Arab American University
Faculty of Graduate Studies
Department of Natural, Engineering & Technology Sciences
Master Program in Data Science and Business Analytics

A Comparative Machine Learning Approach to Forecasting Solar Power Across Diverse Climate Conditions

Abedalraouf M. Sh. Abdalshakor Sharabati
202216489

Supervision Committee:
Dr. Majdi Sabe Mofadi Owda
Dr. Hind Khalid Mohammad Sweis
Dr. Mohammad K Jubran

This Thesis Was Submitted in Partial Fulfillment of the Requirements for the Master Degree in Data Science and Business Analytics

Palestine, October/2025

Arab American University
Faculty of Graduate Studies
Department of Natural, Engineering & Technology Sciences
Master Program in Data Science and Business Analytics

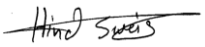
Thesis Approval

A Comparative Machine Learning Approach to Forecasting Solar Power Across Diverse Climate Conditions

Abedalraouf M. Sh. Abdalshakor Sharabati
202216489

This thesis was defended successfully on 1/10/2025 and approved by:

Thesis Committee Members:

	Name	Title	Signature
1.	Dr. Majdi Sabe Mofadi Owda	Main Supervisor	
2.	Dr. Hind Khalid Mohammad Sweis	Members of Supervision Committee	
3.	Dr. Mohammad K Jubran	Members of Supervision Committee	

Palestine, October/2025

Declaration

I declare that, except where explicit reference is made to the contribution of others, this thesis is substantially my own work and has not been submitted for any other degree at the Arab American University or any other institution.

Student Name: Abedalraouf M. Sh. Abdalshakor Sharabati
Student ID: 202216489
Signature: Abedalraouf M. Sh. Abdalshakor Sharabati
Date of Submitting the Final Version of the Thesis: 8/10/2025

Acknowledgments

Completing this study would not have been possible without the expertise and guidance of Dr. Majdi Owda as the main supervisor and Dr. Amani Owda. I would like to express my sincere gratitude to them for their valuable support, expert guidance, and encouragement throughout this research.

Also, I am thankful to the members of the supervision committee, Dr. Hind Sousa and Dr. Mohammad Jabril, for their valuable comments, feedback, and guidance during the thesis evaluation process.

My appreciation to the Arab American University, particularly the Faculty of Graduate Studies, the Professors of the Data Science and Business Analytics Master's program, and the program coordinator, for providing the facilities, academic support, and environment necessary to complete this work. I also acknowledge the use of data sources and research resources.

Special thanks to the Tubas Electricity Company and the Red Eléctrica de España (REE) for providing the data essential for this study. Their cooperation made the practical implementation of this research possible.

I extend my heartfelt thanks to my beloved wife and children for their patience and support. I also thank my great mother, brothers, and sisters for, prayers and encouragement.

And my thanks to everyone who contributed to this work.

A Comparative Machine Learning Approach to Forecasting Solar Power Across Diverse Climate Conditions

Abedalraouf M. Sh. Abdalshakor Sharabati

Supervision Committee:

Dr. Majdi Sabe Mofadi Owda

Dr. Hind Khalid Mohammad Sweis

Dr. Mohammad K Jubran

Abstract

The study mainly compares the efficiency of Machine Learning (ML) models in forecasting solar photovoltaic (PV) power generation under different climatic conditions reported between August 2022 and July 2023. The data were collected from two locations: Tubas in Palestine, with highly variable weather, and Balearic Islands in Spain, with stable conditions. The dataset includes meteorological data, irradiance, temperature, humidity, pressure, and wind speed from NASA's POWER database, and solar power generation data from Tubas Electricity Company and Red Eléctrica de España.

The study involved comprehensive data cleaning, preprocessing, exploratory data analysis, and model development. ML models—Linear Regression, Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Bidirectional Long Short-Term Memory (Bi-LSTM), and a hybrid Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM)—were implemented. Model performance was evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) metrics. Deep learning models, particularly Bi-LSTM and hybrid CNN-LSTM, achieved better performance across both regions. Tubas presented greater forecasting challenges due to fluctuating weather, despite that the deep model remained robust.

Key findings showed that solar irradiance was the most influential predictor in both regions, while temperature, humidity, and wind speed significantly contributed under fluctuating conditions. The study concludes that deep learning models are best suited for solar forecasting under diverse weather conditions. The findings are valuable for enhancing solar energy integration, guiding infrastructure investments.

Keywords: Solar forecasting, machine learning, deep learning, climate variability.

Table of Contents

Declaration I

Acknowledgments II

Abstract III

List of Tables VII

List of Figures VIII

List of Definitions of Abbreviations X

Chapter 1 Introduction 1

1.1. Overview 1

1.2. Importance of the Study 1

1.3. Problem Statement 2

1.4. Objectives of the Study 2

1.5. Research Questions 3

1.6. Boundaries and Limitations 3

1.7. Thesis structure 4

Chapter 2 Literature Review 5

2.1 Introduction 5

2.2 Global Energy Production and Environmental Challenges 6

2.3 The Role of Solar Energy in Addressing Energy and Climate Challenges 8

2.4 Importance of Solar Energy Forecasting 9

2.5 Machine Learning Techniques for Solar Energy Forecasting 10

2.5.1 Traditional, Deep Learning, and Time Series ML Models 10

2.5.2 Hybrid and Ensemble ML Models 11

2.5.3 Advanced ML Techniques for Solar Forecasting 13

2.5.4 Environmental and Climatic Factors in Forecasting Models 13

2.5.5 Regional and Climate-Diverse Case Studies 14

2.6 CO2 Emissions Reduction and Environmental Impact Modeling 20
2.7 Research Gap 20
2.8 Summary 21
Chapter 3 Methodology 22
3.1 Introduction 22
3.2 Used technology 22
3.3 Dataset collection and description 23
3.3.1 Meteorological Data 24
3.3.2 Solar Power Generation Data 24
3.4 Data Cleaning 25
3.4.1 Data Cleaning for Tubas data/sets 27
3.4.2 Data Cleaning for Balearic Islands datasets 31
3.5 Exploratory Data Analysis 34
3.5.1 Descriptive Analysis 35
3.5.2 Visualization 37
3.6 Data Pre-processing 54
3.6.1 Determine Target Feature 54
3.6.2 Feature Scaling 55
3.7 Machine Learning Models 56
3.7.1 Used Models 57
3.7.2 Model Configuration 59
3.7.3 Model Performance Evaluation Criteria 62
3.8 Summary 63
Chapter 4 Results 64
4.1 Introduction 64
4.2 Performance Evaluation Metrics Summary 64

4.3 Comparative Analysis of Models 65
4.3.1 Tubas Dataset Results 65
4.3.2 Balearic Islands Dataset Results 66
4.4 Visual and Graphical Findings 67
4.4.1 Residual Distribution 67
4.4.2 Feature importance Analysis 69
4.4.3 Actual vs. Predicted Power Output 70
4.5 Summary 75
Chapter 5 Discussion 76
5.1 Discussion 76
5.2 Conclusion 79
5.3 Future Work 79
References 81
الملخص 87

List of Tables

Table 2-1 7.5-minute Forecast Interval Results, results from study [27]. 13
Table 2-2 Performance Metrics for Seasonal Solar Irradiance Forecasting, in study [35]. 15
Table 2-3 Comparison of ML Models for Solar Forecasting 17
Table 3-1 Feature Description for Meteorological Data 24
Table 3-2 Feature Description for Solar Power Generation Data (Tubas) 25
Table 3-3 Feature Description for Power Generation Mix Data (Balearic Islands) 26
Table 3-4 Summary of Datasets Used in the Study 27
Table 3-5 Detected Outliers per Feature Using IQR Method for Tubas dataset 29
Table 3-6 Engineered Time-Based Features 30
Table 3-7 Number of Outliers Detected per Feature (IQR Method) for Balearic Island dataset 33
Table 3-8 Descriptive Statistics for the Balearic Region 34
Table 3-9 Descriptive Statistics for Tubas Region 35
Table 3-10 Configuration of Forecasting Models 61
Table 4-1 Performance Evaluation of Models for Solar Power Forecasting 64
Table 4-2 Standard deviation of prediction errors for RF and Tuned RF models in Balearic and Tubas datasets 73

List of Figures

Figure 2-1 Global Energy Consumption by Source (1980–2023). Data source: [6]. 6
Figure 2-2 Global CO2 Emissions from Fossil Fuels (2000–2023). Data source: [20]. .. 7
Figure 2-3 Share of Renewables in Electricity Generation (2000–2023). Data source: [9] 8
Figure 3-1 A Study of Environmental Impact on Renewable Energy Optimization Prediction Mode 22
Figure 3-2 The Applied Steps for Cleaning Raw Datasets 27
Figure 3-3 Total Power Output Comparison by Region 38
Figure 3-4 Distribution of Features in Balearic Islands and Tubas 42
Figure 3-5 Relationships between Solar Power Output and Meteorological Factors in Balearic and Tubas 45
Figure 3-6 Correlation Heatmaps analysis for Balearic and Tubas dataset 46
Figure 3-7 Daily Power Output Trend (Balearic vs. Tubas) 46
Figure 3-8 Normalized Daily Power Output Trend (Balearic vs. Tubas) 46
Figure 3-9 Average Hourly Pattern - Balearic Islands (Balearic vs. Tubas) 47
Figure 3-10 Average Power Output by Hour and Month in Balearic and Tubas 47
Figure 3-11 Monthly Power Output Distribution - Balearic 48
Figure 3-12 Monthly Power Output Distribution - Tubas 49
Figure 3-13 Total Power Output by Season - Balearic and Tubas 50
Figure 3-14 Seasonal Power Output Distributions – Violin Plot Comparative Analysis 51
Figure 3-15 Total Power Output by Weekday - Balearic 51
Figure 3-16 Total Power Output by Weekday - Tubas 52
Figure 3-17 Total Power Output by Day of Month - Balearic 52
Figure 3-18 Total Power Output by Day of Month – Tubas 53
Figure 3-19 Zero vs Non-Zero Power Output - Balearic 54
Figure 3-20 Zero vs Non-Zero Power Output -Tubas 54
Figure 3-21 The Preprocessing Methods Applied on the Dataset 54
Figure 3-22 Modeling Workflow for Solar Power Forecasting 59
Figure 4-1 Predicted vs. Actual Solar Power Output Using Linear Regression – Balearic 71
Figure 4-2 Predicted vs. Actual Solar Power Output Using Linear Regression – Tubas 71
Figure 4-3 Predicted vs. Actual Solar Power Output Using RF – Balearic 72

Figure 4-4 Predicted vs. Actual Solar Power Output Using RF – Tubas 72
Figure 4-5 Predicted vs. Actual Solar Power Output Using Tuned RF – Balearic..... 72
Figure 4-6 Predicted vs. Actual Solar Power Output Using Tuned RF – Tubas 72
Figure 4-7 Predicted vs. Actual Solar Power Output Using XGBoost – Balearic..... 73
Figure 4-8 Predicted vs. Actual Solar Power Output Using XGBoost – Tubas 73
Figure 4-9 Time Series Comparison of Actual and Predicted Output Using Bi-LSTM – Balearic 74
Figure 4-10 Time Series Comparison of Actual and Predicted Output Using Bi-LSTM – Tubas..... 74
Figure 4-11 Time Series Comparison of Actual and Predicted Output Using CNN-LSTM – Balearic 74
Figure 4-12 Time Series Comparison of Actual and Predicted Output Using CNN-LSTM – Tubas 74
Figure 4-13 Residual across All Models and Both Regions 68
Figure 4-14 Feature Importance by F-Score – Balearic 70
Figure 4-15 Feature Importance by F-Score – Tubas 70

List of Definitions of Abbreviations

Abbreviations	Title
ALSTM	Attention-based Long Short-Term Memory
AM	Attention Mechanism
ANN	Artificial Neural Network
ANNs	Artificial Neural Networks
API	Application Programming Interface
AQI	Air Quality Index
ARIMA	AutoRegressive Integrated Moving Average
ARIMAX	ARIMA with eXogenous variables
ATP	Atmospheric Temperature/Pressure
AUC	Area Under the Curve
BPNN	Back Propagation Neural Network
BST	Back-Surface Temperature
Bi-LSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
CNN-LSTM	Convolutional Neural Network – Long Short-Term Memory
CNN-LSTM-AM	CNN-LSTM with Attention Mechanism
CNN-LSTM-RF	CNN-LSTM with Random Forest
CO	Carbon Monoxide
CO_2	Carbon Dioxide
CSV	Comma-Separated Values
DAP	Daily Accumulated Precipitation
DBT	Dry Bulb Temperature
DHI	Diffuse Horizontal Irradiance
DL	Deep Learning
DNI	Direct Normal Irradiance
DT	Decision Tree
DWT	Discrete Wavelet Transform
DY	Day of the Month
EMD	Empirical Mode Decomposition
ERAD	Extra-terrestrial Irradiance

GB	Gradient Boosting
GBM	Gradient Boosting Machine
GCLSTM	Graph-Convolutional LSTM
GCTrafo	Graph-Convolutional Transformer
GHI	Global Horizontal Irradiance
GP	Gaussian Process
GPR	Gaussian Process Regression
GRU	Gated Recurrent Unit
HMLM	Hybrid Machine Learning Model
HR	Hour of the Day
HTCNN	Hierarchical Temporal Convolutional Neural Network
IEA	International Energy Agency
IQR	Interquartile Range
kNN	k-Nearest Neighbors
KW	Kilowatt
LR	Linear Regression
LSTM	Long Short-Term Memory
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
MW	Megawatt
NARXNN	Non-linear AutoRegressive eXogenous Neural Network
NASA	National Aeronautics and Space Administration
NO_2	Nitrogen Dioxide
NRMSE	Normalized Root Mean Square Error
PAI	Panel-of-Array Irradiance
PM	Particulate Matter
PM2.5	Particulate Matter < 2.5 µm
POWER	Prediction Of Worldwide Energy Resources
PV	Photovoltaic

REE	Red Eléctrica de España
RF	Random Forest
RFR	Random Forest Regression
RH	Relative Humidity
RM	Regression Model
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
R^2	Coefficient of Determination
SARIMA	Seasonal ARIMA
SARIMAX	Seasonal ARIMA with Exogenous Variables
SEA	Solar Energy Industries Association
SN	Seasonal Network
SOM	Self-Organizing Map
SO_2	Sulfur Dioxide
STAR	Spatio-Temporal AutoRegressive
STCNN	Spatio-Temporal Convolutional Neural Network
SVM	Support Vector Machine
SVMs	Support Vector Machines
SVR	Support Vector Regression
TCN	Temporal Convolutional Network
UK	United Kingdom
USA	United States of America
WT	Wavelet Transform
XGB	eXtreme Gradient Boosting

Chapter 1 Introduction

1.1. Overview

As part of efforts to reduce carbon dioxide emissions and combat climate change, there is a global trend toward depending on renewable energy as the main power source. This has led to significant interest in renewable energy technologies, particularly solar PV systems. However, one of the major challenges hindering the widespread integration of solar energy into electricity grids is its variability, which is due to its dependence on weather conditions. This study addresses gaps in current solar forecasting research, particularly the lack of generalizability across diverse climate conditions and the limited integration of various environmental parameters in predictive models.

Despite the extensive use of ML and DL for solar PV forecasting, previous studies often focus on local datasets with limited environmental variability. This raises questions about the effectiveness of these models under changing climatic conditions. Furthermore, comprehensive comparative analysis across regions with different meteorological conditions remains insufficient. To bridge this gap, this study develops and evaluates multiple ML models to predict solar energy production in two climatically distinct regions: the Balearic Islands in Spain and Tubas district in Palestine. The research utilizes diverse environmental data to enhance prediction accuracy and model reliability, and compares the models' efficiency for the same parameters to determine the suitability of the model under different environmental conditions.

1.2. Importance of the Study

This study has high importance as it contributes to the growing field of ML-based energy forecasting by testing the adaptability of ML models across varying environmental conditions. It adds value by integrating a wide range of meteorological inputs and conducting performance evaluations under standardized conditions.

Practically, the study's outcomes can be beneficial to multiple stakeholders. The improved forecasting accuracy helps grid operators determine where to invest in transmission upgrades or new substations. Accurate forecasts support decision-makers in

determining the direction of strategic investments in solar energy and contribute to achieving carbon dioxide reduction goals. Additionally, the Comparative insights across different climate conditions, which are carried by this study, help researchers to design appropriate ML models.

By applying these models in real environments, solar energy efficiency and operational planning can be significantly improved, contributing to cleaner energy systems worldwide.

1.3. Problem Statement

Despite the progress made in machine learning-based solar energy forecasting, several key issues remain. Most forecasting models are limited to a single climate region and often ignore broader environmental contexts, such as atmospheric pressure, humidity, and wind speed. These limitations limit the models' adaptability and robustness across regions.

