



MACHINE LEARNING-BASED VIRTUAL SENSOR FOR ESTIMATING PYRANOMETER IRRADIANCE USING LOW-COST TEMPERATURE DATA AND THEORETICAL SOLAR MODELS

Ignacio Sánchez-Gendríz¹, Olavo Cardozo², Filipe Arruda³

Centro de Tecnologias Estratégicas do Nordeste – ¹ignacio.gendríz@cetene.gov.br; ²olavo.cardozo@cetene.gov.br;

³filipe.arruda@cetene.gov.br

RESUMO

O monitoramento da irradiância solar é essencial para o dimensionamento, avaliação e operação de sistemas fotovoltaicos. Embora os piranômetros forneçam medições precisas, seu custo e requisitos de instalação podem limitar seu uso em larga escala, especialmente quando comparados a alternativas de menor custo. Este estudo investiga o uso de modelos de aprendizado de máquina (ML) como sensores virtuais para estimar a irradiância medida por piranômetros, utilizando apenas dados de temperatura obtidos por sensores LM35 de baixo custo e irradiância teórica derivada de modelos de céu claro. Um conjunto de dados de nove dias consecutivos foi processado por meio de interpolação linear e suavização para tratamento de valores ausentes e limpeza dos dados. Quinze modelos de regressão foram treinados e avaliados utilizando a temperatura e a irradiância teórica como variáveis de entrada. Modelos de ensemble e perceptrons multicamadas (MLP) apresentaram desempenho superior em relação às demais abordagens, alcançando coeficientes de determinação (R^2) acima de 0,97. Esses resultados indicam que estimativas baseadas em ML podem se aproximar com alta precisão das leituras de piranômetros, quando treinadas com características de entrada apropriadas. A abordagem proposta representa uma alternativa de baixo custo para a estimativa de irradiância em contextos onde a instalação de piranômetros não seja viável, podendo contribuir para a análise e o monitoramento da energia solar com menor demanda por instrumentação.

Palavras-chave: Energia solar; estimativa de irradiância; aprendizado de máquina; sensores virtuais; monitoramento de baixo custo.

ABSTRACT

Solar irradiance monitoring is essential for the design, evaluation, and operation of photovoltaic systems. While pyranometers provide accurate measurements, their cost and installation requirements may limit their widespread use when compared to lower-cost alternatives. This study investigates the use of machine learning (ML) models as virtual sensors to estimate pyranometer-measured irradiance using only low-cost LM35 temperature data and theoretical irradiance derived from clear-sky models. A nine-day dataset was processed using linear interpolation and smoothing to handle missing values and data cleaning. Fifteen regression models were trained and evaluated using temperature and theoretical irradiance as input features. Ensemble models and multilayer perceptrons (MLP) outperformed other approaches, achieving coefficient of determination (R^2) values above 0.97. These results suggest that ML-based estimations can closely approximate pyranometer readings when trained with appropriate input features. The proposed approach provides a potential low-cost alternative for irradiance estimation in



INTERNATIONAL SYMPOSIUM ON ENERGY

Optimizing renewable energy: advanced strategies for enhanced efficiency and sustainability

scenarios where deploying pyranometers may not be feasible. Such virtual sensors could support solar energy analysis and monitoring with reduced instrumentation requirements.

Keywords: Solar energy; irradiance estimation; machine learning; virtual sensors; low-cost monitoring.

1 INTRODUCTION

Solar energy has emerged as one of the most promising renewable resources for addressing global energy demands while mitigating environmental impacts and fostering economic development (DE JONG et al., 2019; MAKKA; ALABID, 2022). Its strategic relevance lies not only in its potential for large-scale electricity generation but also in its role in advancing sustainable development goals (PAZIKADIN et al., 2020). In Brazil, the Northeast region stands out as a key hub for renewable energy production, particularly photovoltaic (PV) systems, due to the region's high solar irradiance levels and favorable climatic conditions throughout the year (VIEIRA, 2024).

The performance evaluation and operational management of PV systems rely on accurate, continuous monitoring of environmental and electrical variables. The integration of Internet of Things (IoT) technologies into PV monitoring systems enables real-time or near real-time data acquisition, which supports predictive maintenance, fault detection, and intelligent energy management (RAO; SAHOO; YANINE, 2023). Among the key monitored variables are relative humidity, ambient temperature, solar irradiance, current, and voltage (ALOMBAH et al., 2025). From current and voltage, the generated power can be calculated, enabling performance diagnostics and optimization strategies.

