engineering proceedings

MDPI

*Proceeding Paper*

# Big Data Techniques Applied to Forecast Photovoltaic Energy Demand in Spain [†]

**J. Tapia-García [1], L. G. B. Ruiz [2,*], D. Criado-Ramón [1] and M. C. Pegalajar [1]**

[1] Department of Computer Science and Artificial Intelligence, University of Granada, 18014 Granada, Andalucía, Spain; jorgetapia@correo.ugr.es (J.T.-G.); dcriado@ugr.es (D.C.-R.); mcarmen@decsai.ugr.es (M.C.P.)

[2] Department of Software Engineering, University of Granada, 18014 Granada, Andalucía, Spain

[*] Correspondence: bacaruiz@ugr.es

[†] Presented at the 10th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 15–17 July 2024.

**Abstract:** Renewable energies play an important role in our society's development, addressing the challenges presented by climate change. Specifically, in countries like Spain, technologies such as solar energy assume a crucial significance, enabling the generation of clean energy. This study addresses the critical need to accurately predict photovoltaic (PV) energy demand in Spain. By using the data collected from the Spanish Electricity System, four models (Linear Regression, Random Forest, Recurrent Neural Network, and LightGBM) were implemented, with adaptations for Big Data. The LR model proved unsuitable, while the LGBM emerged as the most accurate and timely performer. The incorporation of Big Data adaptations amplifies the significance of our findings, highlighting the effectiveness of the LGBM in forecasting PV energy demand with both accuracy and efficiency.

**Keywords:** photovoltaic; energy demand; renewable energy; big data; forecasting

check for updates

**Citation:** Tapia-García, J.; Ruiz, L.G.B.; Criado-Ramón, D.; Pegalajar, M.C. Big Data Techniques Applied to Forecast Photovoltaic Energy Demand in Spain. *Eng. Proc.* **2024**, *68*, 11. https://doi.org/10.3390/engproc2024068011

Academic Editors: Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, Hector Pomares and Ignacio Rojas

Published: 3 July 2024

## 1. Introduction

Energy plays a pivotal role in the economic development of nations. In this context, photovoltaic (PV) energy has emerged as a crucial component, contributing significantly to global energy dynamics. PV energy, derived from solar radiation through solar cells, has gained immense global significance as a renewable and sustainable source of power. The technology makes use of the sunlight and converts it into electricity, offering a clean and environmentally friendly alternative to traditional energy sources [1]. With its potential to reduce dependence on fossil fuels, mitigate environmental impacts, and foster energy security, PV energy has become a key player in global efforts toward sustainable energy production and addressing climate change. The widespread adoption of PV technology worldwide reflects its importance in diversifying energy portfolios and contributing to a more sustainable future.

In the Spanish context, PV energy holds particular relevance as the nation strives to diversify its energy mix and transition toward cleaner and more sustainable sources [2,3]. Spain has an ideal environment for utilizing solar energy through PV technology, thanks to its abundant sunlight. This adoption of PV energy supports Spain's goals of meeting renewable energy targets and addressing climate change issues. By taking advantage of its solar resources, Spain reduces its reliance on conventional energy sources and contributes to a greener and more secure energy infrastructure. The integration of PV energy in Spain's energy landscape highlights its strategic importance in the nation's pursuit of a sustainable and environmentally conscious future.

Understanding and forecasting PV electricity demand in Spain is increasingly critical as the nation actively embraces renewable energy to meet its growing power needs [4].

With a rising focus on sustainability and a commitment to reducing carbon emissions, Spain's energy landscape is experiencing a transformative shift toward renewable sources, particularly solar energy. Thus, the accurate forecasting of PV electricity demand is vital for efficient energy planning, ensuring a satisfactory integration of solar power into the grid and optimizing the resource allocation [5]. As Spain expands its PV infrastructure, the ability to anticipate and meet the demand for solar electricity becomes paramount for maintaining grid stability, minimizing wastage, and achieving energy efficiency goals. Moreover, precise forecasting supports strategic decision-making, helping policymakers and energy authorities adapt to the dynamic nature of renewable energy sources and contribute to Spain's broader objectives.

