Machine Learning for Forecasting Hydrogen Production from Solar Energy Using K-Nearest Neighbors Regression Model: A Case Study in Petrolina, Brazil

Gabriel O. Almeida[1], Tainan S. Viana[1], Jesse V. G. Thé[2], Maria E. V. Silva[1] and Paulo A. C. Rocha[1]

[1] Federal University of Ceará, Fortaleza (Brazil)

[2] University of Guelph, Guelph (Canada)

Abstract

This paper proposes the application of the kNN regression machine learning algorithm to predict green hydrogen production in the city of Petrolina, Pernambuco. Solar irradiance data were applied to a mathematical model that considers a PEM electrolyzer powered by photovoltaic panels, resulting in a time series of production rates. These values were applied along with temporal information as predictor variables. A bagging approach was applied so that the final forecast response is given by the average of the individual responses of ten models created from random samples of the training set. The process was repeated for different sampling intervals, and the results were evaluated according to performance metrics. The best result was obtained for an interval of 20 minutes, obtaining the values of RMSE, nRMSE and R^2 equal to 75.144 kg km^{-2} , 0.212 and 0.826, respectively. When compared to other papers, the performance of our model is inferior to more robust approaches, such as support vector machines and neural networks, but it presents competitive performance under certain conditions of the prophet model and outperforms another kNN model employed for the same purpose.

Keywords: Solar energy, green hydrogen, times series forecasting, machine learning, kNN regression model

1. Introduction

The global energy landscape is characterized by growing concern about greenhouse gas emissions and their impact on climate change, leading countries to adopt decarbonization strategies. Despite the growing global adoption of renewable energy sources for electricity generation (primarily solar and wind), fossil fuels remain the primary source of energy supply. In this context, hydrogen is emerging as an alternative energy carrier due to its flexibility in serving different sectors: electricity generation, transportation fuel, and industrial input.

Hydrogen can be obtained in several ways, each involving processes and components with different levels of carbonization. Green hydrogen is produced via the electrolysis of water powered by renewable energy sources (solar, wind), enabling production without carbon emissions. Despite its high conversion efficiency compared to other methods, it has higher production costs. This leads to the use of other economically inefficient and unsustainable methodologies, with steam reforming of methane (gray hydrogen) being the main one. The quest to make this fuel viable has driven the development of new technologies and materials that make the electrolysis process cheaper, with the aim of making it scalable (Md Monjour and Zahed, 2025).

The integration of hydrogen with renewable sources brings benefits beyond the environmental. Solar and wind resources have intermittent characteristics related to the dynamics of irradiance and wind speed. To ensure a continuous supply and increase the reliability of the power grid, storage systems are required. This has driven the growth of the battery market, with scaling driving down prices and increasing government support. Installed battery capacity is projected to increase ninefold, reaching 760 GW by 2030 (IEA, 2023). Alternatively, hydrogen can be produced during periods of energy surplus and subsequently used to supply grid demand or for other purposes, such as fuel for the transportation industry or industrial input.

Brazil has several characteristics that favor the adoption of green hydrogen. Its electricity matrix is diversified, with a large share of renewable energy sources. Installed power data collected in March 2025 (ABSOLAR, 2025) indicates a predominance of hydroelectric generation with 110,160 MW (43.0%), followed by solar generation with 69,744 MW (23.7%), and wind generation with 34,050 MW (13.3%). Other factors, such as investment capacity in technological development, institutional strength, and political stability, make the national hydrogen market competitive, boosting industrial development through cleaner production and solving the problem of variability in renewable sources through H2V storage. The production and export of hydrogen and its high-value-added derivatives benefit the country's economy in the international market (PNH2, 2023). To achieve these results, the government

10.18086/swc.2025.04.31 Available at http://proceedings.ises.org

created the National Hydrogen Program (PNH2), bringing together public and private agents to deliberate and establish goals for strengthening hydrogen as an energy vector (MME, 2022). Among the objectives are: consolidating Brazil as the lowest-cost low-carbon hydrogen producer by 2030 and consolidating hubs by 2035.