The main research question addressed in this study is "How can machine learning models be improved to accurately predict solar PV generation under diverse weather conditions?"

The scientific motivation is based on the need for data-driven, environmentally adaptive energy forecasting models. By evaluating the models on real-world datasets from two different climate regions, the study provides a scientific basis for developing more comprehensive and reliable forecasting systems.

1.4. Objectives of the Study

The main objectives of this research are:

- Collect, clean, and preprocess meteorological and solar power generation datasets from Tubas and the Balearic Islands.
- Analyze the trends and relationships between the features in the datasets.
- Employ and train various machine learning models, including linear regression, RF,

XGBoost, Bi-LSTM, and CNN-LSTM, for solar energy forecasting.

- Compare the performance of these models under different climatic conditions using metrics such as MAE, RMSE, and R^2.
- Evaluate the impact of environmental characteristics on forecast accuracy and identify key predictors.
- Highlight the most effective and universally applicable model for solar forecasting.

These objectives are directly linked to the hypothesis of the study: that deep learning models, particularly sequence-based architectures such as Bi-LSTM and CNN-LSTM, outperform conventional models under varying climatic variables.

1.5. Research Questions

The study aims to answer the following key questions:

- What are the effects of different atmospheric characteristics on solar energy generation in diverse climates?
- Which ML models provide the most accurate forecasts across different climate conditions?
- How does model performance differ between stable (Balearic Islands) and variable (Tubas) weather environments?
- Can a generalized forecasting model be developed to perform efficiently across these regions?

1.6. Boundaries and Limitations

The scope of this study is limited by several limitations that guide its focus and applicability. It is limited to solar PV energy forecasting using ML techniques, excluding other renewable sources such as wind or thermal. The datasets cover a full year, from August 2022 to July 2023, allowing for the analysis of seasonal variations in solar energy generation. Spatially, the research is limited to two specific geographical locations: Tubas in Palestine and the Balearic Islands in Spain, which were chosen for their varying

climatic conditions to assess the model's generalizability.

1.7. Thesis structure

This chapter presented an introduction about the study, the importance and significance of the study, the research problem, the study objectives, and the boundaries and limitations of the study. The following chapters in this study are organized as follows:

- **Chapter Two: Literature Review:**

This chapter presents background about the research topic, with a detailed literature review of the previous works related to the proposed study.

- **Chapter Three: Methods and Implementations:**

This chapter presents the description of the proposed models, and the used methodology in detail for implementing the proposed study with all the stages, from the data collection stage to obtaining the model outcomes stage.

- **Chapter Four: Results and Discussion**:

This chapter presents the main findings and the experimental results of the constructed models and discusses these results, the performance, and the accuracy achieved after implementing the models

- **Chapter Five: Conclusion and Future Work:**

This chapter provides a conclusion about the study, including a summary of the main findings, the key achievements, and an evaluation of the limitations of the study. Also, discuss the future works for enhancing or extending the work done in the study and provide recommendations.

Chapter 2 Literature Review

2.1 Introduction

This chapter presents a comprehensive literature review of studies on predicting solar energy production. The integration of ML models in renewable energy forecasting, especially solar energy, has gained high importance in recent years. The motivation is the need to improve the accuracy of energy production forecasts to support grid stability, optimize energy management, facilitate solar energy investment decisions, and contribute to carbon dioxide (CO2) emissions reduction.

Many researchers have used different ML models - from basic ones to advanced models like Artificial Neural Networks (ANNs), Support Vector Machines (SVM), RF, and deep learning methods like LSTM networks to try and predict solar power (Diagne et al., 2013; Voyant et al., 2017). But there's a problem: most only test their models in one country or places with similar weather. That means we don't know if their methods would work somewhere completely different.

In addition, much of the existing research has emphasized a different selection of environmental factors, neglecting the wide variety of meteorological, geographical, and temporal features that influence solar energy output. Given the strong dependence of ML-based solar forecasting on fluctuating environmental factors (Antonanzas et al., 2016), recent research has shifted toward hybrid and ensemble models.

By integrating diverse ML algorithms, these approaches mitigate individual model limitations and enhance overall predictive performance. Despite that, a noticeable gap remains in leveraging these techniques across weather variability and integrating their outcomes to achieve the desired goals (Urraca et al., 2017, 2018).

Additionally, the chapter provides an overview of energy consumption and the need to transition to renewable energy sources, especially to solar energy. It also reviews studies demonstrating how solar energy effectively reduces carbon emissions and contributes to environmental sustainability.

2.2 Global Energy Production and Environmental Challenges

The growing global need for energy has caused significant environmental challenges, mainly due to the widespread use of fossil fuels such as coal, oil, and natural gas. These energy sources have supported the development of industrial and economic evolution, but are also considered the main contributors to greenhouse gas emissions, especially CO2 (in Data, 2023). As shown in Figure 2-1, fossil fuels have a large percentage of global energy consumption between 1980 and 2023, despite their environmental impact.

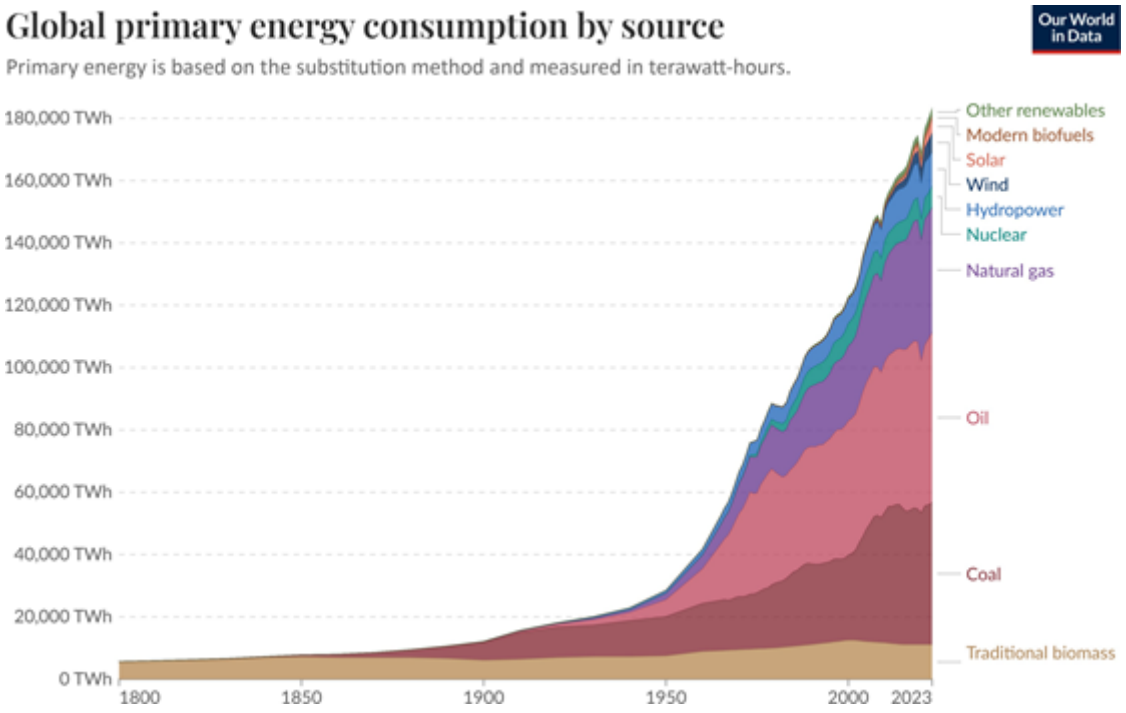


Figure 2-1 Global Energy Consumption by Source (1980–2023). Data source: [6].

Burning fossil fuels produces a large amount of CO2, which leads to climate change. In 2023, global CO2 emissions from fossil fuels reached 37.01 billion metric tons, continuing an upward trend over the past twenty years (Figure 2-2) [20]. This increase is a result of industrial growth, urban expansion, and rising energy demand, especially in fast-growing economic and population nations like China and India (International Energy Agency, 2024a).

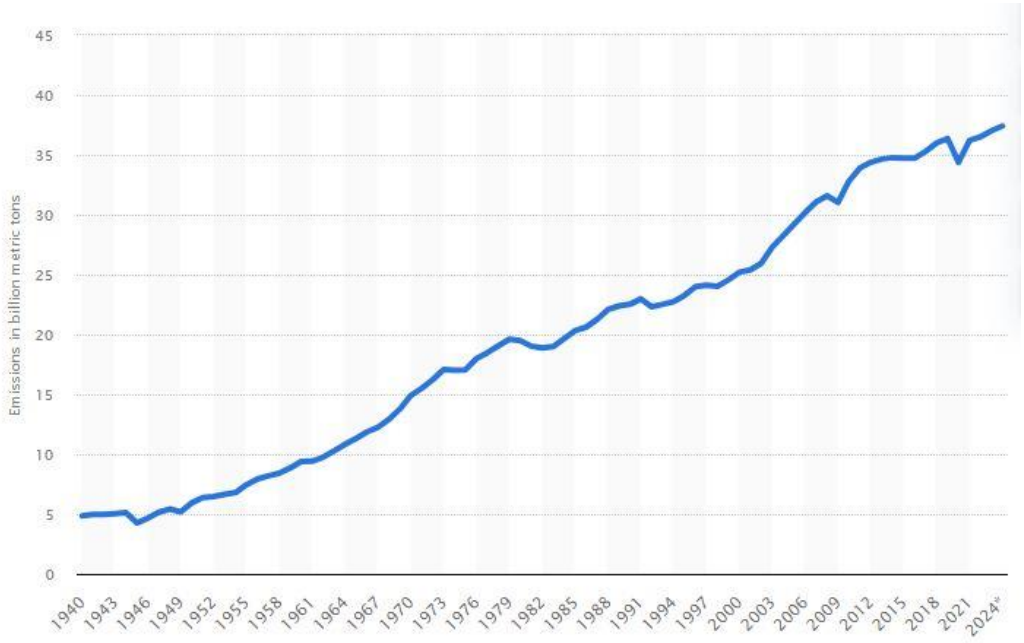


Figure 2-2 Global CO2 Emissions from Fossil Fuels (2000–2023). Data source: [20].

In addition to carbon emissions, fossil fuel extraction and use harm the environment in other ways. Mining, drilling, and hydraulic fracturing can lead to deforestation, water pollution, and land degradation. Furthermore, fossil fuel combustion releases harmful air pollutants such as sulfur dioxide (SO2) and nitrogen oxides (NOx), which contribute to acid rain and poor air quality. According to the Institute for Health Metrics and Evaluation, air pollution caused an estimated 6.7 million premature deaths globally in 2021(Institute for Health Metrics and Evaluation, 2021).

To address these challenges, a transition to renewable energy is crucial. Using clean sources like solar and wind can reduce environmental damage.

The share of renewables in global electricity generation has grown steadily since 2000, reaching over 30% by 2023 (Figure 2-3 Share of Renewables in Electricity Generation (2000–2023). Data source: [9]) (in Data, 2024). But renewable energy still represents a small share of overall energy consumption, which means there is a need to encourage investment in this sector and supportive polices.

In summary, reducing dependence on fossil fuels and increasing the use of renewable energy is essential to mitigate climate change, improve air quality, and ensure long-term environmental sustainability.

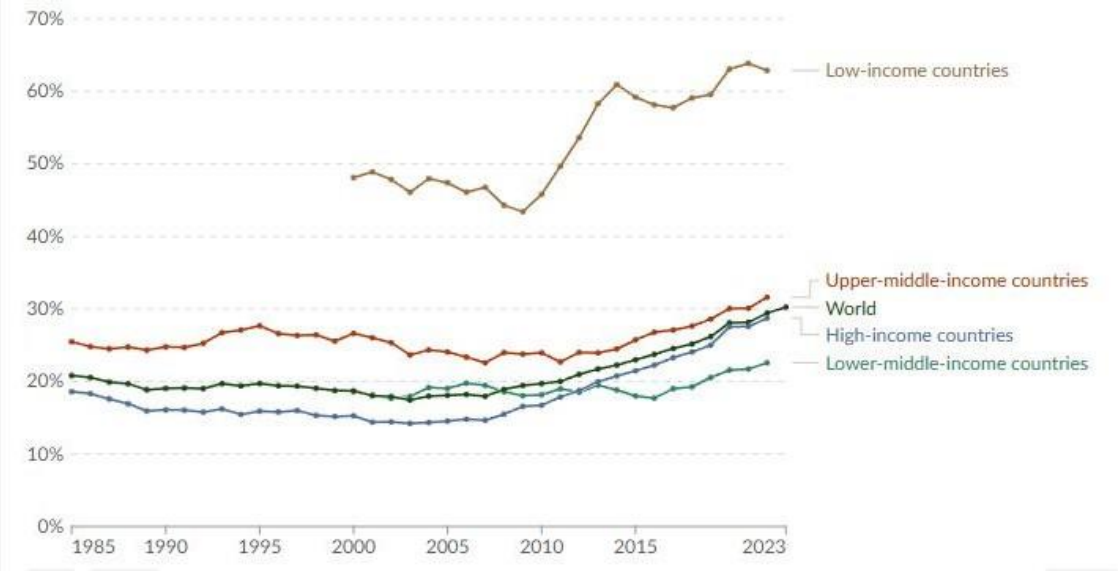


Figure 2-3 Share of Renewables in Electricity Generation (2000–2023). Data source: [9]

2.3 The Role of Solar Energy in Addressing Energy and Climate Challenges

Solar power is one of the most important clean energy sources, helping to address the world's energy and climate challenges. It is an effective way to reduce greenhouse gas emissions as it produces electricity without CO2 emissions (U.S. Energy Information Administration, 2023). Using solar panels instead of fossil fuels helps reduce the pollution that highly contributes to climate change. According to the U.S. Energy Information Administration, solar energy systems generate power with almost no environmental impact once they are installed (U.S. Energy Information Administration, 2023).

In addition, Solar helps countries toward sustainability in energy production, as the countries can reduce their dependence on fossil fuels by investing in solar panels, particularly in remote or rural areas where building traditional power plants or grid connections is more difficult. The Solar Energy Industries Association (SEA) reported that solar power also supports the electric grid, making it stronger and more reliable during emergencies or high demand (Solar Energy Industries Association, 2023).

Another benefit is that solar energy is now one of the cheapest ways to produce

electricity. The cost of solar panels has dropped a lot in the last decade. According to the International Energy Agency (IEA), solar and wind energy made up 95% of the new renewable energy added worldwide in 2023 (International Energy Agency, 2023).

Solar power's role in the world's electricity supply is also growing fast. In 2023, solar generated about 5.4% of the global electricity supply—a big increase compared to previous years (International Energy Agency, 2024b). This growth shows that solar energy is becoming more important for countries trying to reduce emissions and meet energy needs.

2.4 Importance of Solar Energy Forecasting

Solar energy production is unpredictable as it depends on weather conditions or season and time of day. Accurate forecasting helps grid operators manage and plan by predicting how much solar will be available to generate power.

This allows them to balance between supply and demand, which reduces the need for fossil fuel generation (Yang & Kleissl, 2013).

Depending on solar power forecasts, the maintenance of solar power plants can be managed more efficiently. By, for example, planning the maintenance during cloud cover time, as the power generation will be less than at other times, to avoid energy losses during peak sunlight. This improves performance and extends equipment life (Sharma & Zhang, 2024).

Forecasting is also important for energy investments. Solar producers must invest in electricity markets based on insights into how much energy they expect to produce. Also, better forecasts help producers avoid costly mismatches between what they offer and what they deliver. This increases profits (Utopus Insights, 2023).

In short, solar energy forecasting plays an important role in power system management.

2.5 Machine Learning Techniques for Solar Energy Forecasting

In recent years, ML models have become essential for improving the accuracy of solar energy forecasting. These models handle complex patterns in environmental and geographic data that affect energy production. This section presents an overview of different ML models and strategies used in solar energy forecasting.

2.5.1 Traditional, Deep Learning, and Time Series ML Models

Traditional ML models such as SVM, RF, and ANN have the foundation for solar prediction. These models work well when trained on meteorological data like temperature, humidity, and irradiance. In (Essam et al., 2022) the authors found ANN to be highly effective among the tested ML models, ANN, RF, DT, Extreme Gradient Boosting (XGB), and LSTM, achieving an R^2 of 0.9988 and outperforming RF, DT, LSTM, and XGBoost in predicting the output solar power. The ANN model effectively captured non-linear dependencies between environmental variables and solar output.

