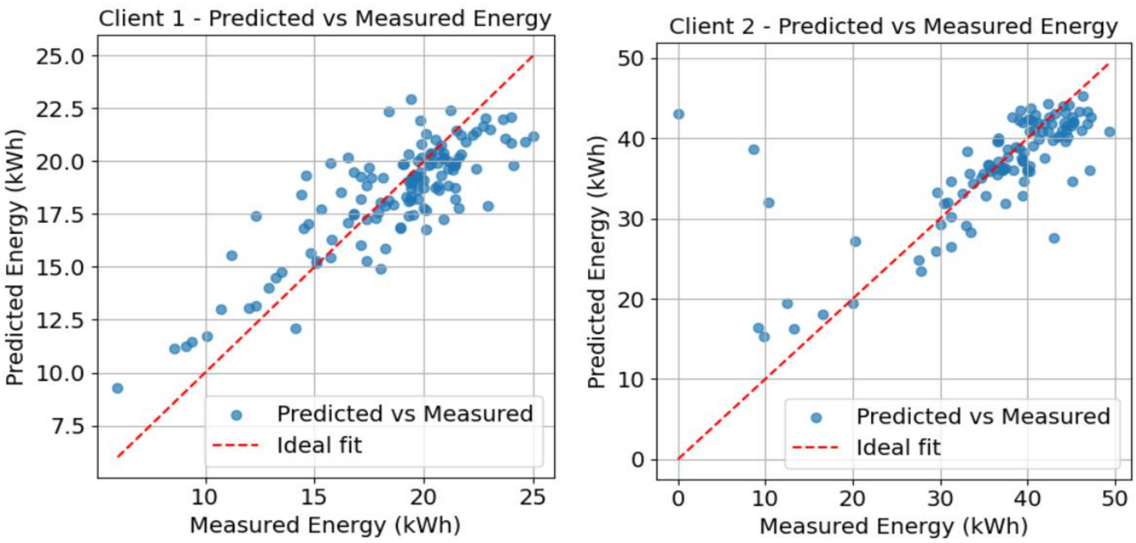

Figure 4: Predicted versus measured energy for both clients.

Figure 6 shows the comparative boxplot of residuals for the two PV systems (Client 1 and Client 2). The residuals represent the difference between measured and predicted energy generation. Client 1 exhibits a narrow residual distribution centered around zero, indicating stable model performance and low bias. In contrast, Client 2 presents a wider spread with more pronounced negative outliers. This suggests the presence of operational anomalies—such as shading events, inverter curtailment, or system downtime—not explicitly captured by the model inputs. These findings indicate that the model is sensitive to irregularities in system operation and that preprocessing steps such as anomaly detection could improve prediction quality.

Despite these challenges, the results confirm the viability of using neural networks for PV estimation in the Amazon region. Even under highly variable meteorological conditions, the model demonstrated its ability to replicate seasonal patterns and provide reliable short-term forecasts. The performance differences between the two clients also highlight the importance of system-specific tuning. For instance, including additional input variables such as cloud cover, panel orientation, shading effects, or maintenance records could potentially improve prediction accuracy, especially in systems where operational variability is high.