Solar irradiance measurement is typically performed using pyranometers, which provide precise data but could involve significant cost and installation requirements. In contrast, low-cost temperature sensors such as the LM35 are inexpensive, easy to deploy, and widely available (DEMIR, 2023). If such sensors could be combined with theoretical irradiance models to accurately predict measured irradiance, they could form the basis of “virtual sensors” — computational models capable of replacing or complementing physical devices. This could reduce instrumentation costs in large-scale PV deployments, particularly in resource-constrained contexts.

Objectives and Contributions

In this work, we propose a low-cost approach for solar energy monitoring by combining temperature data from an LM35 sensor with theoretical irradiance values from a clear-sky model as inputs to machine learning models. Our goal is to assess whether these simple and inexpensive measurements can accurately estimate pyranometer-recorded irradiance, thus reducing system complexity and cost. We train and evaluate 15 regression algorithms, comparing their predictive performance to identify the most effective models for virtual sensor implementation. The findings aim to support the development of cost-efficient, ML-based monitoring solutions for photovoltaic systems, enabling broader deployment and optimized instrumentation.



Manuscript Structure

The remainder of this paper is organized as follows: Section 2 describes the dataset, preprocessing procedures, and ML model training methodology. Section 3 presents and discusses the results, including performance metrics and visual comparisons between measured and predicted irradiance. Finally, Section 4 concludes the study, highlighting the main findings, limitations, and opportunities for future research.

2 METHODOLOGY

Dataset

Continuous experimental data were collected over nine consecutive days from an organic photovoltaic module, with active layer P3HT:PCBM and six cells in series, operating under a fixed voltage, selected based on prior current–voltage (I–V) curve analysis. During the observation period, the module showed higher efficiency under diffuse radiation conditions. The measurements were conducted in Recife, Brazil (latitude 8° 04' 03" S, longitude 34° 55' 00" W).

The dataset included ambient temperature measured by a low-cost LM35 sensor and theoretical irradiance values computed from a clear-sky model using the Python library *pvl* (HOLMGREN; HANSEN; MIKOFSKI, 2018; JENSEN et al., 2023). Pyranometer-measured irradiance was used as the reference target for model training and evaluation.

Preprocessing steps

The raw data were preprocessed to ensure consistency and reduce noise effects. Missing values were handled using linear interpolation, while temporal smoothing was applied to mitigate short-term fluctuations and remove outliers, focusing on preserving the underlying signal trends. All features were temporally aligned to ensure synchronization between sensor measurements and theoretical model outputs.

Machine learning modeling

The dataset was split into training (70%) and testing (30%) sets, following the temporal order of the measurements to preserve the time-series structure and avoid data leakage. Both sets were filtered to retain only instances where the theoretical global irradiance, computed using a clear-sky model, was greater than zero. This step excluded nighttime periods, which do not provide relevant information for irradiance estimation.

The input variables used in the models were: (i) ambient temperature measured by a low-cost LM35 sensor, and (ii) theoretical global irradiance derived from clear-sky modeling for the corresponding minute. The target variable was the measured global irradiance obtained from a pyranometer, serving as the high-precision reference. The application focuses on synchronous estimation — predicting irradiance at the current time step using only information available at that instant — with potential for implementation as a soft sensor (or virtual sensor) to replace expensive hardware in real-time monitoring.

To identify suitable regressors, the training set was used to benchmark 15 classical machine learning algorithms with the LazyPredict Python library (<https://pypi.org/project/lazypredict/>). Although this tool allows evaluation of a much larger set of models, these 15 were intentionally selected to represent a broad diversity of inductive biases and learning strategies. This facilitated a balanced comparison across different algorithmic families. The five top-ranked models from the training phase were then evaluated on the test set to assess their generalization capability.