The author in study (Zulkifly et al., 2021) also compared various traditional models such as SVM, Gaussian Process Regression (GPR), Linear Regression, and Decision Trees (DT). Their study showed that fine-tuned DT achieved the highest performance with a high coefficient of determination R^2 of 95.91%, also showed low average prediction errors with RMSE equal to 5.83%, and maintained a consistent accuracy, indicated by a MAD of 3.7%, demonstrating that simpler models can be highly effective when properly optimized, noting that the authors identified computation time as a differentiator in model selection, Linear Regression models were faster (around 6 seconds) compared to GPR model, which required several hours. Thus, the authors recommended selecting ML models based on both forecasting accuracy and computation time, which was DT, emphasizing computation time as important for real-time forecasting scenarios.

The authors in (Subramanian et al., 2023) compare SVM, RF, and GB models for solar forecasting. The study found the RF model to be the most effective, achieving an Area under Curve (AUC) of 99%, demonstrating high classification reliability.

Deep learning models such as Convolutional Neural Networks (CNN) and LSTM

networks have significantly optimized the performance of forecasting systems. CNNs are well-suited for capturing spatial features from solar images and irradiance maps, while LSTM excels at capturing temporal dependencies in sequential data. In (Gao et al., 2019) the author Showed that LSTM-based models are particularly effective for forecasting 1-hour ahead, handling seasonal fluctuations in large-scale PV power plants. Their model demonstrated robust predictive accuracy across different seasons, as shown by the RMSE, which measures the typical magnitude of prediction errors. The RMSE achieved values ranging from 5.34% in spring to 13.86% in autumn, demonstrating its strength across diverse seasons.

Time series forecasting techniques, including AutoRegressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and ARIMA with exogenous variables (ARIMAX), serve as benchmarks for ML applications. In (Kim et al., 2023) the authors segmented the data by season and compared traditional time series models, including ARIMA, with LSTM model. The LSTM model achieved a lower MAPE, ranging from 5.24% to 6.70%, outperforming ARIMA-based models, which showed MAPE values between 7.17% and 9.85%. These results highlight the effectiveness of sequential learning models like LSTM in improving forecasting accuracy.

2.5.2 Hybrid and Ensemble ML Models

Hybrid models combine different ML techniques, each designed to handle a specific part of solar forecasting. For example, one model (like LSTM) can analyze time-based patterns, another (like CNN) can process spatial details, and a third (like RF or XGBoost) can tune the final predictions. By merging these approaches, hybrid models can better describe both linear and nonlinear relationships in solar forecasting, making them more accurate and reliable under various weather conditions (Nadeem et al., 2024).

In (Kumar et al., 2023), the researchers developed a hybrid ML model named HMLM to improve solar power generation forecasts. The model combines ANN, support vector regression (SVR), and random forest regression (RFR) by using a weighted averaging approach. Each model is independently trained on the same dataset, and their prediction results are combined as a weighted sum to optimize the accuracy. It analyzes environmental factors like temperature, solar radiation, time of day, wind speed, and humidity. Results showed that HMLM performs better than single models like CNN or

ANN, with accuracy between 89.87% and 98.67%. In comparison, CNN and ANN scored 74.54%–82.12%, and 77.72%–85.37%, respectively, noting the hybrid approach was more efficient.

Another hybrid implementation was developed by (Abumohsen, Owda, Owda, & Abumihsan, 2024), which introduced a new hybrid ML model combining CNN, LSTM, and RF models to optimize the accuracy of solar power generation forecasts. The CNN component extracts spatial patterns from solar data, the LSTM captures temporal trends, and the RF refines the predictions to enhance accuracy. The proposed hybrid CNN-LSTM with Random Forest (CNN-LSTM-RF) model achieved a great performance, including an R^2 score of 92%, RMSE of 0.07, and MAE of 0.05, thereby outperforming classical ML methods and individual DL models.

The authors in study (Saini et al., 2023) uses different ML and DL techniques, including CNN, LSTM, SVM, RF, and hybrid models. Results showed that hybrid CNN-LSTM models achieved better results with R^2 99.84% and RMSE value 7.21, the research also explained how important data preprocessing like normalization, wavelet transforms, and Empirical Mode Decomposition (EMD) improved prediction quality.

Ensemble methods have also gained traction because of their ability to manage noisy and nonlinear solar datasets. The authors in the study (Abumohsen, Owda, Owda, Abumihsan, et al., 2024) compared the performance of the XGBoost (XGB) model, Multilayer Perceptron (MLP) model, and Gradient Boosting Machine (GBM) model for solar power forecasting. The results indicated that XGBoost delivered the best performance, achieving an excellent R^2 of 88.66% and low error rates, with a MAE of 0.0653 and an RMSE of 0.10461. The study confirms that XGBoost's methodology is particularly effective in handling the complex and nonlinear patterns characteristic of solar energy data.

The authors in (Zhou et al., 2021) used advanced hybrid deep learning by integrating CNN, LSTM, and attention mechanisms across intervals of 7.5, 15, and 30 minutes. Their model outperformed standalone LSTM and Attention-based Long Short-Term Memory (ALSTM) models, particularly in short-term forecast scenarios. The CNN-LSTM-AM hybrid achieved the lowest Mean Absolute Percentage Error (MAPE) (18.82%) and RMSE (1.30) for short intervals such as 7.5 minutes, as presented in Table 2-1.

Table 2-1 7.5-minute Forecast Interval Results, results from study [27].

Metric	MLP	LSTM	ALSTM	Proposed Model
MAPE (%)	23.11	24.61	23.32	18.82
RMSE	1.32	1.37	1.42	1.30
MAE	0.72	0.78	0.83	0.70

2.5.3 Advanced ML Techniques for Solar Forecasting

In addition to standard ML models, researchers have explored hierarchical and graph-based models to enhance accuracy. The author in (Perera et al., 2024) introduced a new forecasting approach called Hierarchical Temporal Convolutional Neural Networks (HTCNN) for regional solar power predictions. Their two models - HTCNN A1 and A2 analyzed both local solar farm data and broader regional patterns. HTCNN A2 performs better with reduced prediction errors NRMSE by 6.5% compared to standard models, and boosted forecast accuracy by 40.2%.

The author in (Simeunović et al., 2021) proposed Graph-Convolutional LSTM (GCLSTM) and Graph-Convolutional Transformer (GCTrafo) models for multi-site forecasting. These models captured spatio-temporal dependencies between solar PV systems without relying on real-time weather data. The GCLSTM model achieved NRMSE between 8.33%-12.60% for up to four-hour forecasts, while the GCTrafo model showed better performance over extended forecast periods, especially for forecast horizons between 4 and 6 hours, due to its ability to effectively capture long-term temporal dependencies through attention mechanisms.

2.5.4 Environmental and Climatic Factors in Forecasting Models

Solar power forecasting depends basically on environmental factors as inputs. The author in (Yu et al., 2019) used LSTM models with the input factors GHI, dew point, surface pressure, temperature, relative humidity, wind speed, cloud type (via k-means clustering), and clearness-index. The model featured by creating the cloud type clusters and showed significantly better performance under cloudy and mixed weather conditions. Their LSTM model consistently achieved R^2 values above 90% and outperformed RNNs, which achieved R^2 values of 70% -79%.

The authors in (Chuluunsaikhan et al., 2021) examined the role of environmental

pollution in solar forecasting. They integrated air pollutants with environmental factors such as horizontal irradiation, module temperature, humidity, sunshine, solar radiation, cloud amount, temperature, ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), Particulate matter with a diameter of 10 micrometers or less (PM10), Particulate matter with a diameter of 2.5 micrometers or less (PM2.5), month, and hour to build the models Linear Regression, SVR, Multi-Layer Perceptron (MLP), k-Nearest Neighbors (kNN), RF, and GB. RF outperformed other models with a high R^2 value of 98.25% proving that pollution data can enhance model robustness.

The authors in (Shah et al., 2024) applied Air Quality Index (AQI) and other weather data into ML models, ConvLSTM2D model reported the best result was a highly accurate 24-hour forecast with an R^2 of 96.91%, MAE of 0.18, and RMSE of 0.10. This confirms that AQI can contribute in solar power forecasting.

The authors in (Frederiksen & Cai, 2022) used Sloped extra-terrestrial irradiance (ERAD), cloud cover, temperature, and visibility as key features in a Non-linear Autoregressive Exogenous Neural Network (NARXNN). Their model achieved seasonal NRMSE between 3.85% and 6.87%, even under complex UK weather conditions.

2.5.5 Regional and Climate-Diverse Case Studies

Regional and climate-diverse modeling plays a critical role in the ability to generalize the models. The author in (Yu et al., 2019) demonstrated that LSTM models maintained strong predictive power across various regions, especially in mixed cloud conditions as mentioned in section 2.5.4 the approach employed k-means clustering for weather classification and confirmed the ability of ML models like LSTM's to capture environmental fluctuations in climate.

The author focuses (AlSharabi et al., 2025) in a study on the complex climate conditions in Saudi Arabia and uses long-term solar irradiation prediction, using Discrete Wavelet Transform for feature extraction, GP model and SVM. The GP model outperformed SVMs and remained stable under the region's extreme conditions, with R^2 up to 99.30%.

In tropical climates, the authors in (Gayathry et al., 2024) applied SVR, Seasonal Auto-Regressive Integrated Moving Average with Exogenous Factors (SARIMAX), and

LSTM models to different Indian seasons, showing high R^2 for winter (0.97) and summer (0.96), with a reduction in monsoon periods (0.85) Table 2-2. This result suggests that even high-performing models must be customized to handle a variety of climate conditions.

Table 2-2 Performance Metrics for Seasonal Solar Irradiance Forecasting, in study [35].

Season	Model	ME	MAE	RMSE	NRMSE	R^2
Winter	SARIMAX	529.45	32.6	65.1	24.1	0.96
	SVR	LSTM	458.15	33.1	58.1	21.5	0.97
		386.98	35.0	66.0	24.3	0.96
Summer	SARIMAX	624.50	33.0	77.0	29.3	0.95
	SVR	LSTM	545.00	37.0	72.0	27.4	0.96
		676.50	48.0	103.0	39.2	0.91
Monsoon	SARIMAX	413.00	51.0	98.0	60.0	0.83
	SVR	LSTM	501.89	48.0	92.0	56.3	0.85
		518.00	60.0	119.0	72.8	0.75
Post-monsoon	SARIMAX	440.00	46.0	92.3	62.1	0.82
	SVR	LSTM	366.60	45.6	85.6	57.6	0.85
		608.00	54.0	108.0	72.7	0.76

In the UK, the authors in (Frederiksen & Cai, 2022) addressed the challenges of accurate solar power forecasting under highly variable weather conditions in the UK. It notably captured significant intra-daily variations in PV output, demonstrating strong predictive performance across different seasons, with NRMSE values ranging from 3.85% to 6.87% depending on seasonal and location-specific conditions.

The authors in the study (Simeunović et al., 2021) tested GCLSTM and GCTrafo models across different European locations. Their results confirmed that graph-based models can forecast PV output without relying heavily on local weather data, making them ideal for decentralized energy grids.

These studies illustrate the need for adaptable ML models that consider geographic, seasonal, and climatic variability.

A comparative summary of recent studies in solar energy forecasting is presented

in Table 2-3, highlighting the diversity of ML models, evaluation metrics, and reported performance across various datasets and climatic conditions. The table illustrates how hybrid and deep learning models, such as CNN-LSTM, HTCNN, and ConvLSTM2D, often outperform traditional methods in terms of precision (e.g., R^2 values above 0.90 and low RMSE and MAE).

Table 2-3 Comparison of ML Models for Solar Forecasting

Ref.	Country	Features	Models Used	Results	Best Model
(Kumar et al., 2023)	India	temperature, solar radiation, time of day, wind speed, humidity	ANN, SVR, RFR, CNN, (HMLM)	Accuracy: 89.87% - 98.67%	(HMLM)
(Yu et al., 2019)	China	global irradiance, dew point, surface pressure, temperature, relative humidity, wind speed, cloud type (clusters), clearness-index.	LSTM, RNN, CNN, BPNN, SVR, ARIMA	$R^2 > 0.90$	LSTM
(Essam et al., 2022)	USA	Plane of Array Irradiance (PAI), Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), Beam Solar Time (BST), Relative Humidity (RH), Diffuse Horizontal Irradiance (DHI), Dry Bulb Temperature (DBT), Day Angle Parameter (DAP), Atmospheric Pressure (ATP)	ANN, RF, DT, XGB, LSTM	R^2>0.9988 MAE = 0.4693 RMSE = 0.8816	ANN
(Perera et al., 2024)	Australia	Wind speed, Temperature, UV index, Cloud cover, Humidity, Atmospheric pressure, Dew point	HTCNN A1/A2, TCN, CNN, LSTM, SARIMA, SARIMAX, SN	Forecast skill score=40.2%, normalized RMSE reduced by 6.5%	HTCNN A2
(Zhou et al., 2021)	China	unspecified	CNN, LSTM, AM, SOM clustering, ANN, ALSTM	MAPE=18.82% RMSE=1.30 (7.5 min interval)	CNN-LSTM-AM
(Kim et al., 2023)	South Korea	Temperature, humidity, historical irradiance	Holt-Winters, MLP, ARIMA, SARIMA, ARIMAX, SARIMAX, LSTM	MAPE: 5.24%-7.21%	LSTM

(Abumohsen, Owda, Owda, Abumihsan, et al., 2024)	Palestine	Date and Hour, Temperature, Solar Radiation, Humidity, Wind Speed, Atmospheric Pressure, DateTime	MLP, GBM, XGB	R^2=88.66%, RMSE=0.10461 MAE=0.0653	XGB
(Abumohsen, Owda, Owda, & Abumihsan, 2024)	Palestine	Date and Hour, Temperature, Solar Radiation, Humidity, Wind Speed, Atmospheric Pressure, DateTime	CNN, LSTM, RF, GRU, SVR, Bi-LSTM, RNN	R^2=92%, RMSE=0.07, MAE=0.05	CNN-LSTM-RF
(Zulkifly et al., 2021)	Malaysia	Temperature, humidity, irradiance	SVM, GPR, LR, DT	R^2=95.91% RMSE=5.83% MAD=3.7%	Decision Tree (Fine)
(Essam et al., 2022)	Egypt	Irradiance, wind speed, temperature	ANN, RF, DT, XGB, LSTM	R^2=0.9988 RMSE=0.8816 MAE=0.4693	ANN
(AlShahabi et al., 2025)	Saudi Arabia	Global Horizontal Irradiation (GHI), Direct Normal Irradiation (DNI), Diffuse Horizontal Irradiation (DHI), air temperature, relative humidity, wind speed, wind direction, and barometric pressure	SVM, GP, DWT	R^2=99.3%	GP
(Frederiksen & Cai, 2022)	United Kingdom	Sloped extra-terrestrial irradiance (ERAD), cloud-cover, temperature, visibility	NARXNN	Seasonal nRMSE=3.85%-6.8	NARXNN 7%
(Shah et al., 2024)	India	Apparent temperature, air temperature, dew point temperature, wind speed, wind direction, relative	Linear Regression, GB, RF, XGBoost, ConvLSTM2D	ConvLSTM2D: R^2=0.9691, RMSE=0.10, MAE=0.18	ConvLSTM2D

		humidity, and Air Quality Index (AQI)			
(Chuluunsaikhan et al., 2021)	Mongolia	horizontal irradiation, module temperature, humidity, sunshine, solar radiation, cloud amount, temperature, ozone (O3), sulfur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), PM10, PM2.5, month, and hour	LR, SVR, MLP, kNN, RF, GB	RF: R^2=98.25% RMSE=0.89 MAE=0.28	RF
(Gayathry et al., 2024)	India	Hour, temperature, cloud type, relative humidity, solar zenith angle, and historical GHI	SARIMAX, SVR, LSTM	R^2=0.97 (winter) 0.96 (summer) 0.85 (monsoon)	SVR
(Gao et al., 2019)	China	Solar irradiance, air temperature, relative humidity, wind speed	LSTM, BP, LSSVM, WNN	RMSE=5.34%-13.8 MAPE=1.38%-2.01	LSTM
(Subramanian et al., 2023)	India	Temperature, irradiance,wind speed	SVM, RF, GB	AUC=99%	RF
(Saini et al., 2023)	India	solar irradiance, temperature, humidity, wind speed, cloud cover, atmospheric pressure, interplanetary insolation, and time	CNN, LSTM, SVM, RF, hybrid CNN-LSTM, EMD, WT	R^2=99.84% RMSE=7.21	Hybrid CNN-LSTM
(Simeunović et al., 2021)	Europe (multiple)	rolling mean power, global clear sky irradiance, and direct clear sky irradiance	GCLSTM, GCTrafo, STCNN, STAR	GCLSTM: NRMSE=8.33%-12 GCTrafo: NRMSE=10.83%-1	GCLSTM .6(0sh%o,rt horizon), 6G.0C7%Trafo (long horizon)

This overview supports the importance of selecting models based on both technical capabilities and regional climate variability.

