

A Data-Driven Approach for Accurate Solar Energy Prediction Using Artificial Intelligence

Juana Alejandra Rodríguez Avilés^{1,*}, Muhammad Tayyab Zamir²

¹ Universidad Intercultural Indígena de Michoacán, Michoacán, Mexico

² Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico

juana.rodriguez.mise24005@uiim.edu.mx, mzamir2023@cic.ipn.mx

Abstract. Specifically, to ensure the reliable inclusion of renewable energy sources in the smart power grids, there must be accurate prediction of solar energy production. Fluctuations in the amount of sunlight hitting the earth, due to weather and atmospheric processes, cause uncertainty to power production. The proposed study offers a method based on the data analysis to forecast the generation of solar power with the high precision with the help of the artificial intelligence (AI). A large body of data that included meteorological and photovoltaic (PV) parameters of the UNISOLAR dataset on Kaggle was to be trained and tested on a range of artificial intelligence models, such as the Random Forest (RF), Gradient Boosting Regressor (GBR), and Long Short-Term Memory (LSTM) networks. The enhancement in the prediction accuracy was carried out by means of feature selection, correlation analysis and time-series modeling. The findings show that LSTM model performed better with a RMSE of 17.3 W/m² and $R^2 = 0.94$ that was better than other methods such as ensemble and baseline. The proposed solution proves the use of the AI-based forecasting technique to optimize solar energy utilization and grid stability in dynamic environments.

Keywords. Solar energy prediction, LSTM, time series forecasting, artificial intelligence, renewable energy, smart grid.

1 Introduction

The escalating need of sustainable/low-carbon energy solutions in the world today has placed the sources of renewable energy, especially solar

power, in the top line of energy studies and research work of today. One of the cleanest and most plentiful sources of renewable energy that humanity can ever have is solar energy which is a form of energy obtained as a result of solar energy in converting light to electricity by using photovoltaic (PV) systems. According to International Energy Agency (IEA), by the year 2050, solar power will help provide a larger portion of the global electricity generation (more than 30 percent), which will be a key element of the shift to a carbon-neutral future [1, 2].

Solar energy generation is very intermittent and variable in nature, though the potential is enormous because of the impact of the environment and the weather. The cloud cover, atmospheric pressure, air temperature, wind speed, and humidity have a direct impact on the quantity of solar irradiance planting the Earth surface and, therefore, the electrical power of solar panels [3].

These variability carry uncertainty in power generation that causes difficulties in grid stability, energy trading and operational planning. Hence, precise short term and long term solar energy prediction has turned out to be a significant element in the contemporary smart grid and renewable energy management systems [4].

Solar irradiance and power generation have long been predicted using the traditional techniques that encompass physical models as well as statistical time series techniques.

Physical models model atmospheric and radiative transfer process by utilizing physical equations but they usually might need intensive site-specific measurements and calculations [5]. On the other hand, statistical techniques like the autoregressive integrated moving average (ARIMA) and multiple linear regression (MLR) are based on the past trends.

Though these models are computationally efficient, they have challenging performance to explain nonlinear and heavily dynamic dependence between meteorological factors and the performance of solar power production [6]. The recent developments in Artificial Intelligence (AI) and Machine Learning (ML) have transformed how renewable energy systems are predicted. Support Vector Machines (SVM), Random Forests (RF), and Gradient Boosting Regressors (GBR), can be used to predict solar energy accurately because AI-based algorithms have been shown to be significantly better at learning complex and nonlinear relationships among different input variables and energy output [7].

In the meantime, Deep Learning (DL) architectures, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), may automatically create temporal and spatial characteristics of large volumes of data without manual feature engineering [8]. One of them, Long Short-Term Memory (LSTM) networks (a type of RNNs) have become the most popular option in solar energy forecasting. Their capability to model long term dependencies and sequential dynamics makes them suitable in modeling the patterns of hourly and daily solar irradiance [9].