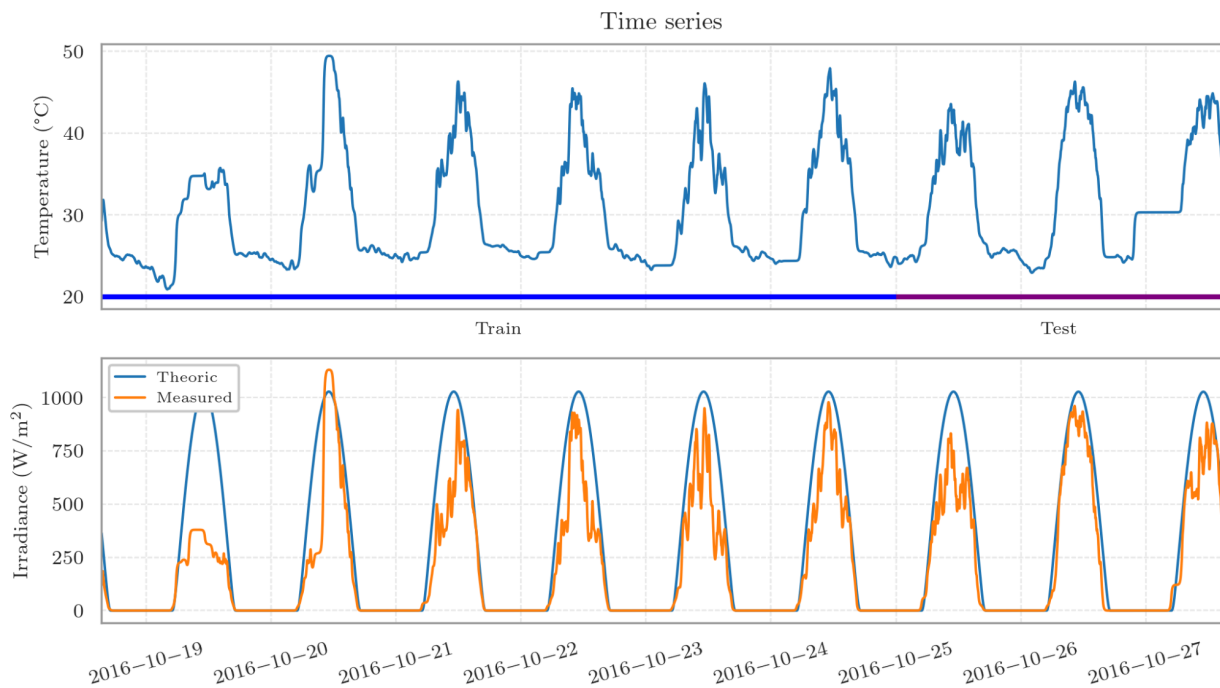
Model performance was evaluated using three widely adopted regression metrics: the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE) (SÁNCHEZ-GENDRIZ; SILVA; GUEDES, 2025). These metrics quantify, respectively, the proportion of variance explained by the model, the magnitude of prediction errors with higher sensitivity to outliers, and the average absolute deviation from the true values.

3 RESULTS AND DISCUSSION

Exploratory analysis of input and target variables

Figure 1 shows the time series of measured temperature, theoretical irradiance, and irradiance measured by the pyranometer, after applying the preprocessing steps described in the Methodology section. Following missing-value imputation and smoothing, the relationship between the theoretical and measured irradiance becomes evident. Moreover, the temperature values acquired from the LM35 sensor exhibit a clear association with the measured irradiance, suggesting that temperature, combined with the theoretical irradiance, can be used to estimate pyranometer measurements through machine learning regression models.

Figure 1 - Time series of measured temperature (upper panel) and irradiance (lower panel), after preprocessing. The lower panel compares theoretical irradiance with pyranometer measurements. The upper panel also shows the training (blue) and testing (purple) periods used for model evaluation.





INTERNATIONAL SYMPOSIUM ON ENERGY

Optimizing renewable energy: advanced strategies for enhanced efficiency and sustainability

The figure also depicts the data split used for model benchmarking. A hold-out cross-validation strategy with a 70/30 split was adopted, as described in the Methodology section. In the upper panel, the training and test periods are indicated in blue and purple, respectively.

ML model performance metrics

To establish a comparative baseline, the LazyPredict Python library was employed to train and evaluate a pool of 15 classical regressors on the development set. These models were chosen to span a broad spectrum of learning paradigms, including Linear regression techniques (e.g., ridge, lasso); Tree-based models (e.g., decision tree, extra tree regressor); Ensemble methods (e.g., bagging, random forest); Kernel-based approaches (e.g., support vector regression) and Neural networks (e.g., multilayer perceptron). This diversity covers a wide range of inductive biases and learning strategies, providing a robust foundation for comparative analysis (SÁNCHEZ-GENDRIZ; SILVA; GUEDES, 2025). The top four models, selected based on development set performance, were further evaluated on the hold-out test set. Table 1 presents the performance metrics for these top-performing regressors. Among the baseline models, the MLPRegressor achieved the highest performance, consistent with the strong approximation capabilities of multilayer perceptrons in modeling complex, non-linear relationships often found in solar energy applications (LEE; LEE, 2023). The subsequent top-ranked models were ensemble-based regressors, and their performance can be attributed to the ensemble learning principle.