One challenge to overcome for large-scale green hydrogen use stems from the stochastic and random nature of renewable energy, making production highly dependent on local and dynamic meteorological factors. Energy and financial planning depend on the balance between energy supply and demand, with the issue of unpredictability being a bottleneck for a complete energy transition. In the literature, several studies propose the use of machine learning (ML) algorithms to improve the forecasting and optimization of these systems. In India, global horizontal irradiance (GHI) data for the cities of Fathergah and Bhadla are used to train a bidirectional long-short-term memory neural network (BiDLSTM) by applying a Completely Empirical Joint Decomposition with White Noise (CEEMDAN) to the data, with the aim of mapping production potential and supporting the adoption of the green economy (Karan et al., 2024). Similarly, Chaoyang et al. (2024) combine a multivariate empirical decomposition (MED) with an Ant Lion Optimizer (ALO) algorithm and bidirectional long short-term memory (BiLSTM) to perform solar irradiance forecasts aimed at integration with electrolyzers in Jiangsu Province, China. In both cases, various optimization steps are integrated with a deep learning algorithm to provide a new model and improve the accuracy of hydrogen production potential. However, the models require a large amount of data, which can limit their applicability and generalizability.

In this context, this paper proposes the use of the k-nearest neighbor regression algorithm to predict hydrogen production in a hypothetical system consisting of a proton exchange membrane (PEM) electrolyzer powered by photovoltaic panels. Solar irradiance data were collected from the city of Petrolina, and a mathematical model was applied to obtain production data. Past production values (lags) and information related to the time and day of the year were used as input variables to capture seasonal dynamics. The results are compared with other similar studies found in the literature.

The remainder of the article is divided as follows: Chapter 2 presents the methodology adopted in the research, starting with the data collection and processing, then the physical characteristics of the electrolysis system, and concluding with the training and evaluation of the kNN model. The chapter presents the results obtained, comparing performance metrics obtained by other authors and analyzing the predicted versus actual relationship over time. Chapter 3 presents the main conclusions drawn from the research.

2. Methodology

2.1. Data Collection

The data were collected from SONDA network portal (SONDA,2025), which provides measurements from radiometric stations spread across various regions of Brazil. This study selected the PTR-11 station in Petrolina, located in the state of Pernambuco (coordinates: -9.0689, -40.3197), at an altitude of 387 m. The region is in the semi-arid region of northeastern Brazil, characterized by a dry and hot semi-arid tropical climate in the north and a hot semi-arid steppe in the south. The average annual precipitation is around 560 mm, marked by scarce and irregular rainfall, with the rainy season occurring between January and April (SONDA, 2024). Proximity to the Equator and low rainfall favor the implementation of solar energy projects, making it suitable for productive analysis of green hydrogen.

Global horizontal irradiance (GHI) data were collected at one-minute sampling intervals. To ensure the machine learning model could learn from the seasonal patterns of the solar resource, one year of data was used for training and another year was selected for the testing phase. The chosen period considered the SONDA portal's validation process, based on the data quality control strategy adopted by Baseline Surface Radiation Networking (WRMC, 2025), which identifies data with suspicious measurement characteristics. The process occurs in different stages, and the next stage can only be continued after approval of the previous stages. The training period was chosen through a balance between more recent data periods and those with small percentages of suspicious or missing data. A summary of the main information about the dataset is presented in Table 1. Missing data is treated by removing these data before application within the machine learning model.

The daily dynamics of solar radiation are characterized by nighttime periods with insignificant radiation levels for power generation. A filter was applied to eliminate data contained in this period, using solar elevation angle values less than 15° as the exclusion criterion (Francisco et al., 2020). The Python library “pvlib” was used to obtain solar elevation values, which were concatenated with the irradiance dataset.

Table 1: Details of the datasets collected from the SONDA portal.