2.6 CO2 Emissions Reduction and Environmental Impact Modeling

In the study (Magazzino et al., 2021), data from China, India, and the USA were used in ML algorithms and analyzed to investigate how renewable energy production, coal consumption, and economic growth affect CO_2 emissions. Using Causal Direction from Dependency (D2C) Algorithm and analyzing the result using a supervised prediction model, they identified a correlation among these variables. The findings showed that increased usage of solar and wind energy leads to a reduction in CO_2 emissions in China and the USA. In contrast, India's continued reliance on coal, despite its growing renewable energy sector, is projected to result in higher CO_2 emissions. These results highlight the necessity of transforming from fossil fuels to renewable energy sources to reduce carbon emissions.

2.7 Research Gap

Considering the studies that have been reviewed, ML methods are widely used for predicting solar energy production. Many models, including traditional techniques like SVM and RF, and some advanced approaches like LSTM networks and hybrid deep learning models achieve high accuracy in specific cases, but there are still several important gaps that limit their comprehensive effectiveness:

- Most studies focus on one geographic location. These models build on datasets from one country or region, which means they may not perform well under different climate conditions. For example, a model trained in a temperate region might not give accurate results when applied in a desert climate. There is a clear need for more generalizable models that can adapt to diverse climate conditions.

- Many models rely on limited environmental data such as temperature, humidity, wind speed, cloud cover, and air quality. These factors have a significant impact on solar radiation, and including them can greatly improve forecasting accuracy. But only a limited number of studies fully integrate these variables into their models. On the other side, there is a lack of comparison between different models because each study may use different variables ((Abumohsen, Owda, Owda, Abumihsan, et al., 2024) uses Temperature, wind speed, and irradiance, (Kumar et al., 2023) used Temperature,

humidity, and wind speed).

2.8 Summary

In summary, although ML has made important improvements in solar energy forecasting, there are still some gaps, including environmental factor diversity and model interpretability. This study will play a role in addressing these gaps by developing a generalized ML model that integrates a wide range of environmental variables across multiple geographic regions and conducts a comparative analysis across different climate conditions, also uses ML algorithms under consistent conditions to identify the most robust approaches.

Chapter 3 Methodology

3.1 Introduction

This chapter contains the work stages of the proposed study of environmental impact on renewable energy optimization, including data collection methods and data description, cleaning, and preparation, in addition to exploratory data analysis and data preprocessing. Then it utilizes the ML model building and the amount of predicted solar power generation using LR, RF, XGBoost, Bi-LSTM, and CNN-LSTM models.

Figure 3-1 shows the main phases of the methodology considered in the investigation, as the second part of this chapter will describe in depth the work packages within each stage.

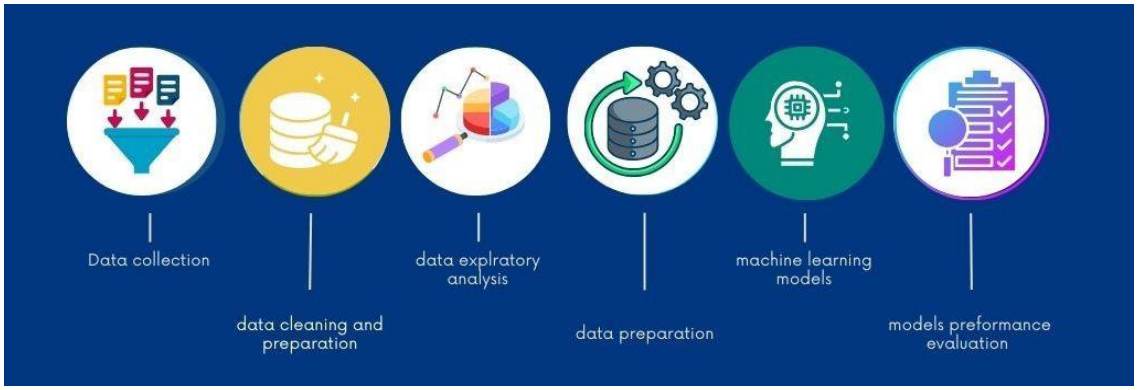


Figure 3-1 A Study of Environmental Impact on Renewable Energy Optimization Prediction Mode

3.2 Used technology

Data preparation, exploration, preprocessing, and modeling activites used for implementing the proposed models presented in this study were accomplished using the Python programming language, executed via the “Jupyter Notebook” web-based platform.

Python is an open-source programming language that is used widely these days in developing data-driven applications and models. It is built on top of an extensive set of powerful libraries for processing and manipulating data, in addition to the modeling

utilities provided by the built-in ML libraries, which make it an advanced, flexible, capable coding option for these purposes (Géron, 2022). The main Python libraries that were used throughout the implementation process of the proposed models in this study are Pandas and Numpy, Seaborn, and Matplotlib, which are two of the most important core libraries in Python.

- Numpy is the fundamental package for scientific computing with Python [63].

- Pandas is a Python package for high-level data manipulation and analysis, which is built on the Numpy package (VanderPlas, 2016).

- Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python (Embarak et al., 2018).

- Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics (Embarak et al., 2018).

- Scikit-learn is a Python library designed to support prediction analysis and ML modeling. It offers a variety of powerful machine-learning algorithms for supervised and unsupervised learning (Hao & Ho, 2019).

- XGBoost: Used to build a scalable gradient boosting regression model. Known for its computational efficiency and accuracy, XGBoost supports regularization and parallel processing (Chen & Guestrin, 2016).

- TensorFlow (Keras API): Deep learning models, including CNN-LSTM and Bi-LSTM, were built using the Keras API. The models used 1D convolutional layers, pooling layers, LSTM units, and bidirectional processing for sequence modeling, along with Dropout and Adam optimization. TensorFlow allowed for flexible architecture definition and efficient training (Abadi et al., 2016).

3.3 Dataset collection and description

The dataset used in this study was collected from multiple sources to evaluate and compare solar power generation forecasts between Balearic Islands in Spain and the

Tubas district in Palestine based on various meteorological factors. The datasets were obtained covering the period from January 1, 2022, to July 31, 2023. The following describes the collected data:

3.3.1 Meteorological Data

Meteorological data were collected from the NASA POWER (Prediction Of Worldwide Energy Resources) database, which provides weather parameters for a single point location with available hourly temporal level (of Worldwide Energy Resources (POWER) Project, 2025). The extracted dataset contains five parameters, which are Sky Surface Shortwave Downward Irradiance, Temperature, Specific Humidity, surface pressure, and wind speed Table 3-1, for the locations:

- Tubas, Palestine at Latitude 32.3211 Longitude 35.3597.
- Balearic Islands, Spain at Latitude 39.5696, Longitude 2.6502.

Table 3-1 Feature Description for Meteorological Data

Variable	Description	Type	Unit	Data Nature
YEAR	Year of observation	Integer	-	Categorical
MO	Month of observation	Integer	-	Categorical
DY	Day of observation	Integer	-	Categorical
HR	Hour of observation (local time)	Integer	-	Categorical
ALLSKY_SFC_SW_DWN	All-sky surface shortwave downward irradiance	Float	Wh/m^2	Continuous
T2M	Air temperature at 2 meters	Float	°C	Continuous
QV2M	Specific humidity at 2 meters	Float	g/kg	Continuous
PS	Surface atmospheric pressure	Float	kPa	Continuous
WS10M	Wind speed at 10 meters	Float	m/s	Continuous

3.3.2 Solar Power Generation Data

The solar power generation data serves to evaluate the performance of solar energy

forecasting models. Its hourly temporal allows the analysis and aligns well with the environmental data collected from NASA's POWER database

- Solar Power Generation Data for Tubas area: The solar power generation dataset for Tubas was obtained directly from Tubas Electricity Company in Palestine. This dataset records the actual generated energy from the deployed PV system, which has a capacity of 8 MW and an annual production rate of 15 million kWh (Palestine Investment Fund, 2025). The data include 4,437 hourly continuous records Table 3-2.

Table 3-2 Feature Description for Solar Power Generation Data (Tubas)

Variable	Description	Type	Unit	Data Nature
Date	Timestamp indicating date and hour of measurement	Datetime	-	Time
ActivePower	Active power output of the solar PV system	Float	kW	Continuous

- Solar Power Generation Data for Balearic Islands area: The solar power generation dataset for the Balearic Islands, Spain (with installed solar PV capacity of 168 MW (Donoso et al., 2023)) was obtained from the Spanish electricity grid operator (de España (REE), 2025). It includes detailed records of the generation mix in megawatts (MW) at 5-minute intervals. The dataset consists of 18 columns and 10204 entries, each reflecting the contribution of various generation sources to the regional power grid Table 3-3.

Table 3-4 provides a summary of the datasets used in this study, including their geographic origin, temporal resolution, data sources, period, and key features.

3.4 Data Cleaning

Cleaning the data is considered a crucial stage in the data utilization process, which has a significant impact on exploratory data analysis and on the modeling phases. As a requirement for the study, we have to deal with two datasets, one for tubas and the other for the Balearic Islands. To do that, the datasets are handled separately based on the source of each one.

Table 3-3 Feature Description for Power Generation Mix Data (Balearic Islands)

Variable	Description	Type	Unit	Data Nature
Hour	Timestamp at 5-minute intervals	Datetime	-	Time
Solar PV	Electricity generated by solar PV sources	Float	MW	Continuous
Wind	Electricity generated by wind power	Float	MW	Continuous
Coal	Electricity generated by coal	Float	MW	Continuous
Combined cycle	Electricity generated by combined cycle plants	Float	MW	Continuous
Cogeneration	Electricity generated by cogeneration units	Float	MW	Continuous
Diesel engines	Electricity generated by diesel engines	Float	MW	Continuous
Gas turbine	Electricity generated by gas turbines	Float	MW	Continuous
Energy links	Electricity is exchanged through inter-island and mainland links	Float	MW	Continuous
Renewable wastes	Electricity from renewable waste sources	Float	MW	Continuous
Non-renewable wastes	Electricity from non-renewable waste sources	Float	MW	Continuous
Auxiliary generation	Backup or auxiliary electricity generation	Float	MW	Continuous

Table 3-4 Summary of Datasets Used in the Study

Dataset Type	Location	Resolution	Source	Period	Key Features
Meteorological Data	Tubas, Palestine	Hourly	NASA POWER	Jun, 2022 –Jul 2023	Solar irradiance, temperature, humidity, pressure, wind speed
Meteorological Data	Balearic Islands, Spain	Hourly	NASA POWER	Jun, 2022 –Jul 2023	Solar irradiance, temperature, humidity, pressure, wind speed
Solar PV Generation	Tubas, Palestine	Hourly	Tubas Electricity Company	Jun, 2022 –Jul 2023	Timestamp, active power output (kW)
Power Generation Mix	Balearic Islands, Spain	5-minute.	Red Eléctrica de España (REE)	Jun, 2022 –Jul 2023	Timestamp, solar PV (MW), wind, coal, thermal, inter-island links, renewable/non-renewable wastes

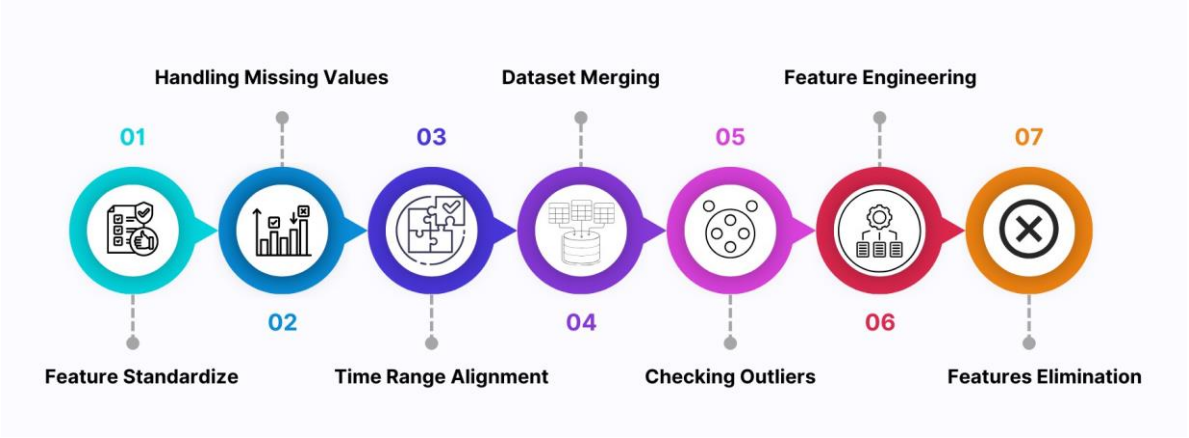


Figure 3-2 The Applied Steps for Cleaning Raw Datasets

3.4.1 Data Cleaning for Tubas data/sets

This section describes the phases involved in preparing and refining the

meteorological and solar PV power generation datasets collected for the Tubas region (Figure 3-2).

3.4.1.1 Feature Standardize

To merge the table of meteorological power data, we have to unify the datetime format in both datasets, the meteorological data provided date and time as four space columns: YEAR, MO, DY, and HR. These were combined into one datetime column. Also, the solar power dataset contained a timestamp as a string under the Date column, which was transformed into a similar format in the meteorological data.

3.4.1.2 Handling Missing Values

The issue of missing values is considered a common problem in any collected dataset used to build data-driven models. Missing values have to be checked and handled as a first step before proceeding to any other task within the data science process.

By checking the missing values, there is any missing value in both datasets, meteorological and power, which back that the source of the data is a system not allowing missing values.

NASA's dataset sometimes uses -999 to denote missing or invalid data, as mentioned in the extracted dataset. Also, checking the availability of -999 value confirms that there is no presence of this value nor NaN value.

3.4.1.3 Time Range Alignment

To prevent misalignment during merging and ensure that each row in the meteorological data could be associated with valid energy production data, the meteorological data was filtered to match the time range of the solar power dataset. This was done using the minimum and maximum timestamps from the solar data as the start and end periods. On the other hand, the solar power dataset was not recorded at hourly intervals, which causes an inconsistency issue. So, resampling to hourly frequency was used, which creates a row for every hour in the time range.

To analyze seasonal and monthly patterns in solar power generation, a consistent

one-year time range was selected across all datasets. This period was from August 1, 2022, to July 31, 2023. Consequently, all records falling outside this interval were dropped and excluded from the analysis.

3.4.1.4 Dataset Merging

With consistent data in both meteorological and power tables, the datasets were merged using an inner join on the datetime column. This ensured that the data is ready for analysis and model building.

3.4.1.5 Checking Outliers

Maintaining data consistency is considered an important step in building powerful prediction models. Outliers are one of the common issues in data inconsistency (Ackerman et al., 2020). In a general definition, a value is recognized as an outlier if it has a significant and obvious deviation from other observations among the processed subjects (Aguinis et al., 2013).

To detect outliers, the interquartile range (IQR) method was used. For each numeric feature, values outside 1.5 * IQR from the 25th and 75th percentiles were identified as outliers (Acharya, 2024) Table 3-5 shows the detected outliers, the majority of outliers occurred in the wind_speed feature, while no outliers were detected in the power_output target. These data points were retained because they likely represent meaningful physical variability rather than noise or error. Given the nature of meteorological data, extreme conditions can be realistic and critical to model accuracy.

Table 3-5 Detected Outliers per Feature Using IQR Method for Tubas dataset

Feature	Outlier Count
solar_irradiance	0
temperature	0
humidity	13
pressure	7
wind_speed	268
power_output	0

3.4.1.6 Feature Engineering

Feature engineering is a process of extracting informative features from the raw data to improve model accuracy. Feature engineering often contributes more to the success of a model than the choice of algorithm itself (Géron, 2022). Time-based variables are particularly valuable in energy forecasting, where the generated power often follows daily and seasonal cycles.

The datetime feature was extracted, and the features in Table 3-6 were added to help improve the model's learning.

Table 3-6 Engineered Time-Based Features

Variable	Description	Feature Type
year	Calendar year of the observation	Numerical
month	Calendar month (1–12)	Numerical
week	ISO week number of the year	Numerical
weekday	Day name of the week (e.g., Monday)	Categorical
daynum	Day of the month (1–31)	Numerical
dayofyear	Day number in the year (1–366)	Numerical
hour	Hour of the day (0–23)	Numerical

3.4.1.7 Features Elimination

The process of selecting the final feature list from the dataset was demonstrated by eliminating useless features. The elimination of features was applied to variables that are irrelevant to the purpose of this study, but there is no irrelevant feature in the obtained dataset, and all are useful for modeling, so none of them was removed.

Also, low variance and high correlation techniques were applied to ensure that each feature has a significant input to the model:

Low Variance features: The Low variance filter is the process of removing the variables that are less contributing to/affecting the target variable (Brownlee, 2020). Low-variance features are recommended as a prior step in preprocessing (Brownlee, 2020). To

accomplish this, the threshold level was set to 0.01, and any feature with variance less than 0.01 was considered useless, but no such features were found in the dataset.