Table 1 - Performance metrics on the test set. Results correspond to the top four regressors identified using the LazyPredict Python library.

Model	R ²	Rel_RMSE	Rel_MAE	RMSE	MAE
MLP	0.98	0.04	0.02	41.24	23.58
HistGradientBoosting	0.98	0.04	0.02	43.04	25.11
GradientBoosting	0.98	0.05	0.03	49.82	29.08
RandomForest	0.97	0.05	0.03	52.91	31.13

In this approach, predictions from multiple learners, each trained on different data subsets or with varied configurations, are combined to improve robustness, reduce overfitting, and enhance generalization. All models presented in Table 1 achieved strong performance metrics, as evidenced by R² values above 0.97. The relative error metrics, Rel_RMSE and Rel_MAE, were calculated by normalizing RMSE and MAE against the maximum theoretical irradiance observed during the measurement period. These correspond to errors of approximately 4% and 2%, respectively, relative to the maximum expected values. The following section presents a comparison between the measured irradiance from the reference sensor and the values predicted by the models.

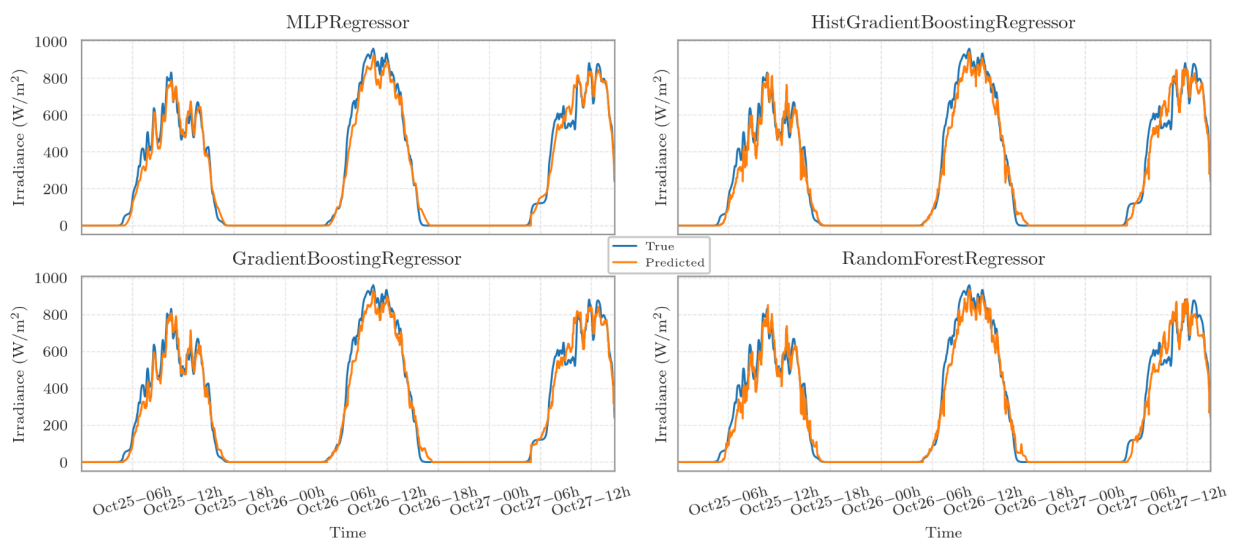
Predicted vs. measured irradiance for the top four ML regressors

Figure 2 shows the time series comparison between the measured irradiance values and the predictions generated by the four best-performing machine learning models. Across all cases, the predicted curves closely match the measured data, reflecting



the strong predictive capabilities indicated by the performance metrics in Table 1. To ensure physical consistency, a post-processing rule was applied to the model outputs: whenever the theoretical irradiance from the clear-sky model is zero (nighttime), the predicted irradiance is also set to zero. This step prevents rare but possible inconsistencies, such as nonzero predictions in periods without solar incidence, which could arise from interpolated input data or other artifacts in the preprocessing stage.

Figure 2 - Time series comparison between measured and predicted irradiance values for the top four regressors: MLPRegressor, HistGradientBoostingRegressor, GradientBoostingRegressor, and RandomForestRegressor.



Reproducibility and Artifact Availability

Reproducibility is crucial in scientific research, as it ensures that results can be verified and extended. Since experimental findings guide policy, planning, and future studies, we provide a computational artifact to enable transparent and consistent analysis reproduction. The artifact is a self-contained Docker image with data preprocessing, machine learning experiments in Jupyter Notebook, and model evaluation.