The accurate prediction and management of the PV energy demand depends on deploying precise models that play an essential role in the efficiency of energy systems [6]. These models are important tools to anticipate variations in PV energy demand, ensure optimal grid integration, and facilitate effective energy management strategies. Precise forecasting helps align energy production with demand, minimizing the risk of grid underutilization or overloading [7]. Additionally, it enables proactive planning for energy storage and distribution, contributing to enhanced grid stability and resilience. In the context of PV energy, where generation is subject to weather conditions, the reliability of models becomes critical in examining the variability inherent in solar power production. Therefore, using accurate models is fundamental to successfully predicting and managing PV energy demand.

As a consequence, the importance of Big Data techniques in enhancing the accuracy of energy demand predictions cannot be ignored. In the PV energy field, where vast and diverse datasets are commonly processed, Big Data techniques offer a powerful solution to extract meaningful information. The ability to process large volumes of data in real time allows for a more comprehensive understanding of dynamic factors influencing energy demand, including weather patterns, economic indicators, and consumer behavior [1,7–9]. Big Data approaches facilitate the identification of complex relationships and trends that may go unnoticed with traditional methods. This enhanced understanding, in turn, leads to more accurate and responsive energy demand predictions.

Big Data approaches offer significant advantages in terms of scalability and efficiency when it comes to handling large datasets. These techniques are highly proficient in managing vast amounts of information and provide a flexible infrastructure that can easily accommodate the increasing datasets that are characteristic in energy demand prediction. This scalability allows for the analytical capabilities of the system to adapt flawlessly as datasets expand, without affecting performance.

Moreover, the efficiency of Big Data approaches results from their distributed processing abilities that enable parallel computation across multiple nodes. This parallelization accelerates data processing times [10–12]. The efficiency gains are especially critical in the context of energy demand forecasting, where real-time or near-real-time analysis is essential for effective decision-making.

Our study aims to propose and implement predictive models to improve the accuracy of PV energy demand predictions in Spain, considering the dynamic nature of renewable energy sources and the critical need for reliable forecasts in energy planning. To achieve this, our proposed methodology employs advanced predictive modeling techniques, leveraging the power of Big Data. We focus on the Linear Regression (LR), Random Forest (RF), Light Gradient Boosting Machine (LGBM), and Recurrent Neural Network (RNN) models, each designed to handle large datasets. By utilizing these models and the capabilities of Big Data, we seek to provide a robust framework for forecasting PV energy demand, contributing to more informed decision-making in the field of sustainable energy management in Spain.

## 2. Methodology

This section introduces the proposed methodology followed for predicting PV energy demand in Spain. The following subsections detail the dataset used and the four models applied in this study.

### *2.1. Dataset*

The dataset utilized in this study was sourced from the Spanish Electricity System (SES). The SES plays a pivotal role in electric storage and the management of high-tension energy transportation. Its responsibility includes real-time adjustments to energy production to ensure a balance between scheduled production and demand.

The data are publicly accessible on the SES website [13], providing users with the ability to retrieve various types of electric energy information. Employing scraping techniques, we collected the data. The resulting dataset comprises columns for the hour, date, and solar PV energy spanning from 2017 to 2022. The data granularity is set at 10 min intervals. The dataset has a period of 5 years and a couple of months, encompassing data up to 1 May 2022. In total, we collected 280.368 samples.

In the data preprocessing stage, we addressed issues such as repeated values and missing data by cleaning the dataset. Additionally, we normalized the data to ensure consistency. To prepare the PV energy demand data [9], we employed a sliding window approach, a method that involves analyzing the data in sequential segments. This technique allows for a systematic and continuous analysis of the dataset. Once the data were appropriately prepared, we proceeded to set up four models for learning and predicting the PV energy demand, as detailed in the subsequent sections.

### *2.2. Linear Regression*

The initial model implemented was Linear Regression (LR), utilizing the Scikit-Learn implementation. LR served as the baseline model for the comparative analysis with the subsequent models.

LR is a statistical technique used for modeling the relationship between a dependent variable and one or more independent variables. In this context, LR aims to establish a linear relationship between the features of the dataset and the solar PV energy demand [14]. LR is a simple and widely used model that aims to establish a linear relationship between variables. Despite its simplicity, LR has demonstrated effectiveness in various scenarios, often yielding good results, though there was room for improvement. This basic model serves as a valuable benchmark, offering insights into the complexity of the problem at hand and providing an initial assessment of whether linear solutions are sufficient to address the intricacies of the dataset.

