Prediction of Photovoltaic Energy Time Series: a Comparative Study between LSTM, Hybrid and XGBoost Models

Noel Freire-Mahía, Agustín García Fischer, Antonio Díaz-Longueira, Esteban Jove, and Héctor Quintián

University of A Coruña, Department of Industrial Engineering, CTC, CITIC, Ferrol, A Coruña, Spain
Correspondence: noel.freire@udc.es

DOI: *https://doi.org/10.17979/spu.23.c40*

Abstract: In recent years significant progress has been made in the field of renewable energy, with photovoltaics standing out in particular. This is partly because the users usually try to use clean energy to protect the planet, as well as seeking energy independence and saving money. However, one of the main disadvantages of implementing photovoltaic systems is knowing how much energy will be generated and how to manage it. For this reason, multiple models have been created that are capable of making accurate predictions. This article uses historical data from a simulated installation with PVGIS located at the epef, based on which some of the most prominent methods for making these predictions are presented and analysed, specifically LSTM, XGBoost and hybrid models. These predict the power generated at two stages with acceptable accuracy, enabling the user to improve energy planning. Finally, a comparison will be made, concluding with a brief recommendation on which one to use depending on the context.

1 Introduction

Today, energy production is a key factor in developing countries. In the last century, fossil fuels such as oil and coal were used to produce energy, but in recent years many international organisations have been trying to promote long-term sustainable energy sources such as wind, solar and hydroelectric power. The main reasons for this are the reduction of greenhouse gases and the availability of an "unlimited" energy source compared to non-renewable sources (Rusilowati et al., 2024),

In Spain, for the first time in the country's history, non-renewable energy production surpassed traditional sources in 2023. One of the main drivers of this trend was and continues to be photovoltaic solar energy, as production has increased by 300% since 2020 (transmission system operator, 2025). This is because Spain has a large number of hours of sunshine per year (2800 hours / year (Gil et al., 2015)) compared to other European countries. Other reasons include national funding programmes which contribute to the economic factor by reducing installation costs, such as a grant scheme created by the Ministry for Ecological Transition and Demographic Challenge (MITECO) (Royal Decree 477/2021, 2021) , or European programmes, such as NextGenerationEU (Sanahuja Perales, 2022), and, finally, the advantage that each user can disconnect from the electricity grid thanks to self-consumption with their own installation.

However, there are also some significant disadvantages, such as efficiency, since only 25% of sunlight is converted into electricity, and the depreciation cost. But the main problem is the irregularity of production, since the user cannot know how much energy will be converted in a few hours. To address this challenge, several predictive models are created to estimate

power production, allowing user to manage it properly (store it, use it, etc.). Many articles deal with similar topics, but they use a single model without comparison ((Campos et al., 2024) or (Park et al., 2021)) or the data extraction is from their own installation (Nithya et al., 2024). This article has some distinctive features, as it compares several methods to see which one best predicts the energy generated in 1-2 hours in an educational building based on PVGIS data.

The structure of this document is as follows: after this introduction, the Materials and Methods section is presented. Next, the Development section describes the steps followed to develop the forecasting models. Finally, the obtained results are presented, and the conclusions and future work are discussed.

2 Materials

The dataset was obtained from Photovoltaic Geographical Information System (PVGIS), which "provides information on solar radiation and photovoltaic system performance for any location in the world, except the North and South Poles" (European Union, 2025). This tool provides a large amount of detailed information on various aspects, most notably the power generated. In fact, the main objective of this wwebsite is to assess the viability of a solar photovoltaic system in a specific location and to show the potential of photovoltaic energy throughout Europe (Šúri et al., 2005). The program registered data from Polytechnic Engineering School of Ferrol, (University of La Coruña), located in Galicia, Spain (Latitude 43.482° N. Longitude -8.223° W). This location was selected because it was where the MestreTIC contract practices were carried out, as well as being in an academic environment capable of providing the ideal context for research and learning. The solar panels are made of crystalline silicon and fixed to the roof with a 35 ° inclination to optimize radiation capture. The installation has 1 kWp of power with a system loss percentage of 14%. The dataset contained information registered from January 2017 to December 2020.

Once the above configuration has been selected, the CSV is downloaded with the data, which is composed of the power obtained with the photovoltaic panels in one hour (W), the total wind speed at 10 metres from the ground (m/s), air temperature at 2 metres (degrees Celsius), height of the sun (degrees), the global irradiance on the panel (W/m2), the date and time at which these data were taken. The model will work with a volume of 35064 data for each covariate.

3 Methods

Now the available models for time series forecasting (TSF) will be explained below.

LSTM

Long short-term memory is a special type of recurrent neural network (RNN) designed to learn long-term dependencies in sequences. It is widely used for TSF because it is based on a structure with "gates" (input, forget, and output) that allow different sequences to be processed and information to be retained or discarded at each time step (Smagulova and James, 2019).