Data	Period	Unit	Maximum	Mean	Samples
Train	2013	W m^{-2}	1425	223	525600
Test	2014	W m^{-2}	1438	223	525600

2.2. Hydrogen production model

Hydrogen is obtained via the process of water electrolysis using a Proton Exchange Membrane (PEM) powered by a photovoltaic panel. The same mathematical modeling presented in Syed et al. (2021) was applied, where electrical energy delivered by the PV system (E_{PV}), in kWh km^{-2}, can be determined in terms of the global horizontal irradiance (G), the module efficiency (η_{PV}) and the Power Conditioning efficiency (η_{PC}), as described in eq. 1 (Saumia et al.,2017). Eq. 2 defines the amount of hydrogen produced (M_{H_2}), in kg km^{-2}, calculated from the electrical energy (E_{PV}), the efficiency of the electrolysis process (η_{ele}), and the Higher Heating Value of hydrogen (HHV_{H_2}), in kWh kg^{-1}. A power-to-energy conversion was used to enable the application of the collected irradiance data (W m^{-2}) in G (kWh km^{-2}) in eq. 01, considering the one-minute interval between consecutive samples. The value of the conversion constant (c), as well as other constants used, are presented in Tab. 2.

$$E_{PV} = G\ \eta_{PV}\ \eta_{PC} \quad \text{(eq. 1)}$$

$$M_{H_2} = \frac{E_{PV}\ \eta_{ele}}{HHV_{H_2}} \quad \text{(eq. 2)}$$

Table 2: Adopted parameter values

c	η_{PV}	η_{PC}	η_{ele}	HHV_{H_2}
16.67	0.17	0.85	0.75	53

2.3. K-Nearest Neighbors Regressor

The kNN regressor is a nonparametric method that allows for estimating data points based on the closest observations. This is presented mathematically by eq. 3, in which the estimate for a new point x_0 is obtained by averaging the observations of the k-nearest neighboring points x_i , belonging to the set $\mathcal{N}_0$ (Gareth et al., 2023).

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i\, \epsilon\, \mathcal{N}_0} y_i \quad \text{(eq. 3)}$$

To implement the model, the Python scikit-learn library was applied, using the object creation "KNeighborsRegressor()", allowing data fit and parameter optimization. The following variables were chosen as model inputs: normalized time of day; normalized day of year; and lagged hydrogen production (lags). To ensure that the order of magnitude of the features does not lead to a misinterpretation of the importance of each of them, a "StandardScaler()" function was used to rescale the inputs with a zero mean and unit standard deviation. The number of production lags considered in the model are determined iteratively, adopting the configuration that results in the smallest errors.

Figure 1 presents a general flowchart of the process, where ten distinct models are built from bootstrapping sampling of the training data, and the average of the individual predictions is used as the final prediction. This approach is called bagging and has the effect of reducing variance and improving the results (Zhou, 2012). The determination of the best hyperparameter K is determined by a self-tuning process in which incremental quantities are tested until the error stabilizes. At the end of the process, the model is applied to the test set and error metrics are used to evaluate performance. The process was repeated for different time resolutions, via one-minute base frequency resampling.

2.4. Performance metrics

To evaluate the model's performance, a comparison was made between the actual and predicted values. The following metrics were applied: Root Mean Square Error (RMSE); Normalized Root Mean Square Error (nRMSE); Mean Absolute Error (MAE); Normalized Mean Absolute Error (nMAE); and coefficient of determination R^2. The results are presented in eq. 4-9. The actual values, predicted values, mean value and sample size are given from: y_i, $\hat{y}_i$, $\bar{y}$, N, respectively.

$$RMSE = \sqrt{\frac{1}{N}\sum(\hat{y}_i - y_i)^2} \quad \text{(eq. 4)}$$

$$nRMSE = \frac{RMSE}{\bar{y}} \quad \text{(eq. 5)}$$

$$MAE = \frac{1}{N}\sum|\hat{y}_i - y_i| \quad \text{(eq. 6)}$$

$$nMAE = \frac{RMSE}{\bar{y}} \quad \text{(eq. 7)}$$

$$MAPE = \frac{1}{N}\sum\frac{|\hat{y}_i - y_i|}{y_i} \quad \text{(eq. 8)}$$

$$R^2 = \sum\frac{\hat{y}_i - \bar{y}}{y_i - \bar{y}} \quad \text{(eq. 9)}$$

3. Results and discussion

The models created for the different time resolutions were compared using error metrics for the test set. The results are presented in Table 3. It is observed that as time grows, the RMSE and MAE also increase. This occurs because these metrics are associated with the order of magnitude of the data sets, where larger time intervals imply higher production levels for each sample. Normalizing these values in relation to their averages provides better information for comparing performance, , where it is observed that nRMSE and nMAE reach their minimum value when the time resolution is 20 min, obtaining values of 0.212 and 0.149, respectively. The value of the coefficient of determination R^2 for this case is 0.826 .