### *2.3. Random Forest*

Random Forest (RF) is a versatile and widely employed ensemble learning technique. In our study, we utilized the RF version from MLlib, specifically designed for Big Data applications.

RF is more complex than LR, employing an ensemble of decision trees to enhance predictive accuracy. RF introduces parameters such as the number of estimators (trees) and depth, providing flexibility to adapt to various data complexities [8]. This complexity allows for RF to capture nonlinear relationships in the data, making it a suitable choice for scenarios where linear solutions may fall short. The decision to incorporate RF into our analysis derives from its capacity to handle hidden relationships within the data, making it an appropriate choice for predicting the PV energy demand in our study.

### *2.4. Light Gradient Boosting Machine*

Light Gradient Boosting Machine (LGBM) is a high-performance gradient-boosting framework designed for efficiency and scalability [15]. In our study, we employed the LGBM version from SynapseML, developed for Big Data applications, as well as MLlib.

SynapseML is a machine learning library that provides optimized algorithms for distributed computing environments, enhancing the efficiency of model training and prediction.

Compared to RF, LGBM introduces parameters such as the number of estimators (trees) and leaves. These parameters contribute to its adaptability and effectiveness in handling large datasets, making LGBM a favorable choice for scenarios where computational efficiency is crucial. The efficiency of LGBM in handling large datasets justifies its use in predicting PV energy demand while ensuring computational speed and scalability.

### *2.5. Recurrent Neural Network*

Lastly, a Recurrent Neural Network (RNN) is a type of neural network designed to process sequential data by maintaining a memory of previous inputs [16]. In our study, we utilized the Big Data-adapted version of an RNN from Keras. Keras is an open-source deep-learning library that facilitates the construction of neural networks in a user-friendly and modular manner.

In comparison to RF and LGBM, RNN introduces parameters such as the number of layers and neurons. These parameters enable RNN to capture temporal dependencies within sequential data, making it suitable for time-series prediction tasks.

We incorporated the RNN into our analysis because of its ability to model sequential dependencies in data, making it well suited for our problem as it addresses data over time.

## 3. Experiments

We carried out a series of experiments to evaluate the feasibility of Big Data-based solutions. To establish a baseline model, we selected LR for comparison with other models, even though its performance is not expected to be very robust. However, LR's simplicity provides insights into the behavior of other models.

LR, which has only one parameter for training, namely, the number of iterations, was used as our foundation model. We tested RF from the MLlib library with different numbers of trees, ranging from 25 to 100, and different max depths of the trees, ranging from 4 to 8. For LGBM, which was implemented using the SynapseML library, we conducted experiments with different numbers of trees, ranging from 25 to 100, and numbers of leaves, ranging from 20 to 40. We subjected the RNN to experimentation involving 1 to 3 layers and 10 to 50 neurons. We designed these comprehensive experiments to explore the performance and behavior of each model in terms of both the accuracy and time cost.

## 4. Results

Table 1 presents the results of the four distinct models applied in our study: LR using MLlib, RF using MLlib, LGBM using SynapseML, and using Keras. For each model, the table displays the information in five rows. The first column denotes the model's name, the second column signifies the window size utilized in the model, and the third and fourth columns represent the specific parameters relevant to each model, as detailed in the preceding section. The fifth and sixth columns provide the errors in terms of the RMSE and MAE, respectively. Finally, the time required to train each model is presented in the last column.

We can now evaluate the models' performance, taking into account both the predictive accuracy and computational cost. The top-performing models, derived from Linear Regression, exhibited closely comparable results. LR was tested with different numbers of iterations, as it does not involve additional parameters. The most favorable results, though not good enough, were observed with a window size of 288 and 150 iterations. These outcomes closely resembled those obtained with a smaller window size and a reduced number of iterations, showing only a marginal improvement of approximately half a unit.

Note that we tested other linear regressors on Scikit-Learn with alternative gradients and settings in order to confirm whether there was an execution mistake or error. However, the consistency across these models points toward the conclusion that the LR model may not be optimally suited for this problem. In this study, the LR model was employed as a baseline reference in our analysis for the rest of the models.