This approach removes uncertainties about software versions, dependencies, or environments as it allows others to replicate the results exactly as presented. Permanently available on Zenodo¹, it ensures readers, reviewers, and researchers can access the same setup and verify our outcomes. While technical details are not the focus, the artifact adds transparency and strengthens our findings.

4 CONCLUSIONS

The results of this study suggest that machine learning (ML) models trained with low-cost ambient temperature measurements and theoretical irradiance values from a clear-sky computational model can serve as effective predictors of real irradiance measured by a pyranometer. The selected regressors achieved coefficients of determination (R^2) above 0.97 during the study period, and the close agreement between predicted and measured time series supports the predictive capabilities indicated by

¹ The artifact can be accessed at [Zenodo, DOI: 10.5281/zenodo.16983569].



INTERNATIONAL SYMPOSIUM ON ENERGY

Optimizing renewable energy: advanced strategies for enhanced efficiency and sustainability

the performance metrics. A key limitation of this work is the restricted duration of the dataset, which covers approximately 10 consecutive days. While valuable for illustrating the feasibility of training ML models with the proposed inputs, a more rigorous evaluation over extended periods is needed. Such an approach would enable the assessment of model robustness under seasonal variations and potential data or concept drift affecting the relationship between predictors and target variables.

Future research should include temporal cross-validation strategies, incorporate additional environmental variables such as humidity, and explore longer-term monitoring campaigns. Overall, the study indicates the potential of combining IoT-based monitoring with ML models as a cost-effective and scalable approach for solar energy applications, including performance assessment and the optimization of photovoltaic instrumentation deployment.

REFERENCES

- ALOMBAH, Njimboh Henry et al. Advanced IoT-based monitoring system for real-time photovoltaic performance evaluation: Conception, development and experimental validation. *Scientific African*, v. 28, p. e02763, jun. 2025. DOI: 10.1016/j.sciaf.2025.e02763.
- DE JONG, Pieter et al. Estimating the impact of climate change on wind and solar energy in Brazil using a South American regional climate model. *Renewable Energy*, v. 141, p. 390–401, out. 2019. DOI: 10.1016/j.renene.2019.03.086.
- DEMIR, Batikan Erdem. A new low-cost Internet of Things-based monitoring system design for stand-alone solar photovoltaic plant and power estimation. *Applied Sciences*, v. 13, n. 24, p. 13072, dez. 2023. DOI: 10.3390/app132413072.
- HOLMGREN, William F.; HANSEN, Clifford W.; MIKOFSKI, Mark A. pvlib python: a Python package for modeling solar energy systems. *Journal of Open Source Software*, v. 3, n. 29, p. 884, set. 2018. DOI: 10.21105/joss.00884.
- JENSEN, Adam R. et al. pvlib iotools—Open-source Python functions for seamless access to solar irradiance data. *Solar Energy*, v. 266, p. 112092, dez. 2023. DOI: 10.1016/j.solener.2023.112092.
- LEE, Younghyun; LEE, Jonghwan. Solar power prediction modeling based on artificial neural networks under partial shading. *Applied Sciences*, v. 13, n. 18, p. 10013, set. 2023. DOI: 10.3390/app131810013.
- MAKA, Ali O. M.; ALABID, Jamal M. Solar energy technology and its roles in sustainable development. *Clean Energy*, Oxford University Press, v. 6, n. 3, p. 476–483, jun. 2022. DOI: 10.1093/ce/zkac023.
- PAZIKADIN, Abdul Rahim et al. Solar irradiance measurement instrumentation and power solar generation forecasting based on Artificial Neural Networks (ANN): A review of five years research trend. *Science of The Total Environment*, v. 715, p. 136848, mai. 2020. DOI: 10.1016/j.scitotenv.2020.136848.
- RAO, Challa Krishna; SAHOO, Sarat Kumar; YANINE, Franco Fernando. An IoT-based intelligent smart energy monitoring system for solar PV power generation. *Energy Harvesting and Systems*, v. 11, n. 1, Walter de Gruyter GmbH, 2023. DOI: 10.1515/ehs-2023-0015.
- SÁNCHEZ-GENDRIZ, Ignacio; SILVA, Ivanovitch; GUEDES, Luiz Affonso. Kolmogorov-Arnold Networks for interpretable analysis of water quality time-series data. *J*, v. 8, n. 3, p. 24, jul. 2025. DOI: 10.3390/j8030024.
- VIEIRA, B. A. Sustainable potential of renewable energies in the Brazilian Northeast. *Revista de Gestão – RGSA*, v. 18, n. 12, e010554, 2024. DOI: 10.24857/rgsa.v18n12-195.