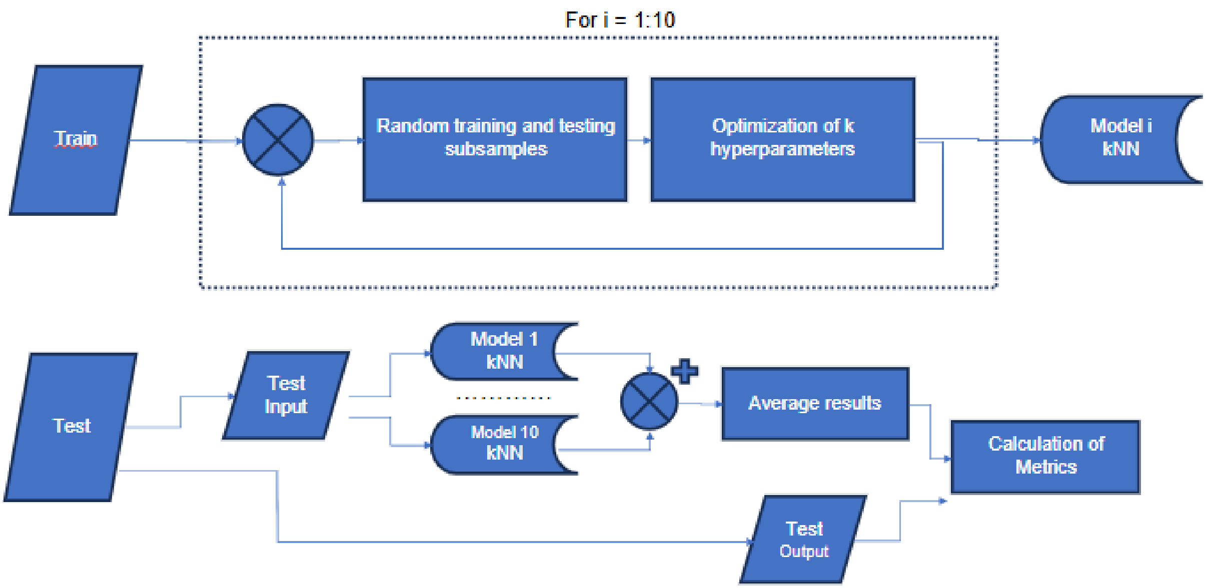

Figure 1: Model training and testing flowchart

Table 3: Comparison of models based on time resolution.

TIME RESOLUTION (min)	RMSE	nRMSE	MAE	nMAE	MAPE	R^2
10	41.310	0.231	27.916	0.156	0.236	0.799
15	59.341	0.223	41.133	0.154	0.241	0.811
20	75.144	0.212	52.946	0.149	0.239	0.826
25	94.984	0.216	67.073	0.152	0.246	0.821
30	111.515	0.213	79.532	0.152	0.257	0.827

A comparison with the performance achieved by other authors in studies on hydrogen production forecasting is summarized in Table 4. The time resolution selected was the one that performed best (20 min) and the metrics applied were the coefficient of determination R^2 and the Mean Absolute Percentage Error (MAPE). The peculiarities adopted by each study should be considered, as the type of machine learning algorithm and the dataset used for training and testing.

The coefficient of determination (R^2) found in this work was lower than that found by other more robust learning algorithms. For example, Qian et al. (2025) uses a long-short-term neural network (LSTM) model to identify long-term dependencies for a 9-year dataset of solar radiation in Lanzhou City, China. A larger set of meteorological variables were used as predictor variables, such as temperature, wind speed and precipitation. The data was separated according to summer and winter season, in which better performance was obtained in the latter. Another example of neural network application occurs in the article by Alhussan et al.(2023) who used meteorological data to forecast hydrogen production in Hawaii. applying the BER metaheuristic search algorithm and the particle swarm optimization (PSO) algorithm to determine dynamic hyperparameters. Comparison with this work must consider the processing time and the amount of data required for the model to function efficiently.