XGBoost

eXtreme Gradient Boosting (XGBoost) is a model consisting of a set of sequentially trained decision trees. This definition is similar to a standard random forest, but there are quite notable differences between the two methods.

1. **Random Forest → Bagging**. This consists of creating several unrelated (independent) decision trees, trained with a random sample with replacement (bootstrap) (Altman and Krzywinski, 2017).

2. **XGBRegressor → Boosting**. This technique means that instead of creating trees with random weights, they are created sequentially and each tree attempts to correct the errors of the previous one. This is achieved thanks to the gradient (derived from the loss function) that indicates the direction and magnitude of the error. It also has optimisation techniques such as EarlyStopping, which makes it more efficient (Chen and Guestrin, 2016).

Hybrid. CNN-LSTM-AM

The proposed state-of-the-art hybrid architecture consists of three main layers, each with a specific function. The CNN layer is responsible for learning repetitive patterns or sudden changes within small time windows, so that if a small sudden change occurs that LSTM networks cannot detect, this layer is able to identify it and adapt predictions to these variations. The LSTM layer, described in detail in section 3, is designed to learn long-term dependencies within the sequence. Finally, the attention mechanism allows the model to focus on the most relevant parts of the sequence when making predictions (Vaswani et al., 2017).

Together, this forms an architecture capable of capturing both the global behaviour of the series and local details, as well as identifying which parts of the input sequence are most relevant to the prediction.

4 Development

This section will explain the steps that were followed to ensure the correct processing of data and the forecasting models. This is presented in the figure 1.

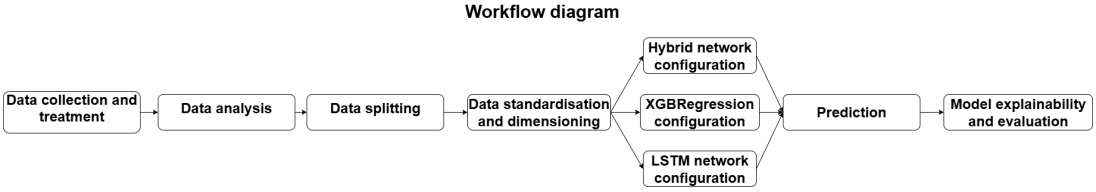

Figure 1: Flowchart of the project.

4.1 Data collection and treatment

Initially, data is obtained from PVGIS and the date variable is processed to adapt it to a more easily interpretable format. After that, the date is readapted to a more suitable format, so a series of operations are performed in Excel to convert it from this format "20200613:0210" to this other "2020-06-13 02:01".

4.2 Data analysis

The first step in the Python environment is to obtain the relationship between all the variables and thus obtain a little more information about the data set used. To do this, a correlation matrix was created.

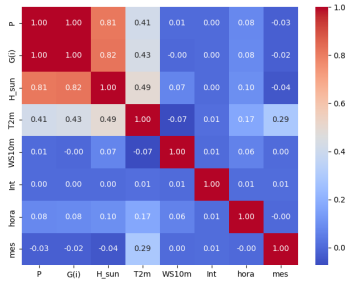

Figure 2: Correlation matrix of the variables in the dataset

As can be seen in the figure 2, there is a strong relationship between power, radiation, and the height of the sun. This information is used as evidence to verify whether the data set is correct, since if there is no correlation, it means data is erroneous.

Next, a fast Fourier transform (FFT) is calculated to extract the most relevant frequencies of the variable to be predicted [(Lucarini et al., 2021)].

The highest frequency peak has a approximate value of 24 hours ($\frac{1}{0.04166667} = 23.9995$).

The second highest frequency is equivalent to 12 hours ($\frac{1}{0.0833333} = 12.000005$).

Finally, the third peak has a value of 8766 hours, which is equivalent to 1 year ($\frac{1}{0.0001140771 \cdot 24 \cdot 365.25} = 1.0000001$).

With these results, it was decided to work with data from the 24 hours prior to the desired prediction. This means that the model will have 192 input data points ($24 \cdot 8 = 192$) to obtain a power prediction 2 hours in advance.

4.3 Data splitting

The next step is to divide the dataset into three groups. The largest group is the training set, as it is necessary for adjusting the model's weight. The rest of the dataset is divided into a validation set, which is used to maximise the model's performance in the training phase, and a test set, which is used to check the quality of the model with unseen data.

The percentages for each group of data sets are: 80% training, 10% validation and 10% test

4.4 Data Standardisation and dimensioning

Once the dataset has been split, StandardScaler is used to scale the training set. Since there are 8 variables, 8 scalers are needed. Next, we save all the scalers and use them with the other 2 groups. This is done to try to replicate a real-world scenario, since if a new maximum is reached in any variable, the scaler will not be updated instantly.