High Correlation: Highly correlated features provide redundant information and can cause multicollinearity. A Pearson correlation matrix was computed for all numeric features, and no feature had correlation values exceeding 0.95, so no need to remove any features due to multicollinearity.

3.4.2 Data Cleaning for Balearic Islands datasets

This section describes the phases involved in preparing and refining the meteorological and solar power generation datasets collected for the Balearic region. For the Balearic regions, the dataset containing meteorological data was obtained from NASA POWER (of Worldwide Energy Resources (POWER) Project, 2025) recorded hourly, and solar power obtained from REE (de España (REE), 2025), recorded every 5 minutes, and contains a rich feature Table 3.3

3.4.2.1 Data Import

The solar power data, the REE platform provides separate daily excel files that require additional merging, pre-cleaning before being combined with meteorological data:

Data from all daily reports related to solar energy has been collected and compiled into a single table.

Duplicate observations were identified and removed based on the Hour column (datetime).

As the data of solar power contains the amount of power generated from all systems in the Balearic Islands, the irrelevant feature, which is not useful or related to the study object, was removed, like Wind, Coal, Combined cycle, Cogeneration, Diesel engines, Gas turbine, Balearic-Peninsula link, Other special regime, Thermal renewable, Mallorca-Menorca link, Mallorca-Ibiza link, Other renewables, Auxiliary generation, Ibiza-Formentera link, Non-renewable wastes, and Renewable wastes.

The solar power data records every 5 minutes, and to integrate the data with hourly meteorological data, the solar power data is aggregated to be hourly by summing the data

generated for every hour.

These steps produced a single, clean solar power dataset has a structure compatible to integrate with the meteorological dataset.

3.4.2.2 Feature Standardize

This step follows the same procedure as detailed in Section 3.4.1.1 for the Tubas dataset. The meteorological data contained four columns: YEAR, MO, DY, and HR, which were combined into a single datetime column. Similarly, the solar power dataset's timestamp (recorded under Hour) was transformed into the same datetime format to make the integration.

3.4.2.3 Handling Missing Values

The process for identifying and handling missing values is also the same approach described for the Tubas data. As confirmed through inspection and there were no missing values (neither NaN nor placeholders like -999) in either dataset.

3.4.2.4 Time Range Alignment

To ensure accurate integration of datasets, the meteorological data was temporally aligned with the solar power data, following the same logic applied in the Tubas case (see Section 3.4.1.3).

3.4.2.5 Dataset Merging

The merging process was identical to that used in 3.4.1.4. Once the datetime columns in both the meteorological and solar power datasets were standardized, the two tables were merged using an inner join on the datetime field. This operation ensured that only rows with corresponding entries in both datasets were retained. As a result, the merged dataset was fully aligned and ready for subsequent analysis and modeling.

3.4.2.6 Checking Outliers

As mentioned in 3.4.1.5, detecting outliers is very crucial to build an accurate

model; additionally, outliers might distort descriptive statistics. The IQR method is also applied to detect outliers for each numeric feature, the findings are shown in Table 3-7.

Table 3-7 Number of Outliers Detected per Feature (IQR Method) for Balearic Island dataset

Feature	Outlier Count
solar_irradiance	0
temperature	0
humidity	0
pressure	672
wind_speed	328
power_output	107

3.4.2.7 Feature Engineering

As in 3.4.1.6, feature engineering was applied to improve the learning efficiency of the models. Time-based features are crucial in the forecasting of renewable energy, where solar generation is known to follow hourly and seasonal patterns (Géron, 2022).

From the datetime field, the same set of features listed in Table 3.6 was extracted. These include indicators for year, month, week, weekday, day of year, and hour.

3.4.2.8 Feature Elimination

The feature elimination process was applied to ensure that all input variables provided meaningful and independent contributions to the model. This step followed the same approach as previously used for the Tubas dataset and included two criteria: low variance filtering and high correlation analysis.

Low Variance filtering: A variance threshold of 0.01 was used to detect characteristics with minimal variability. No variables show value below this threshold, indicating that all features are informative and should be considered.

High Correlation Analysis: A Pearson correlation matrix was computed among all numeric variables, including the target power_output. Features with correlation

coefficients greater than 0.95 were considered redundant. The analysis revealed no such highly correlated pairs.

As a result, no features were eliminated at this stage. All available environmental features were retained for further modeling and analysis, and further analysis will proceed during exploratory data analysis.

3.5 Exploratory Data Analysis

In data-driven models, exploratory data analysis using statistical or graphical methods plays a crucial role in investigating the dataset (Owda et al., 2023). Additionally, it can be helpful to obtain a brief overview, identify general trends, and make informed assumptions to build upon before proceeding to the phase of ML modeling.

The data analysis process encompasses univariate analysis, which is used to study a single feature within the dataset, and multivariate analysis to explore the relationship between two or more features. An important part to consider in the process of exploring the variables.

Table 3-8 Descriptive Statistics for the Balearic Region

	solar_ irradiance	temperature	humidity	pressure	wind_speed	power_ output
count	8759.00	8759.00	8759.00	8759.00	8759.00	8759.00
mean	197.24	20.33	11.12	101.28	4.21	455.66
std	272.46	5.92	3.65	0.61	2.60	646.63
min	0.00	6.78	3.91	99.18	0.03	0.00
25%	0.00	15.52	8.26	100.93	2.36	0.00
50%	7.72	19.77	10.73	101.22	3.69	37.70
75%	371.40	25.42	13.88	101.57	5.46	866.50
max	974.40	34.41	22.58	103.34	17.41	3478.00

In the dataset is the selection of exploration techniques, which should be determined wisely according to various considerations, such as the type of variable to explore, the desired information of interest that must be evaluated, and the audience who will use or

look at the results of exploration (Deming et al., 2018).

In this study, descriptive and visualization techniques using a graphic approach were used, which provided comprehensive insight into every feature in the dataset and identified the different relationships between variables in the dataset.

The analysis is compatible with the study's goal, as it uses a comparative approach to make comparisons and gain insight into the differences between two areas in the efficiency of power generation, the differences in meteorological and environmental characteristics, and their relations to each other.

3.5.1 Descriptive Analysis

In this descriptive statistical analysis, the meteorological and power generation variables for the Balearic Table 3-8 and Tubas Table 3-9 regions are analyzed. The Balearic dataset has 8759 observations, while the Tubas dataset contains 3916 entries, indicating a more temporal coverage in Balearic. Both datasets share the same features, such as solar irradiance, temperature, humidity, pressure, wind speed, and the amount of power output.

- Solar Irradiance: a direct measure of sunlight intensity, is the most important characteristic for PV energy production (Chaaban, 2023).

The average solar irradiance in Tubas region is 230.37 W/m², with values ranging from 0.0 to 1023.97 W/m².

The average solar irradiance in Balearic Islands is 197.24 W/m², with values ranging from 0.0 to 974.40 W/m².

Table 3-9 Descriptive Statistics for Tubas Region

	solar_ irradiance	temperature	humidity	pressure	wind_speed	power_output
count	3916.00	3916.00	3916.00	3916.00	3916.00	3916.00
mean	230.37	22.22	9.09	99.43	2.21	201.75
Std	306.84	8.78	3.28	0.50	1.21	279.28
Min	0.00	1.69	3.34	98.29	0.01	0.00
25%	0.00	15.81	6.53	99.04	1.36	0.00

50%	11.12	21.83	8.42	99.40	1.92	3.29
75%	437.95	28.17	11.20	99.77	2.83	391.61
Max	1032.97	46.36	18.22	101.04	7.06	978.85

Irradiance in Balearic and Tubas exhibits a high standard deviation (272.46 and 306.84, respectively), indicating seasonal variations, as expected due to weather fluctuations. The 25th percentile is 0.0 in both cases, meaning that more than 25 percent of the readings occurred in areas with no solar radiation. This is also because a large portion of the data points represent nighttime or periods of winter or heavy cloud cover. These results indicate that radiation alone is insufficient to determine energy production without taking into account the time factor.

Temperature: Temperature also affects the efficiency of PV energy, as high temperatures can reduce the efficiency of solar panels (Shaker et al., 2024).

A slightly higher mean temperature is observed in Tubas (22.22°C) compared to the Balearic Islands (20.33°C).

Tubas experience a wider range of temperatures, ranging from 1.69°C to 46.36°C, compared to the Balearic Islands, which range from 6.78°C to 34.41°C. This indicates harsh summer conditions in Tubas and milder conditions in the Balearic Islands.

Humidity: Humidity affects atmospheric clarity and, consequently, solar radiation penetration.

The Balearic Islands exhibit a higher average humidity of 11.12% , compared to 9.09% in Tubas.

The standard deviation of humidity is also slightly higher in the Balearic Islands, 3.65 compared to 3.28 in Tubas, reflecting more variable humidity levels.

Both locations have a relatively low humidity range.

Wind speed: Wind speed can indirectly affect the performance of PV panels by helping to cool the panels. The Balearic Islands record higher average wind speeds (4.21 m/s) and a wider spread (maximum: 17.41 m/s). On the other hand, Tubas shows lower average wind speeds of 2.21 m/s and a lower spread (maximum: 7.06 m/s), as well as

lower variability with a standard deviation of 1.21 compared to 2.60 in the Balearic Islands.

Power Output: The amount of generated power is the dependent variable and the target variable of the ML model.

The average power output in Balearic Islands is 455.66 MW, ranging from 0.0 to 3478.00 with a high standard deviation of 646.63, indicating a wide range of production levels. Tubas also has a high standard deviation (279.28)

In both regions, 50% of the observations record power output is very low (50th percentile in Balearic: 37.70 MW, Tubas: 3.29 KW), suggesting that a wide range of periods have low power production. This highlights the strong skewness of the power output distribution. The standard deviation for power output is notably large in both datasets, suggesting significant fluctuation across different times of the year.

The average solar irradiance and temperature were observed to be higher in the Balearic dataset. However, the average of humidity, pressure, and wind speed is fewer a little, showing differences in the meteorological dependencies in each region. Temperature values in both regions follow seasonal patterns, pressure remained relatively stable in both regions as standard deviation equal 0.61 and 0.50 in Balearic and Tubas respectively, while wind speed in Tubas exhibited a slightly lower mean with less variation compared to Balearic.

Regarding the target variable, power output, both datasets show highly skewed distributions, with a considerable number of zero values may indicate a relation with repeated low values in solar irradiance. These comparative insights refer to the influence of geographic and meteorological factors on solar power generation.

3.5.2 Visualization

The illustrated figures below explore the comparison characteristics of the dataset's features in Balearic and Tubas.

3.5.2.1 Distributions Comparison

Figure 3-3 shows that Balearic generated significantly more total solar power than

Tubas over the observed period (Total Power Output - Balearic: 3991102.40 MW and in Tubas: 790061.63 KW). This reflects the larger system size and higher installed capacity in Balearic Islands as mentioned in section 3.3.2.

Figure 3-4 present the distribution of power output and the other variables through the data set in Balearic Islands and Tubas.

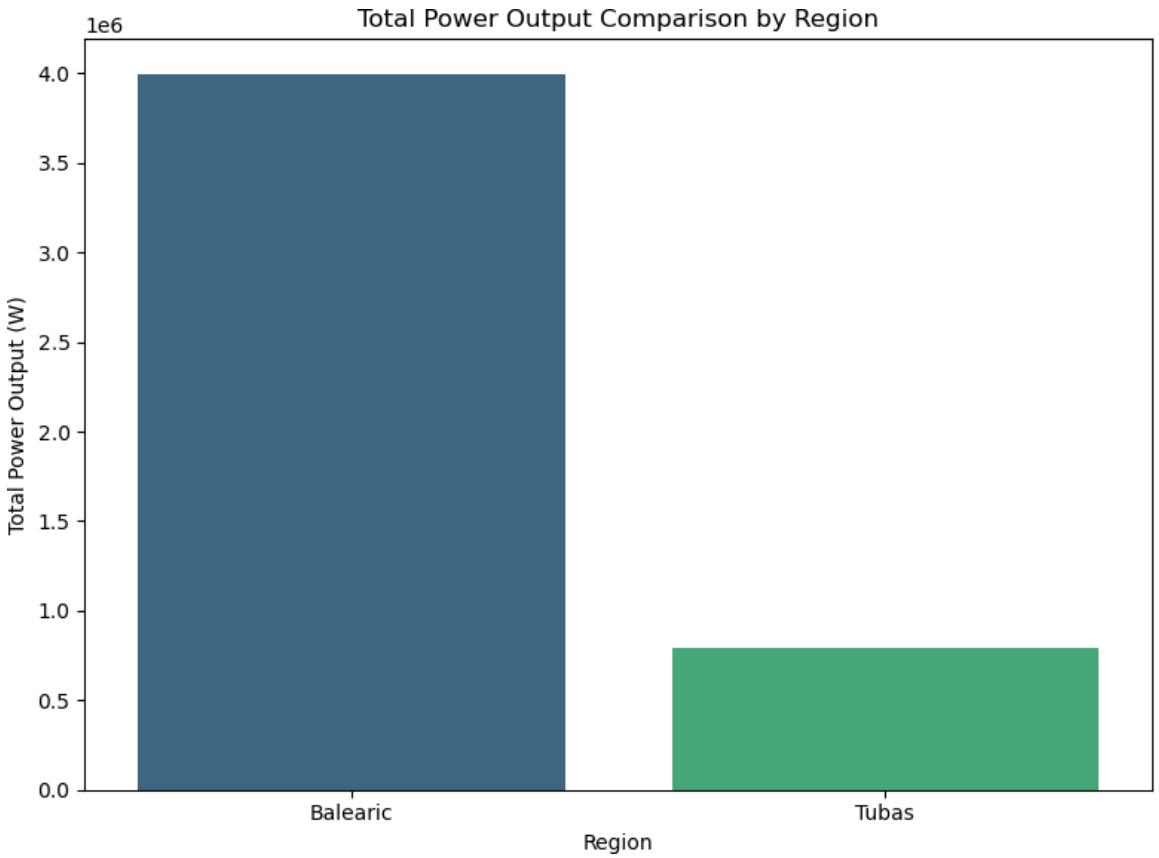


Figure 3-3 Total Power Output Comparison by Region

Power generation is heavily skewed in both regions, where little or no power is generated during most hours. The Balearic Islands have higher production records, indicating better solar panel deployment or more consistent sunlight.

Both regions have many low (or zero) irradiance hours, typically at night. Tubas show a slightly wider peak at higher irradiance values, indicating stronger sunlight during daylight hours.

Tubas are characterized by temperature fluctuations and higher maximum temperatures compared to the Balearic Islands. The Balearic Islands have a more even

distribution of temperatures, indicating a stable climate.

Tubas' humidity curve is peaked, with a focus on the lower humidity range, indicating consistently dry conditions, while the distribution of the Balearic Islands is wider and flatter, indicating more variability in atmospheric humidity. The lack of overlap in the middle and upper humidity levels shows that Balearic experiences much higher humidity values than Tubas.

The two regions show separated peaks, with no overlap between distributions. This difference indicates distinct atmospheric conditions, likely due to differences in elevations (Balearic is higher, near sea level).

Tubas have a high, narrow peak at low wind speeds, exhibiting calm conditions. Balearic Islands distribution is flatter, reaching speeds of over 17 meters per second, indicating stronger winds. The right-skewed shape for Balearic indicates the presence of occasional high-wind days.

3.5.2.2 Bivariate Analysis

Figure 3-5 illustrate the strength or variability of the relationship between power output and the other variables through the data set and shows potential differences across regions (Balearic vs. Tubas).

The scatter plot clearly shows a positive nonlinear relationship between solar irradiance and power output for both regions. As irradiance increases, power output also increases. However, the trend saturates at higher irradiance levels, especially for Tubas. Balearic Islands show a wider spread and higher maximum outputs, indicating more panels or better system efficiency.

The relationship between temperature and energy output appears less clear and more dispersed, but it still shows a clear trend, In the Balearic Islands, there is an increase in energy production between 15 and 25°C, with a decline as temperatures exceed 30°C may due to the reduced efficiency of panels at higher temperatures. In Tubas, energy production increases consistently with temperature, but with less fluctuation.

A weak to moderate inverse trend between humidity and energy output is observed in the Balearic Islands. Energy output is more dispersed and slightly higher at low

humidity levels, gradually decreasing to above 15%. The Tubas region shows greater dispersion at low humidity levels, with a slight decrease in output as humidity rises. This suggests that high humidity may negatively impact solar performance, likely due to increased atmospheric humidity scattering sunlight.

The pressure diagram shows limited correlation. The Balearic Islands' production is relatively high across a narrow pressure range (101-102.5 kPa), suggesting that pressure is not a strong limiting factor in this environment. Tubas' energy values are concentrated at lower pressures (around 99-100 kPa), but also without a clear trend. The absence of a gradient suggests that pressure is not an influential factor in energy production at either region.