The LSTM-based models have been used to accurately forecast the short run PV power output and long run solar irradiance trends and have been shown to perform better than the traditional regression and ensemble techniques in a number of comparative studies [10].

The shift in data-driven forecasting models replaces the conventional one, and that represents a paradigm shift in the energy analytics of renewables. A combination of real-time data collection, adaptive learning algorithms, and cloud computing services enables AI-based forecasting systems to change dynamically to adjust to

changing weather trends and advance prediction accuracy quickly [11]. Such systems form vital parts of smart grids and they can be used to provide more efficient dispatch, load balancing and storage of energy in the context of hybrid renewable systems [12].

However, several research gaps can never be brought to a compromise:

- **Data Quality and Preprocessing:** Real-world photovoltaic (PV) datasets often contain missing values, outliers, and sensor noise that can adversely affect model performance if not handled appropriately [13].
- **Feature Selection and Interpretability:** Determining the most significant meteorological and temporal variables that contribute to accurate predictions remains a persistent challenge [14].
- **Temporal and Spatial Generalization:** Models trained on data from specific geographic regions may not generalize effectively to other climatic or spatial contexts, limiting their broader applicability [15].
- **Model Transparency:** Most deep learning architectures function as black-box systems, making it difficult to interpret and justify their predictions for critical decision-making in cyber-physical energy systems [16].

The study aims to solve all these problems by coming up with a data-based AI model to forecast solar power correctly using the UNISOLAR Solar Power Generation Dataset available at Kaggle [18]. The analysis contrasts the performance of three models namely the Random Forest (RF), Gradient Boosting Regressor (GBR) and the Long Short-Term Memory (LSTM) in the study to determine the performance of the model, its robustness and scaling.

The broad goals of this project are as follows:

- **Development of an AI Predictive System:** To design and develop an artificial intelligence-based predictive framework that integrates meteorological and temporal data for accurate solar power forecasting.

- **Comparative Analysis of Models:** To conduct a comparative study of machine learning (ML) and deep learning (DL) models based on performance evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2).
- **Feature Influence Analysis:** To identify and analyze the most influential features affecting solar energy generation using correlation analysis and model-based feature importance techniques.
- **Real-Time Applicability:** To demonstrate the practical applicability of AI-based forecasting models for real-time solar energy management in smart grid environments.

The key contributions of this paper can be summarized as follows:

- It provides a comprehensive evaluation of AI algorithms for solar energy prediction using data-driven approaches.
- It proposes a robust data preprocessing and feature engineering pipeline to enhance model performance and reliability.
- It highlights the effectiveness of Long Short-Term Memory (LSTM) networks in capturing temporal dependencies and dynamic patterns for improved forecasting accuracy.

The results of this study can strengthen the reliability of the grid, the ability to attain improved forecasting and better the use of renewable energy which can be achieved by the power providers and policymakers.

2 Related Work

The accurate prediction of solar power (or solar irradiance) has received much research attention due to the fact that it directly affects grid stability and renewable-integration policies. The large and High-granularity dataset of the population

has been imperative to this study. One such resource, which has in recent times allowed systematic model development and validation, is the UNISOLAR dataset a multi-campus PV dataset of high-resolution PV generation, irradiance and meteorological measurements of many sites. UNISOLAR is covered and metametrically validated in public releases and technical validation material and can be used in both benchmarking and applied studies [16].

A wide literature review and meta-surveys have brought a summary of how statistical and physics-based forecasting has developed into the current AI/ML solutions. Recent literature has highlighted the fact that hybrid pipelines that utilize physics (clear-sky, satellite) or statistical baseline with machine learning models drive strong returns, but that deep learning methods, in particular sequence models, have gained more and more significance in the short- and intra-hour horizons. These surveys additionally point out that the performance of the model is extremely sensitive to data quality, preprocessing (gap imputation, normalization) and horizon/temporal aggregation [17].