4.5 Networks configuration

Every model have his own configuration, so these cannot be compared with each others.Therefore, three tables are created to show the configuration of each model.

Neurons in the LSTM layer	60
Optimizer	Adam
Epoch	EarlyStopping=10 (200 max)
Dropout	0.2
Learning Rate	1×10^{-4}
Model input data	Scaled by SandartScaler

Table 1: LSTM model configuration

Max depth	5
Colsample	0.7
Nº of estimators	EarlyStopping=20 (700 max)
Subsample	0.7
Learning Rate	1×10^{-2}
Model input data	Scaled by SandartScaler

Table 2: XGBoost model configuration

Neurons in the CNN layers	[128,64,32]
Neurons in the LSTM layers	[200,100,50,40,20]
Optimizer	Adam
Epochs	EarlyStopping=10 (1000 max)
Dropout	0.3
Model input data	Scaled by SandartScaler

Table 3: Hybrid model configuration

5 Results

With the settings established above, the results obtained are as follows.

Model	**Errors (Scaled)**				**Errors (Watts)**	
	Train_Loss(MSE)	Val_Loss(MSE)	Train_Mae	Val_MAE	RMSE_t+1	RMSE_t+2
LSTM	0.1140	0.1150	0.1816	0.1679	73.12	83.23
XGBoost	0.0903	0.1191	0.1378	0.1666	70.65	87.50
Hybrid	0.1205	0.1189	0.1793	0.1744	74.55	85.40

Table 4: Performance indexes for different models

Looking at these results, it appears that the hybrid model, among the three options, performed worst under these conditions. To verify this, a test was performed on the Figure 3.

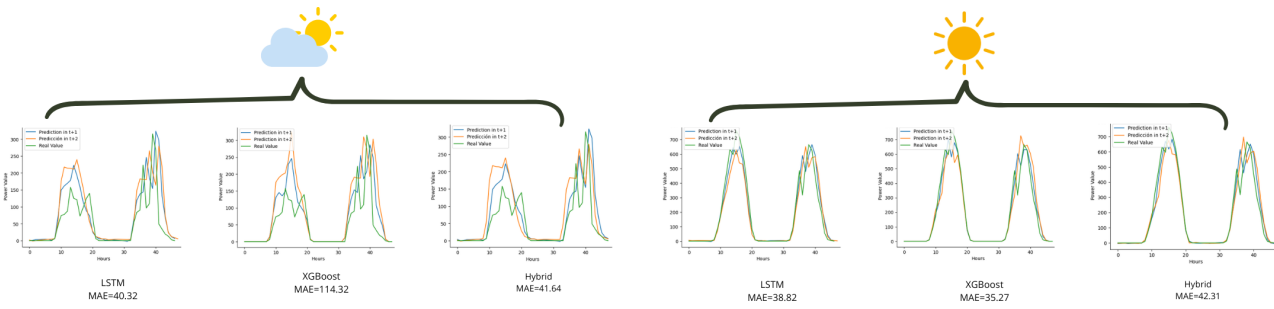

Figure 3: MAE for each model in relation to the type of day

Here we can see the current value (green line) and the prediction made one (blue line) or two (orange line) steps earlier. Below each graph we can see the model used to obtain that prediction and the MAE of this data set (48 h) without scaling. This is used to see if the quality of the forecast between models changes depending on the day.

In this case if the day is cloudy or rainy in summer, hybrid and LSTM models are able to recall similar days from other years and have a very small error compared to the XGBoost model, as the latter works with trees, so it only recalls the most common days, such as sunny days in summer. This statement is clearer in the following figure, as these days are more typical, so XGBoost is more accurate than the other models.

On the other side, if explainability is sought and black boxes are to be avoided, the graphs shown on Figure (4) shows us the feature importance of each model.

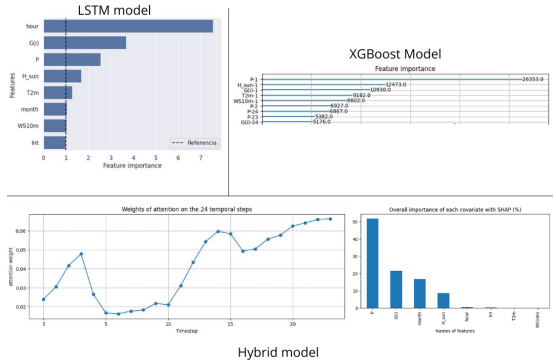

Figure 4: Covariates and timestamp importance of each model

The order of the top three most important covariates may change, but there are always two identical in the top three positions, Power and global irradiance. It can also be observed that in the first three positions of each model there is a fixed time effect covariate: in LSTM, the time (Top 1); in XGBoost, solar radiation H_sun (Top 2); and in the hybrid model, the month (Top 3).