**Table 1.** Best-five results obtained for each model according to three parameters (windows size) and two different parameters depending on the specific model.

| Model | $w$ | $P_1$ | $P_2$ | RMSE | MAE | Time |
|---|---|---|---|---|---|---|
| LR | | | | | | |
| | 36 | 50 | - | 4186.062 | 2586.708 | 1.098 |
| | 36 | 150 | - | 4186.059 | 2586.707 | 3.386 |
| | 72 | 50 | - | 4186.054 | 2586.753 | 2.953 |
| | 72 | 150 | - | 4186.053 | 2586.745 | **1.052** |
| | 288 | 150 | - | **4185.884** | **2586.648** | 2.085 |
| RF | | | | | | |
| | 144 | 75 | 8 | 400.390 | 179.408 | **215.54** |
| | 144 | 100 | 8 | 402.533 | 180.446 | 291.603 |
| | 288 | 50 | 8 | 400.253 | 176.163 | 207.843 |
| | 288 | 75 | 8 | **398.695** | **175.223** | 308.564 |
| | 288 | 100 | 8 | 401.334 | 176.549 | 411.128 |
| LGBM | | | | | | |
| | 144 | 100 | 20 | **323.295** | **109.411** | **32.174** |
| | 144 | 100 | 30 | 325.137 | 109.545 | 36.312 |
| | 144 | 100 | 40 | 325.146 | 108.886 | 37.352 |
| | 144 | 150 | 20 | 324.403 | 108.833 | 37.662 |
| | 144 | 150 | 40 | 325.197 | 108.655 | 42.159 |
| RNN | | | | | | |
| | 36 | 1 | 10 | 486.87 | 300.158 | 101.38 |
| | 36 | 2 | 10 | **486.866** | **300.154** | 94.73 |
| | 36 | 1 | 40 | 630.801 | 447.929 | 65.08 |
| | 36 | 2 | 40 | 630.785 | 447.911 | 64.86 |
| | 144 | 2 | 20 | 682.339 | 411.123 | **62.78** |

RF achieved ten times better accuracy than LR. The time cost of RF was slightly higher than LR, as anticipated. The optimal error for RF was achieved with 75 trees, a max depth of eight, and a window size of 288. While not as accurate as with 288 predictors, RF with 144 predictors demonstrated the fastest performance.

LGBM demonstrated stable behavior, with consistently similar results across various configurations. The optimal performance was achieved with a window size of 144, 100 estimators, and 20 leaves, resulting in an RMSE of 323.295 and an MAE of 109.411. Interestingly, this configuration also generated the fastest model.

Finally, we can analyze the RNN. It is noteworthy that the RNN exhibited comparatively higher errors, with the performance varying considerably across different configurations. Discrepancies of up to 200 units in the RMSE and MAE suggest potential overfitting. Additionally, there is an almost double time difference between some configurations. It may be caused by the complexity of the settings, involving more neurons and layers. The optimal results were achieved with a window size of 36, two layers, and 10 neurons per layer. However, it did not prove to be the fastest configuration.

Table 2 collects the optimal errors attained by each model. The ranking order is as follows: LGBM, RF, the RNN, and LR. LGBM achieved the top positions for both the RMSE and MAE, with values of 323.295 and 324.635, respectively. RF, while securing the second position, exhibited an approximately 24% higher error compared to LGBM. Furthermore, the RNN demonstrated a performance approximately 22% worse than RF. It is impressive to see a consistent improvement trend between the models until reaching LR, which displayed an astonishing 760% higher error than the third-worst model, the RNN.

Table 3 below presents a comparison of the time required to train the best of each model. As expected, the simplicity of LR results in the shortest training time. LGBM took a bit longer, but it achieved the highest predictive accuracy among all the models. RF and the RNN are slower than LR and LGBM. However, LGBM is distinguished for its ability to balance high accuracy with a relatively swift training time, making it the best choice overall.