There is a slight positive relationship between wind speed and energy production in the Balearic Islands, particularly in the 2-6 meters per second range. Energy production in the Balearic Islands tend to increase slightly with wind speed, possibly due to the cooling effects of panels that help maintain efficiency during sunny hours. In Tubas, wind speeds are low and concentrated below 4 meters per second, meaning there is little variation in wind characteristics, which explains the lack of wind influence on energy production.

Figure 3-6 shows a correlation analysis to quantify the strength and direction of the relationship within the dataset's features.

In both regions, solar radiation showed the strongest positive correlation with power output ($r = 0.76$ and $r = 0.80$ in Tubas and the Balearic Islands, respectively), indicating that it is a key factor in PV power generation rates. However, the remaining factors showed different effects between the two regions.

In Tubas, both temperature ($r = 0.59$) and wind speed ($r = 0.39$) were moderately correlated with power output, suggesting that, in addition to sunlight, heat and wind play supporting roles in generation efficiency. On the contrary, humidity ($r = -0.25$) and pressure ($r = -0.15$) showed negative correlations, suggesting that drier, lower-pressure conditions may be more helpful to solar power production in this inland region.

In the Balearic Islands, the contribution of other attributes was relatively limited. Except for solar radiation, temperature showed only a weak correlation with energy output ($r = 0.21$), while all other variables correlated close to zero. This suggests that the

coastal climate of the Balearic Islands exhibits more consistent environmental conditions, with solar radiation alone largely determining the variability of output.

These observations confirm that while solar radiation is globally important, the importance of these characteristics varies by location, highlighting the need for site-specific modeling strategies in solar energy forecasting.

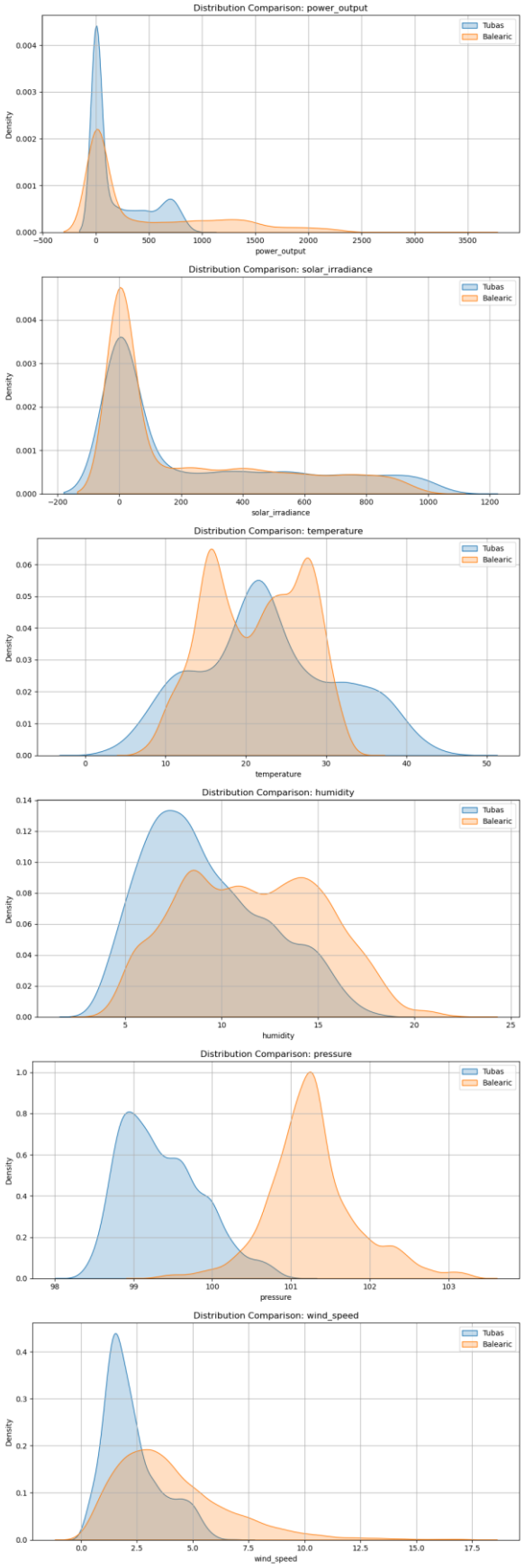


Figure 3-4 Distribution of Features in Balearic Islands and Tubas

Furthermore, the strong correlation between variables in the Balearic Islands between temperature and humidity (r = 0.89) may indicate the presence of multicollinearity issues, which should be carefully addressed in model training. Furthermore, the positive effect of wind speed in Tubas (r = 0.39), compared to its slightly negative effect in the Balearic Islands (close to zero), reinforces the idea that environmental interactions depend on region. These results confirm the need to consider the interdependence of meteorological factors to improve the accuracy of predictive solar energy models.

3.5.2.3 Time Series Trend Analysis

In this subsection, a temporal trend analysis is conducted to explore the temporal behavior of solar energy production in the Balearic and Tubas regions. By studying the evolution of energy generation over time, we aim to uncover patterns related to seasonality, as well as months, hours, and days.

Figures 3.7 and 3.8 show the average daily raw energy production and the normalized output trends respectively.

In the raw trend Figure 3-7, the Balearic Islands consistently exhibit significantly higher energy production than Tubas, reflecting differences in system size. The Balearic Islands curve also exhibits stronger seasonal variation, with significant peaks during the summer months and drops in winter, while Tubas maintains a relatively flatter pattern with declining production throughout the year. However, due to the large variation in generation capacity, direct visual comparison of time trend patterns is limited.

To address this, energy outputs were normalized using a minimum and maximum scale, as shown in Figure 3-8. This transformation enables comparison of relative seasonal behavior and temporal dynamics regardless of system size. When normalized, the two regions exhibit similar seasonal patterns, with increases in energy generation beginning in early spring and peaking around summer, followed by declines as winter approaches. The Balearic Islands trend shows smoother seasonal arcs, while the Tubas signal is more irregular, indicating possible data variability or local weather effects.

Figure 3-9 shows the average hourly patterns of solar radiation and solar power output in Tubas and the Balearic Islands. In both regions, the curves show the expected shape of the output rate of solar power systems: irradiance increases after sunrise, peaks at midday, and then decreases with sunset. The timing and shape of the power output curves follow the level of the irradiance curves, reflecting the direct dependence of PV performance on available solar radiation.

In Tubas, power output begins to rise with the onset of solar radiation around 7:00 a.m., peaks between 11:00 a.m. and 1:00 p.m., and then gradually declines until 5:00 p.m. It should be noted that the power output curve lags slightly behind the irradiance peak, which may be due to efficiency losses due to temperature or inverter response behavior. The alignment between the two curves is relatively accurate, but the irradiance reaches higher relative values. In contrast, the Balearic Islands exhibit a more pronounced power output curve, peaking between 12:00 and 2:00 PM. interestingly, the irradiance curve in the Balearic Islands exhibits a flatter midday plateau, consistent with a strong and sustained power output response.

These visual patterns reinforce the role of irradiance as a key driver of solar power generation and reveal the importance of site-specific calibration and performance modeling for accurate solar forecasts.

Figure 3-10 provides a comprehensive overview of the hourly-monthly behavior of solar energy production in both the Tubas and Balearic. These Heatmaps show the average hourly energy production in all months, providing insight into the power generation performance during the day-month intervals.

In the Balearic Islands, the heat map shows a clear and intense daily pattern of power production from around 10:00 to 16:00, with peak production occurring between 11:00 and 13:00. The highest power production occurs from April to September. Production decreases significantly during the winter months.

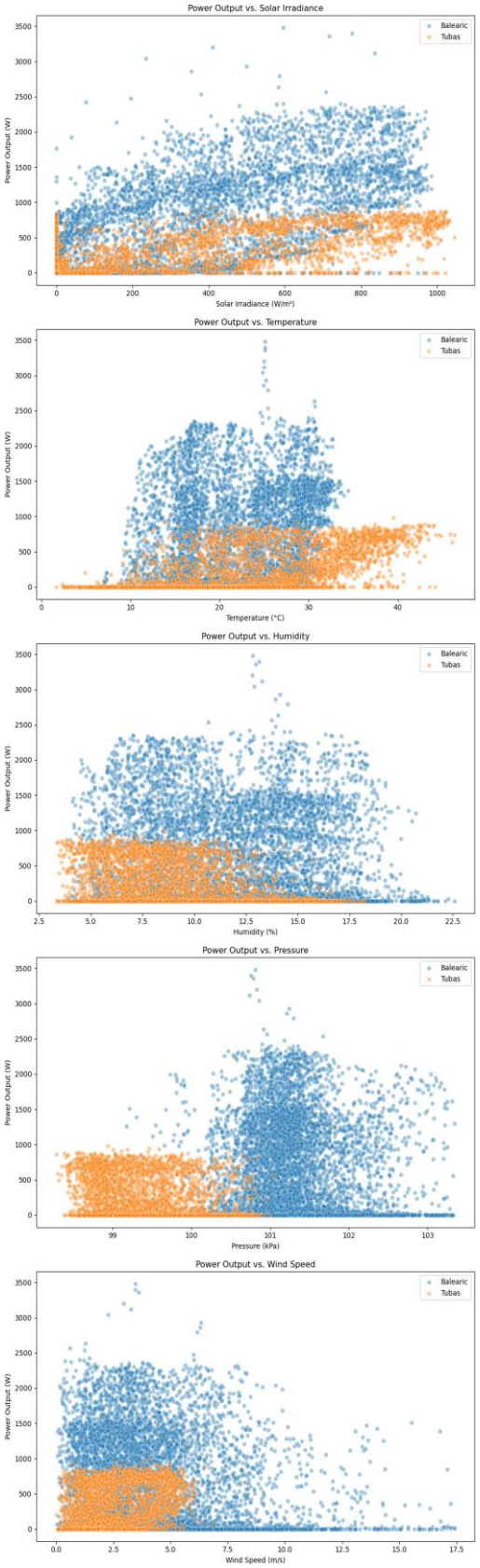


Figure 3-5 Relationships between Solar Power Output and Meteorological Factors in Balearic and Tubas

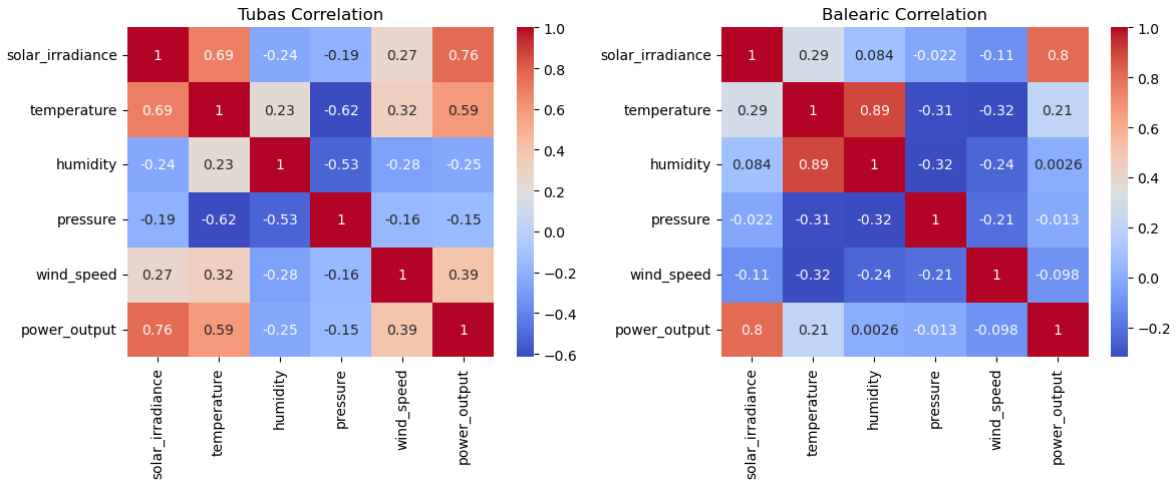


Figure 3-6 Correlation Heatmaps analysis for Balearic and Tubas dataset

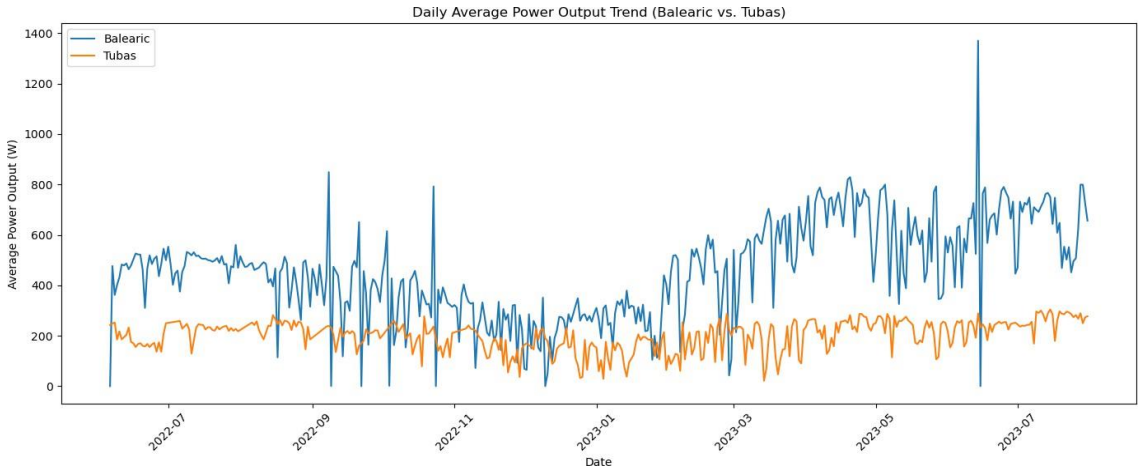


Figure 3-7 Daily Power Output Trend (Balearic vs. Tubas)

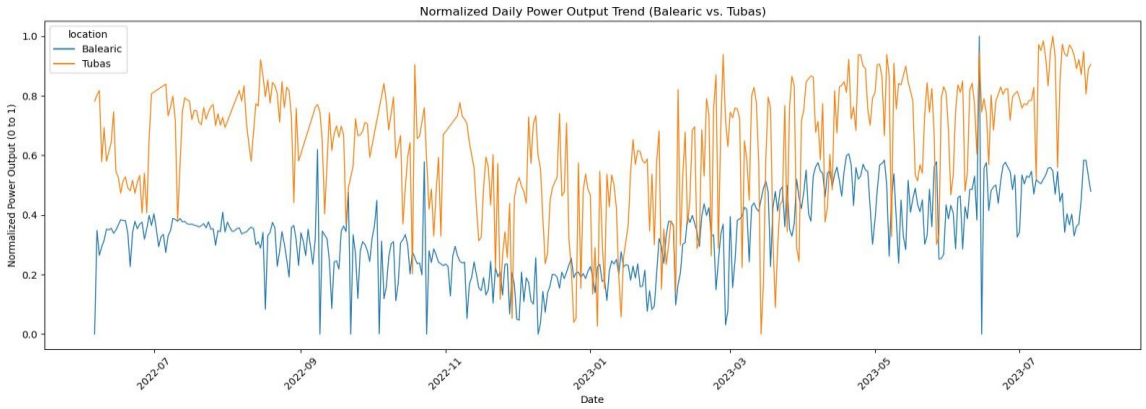


Figure 3-8 Normalized Daily Power Output Trend (Balearic vs. Tubas)

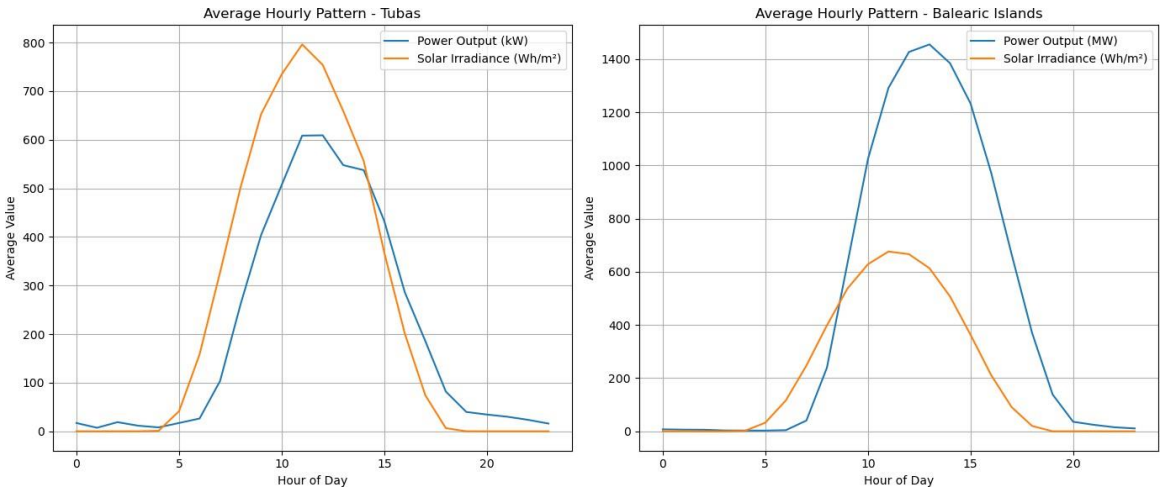


Figure 3-9 Average Hourly Pattern - Balearic Islands (Balearic vs. Tubas)

In Tubas, the pattern is similar, with peak between 10:00 and 15:00, with lowest power generation during winter months. In October, readings appear inconsistent with other data, as energy production occurs at all hours of the day and is higher than usual at night, likely indicating an error in the recorded data. The different generation patterns between Tubas and the Balearic Islands reinforce the need to evaluate the performance of ML models in different climate regions to ensure their generalization.