6 Conclusions and Future Works

Once the results are displayed, some conclusions can be drawn. First, there is no clearly better model, as it depends on the type of day, so one option would be to create groups of days and try to predict what type will have. This information could be used to select the right model for each day. Another statement is that the complex model is not always the best; sometimes it results in overfitting or does not fit the data, such as the hybrid model in this case.

In terms of overall results, we can see that the main prediction problems occur on changing days, as we can see in Figure 3. To resolve this, some future work is proposed:

1. **Consider data available at a higher frequency**: A practical way to reduce these errors would be to collect meteorological and energy data at intervals of one to ten minutes. This higher resolution would smooth out sudden variations in actual energy values and, consequently, improve the accuracy of predictions.
2. **Test Large Time Series Models (LTSM)**: Conduct tests with a large model specialised in energy prediction or one that works with internal climate data.

Acknowledgements

This activity is carried out in execution of the Strategic Project "Critical infrastructures cybersecure through intelligent modeling of attacks, vulnerabilities and increased security of their IoT devices for the water supply sector" (C061/23), the result of a collaboration agreement signed between the National Institute of Cybersecurity (INCIBE) and the University of A Coruña. This initiative is carried out within the framework of the funds of the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation), the project of the Government of Spain that outlines the roadmap for the modernization of the Spanish economy, the recovery of economic growth and job creation, for the solid, inclusive and resilient economic reconstruction after the COVID19 crisis, and to respond to the challenges of the next decade.

CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01)

Xunta de Galicia. Grants for the consolidation and structuring of competitive research units, GPC (ED431B 2023/49).

This research is co-financed by the Interreg Atlantic Area Programme through the European Regional Development Fund, EAPA_0019/2022 SAtComm project

Antonio Díaz-Longueira's research was supported by the Xunta de Galicia (Regional Government of Galicia) through grants to Ph.D. (http://gain.xunta.gal), under the "Axudas á etapa predoutoral" grant with reference: ED481A-2023-072.

Bibliography

N. Altman and M. Krzywinski. Ensemble methods: bagging and random forests. *Nature Methods*, 14(10):933–935, 2017.

F. D. Campos, T. C. Sousa, and R. S. Barbosa. Short-term forecast of photovoltaic solar energy production using lstm. *Energies*, 17(11):2582, 2024.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

European Union. Photovoltaic geographical information system (pvgis, 2025. URL *https://joint-research-centre.ec.europa.eu/photovoltaic-geographical-information-system-pvgis_en*.

V. Gil, M. A. Gaertner, E. Sanchez, C. Gallardo, E. Hagel, C. Tejeda, and M. de Castro. Analysis of interannual variability of sunshine hours and precipitation over peninsular spain. *Renewable Energy*, 83:680–689, 2015.

S. Lucarini, M. V. Upadhyay, and J. Segurado. Fft based approaches in micromechanics: fundamentals, methods and applications. *Modelling and Simulation in Materials Science and Engineering*, 30(2):023002, 2021.

C. Nithya, J. P. Roselyn, and D. Devaraj. Real-time solar pv generation in a building using lstm-based time series forecasting. *Discover Electronics*, 1(1):19, 2024.

M. K. Park, J. M. Lee, W. H. Kang, J. M. Choi, and K. H. Lee. Predictive model for pv power generation using rnn (lstm). *Journal of Mechanical Science and Technology*, 35(2):795–803, 2021.

Royal Decree 477/2021. Real decreto 477/2021, de 29 de junio, por el que se aprueba la concesión directa a las comunidades autónomas y a las ciudades de ceuta y melilla de ayudas para la ejecución de diversos programas de incentivos ligados al autoconsumo y al almacenamiento, con fuentes de energía renovable, así como a la implantación de sistemas térmicos renovables en el sector residencial, en el marco del plan de recuperación, transformación y resiliencia. *h*ttps://www.boe.es/eli/es/rd/2021/06/29/477, 2021.

U. Rusilowati, H. R. Ngemba, R. W. Anugrah, A. Fitriani, and E. D. Astuti. Leveraging ai for superior efficiency in energy use and development of renewable resources such as solar energy, wind, and bioenergy. *International Transactions on Artificial Intelligence*, 2(2):114–120, 2024.

J. A. Sanahuja Perales. El pacto verde, nextgenerationeu y la nueva europa geopolítica. Technical report, Fundación Carolina, 2022.

K. Smagulova and A. P. James. A survey on lstm memristive neural network architectures and applications. *The European Physical Journal Special Topics*, 228(10):2313–2324, 2019.

M. Šúri, T. A. Huld, and E. D. Dunlop. Pv-gis: a web-based solar radiation database for the calculation of pv potential in europe. *International Journal of Sustainable Energy*, 24(2):55–67, 2005.

transmission system operator. Electrical network (api), 2025. URL *https://www.ree.es/en/datos/generation*.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.