**Table 2.** A summary of the best errors obtained.

| Model | RMSE | |
|---|---|---|
| | **Best** | **Mean** |
| LGBM | **323.295** | **324.635** |
| RF | 398.695 | 400.641 |
| RNN | 486.866 | 583.523 |
| LR | 4185.000 | 4158.000 |

**Table 3.** A summary of the time cost of each model for the most accurate model (second column) and the mean of the experiments.

| Model | Times (s) | |
|---|---|---|
| | **Best** | **Mean** |
| LR | **1.05** | **1.94** |
| LGBM | 32.17 | 37.13 |
| RF | 71.85 | 71.95 |
| RNN | 94.73 | 77.76 |

After evaluating the performance of our photovoltaic demand prediction models, we now turn to some visualizations to gain a more intuitive understanding of their behavior. Figure 1 presents the actual demand alongside the predictions from each model: LGBM, RF, and the RNN. Bear in mind that, to improve clarity and better focus, we excluded the LR model visualization as its values would make the overall trend harder to interpret. Having said that, we can first examine the long-term trend and observe seasonal variations. Apparently, all four models seem to capture the overall trend of the data, though none perfectly match the actual demand curve. LGBM appears to track the actual demand a bit closer and the RNN deviates the most from the actual values.

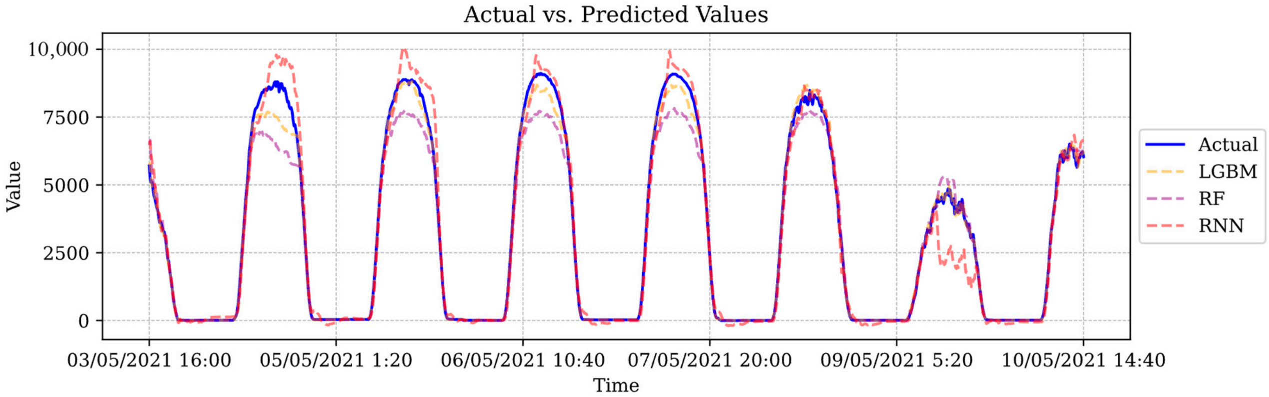


**Figure 1.** Overall trend of photovoltaic demand over a long period.

Figure 2 zooms in on a specific cycle of the PV demand in order to highlight the cyclical nature of this series. As mentioned before, the actual demand follows a cyclical pattern with a clear peak. While the models align well with the actual demand during the flatter, i.e., lower, portion of the cycle, their performance weakens as the demand approaches the peak. This is evident by the increasing spread in the prediction compared to the tighter grouping at the bottom. As the demand increases, the models struggle to maintain the same level of accuracy.