Other authors achieved higher coefficients of determination with the Prophet, Stochastic Gradient Descent, Support Vector Machine (SVM) and SARIMAX models. The prophet's time series model accommodates trends and seasonality and can be adapted for solar radiation forecasting. The results of Syed et al. (2021) demonstrate the superiority of the model in predicting hydrogen production compared to Stochastic Gradient Descent (SGD) and Seasonality Auto-Regressive Integrated Moving Average (SARIMAX) . Guishi et al. (2023) apply the Prophet model to different areas of China, where they observe better performance for data in the MPZ zone. In this region, the Prophet model outperforms the performance found in this work. However, for the TMZ zone, the coefficient of determination is lower, reaching the value of 0.628 and indicating a competitive performance of the developed kNN. The results found by Hassan et al. (2025) for the application of kNN in the productive prediction of a solid oxide electrolysis cell (SOEC) achieved a coefficient of determination of 0.765. The authors used only a single model without any bagging strategy with a number of neighbors K = 5.

Figure 2 presents a scatterplot of predicted and actual hydrogen production values in kg km^{-2}. A linear approximation of the data was applied, represented by the blue line. The dashed black line represents the ideal case in which the predictions equal the actual values. The proximity between the two lines suggests the effectiveness of the model in making predictions, which are more assertive around the average values than at the extremes of the data points.

Figure 3 shows the model's behavior over a week of the test set, where the predicted curve follows the behavior of the actual values. It should be noted that the production dynamics are closely related to the behavior of solar irradiance, reaching maximum values near solar noon. Time scale times are described relative to the reference time zone (GMT). Figure 4 shows a comparison between three days with distinct profiles of accumulated hydrogen production: maximum daily production with a value of 14861 kg km^{-2} occurred on 11/30; minimum daily production with a rate of 3987 kg km^{-2} on 12/17; intermediate production values occurred on 07/23 with an accumulated value of 10827 kg km^{-2}. These differences can be explained by the behavior of meteorological variables that directly affect irradiance levels (SONDA, 2014). On the day of minimum production, there were two rainy periods with a maximum accumulated precipitation of 0.5 mm, limiting global radiation levels to a maximum of 585 kWh m^{-2}. In the period with highest production, the irradiance curve presents a smooth growth curve with almost no influence from clouds, reaching a maximum of 1020 kWh m^{-2}. The information about accumulated hydrogen production can be used to determine the number of photovoltaic panels and the available area needed to meet energy demand.

Table 4: Comparison between the results of this work and others found in the literature.

Study	Region	ML Algorithm	R^2	MAPE (%)
This work	Petrolina, Brazil	kNN regressor	0.826	23.9
Hassan et. Al (2025)	Developed Dataset	kNN regressor	0.765	-
Qian et. al (2025)	Lanzhou, China	LSTM	0.968 (summer) 0.980 (winter)	0.610 (summer) 0.552 (winter)
Syed et. al (2021)	Islamabad, Pakistan	Prophet	0.983	3.71
		SGD	0.969	3.58
		SARIMAX	0.966	9.08
Guishi et. al (2023)	China (TMZ)	SVM	0.960	-
		Prophet	0.628	-
	China (MPZ)	SVM	0.968	-
		Prophet	0.855	-
Alhussan et. Al (2023)	Hawaii, EUA	RNN	0.99	0.6