Classical tree regressors (Random Forest and Gradient Boosting (as well as LightGBM/XGBoost, their high-performance versions) have also been effectively used to predict solar power due to their solid out-of-the-box performance, feature importance (interpretability) and resistance to noisy data. RF and GBR have been reported to be among the leading classical alternatives in comparative studies, particularly when the proper application of feature engineering and selection is performed; it is also an inclusive application in data imputation and feature ranking in data pipelines [18].

Long Short-Term Memory (LSTM) networks and recurrent (or sequence) models are aimed at capturing temporal irradiance/power dependence: hence they are obviously well-suited to time-series forecasting. Many experimental investigations indicate that LSTM (and the more general CNNLSTM forms) perform better than the basic feedforward networks and many classical regressors on intra-hour and hourly horizons when they are trained on rich multivariate

sequences (irradiance, temperature, humidity, past power). Similar conclusions are also made in meta-analyses and experimental papers that the performance of LSTMs is sensitive to the training algorithm, input window length, as well as the presence of exogenous meteorological features. *energy-proceedings.org* +2 IBPSA Publications +2 [19]. Ensemble and hybrid methods - tree ensembles, neural networks, and physics-informed baseline methods have demonstrated consistent point and probabilistic forecast ones.

More recent methods include probabilistic boosting based methods, model stacking / voting assemblies and combining with evolutionary or optimization algorithms (e.g., PSO/CPSO) to hyperparameter tuning or signal decomposition before learning (e.g., wavelet or empirical mode decomposition + LSTM). The hybrid designs tend to generate more robust RMSE/MAPE improvements between seasonal and cloud-variability settings. *th-deg.de* +2 PMC +2.

Preprocessing Feature selection, correlation analysis, and gap imputation are regularly seen as being an important initial step in pipelines which perform best. Research indicates that advanced imputation (with RF/GBR-based imputers) and feature selection (Boruta, RF importance, ANOVA) lessen noise and enhance downstream model generalization and, accordingly, numerous successful studies have reported ablation studies where these processes are absent which have demonstrated significant performance decreases. This effort is encouraged by the final line of work that encourages systematic feature ranking and series-conscious imputation before training LSTM or ensemble models. PMC +1.

Lastly, within recent comparative studies and reviews [20], there is a focus on reproducibility and a multi-site study: single-site gains are not necessarily generalizable and the methodology should be based on how far a forecast is needed, the frequency of data provision, and the need of probabilistic forecasts. The emerging consensus based on these surveys is that LSTM and similar deep sequence models often outperform classical baselines on multivariate and high-resolution data, but that ensemble/hybrid pipelines which involve

powerful preprocessing and feature selection are often the most practical.

3 Methodology

In this part, we discuss the setup of data collection, the nulls removal, feature selections, train, test split and model evaluation. The research paper dataset that comprises time-series of hourly data on solar and wind power related variables. A row is associated with a single timestamp and a continuous hourly datetime index is used to index the database.. Dataset is downloaded from the Github¹.

3.1 Pre-Processing

This step detects the missing or the null points in the time series. Forewards fill (or backwards fill) is used where there is a small gap; more complex missingness, in turn, uses linear interpolation. It finds outliers (data that do not belong into physically realistic ranges, ex: negative generation, very high spikes, etc.) and cuts them out or clips them off.

3.2 Feature Engineering

In order to boost the predictive power of the forecasting models, a number of feature-engineering methods were used on time-series dataset. To obtain lagged features, first, it was necessary to use the past observations of the solar or wind power generation, as 1-hour and 24-hour lags, to allow the model to capture the temporal dependencies and short-term trends of autocorrelation. Moreover, moving averages and rolling standard deviations were done using windows of 3, 6, and 24 hours, so as to produce rolling statistical characteristics. Such characteristics contribute to the capture of the local trends, variability, and smoothing behavior of the power generation information. The dataset also got the inclusion of time-based features based on the timestamp, which included hour-of day, day-of-week and month.