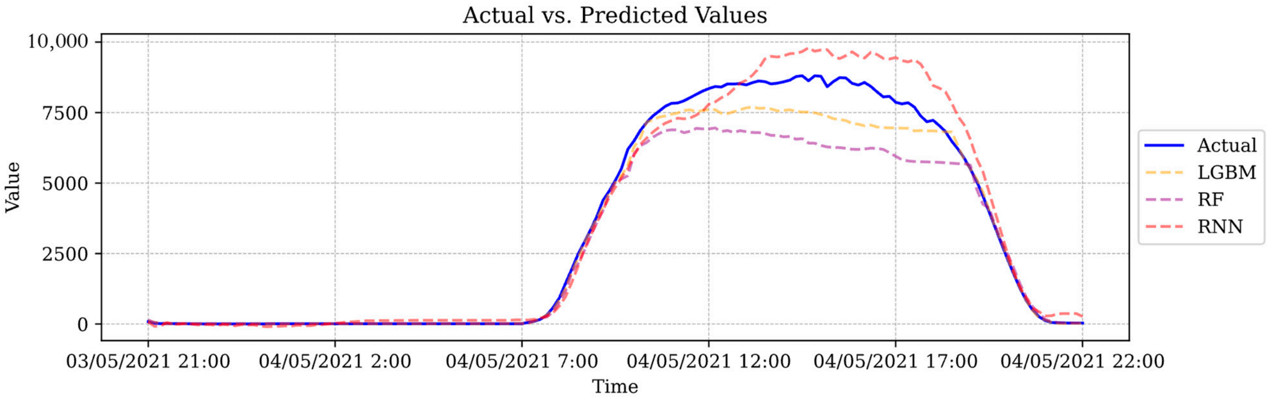
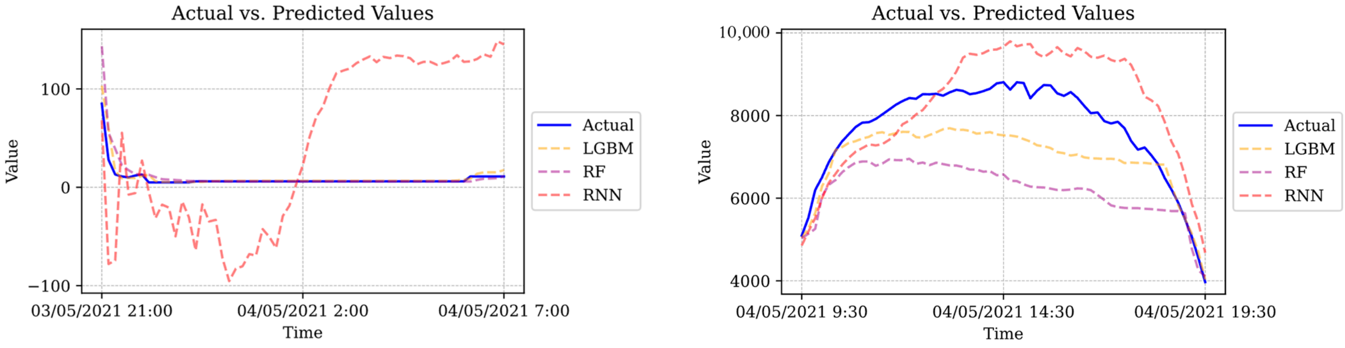


**Figure 2.** The cyclical period of the PV demand time series.

Finally, Figure 3 splits the previous cycle into two parts: the flatter Figure 3a and the peak Figure 3b. The first figure has a period with the lowest range of demand values compared to the entire series. While the overall trend appears flat, a closer look reveals some fluctuations. Here, the RNN exhibits the most variation in its predictions compared to the other models, which seem to better match the flatter pattern of the actual demand. If we focus on the second figure, the RNN consistently overestimates the demand. In contrast, the other models underestimate it. LGBM stands out as the closest predictor in this timeframe, with its line tracking the actual demand more closely than its counterparts.



(**a**) (**b**)

**Figure 3.** The lower (**a**) and upper (**b**) portions of the PV demand cycle.

## 5. Conclusions

In conclusion, our study has effectively met its objectives by successfully acquiring photovoltaic solar energy production data from the Spanish Electric Network. Through the implementation of various predictive models using Big Data-oriented techniques, we achieved generally acceptable outcomes. This integration of Big Data techniques played a crucial role in enhancing both the time efficiency and accuracy of our predictions.

Among the models employed, LGBM emerged as the best performer, demonstrating superior accuracy and efficiency. This highlights its effectiveness in handling the complexities of our dataset. On the other hand, the LR model faced challenges in delivering accurate results, and its time cost was notably prolonged. This emphasizes the limitations of LR in capturing the patterns present in the data. Our findings stress the importance of leveraging advanced predictive modeling techniques, such as LGBM, for accurate and timely predictions in the realm of photovoltaic energy demand forecasting in Spain.

By incorporating Big Data-oriented techniques, our study not only contributes a valuable understanding of forecasted trends but also sets a precedent for utilizing innovative approaches in energy demand prediction.