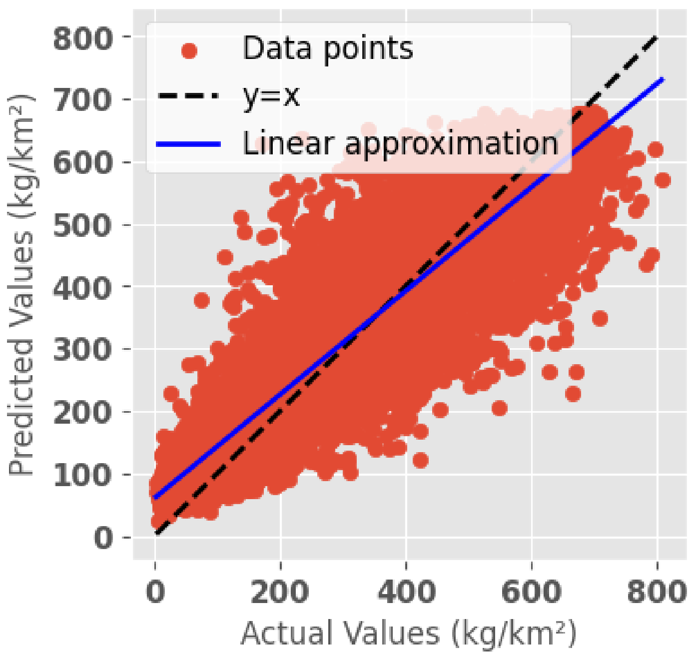

Figure 2: Comparison between actual and predicted values

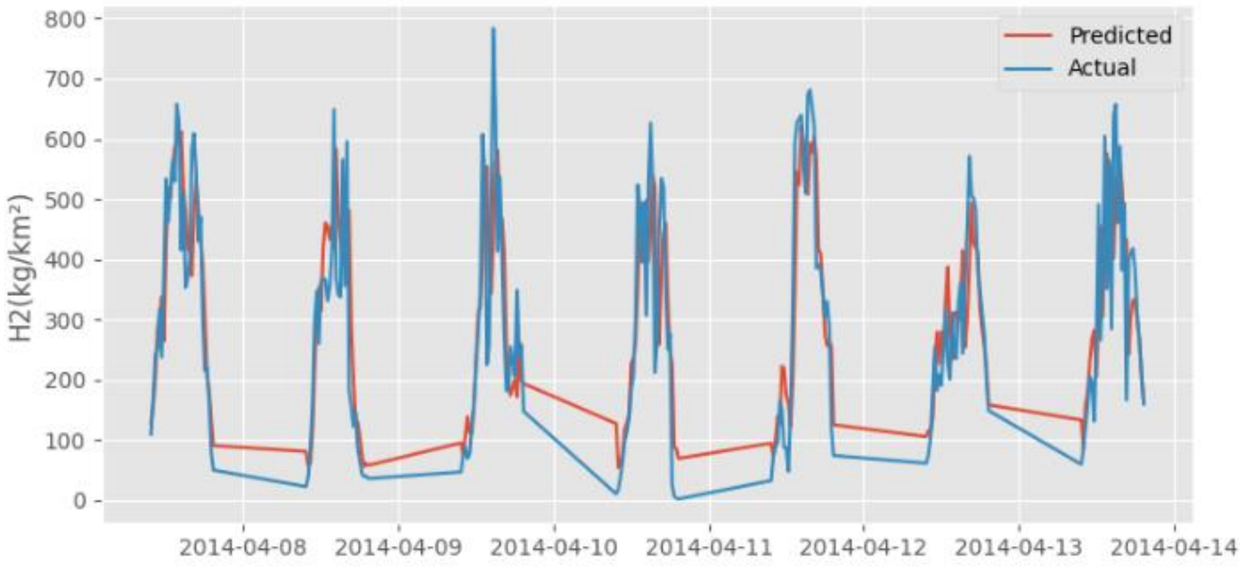

Figure 3: Forecast model behavior for one week.

These graphical results can be integrated into monitoring interfaces, where sampling intervals are selected according to the need to meet short- or long-term demands. This provides greater reliability for the production system, guiding operation. For example, when favorable production conditions are expected, portions of the photovoltaic systems can be used exclusively for the electrolysis process, while the remainder can be used to meet electricity demand. The hydrogen produced can be used to generate power during peak periods or as a commercial input. The price of electricity and hydrogen is relevant information to determine the most cost-effective configuration.