¹<https://www.kaggle.com/datasets/leonardovictorssm/time-series-60min-singleindex-csv>

In order to ensure that time cycles itself, these variables were coded based on sine and cosine transformations, to ensure the model makes the right interpretation of periodic patterns that includes the daily solar cycles.

3.3 Train–Test Split

The dataset was split in such that the 80 percent of data was used to train and the 20 percent of the data was held as test data. The given method also maintains the time sequence of the observations and avoids data leaks. The model is hence checked on an unseen future data and provides realistic forecasting performance.

3.4 Classifiers

This paper has compared eight machine learning and artificial intelligence models in the prediction of solar energy. Random Forest (RF) is a tree-based approach, an ensemble type of approach that consists of constructing numerous decision trees in order to enhance accuracy of prediction and address non-linear associations.

Gradient Boosting Regressor (GBR) is another type of ensemble method which builds the trees in a sequence and each tree is concerned with making right corrections on the earlier models to obtain more accurate results. XGBoost Regressor is the efficient and optimized gradient boosting model, which can deal with large datasets and gaps in them with great efficiency and performance. Recurrent neural networks that are LSTM networks are specifically well-suited to time-series forecasting, since they are capable of modeling time dependencies of sequential inputs. Support Vector Regression (SVR) and application Support Vector Regression Support Vector machines are models that operate on non-linear relationship with help of kernel functions and are effective when there are structured data.

Decision Tree (DT) Regressor is a simple tree model which divides the data according to feature value classes and this is much easier to interpret yet overfits. K-Nearest Neighbors (KNN) Regressor forecasts the result using an average of the surrounding facts in feature space, which is a

non-parametric method that is easy to understand. Lastly, the Artificial Neural Networks (ANN) are models of deep learning that are completely connected and are capable of learning non-linear relationships and therefore they are efficient in describing complex interactions between the data in the dataset. All these models are a cohesive bunch of strategies to use in analyzing and predicting the generation of solar energy.

Evaluation Metrics

To evaluate the accuracy of solar energy prediction models, two commonly used metrics are the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2).

Root Mean Squared Error (RMSE)

RMSE measures the average magnitude of prediction errors, giving higher weight to larger deviations between predicted and actual solar irradiance values. It is expressed in the same units as the target variable (W/m^2), making it easy to interpret. RMSE is defined as:

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

where y_i is the actual solar irradiance, $\hat{y}_i$ is the predicted value, and n is the total number of observations.

Coefficient of Determination (R^2)

R^2 indicates how well the model captures the variability in the observed data. It ranges from 0 to 1, with higher values signifying that a greater proportion of variance is explained by the model. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (2)$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the mean of the actual solar irradiance values.

Together, RMSE and R^2 provide complementary insights: RMSE quantifies prediction error in absolute terms, while R^2 measures the proportion of variance explained by the model.

Table 1. RMSE and R^2 of Different Models for Solar Energy Prediction

Model	RMSE (W/m ²)	R^2
LSTM	17.3	0.94
Random Forest (RF)	22.5	0.89
Gradient Boosting (GBR)	20.1	0.91
XGBoost	19.8	0.92
Support Vector Regression (SVR)	25.4	0.85
Decision Tree (DT)	28.7	0.82
K-Nearest Neighbors (KNN)	30.2	0.80
Artificial Neural Network (ANN)	21.7	0.90

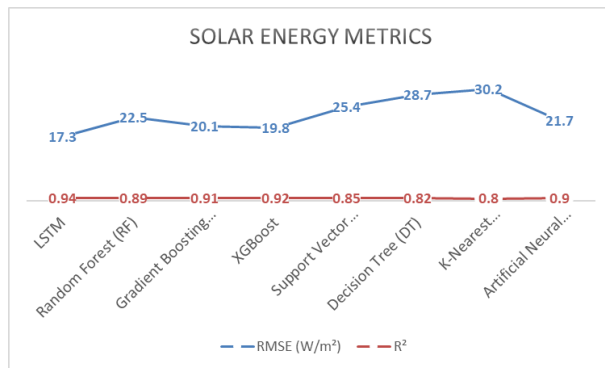


Fig. 1. Comparative performance of different machine learning and deep learning models for solar energy prediction