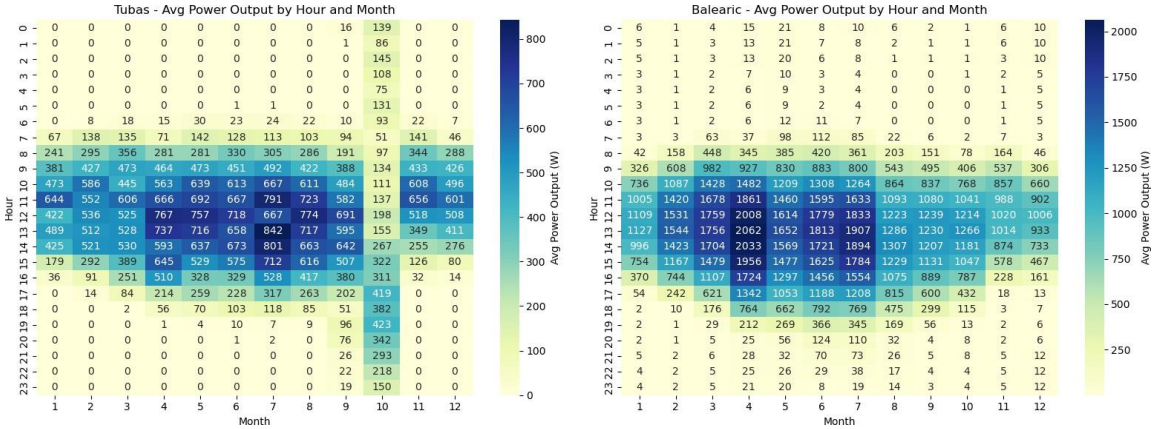


Figure 3-10 Average Power Output by Hour and Month in Balearic and Tubas

Boxplots shown in Figure 3-11 and Figure 3-12 were used to visualize the monthly distributions of energy production in the Baleari.c and Tubas regions, providing insights into monthly behavior and the presence of outliers. These plots summarize how the data is distributed throughout the months.

In the Balearic region, energy production shows a strong seasonal pattern. Energy generation is lowest in January and December, increases gradually from March, and peaks between April and July. These high-production months are characterized by high medians and wide IQRs, indicating greater stability in the high level of energy production. From August, energy production declines, with the period from October to December showing compressed IQRs and low medians. It is worth noting that outliers are more frequent in these low-production months, particularly in October and November. This is a statistical effect; during months when energy production is typically low, the presence of some sunny days allows for relatively high production during that period and is therefore classified as an outlier. In contrast, the Tubas region has a more consistent and regular pattern of energy production throughout the months of the year. While a little increase in energy generation is observed from April to August, the medians and IQRs in most months remain relatively narrow. These differences highlight the need for specific condition modeling approaches.

The Figure 3-13 shows the seasonal distribution of total annual solar energy production in the Balearic Islands and Tubas regions. Each section represents the proportion of annual energy generation contributed by each meteorological season (National Oceanic and Atmospheric Administration, 2023), winter, spring, summer, and autumn, highlighting how seasonal conditions affect the overall performance of solar PV systems.

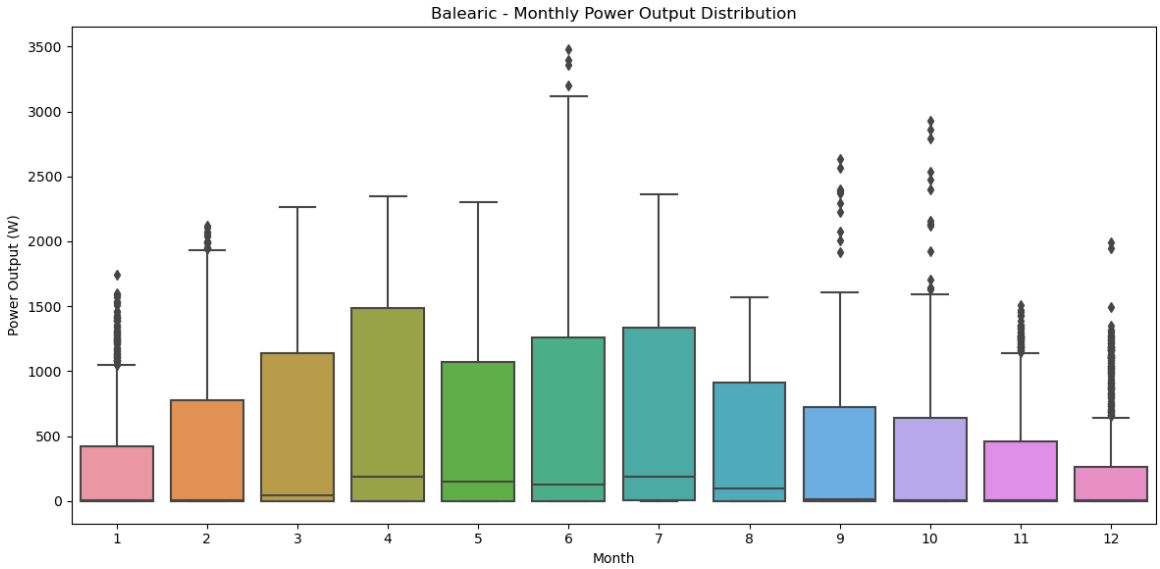


Figure 3-11 Monthly Power Output Distribution - Balearic

In the Balearic Islands, spring (33.8%) produces the largest share of energy,

followed by summer (31.8%). Together, these two seasons account for more than 65% of the total production, as a result of high solar radiation and long daylight hours during these seasons. Autumn contributes 18.2%, while winter represents the smallest share at 16.2%, corresponding to low sun radiation and short days.

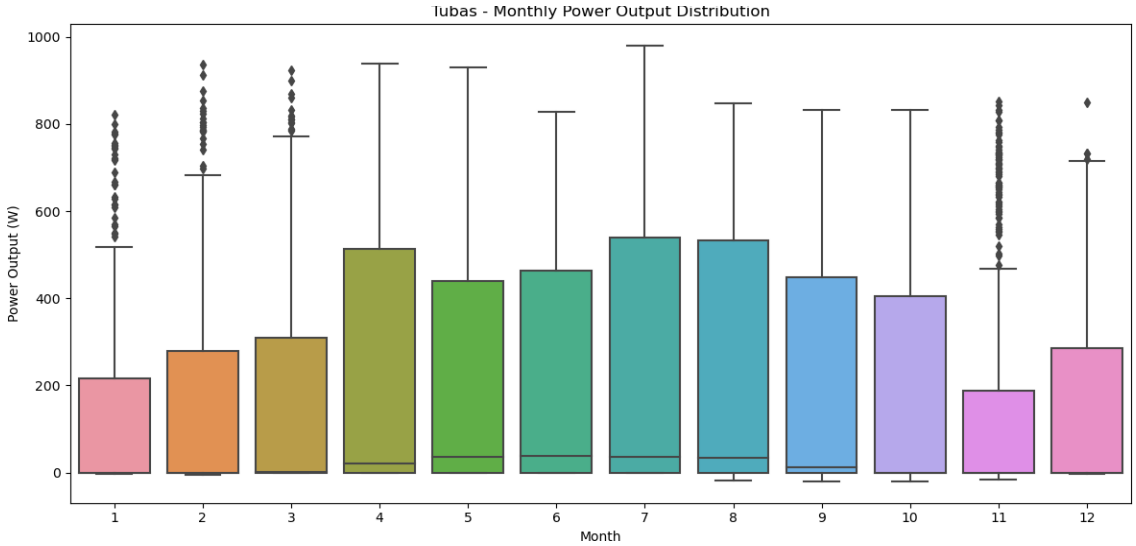


Figure 3-12 Monthly Power Output Distribution - Tubas

Interestingly, Tubas show a different seasonal pattern. Autumn dominates with 34.4% of annual production, followed by summer (31.0%). Spring contributes 20.2% , while winter remains the least contributor at 14.4%. The higher fall contribution in Tubas may reflect clearer sky or more suitable sunshine, as high temperatures (as in summer) negatively impact solar panel efficiency.

The violin plot in Figure 3-14 shows a seasonal comparison of energy production distributions in both regions, providing a view of intensity patterns.

In winter, shows very limited production in both regions, with densities closely concentrated near zero, although the Balearic Islands maintain a slight edge in the upper distribution.

In summer, both regions have the widest energy production distribution, but the Balearic Islands' production intensity extends much higher, indicating more frequent high-productivity days. In Tubas, maintain a more compact distribution, indicating lower but more consistent production.

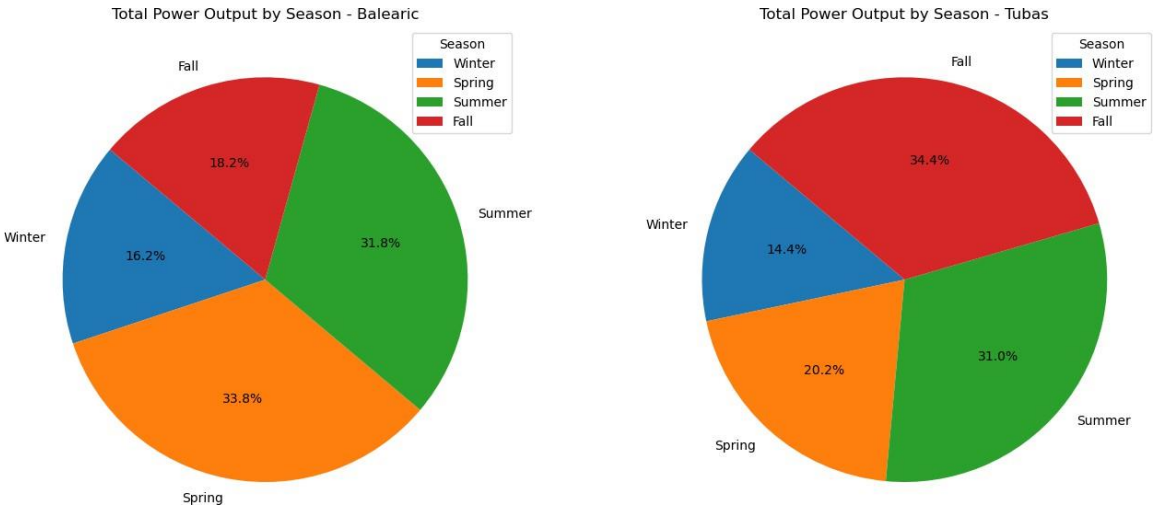


Figure 3-13 Total Power Output by Season - Balearic and Tubas

During spring, the Balearic Islands again exhibit widespread with higher averages and extended intensity tails. Tubas has production behavior similar to summer, but with slightly lower upper values, indicating less radiation or shorter daylight hours.

In autumn, the distribution in Tubas shifts upward, consistent with previous pie chart results indicating that autumn contributes the most energy to this region. The energy distribution in the Balearic Islands remains moderate in autumn, but lower than in spring and summer.

Overall, the violin plots reinforce the idea that the Balearic Islands experience stronger seasonal variability, while Tubas offer more stable and seasonally balanced production. This visualization highlights how the interaction between seasonal radiation and system efficiency affects energy production in different climates.

The weekday power output plots in Figure 3-15 and Figure 3-16 for Balearic and Tubas show no significant variation over the days of the week, indicating that solar generation is largely independent of human activity related to the weekdays.

The plots in Figure 3-17 and Figure 3-18 show total power output aggregated by each day of the month for both Balearic and Tubas. In both regions, the distributions appear relatively uniform across most days, but a notable drop is observed on the 31st day, which is likely because not all months contain 31 days, resulting in fewer data points and reduced output. This suggests that there is no relation between the amount of power production and the day of the month.

To gain insight into periods of inactivity or potential power generation, the power

output data was classified into two categories: zero and nonzero. The results in Figure 3-19 and Figure 3-20 reveal notable but distinct patterns. In the Balearic dataset, a nonzero energy output is observed in approximately 63% of recorded hours, with zero output accounting for the remaining 37%.

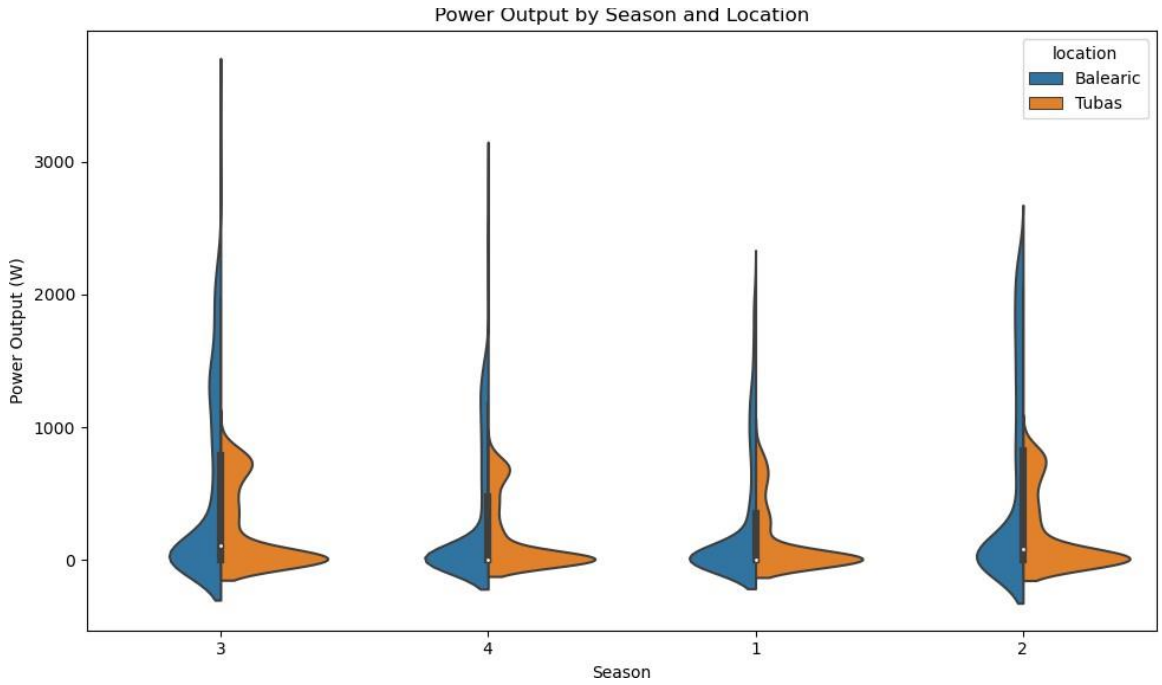


Figure 3-14 Seasonal Power Output Distributions – Violin Plot Comparative Analysis

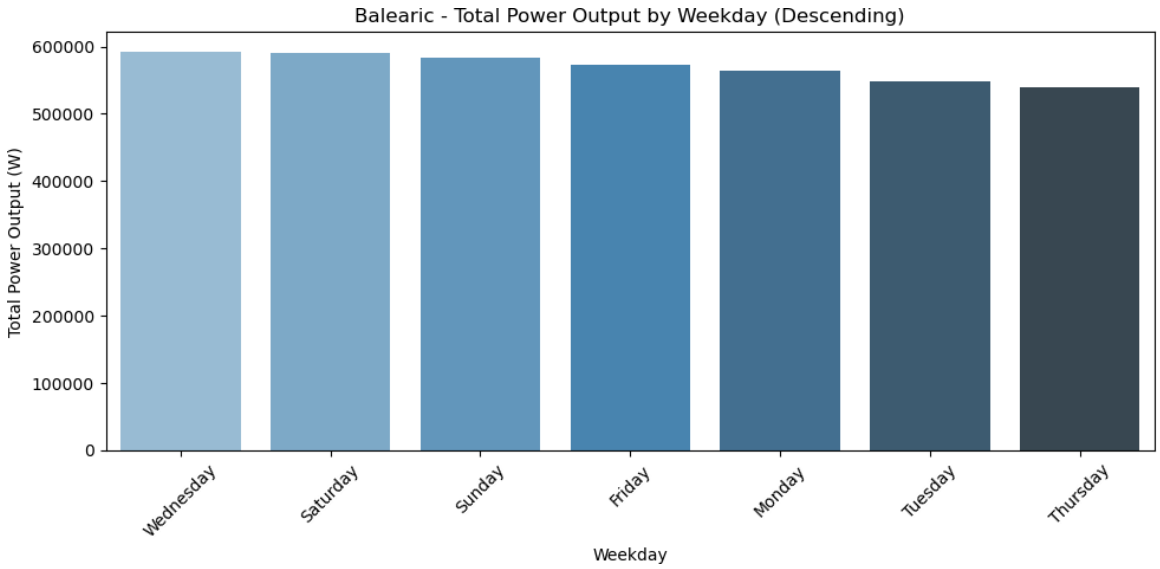


Figure 3-15 Total Power Output by Weekday - Balearic

In contrast, the Tubas dataset exhibits a different distribution: nonzero output accounts for approximately 53%, while zero output occurs in approximately 47% of

observations. This is attributed to seasonal data gaps or local operational issues.