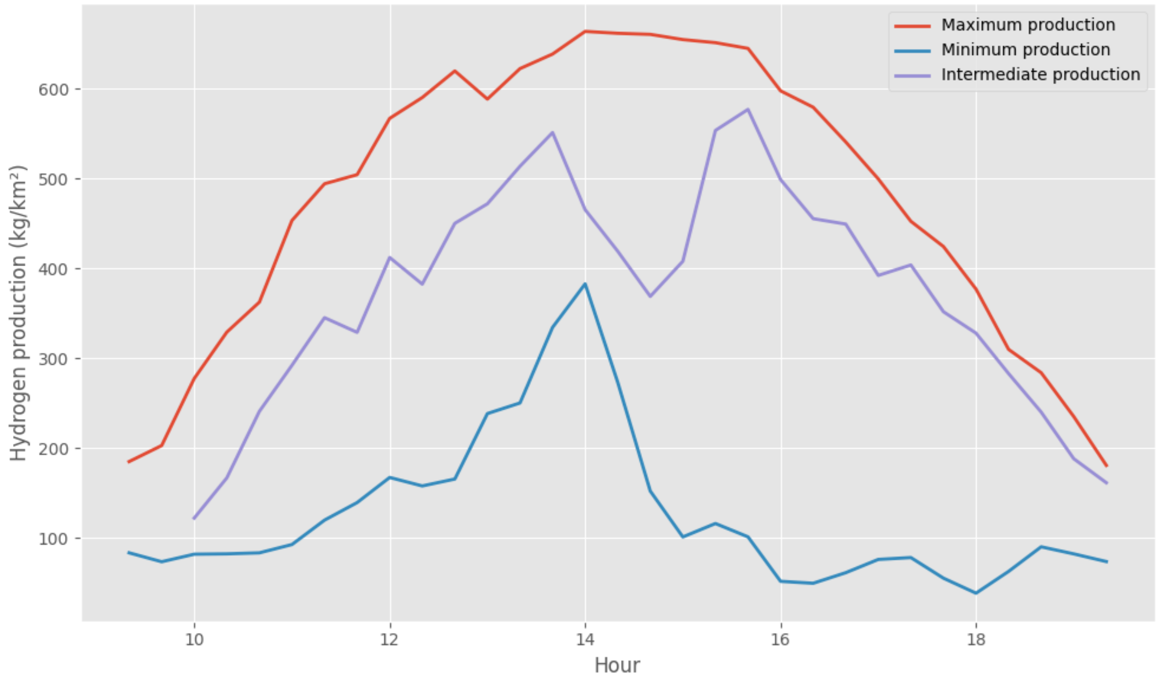

Figure 4: Forecast model for three different conditions.

4. Conclusion

In this paper, the k-nearest neighbors (KNN) machine learning algorithm was used to predict hydrogen production in a system consisting of a proton exchange membrane (PEM) electrolyzer powered by a photovoltaic panel. Irradiance data from 2013 to 2014 were collected from the SONDA portal to fine-tune the model. The first year was used for training and the second for testing. A bagging approach with ten distinct models was adopted to reduce variance and improve results.

To evaluate the model's performance, a set of performance metrics were used to compare predicted and actual values. A resampling process for different time intervals was employed to verify consequent improvements in performance. The results indicated that increasing the time interval leads to a significant increase in non-normalized measures (RMSE, MAE), justified by the fact that the mean value of each sample represents the production over that time interval. The time window that presented the best performance was 20 min, obtaining an RMSE of 75.144 kg km^{-2}, an nRMSE of 0.212, an MAE of 52.946 kg km^{-2}, an nMAE of 0.149, a MAPE of 23.9%, and an R^2 of 0.82. Comparison with other similar studies shows that more robust algorithms, such as LSTM neural networks and support vector machines (SVMs), demonstrated greater accuracy in predicting hydrogen production. However, the results perform competitively when compared to studies using the Prophet model or the same type of algorithm.

The performance of production forecasts over time was analyzed for a weekly and daily window, demonstrating the algorithm's ability to understand seasonal patterns related to irradiance dynamics, characterized by a production peak around solar noon. These graphs can be integrated into monitoring interfaces for different sampling windows, assisting system operators in short- and long-term decision-making

5. Acknowledgments

This study was financed in part by the FUNCAP - Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (Research and Innovation Network on Renewable Energy, Rede VERDES, Grant No. 07548003/2023), CNPq - Conselho Nacional de Pesquisa, Universal - Grant No. 408431/2023-7 and CNPq - Conselho Nacional de Pesquisa, Bolsa de Produtividade em Pesquisa Grant. No. 303585/2022-6 and 307096/2021-1.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001

6. References

ABSOLAR. Associação Brasileira de Energia Solar Fotovoltaica. Energia Solar Fotovoltaica no Brasil: Infográfico ABSOLAR. 2025. n° 83. URL: https://www.absolar.org.br/mercado/infografico/ .