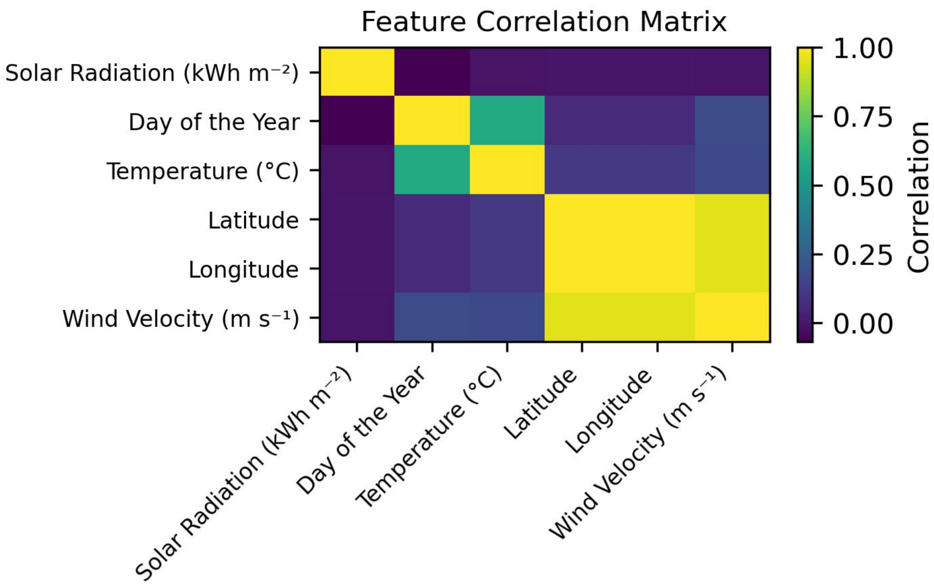

Figure 5: Feature correlation matrix for the input variables used in the photovoltaic energy prediction model, showing moderate correlation of solar radiation with energy output and low multicollinearity among variables.

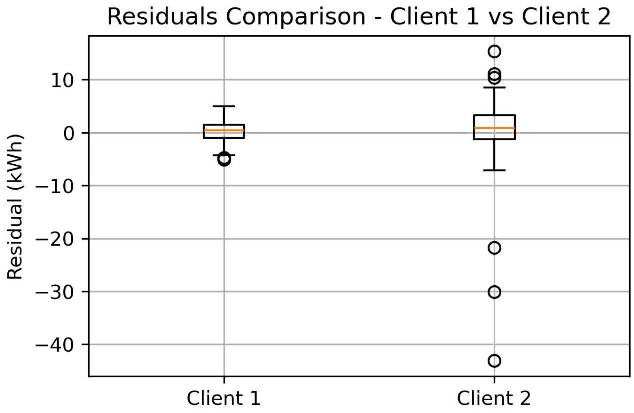

Figure 6: Comparison of residual distributions between Client 1 and Client 2. Client 2 exhibits a wider error range and negative outliers, indicating possible operational anomalies affecting prediction performance.

Another important finding is that neural networks, while effective, are not infallible when confronted with incomplete or noisy data. In practical applications, this underlines the necessity of integrating data quality monitoring into forecasting frameworks. Detecting and correcting anomalies in real time—such as sudden inverter shutdowns or abnormal drops in generation—could significantly enhance the reliability of ANN-based models. This suggests a promising avenue for future work: combining neural networks with fault detection and diagnosis systems, thereby ensuring not only accurate forecasting but also improved operational resilience.

In summary, the results indicate that the proposed ANN model is a robust tool for predicting daily PV generation, especially in equatorial regions with high solar potential but significant environmental variability. While the model performs well under stable system conditions, its accuracy may decline when operational anomalies are present. This reinforces the need for client-specific adjustments and possibly the integration of complementary features. These findings contribute to the broader effort of advancing data-driven forecasting tools to support renewable energy planning and grid integration in regions such as the Amazon. The results indicate that the ANN successfully captures the seasonal behavior of PV generation, a key requirement for forecasting applications based on climatic recurrence. Although no explicit time-lag was implemented in the model, its ability to generalize seasonal trends suggests that the learned patterns may be extrapolated to similar periods in subsequent years, supporting its use in seasonal forecasting scenarios.

4. Conclusions

This study demonstrated the potential of artificial neural networks (ANNs) for estimating the daily energy generation of photovoltaic (PV) systems in the Amazon region using real operational data. By leveraging environmental variables such as solar radiation, temperature, and wind velocity, along with geographical parameters, the model effectively captured the nonlinear relationships governing PV system performance. The results confirm the viability of ANN-based estimation frameworks for analyzing distributed solar energy generation in equatorial climates.