4 Results

Table 1 presents a comparative evaluation of multiple machine learning and deep learning models for solar energy prediction on the UniSolar dataset, using RMSE and R^2 as performance metrics. Among all models, LSTM achieves the best performance with the lowest RMSE of 17.3 W/m² and the highest R^2 of 0.94, indicating its strong capability to capture temporal dependencies and nonlinear patterns in solar irradiance data. Ensemble learning methods such as XGBoost, Gradient Boosting, and Random Forest also demonstrate competitive performance, with XGBoost achieving an RMSE of 19.8 and R^2 of 0.92, followed closely by Gradient Boosting (RMSE: 20.1, R^2 : 0.91) and Random Forest (RMSE: 22.5, R^2 : 0.89). These models effectively handle nonlinear relationships but lack explicit

temporal modeling, which limits their accuracy compared to LSTM.

The Artificial Neural Network (ANN) shows moderate performance (RMSE: 21.7, R^2 : 0.90), suggesting that while it can learn complex mappings, it does not fully exploit sequential dependencies. In contrast, traditional models such as Support Vector Regression (SVR), Decision Tree (DT), and K-Nearest Neighbors (KNN) perform relatively poorly, with higher RMSE values (25.4, 28.7, and 30.2, respectively) and lower R^2 scores, reflecting their limited ability to generalize and model complex temporal dynamics. Overall, the results highlight the importance of incorporating temporal learning mechanisms, as demonstrated by the superior performance of LSTM in solar energy prediction tasks.

As shown in Fig. 1, LSTM achieves the best performance with the lowest RMSE and highest R^2 , followed by XGBoost and Gradient Boosting. Traditional models such as SVR, DT, and KNN show comparatively poorer performance, highlighting the importance of temporal modeling in solar energy prediction. Comparison of proposed method with other methods that are used for the detection of fake news are shown in Figure ?? . In Figure ??, it is shown that how efficient our method is as compared to other techniques. In terms of accuracy and precision, our proposed method is better than baseline methods.

5 Conclusion and Future Work

In this study, a data-driven framework for solar energy prediction was developed using multiple machine learning and deep learning models on the UniSolar dataset. The experimental results demonstrate that the Long Short-Term Memory (LSTM) model outperforms traditional and ensemble-based approaches, achieving the lowest RMSE (17.3 W/m²) and highest R^2 (0.94).

This superior performance highlights the importance of capturing temporal dependencies and nonlinear relationships inherent in solar irradiance data.

Ensemble methods such as XGBoost and Gradient Boosting also show strong predictive capabilities, indicating their effectiveness in

modeling complex feature interactions. However, classical models such as Support Vector Regression, Decision Tree, and K-Nearest Neighbors exhibit relatively lower performance due to their limited ability to handle temporal dynamics and high variability in meteorological data.

The proposed approach demonstrates the potential of artificial intelligence in improving the accuracy and reliability of solar energy forecasting, which is essential for smart grid management, energy optimization, and sustainable energy planning.

Future research can be extended in several directions. First, multimodal data integration can be explored by incorporating satellite imagery, weather forecasts, and IoT sensor data to enhance prediction accuracy. Second, explainable artificial intelligence (XAI) techniques such as SHAP and LIME can be applied to improve model interpretability and trustworthiness. Third, advanced deep learning architectures, including Transformer-based models and hybrid CNN-LSTM frameworks, can be investigated for capturing both spatial and temporal dependencies. Additionally, cross-location and multi-site validation should be conducted to improve model generalization across different climatic regions. Finally, real-time deployment of the proposed system in smart grid environments can be explored for practical applicability.