Alhussan, A., El-kenawy, E., Saeed, M., Ibrahim, A., Abdelhameed, A., Eid, M., Elsaid, M., Khafaga, D., Abualigah, L., Elbaksawi, O., 2023. Green hydrogen production ensemble forecasting based on hybrid dynamic optimization algorithm. Frontiers in Energy Research. 11. 01-15. 10.3389/fenrg.2023.1221006.

Chaoyang, Z., Mengxia, W., Mengxing, G., Jinxin, D., Qipei, D., Wei, W., Yunxiang, Z., 2024. Enhancing solar irradiance forecasting for hydrogen production: The MEMD-ALO-BiLSTM hybrid machine learning model, Computers and Electrical Engineering, Volume 120, Part B, 2024, 109747, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2024.109747.

Francisco J., R.B, Clara, A.B, Javier, H.T., Ricardo, A.M., Inés, G.L., David, P.V., 2020. A short-term solar radiation forecasting system for the Iberian Peninsula. Part 1: Models description and performance assessment. Solar Energy. 195, 396-412.

Gareth, J., Daniela, W., Trevor, H., Robert, T., Jonathan, T., 2023. An Introduction to Statistical Learning with Applications in Python, first ed. Springer.

Guishi, C., Ercheng, L., Ying, Z, Yihao, Y., Binbin, C., Youcheng, C., Xiaoqiang,W., Chnangqing, D., 2023. Analysis and prediction of green hydrogen production potential by photovoltaic-powered water electrolysis using machine learning in China, Energy, Volume 284,2023,129302,ISSN 0360-5442.

Hassan, R., Kazemi, M.R., 2025. Accurate prediction of green hydrogen production based on solid oxide electrolysis cell via soft computing algorithms. *Sci Rep* **15**, 35464 (2025). https://doi.org/10.1038/s41598-025-19316-9

IEA. International Energy Agency: Batteries and Secure Energy Transitions. 2023. URL: https://www.iea.org/reports/batteries-and-secure-energy-transitions.

Karan, S., Bijaya, K.P., Tushar, S., Rita, N., 2024. Deep learning solar forecasting for green hydrogen production in India: A case study, International Journal of Hydrogen Energy, Volume 50, Part B, 2024, Pages 334-351, ISSN 0360-3199, https://doi.org/10.1016/j.ijhydene.2023.08.323.

Md Monjur, H.B., Zahed, S., 2025. Hydrogen as an alternative fuel: A comprehensive review of challenges and opportunities in production, storage, and transportation, International Journal of Hydrogen Energy, Volume 102, 2025, Pages 1026-1044

MME. Ministério de Minas e Energia, 2021. Programa Nacional de Hidrogênio (PNH2). URL: https://www.gov.br/mme/pt-br/programa-nacional-do-hidrogenio-1

PNH2. Programa Nacional de Hidrogênio: Plano de Trabalho Trienal 2023-2025. 2023. URL: https://www.gov.br/mme/pt-br/programa-nacional-do-hidrogenio-1/planodetrabalhotrienalpnh2.pdf

Qian, H., Mingbin, Z., Li, S., Li, X., Wang, Z., 2025. Machine Learning Prediction of Photovoltaic Hydrogen Production Capacity Using Long Short-Term Memory Model. Energies. 18. 543. 10.3390/en18030543.

Saumia, R., Belkhir, N., Noureddine, S., Javier, D., Abderahman, G., 2017. Prospects of hydrogen production potential from renewable resources in Algeria. International Journal of Hydrogen Energy. 42, 1383-1395.

SONDA, 2024. Petrolina station - local climatology. URL: https://sonda.ccst.inpe.br/estacoes/petrolina_clima.html . . Data accessed: August 27,2025

SONDA, 2025. Petrolina's station: environmental. URL: https://sonda.ccst.inpe.br/basedados/petrolina.html. Data accessed: March 27,2025.

Syed, A.H., Muhammad, S., Saeed, I., 2021. Forecasting hydrogen production potential in islamabad from solar energy using water electrolysis, International Journal of Hydrogen Energy, Volume 46, Issue 2, 2021,Pages 1671-1681, ISSN 0360-3199.

WRMC, 2025. World Monitoring Center – Baseline Surface Radiation Network. URL: https://bsrn.awi.de/. Data accessed: August 27,2025.

Zhou, Z., 2012. Ensemble Methods: Foundations and Algorithms. Taylor and Francis Group, LLC.