**Author Contributions:** J.T.-G., L.G.B.R., D.C.-R. and M.C.P. have contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

**Funding:** We acknowledge the financial support from the Ministerio de Ciencia e Innovación (Spain) (Grant PID2020-112495RB-C21 funded by MCIN/AEI/10.13039/501100011033).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data are contained within the article.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

| | |
|---|---|
| LGBM | Light Gradient Boosting Machine |
| LR | Linear Regression |
| PV | Photovoltaic |
| RF | Random Forest |
| RNN | Recurrent Neural Network |
| SES | Spanish Electricity System |

## References

1. Qiu, T.; Wang, L.; Lu, Y.; Zhang, M.; Qin, W.; Wang, S.; Wang, L. Potential assessment of photovoltaic power generation in China. *Renew. Sustain. Energy Rev.* **2022**, *154*, 111900. [CrossRef]
2. Cabello-López, T.; Carranza-García, M.; Riquelme, J.C.; García-Gutiérrez, J. Forecasting solar energy production in Spain: A comparison of univariate and multivariate models at the national level. *Appl. Energy* **2023**, *350*, 121645. [CrossRef]
3. Sánchez-Durán, R.; Barbancho, J.; Luque, J. Solar energy production for a decarbonization scenario in Spain. *Sustainability* **2019**, *11*, 7112. [CrossRef]
4. Gomez-Exposito, A.; Arcos-Vargas, A.; Gutierrez-Garcia, F. On the potential contribution of rooftop pv to a sustainable electricity mix: The case of Spain. *Renew. Sustain. Energy Rev.* **2020**, *132*, 110074. [CrossRef]
5. Auguadra, M.; Ribó-Pérez, D.; Gómez-Navarro, T. Planning the deployment of energy storage systems to integrate high shares of renewables: The spain case study. *Energy* **2023**, *264*, 126275. [CrossRef]
6. González-Peña, D.; García-Ruiz, I.; Díez-Mediavilla, M.; Dieste-Velasco, M.I.; Alonso-Tristán, C. Photovoltaic prediction software: Evaluation with real data from northern spain. *Appl. Sci.* **2021**, *11*, 5025. [CrossRef]
7. Grigoryan, H. Electricity consumption prediction using energy data, socio-economic and weather indicators. A case study of Spain, 2021. In Proceedings of the 9th International Conference on Control, Mechatronics and Automation (ICCMA), Belval, Luxembourg, 11–14 November 2021; IEEE: New York, NY, USA, 2021; pp. 158–164.
8. Sadorsky, P. A random forests approach to predicting clean energy stock prices. *J. Risk Financ. Manag.* **2021**, *14*, 48. [CrossRef]
9. Xiao, Z.; Gang, W.; Yuan, J.; Chen, Z.; Li, J.; Wang, X.; Feng, X. Impacts of data preprocessing and selection on energy consumption prediction model of hvac systems based on deep learning. *Energy Build.* **2022**, *258*, 111832. [CrossRef]
10. Pegalajar, M.; Ruíz, L.G.B.; Cuéllar, M.P.; Rueda, R. Analysis and enhanced prediction of the spanish electricity network through big data and machine learning techniques. *Int. J. Approx. Reason.* **2021**, *133*, 48–59. [CrossRef]
11. Barja-Martinez, S.; Aragüés-Peñalba, M.; Munné-Collado, Í.; Lloret-Gallego, P.; Bullich-Massagué, E.; Villafafila-Robles, R. Artificial intelligence techniques for enabling big data services in distribution networks: A review. *Renew. Sustain. Energy Rev.* **2021**, *150*, 111459. [CrossRef]
12. de Freitas Viscondi, G.; Alves-Souza, S.N. A systematic literature review on big data for solar photovoltaic electricity generation forecasting. *Sustain. Energy Technol. Assess.* **2019**, *31*, 54–63. [CrossRef]
13. Spanish Electricity System. Energy demand of the spanish electricity system. *Visiona* **2022**. Available online: https://demanda.ree.es/visiona/home (accessed on 1 September 2022).
14. Ciulla, G.; D'Amico, A. Building energy performance forecasting: A multiple linear regression approach. *Appl. Energy* **2019**, *253*, 113500. [CrossRef]
15. Liang, S.; Deng, T.; Huang, A.; Liu, N.; Jiang, X. Energy consumption prediction using the gru-mmattention-lightgbm model with features of prophet decomposition. *PLoS ONE* **2023**, *18*, e0277085. [CrossRef]
16. Ozcan, A.; Catal, C.; Kasif, A. Energy load forecasting using a dual-stage attention-based recurrent neural network. *Sensors* **2021**, *21*, 7115. [CrossRef] [PubMed]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.