The model achieved satisfactory performance for Client 1 ($R^2 = 0.68$, RMSE = 1.97), while results for Client 2 ($R^2 = 0.49$, RMSE = 6.63) revealed the influence of system anomalies and operational irregularities that cannot be explained by meteorological factors alone. These findings highlight that neural networks can reliably estimate energy generation under stable conditions but may show reduced accuracy when data inconsistencies or hardware issues are present.

Even though the model was configured for same-day estimation, the strong seasonal regularity observed in Amazonian solar resource profiles enables the proposed ANN to be used for seasonal forecasting purposes. The model can infer future generation patterns by reproducing the recurrent climatic behavior of the region, making it suitable for performance prediction and planning in distributed PV systems.

From a practical standpoint, the proposed model can assist in performance analysis, system benchmarking, and energy planning for distributed PV installations. Future work should explore hybrid models that incorporate temporal dependencies, cloud cover data, and operational parameters (e.g., panel orientation, inverter status) to extend the estimation framework into a robust forecasting tool. By integrating these enhancements, neural networks can become powerful instruments for renewable energy management and grid integration in tropical and highly variable environments such as the Amazon region.

5. Acknowledgments

We would like to thank the Pro-Rectory of Student Affairs (PROAES/UFRA) for granting an Academic Scholarship for this research.

6. References

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F. J., Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. Sol. Energy. 136, 78–111.

Campos, M.S., Alcantara, L.D.S., 2016. Interpretação dos efeitos de tempo nublado e chuvoso sobre a radiação solar em Belém/PA para uso em sistemas fotovoltaicos. Rev. Bras. Meteorol. 31(4), 570–579.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 315–323.

Haykin, S., 2009. Neural Networks and Learning Machines, third ed. Pearson, Upper Saddle River, NJ.

Kingma, D.P., Ba, J., 2015. Adam: A Method for Stochastic Optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR).

Lima, M.M., Brito, I.R., Silva, B.M.S., Ferreira, P.V.M., Campos, P.S.S., Marques, G.T., Morais, E.C., Andrade, E.S.S., Almeida, J.F.S., Chase, O.A., 2020. Economic and energy potential evaluation of residential photovoltaic power generation in Belém – PA. Braz. J. Dev. 6(1), 1838–1850.

Marques, A.L.F., Teixeira, M.J., Almeida, F.V., Corrêa, P.L.P., 2023. Application of data science in the prediction of solar energy for the Amazon basin: a study case. Clean Energy. 7(6), 1344–1355.

Mellit, A., Kalogirou, S.A., 2008. Artificial intelligence techniques for photovoltaic applications: A review. Prog. Energy Combust. Sci. 34(5), 574–632.

Mohandes, M., Rehman, S., Halawani, T.O., 1998. Estimation of global solar radiation using artificial neural networks. Renew. Energy. 31(5), 719–728.

Moreira, M.O., Kaizer, B.M., Ohishi, T., Bonatto, B.D., Souza, A.C.Z., Balestrassi, P.P., 2023. Multivariate strategy using artificial neural networks for seasonal photovoltaic generation forecasting. Energies. 16(1), 369.

Sahraei, A., Chamorro, A., Kraft, P., Breuer, L., 2021. Application of machine learning models to predict maximum event water fractions in streamflow. Front. Water. 3, 652100.

Souza, W.A., Alonso, A.M.S., Bernardino, L.G.R., Castoldi, M.F., Nascimento, C.F., Marafão, F.P., 2024. Deep Learning Models to Estimate and Predict the Solar Irradiation in Brazil. In: Jørgensen, B.N., da Silva, L.C.P., Ma, Z. (Eds.), Energy Informatics. EI.A 2023. Lecture Notes in Computer Science, vol 14468. Springer, Cham, pp. 63-82.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.L., Paoli, C., Motte, F., Fouilloy, A., 2017. Machine learning methods for solar radiation forecasting: A review. Renew. Energy. 105, 569–582.

Yona, A., Senjyu, T., Saber, A.Y., Funabashi, T., 2008. Application of neural network to 24-hour-ahead generating power forecasting for PV system. In: IEEE Power and Energy Society General Meeting.