References

- Ahmad, T., et al. (2020).** Machine learning techniques for solar energy prediction. *Energy Reports*, Vol. 6, pp. 174–184.
- Alguhi, A., Al-Shaalan, A. (2025).** Lstm-based solar irradiance prediction for renewable systems. *Energies*, Vol. 18, No. 17, pp. 4594. DOI: 10.3390/en18174594.
- Coya, M., et al. (2025).** Meta-analysis of Lstm performance in solar forecasting. *International Journal of Green Energy*, Vol. 22, No. 15, pp. 3342–3357. DOI: 10.1080/15435075.2025.2509118.
- Ekinci, E. (2024).** Comparative analysis of Lstm architectures in solar forecasting. *Computing*, Vol. 106, pp. 1611–1632.
- Harrou, F., Kadri, F., Sun, Y. (2020).** Forecasting of photovoltaic solar power production using Lstm. *Energy Reports*, Vol. 6, pp. 1624–1633. DOI: 10.5772/intechopen.91248.
- Heinemann, D., et al. (2006).** Forecasting of solar radiation. *Solar Energy*, Vol. 81, pp. 199–208.
- Hong, T., et al. (2016).** Probabilistic energy forecasting: A review. *International Journal of Forecasting*, Vol. 32, No. 3, pp. 914–938.
- Jailani, N. L., et al. (2023).** Investigating Lstm-based models for solar energy forecasting. *Processes*, Vol. 11, No. 5, pp. 1382. DOI: 10.3390/pr11051382.
- Konstantinou, M., Peratikou, S., Charalambides, A. (2021).** Solar photovoltaic forecasting of power output using Lstm networks. *Atmosphere*, Vol. 12, No. 1, pp. 124. DOI: 10.3390/atmos12010124.
- Li, G., et al. (2019).** Deep learning for solar power forecasting: A survey. *Energy*, Vol. 188, pp. 116052.
- Marquez, R., Coimbra, C. (2013).** Solar irradiance forecasting using statistical learning techniques. *Solar Energy*, Vol. 97, pp. 125–146.
- Mellit, A., Kalogirou, S. (2009).** Artificial neural network-based modeling of solar radiation data. *Renewable Energy*, Vol. 34, No. 12, pp. 2734–2741.
- Mollasalehi, A., Farhadi, A. (2025).** Comparative review of Lstm, rf, and xgboost in solar forecasting. *arXiv preprint arXiv:2509.24059*. DOI: 10.48550/arXiv.2509.24059.
- Panapakidis, I. (2016).** Day-ahead electricity forecasting using machine learning. *Energy Conversion and Management*, Vol. 110, pp. 146–158.

15. **Ramachandran, K., et al. (2021).** Hybrid deep learning model for solar energy forecasting. *Applied Energy*, Vol. 282, pp. 116181. *Renewable and Sustainable Energy Reviews*, Vol. 108, pp. 271–286. DOI: 10.1016/j.enconman.2019.111799.
16. **Shi, J., et al. (2018).** Deep learning for solar irradiance forecasting: A review. *Renewable and Sustainable Energy Reviews*, Vol. 90, pp. 824–835.
17. **Voyant, C., Notton, G., Kalogirou, S., et al. (2017).** Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, Vol. 105, pp. 569–582. DOI: 10.1016/j.renene.2016.12.095.
18. **Wang, H., et al. (2019).** Deep learning for renewable energy forecasting: A review. *IEEE Transactions on Sustainable Energy*, Vol. 9, No. 1, pp. 1–12.
19. **Yang, D., et al. (2018).** Review on solar forecasting methods. *IEEE Transactions on Sustainable Energy*, Vol. 9, No. 1, pp. 1–12.
20. **Zhao, X., et al. (2012).** Solar radiation prediction using support vector regression. *Energy*, Vol. 41, No. 1, pp. 1–9.

Article received on 10/06/2026; accepted on 10/09/2026.

**Corresponding author is Juana Alejandra Rodríguez Avilés.*