These results suggest that the Balearic system offers greater continuity in solar energy production, which is useful for stable forecasting. Meanwhile, this distribution suggests that both groups reflect realistic solar-generation behavior. The presence of substantially zero output values suggests that any predictive modeling must explicitly address these periods.

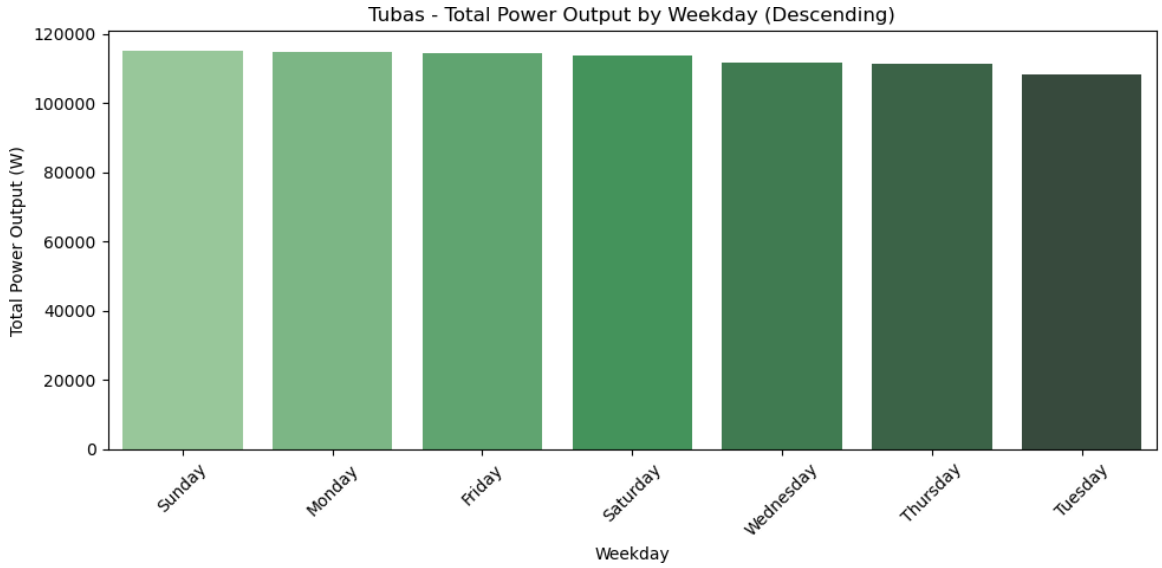


Figure 3-16 Total Power Output by Weekday - Tubas

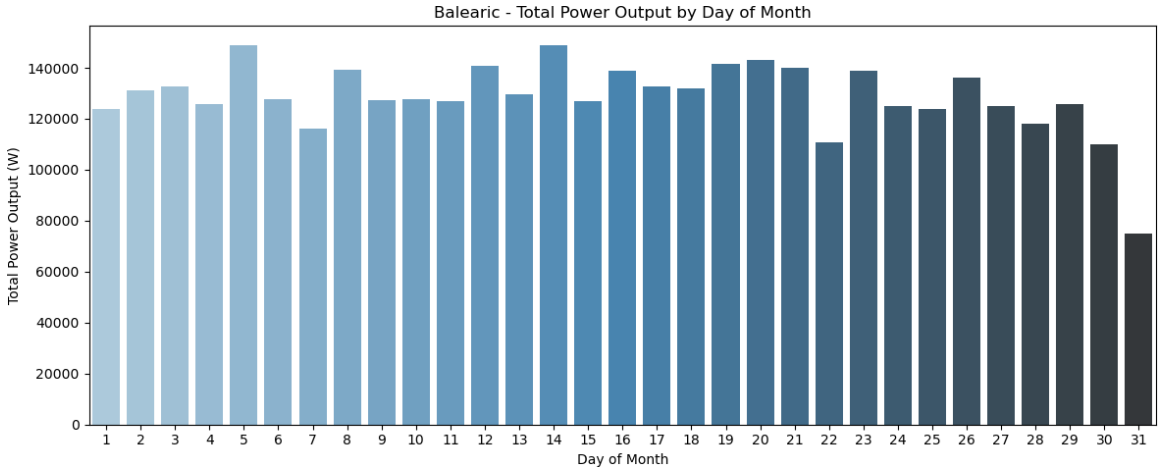


Figure 3-17 Total Power Output by Day of Month - Balearic

In addition, the similarity in the distribution of output categories between Tubas and Balearic suggests that differences in environmental impact or system scale are more

pronounced for continuous output values than for simple generation conditions. Therefore, further analysis should focus on how radiation and temperature influence non-zero values.

The exploratory data analysis provided a comprehensive understanding of the behavior of solar power generation in both the Balearic and Tubas regions. Through descriptive statistics and a wide range of visualizations, the analysis revealed clear differences in weather patterns and seasonal production. In particular, the Balearic region exhibited power production associated with stronger seasonal solar patterns, while Tubas maintained more stable power generation. These insights reveal the importance of regional characteristics in solar power performance and highlight the need to design ML models to consider environmental variability. The results of environmental data analysis provide an important foundation for feature selection and support the comparative evaluation of predictive models in the next phase of this study.

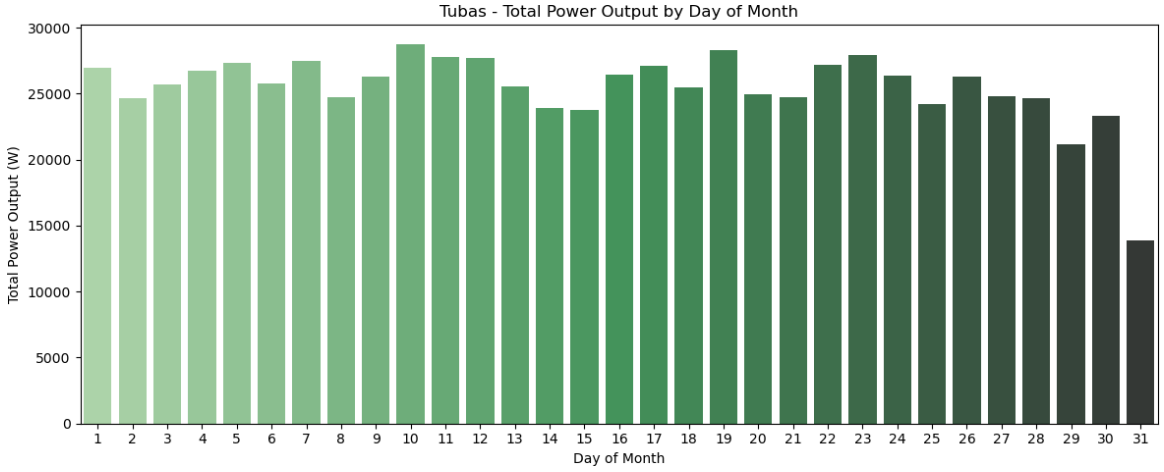


Figure 3-18 Total Power Output by Day of Month – Tubas

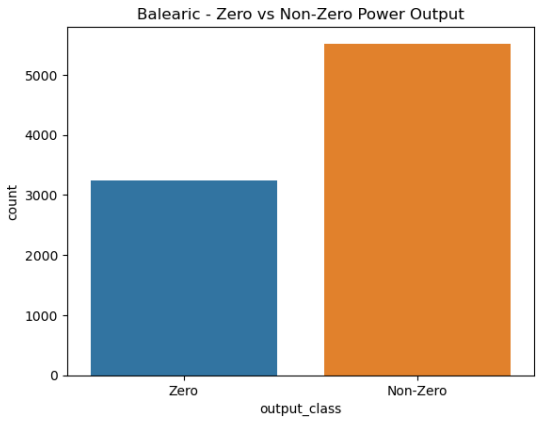


Figure 3-19 Zero vs Non-Zero Power Output - Balearic

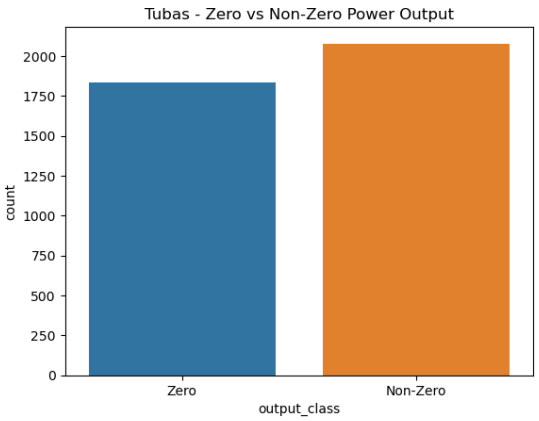


Figure 3-20 Zero vs Non-Zero Power Output -Tubas

3.6 Data Pre-processing

Data preprocessing is an essential component before moving to the model training phase.

Poor data quality not only impacts model performance but can also hinder the ML-based model-building process. Therefore, properly applying data preparation preprocessing techniques ensures the best results using clean and standardized data. Since data quality significantly impacts the effectiveness of predictive models. Referring to Figure 3-1 presented in the first section of this chapter, the steps involved in the fourth stage of data preprocessing are illustrated in this section and illustrated in Figure 3-21

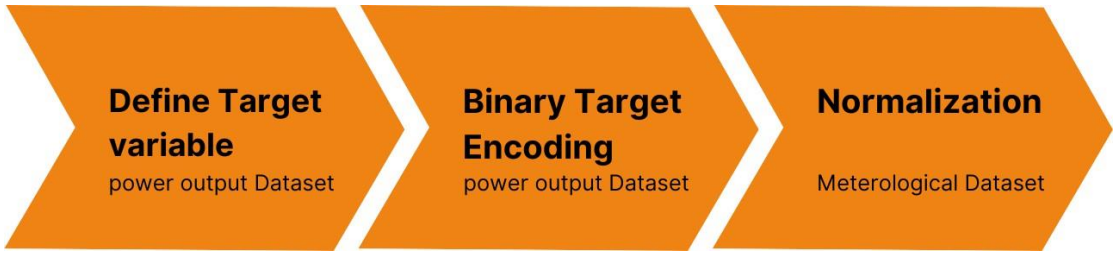


Figure 3-21 The Preprocessing Methods Applied on the Dataset

3.6.1 Determine Target Feature

The first and most critical step in the data preprocessing process is identifying the target variable that ML models will predict. In the context of solar energy forecasting, the

chosen target variable is "power output," which represents the actual amount of electrical energy generated by solar PV systems over time. This continuous property is the primary prediction target in all subsequent modeling tasks.

Due to the nature of solar energy systems, the "power output" variable exhibits an inflated distribution at zero, with long periods of zero generation, especially during the night or in low irradiance or cloud conditions. To accommodate this property and improve model interpretability and performance, a secondary binary target property called "power binary" was derived from the primary target. This property is defined to assume a value of 1 when "power output" is greater than zero, and 0 when there is no power generation. This transformation enables a two-stage modeling approach, where the classification stage predicts the presence of power generation, and the regression stage estimates the actual amount when generation is non-zero. The process of defining this binary feature not only aligns the data with the modeling objectives but also guides other preprocessing decisions, such as addressing class imbalances (Graphite Note, 2024).

$$\text{power\_status} = \begin{cases} 1 & \text{if power\_output} > 0 \\ 0 & \text{otherwise} \end{cases}$$

3.6.2 Feature Scaling

In multivariate ML tasks, particularly those involving regression or neural network models, the scale and distribution of input features significantly influence model and predictive performance (Ahsan et al., 2021). To ensure that all continuous input variables contribute proportionally to the learning process, feature scaling was applied using Z-score normalization. This technique transforms each feature so that its mean is zero and its standard deviation is one, thus centering the data and standardizing its variance:

$$z = \frac{x - \mu}{\sigma}$$

Where x is the original value of a feature, μ is the mean of the feature, and σ is the standard deviation. This transformation ensures that each scaled feature z has zero mean and unit variance.

This standardization was applied to all continuous meteorological variables used as

input features, including solar radiation, temperature, humidity, pressure, and wind speed. These variables differ in their original scales (e.g., irradiance in W/m² versus humidity in kPa), and without normalization, models that rely on gradient descent optimization distance metrics may be biased toward higher or lower values, leading to less model performance.

Furthermore, normalized features improve numerical adaptation to the optimization environment, and enhance the model's ability to generalize across data distributions (Rahmad Ramadhan & Anne Mudya, 2024). Additionally, is less sensitive to outliers or extreme values, making it more suitable for datasets with natural variability (Rahmad Ramadhan & Anne Mudya, 2024), such as meteorological data.

3.7 Machine Learning Models

This section describes the workflow and the ML models implemented for predicting solar power output based on meteorological and historical energy data under different meteorological conditions. The models were selected based on their ability to handle time-series data and their proven track record in regression tasks.

The techniques utilized in this phase and their results are highly dependent on the preprocessing steps presented in the previous section.

Figure 3-22 highlights the core phases of the entire process for constructing the proposed models from a ML perspective. This includes the essential data preprocessing phase before model training, outlined in the previous Section. The workflow is initiated by loading the datasets of the Balearic Islands and Tubas, each undergoing the same preprocessing techniques. This phase involves three critical steps: defining and encoding the target variable and normalizing input features.

Following preprocessing, five distinct models are constructed: LR, RF, XGBoost, Bi-LSTM, and the hybrid model CNN-LSTM. Each model is tailored with specific preprocessing adjustments. All models (except RF and XGBoost) apply feature selection using the F-score method.

Additionally, the Bi-LSTM and CNN-LSTM models include a reshaping step, transforming data into a three-dimensional format for sequential learning.

A key preprocessing technique is the treatment of zero-inflated data, i.e., instances where solar power output equals zero, typically during nighttime. These instances were removed from tree-based models (RF, LR, XGBoost) to focus on active generation periods, but retained in deep sequence models (CNN-LSTM and Bi-LSTM) to preserve temporal continuity.-

Subsequently, the modeling process follows a consistent structure: an 80:20 train-test split is performed, the models are trained on the training subset, predictions are made on the unseen test data, and the outputs are evaluated using standard regression metrics including MAE, RMSE, and R^2. This workflow ensures a comprehensive and fair performance comparison across traditional ML and deep learning approaches, enabling a clear evaluation of solar forecasting capability across the different meteorological conditions.

3.7.1 Used Models

The study employs a combination of classical ML and deep learning techniques to predict PV power generation with high accuracy. The aim is to assess the forecasting of each model type when applied to two different meteorological conditions (Balearic Islands and Tubas).

The following models were selected based on their theoretical suitability for regression tasks, time-series forecasting capability, and proven performance in related literature:

- Linear Regression: A fundamental statistical model used here as a baseline. It assumes a linear relationship between input meteorological variables and solar energy output. Despite its simplicity, this model provides a valuable reference for evaluating the added value of more complex methods (Chicco et al., 2014).

- Random Forest: A powerful ensemble model based on decision trees. RF is useful in handling high-dimensional datasets and nonlinear relationships. It has a built-in feature importance selection that enables it to weigh the relevance of each predictor during training, which is crucial in datasets where only some meteorological variables are strongly correlated with output (Breiman, 2001). Also, a tuned RF model was employed, where hyperparameters were optimized using GridSearchCV with 36 different model

configurations being evaluated using 3-fold cross-validation to improve forecasting accuracy.

- Extreme Gradient Boosting: An advanced ensemble model that builds trees sequentially to minimize residual errors with the same property of the RF model in feature selection. XGBoost enhances model generalization (Chen & Guestrin, 2016). It is important to deal with variability across time periods and capture complex dependencies in the data.

Convolutional Neural Network - Long Short-Term Memory: In time-series solar forecasting, LSTM-based models have shown improved accuracy due to their memory of sequential dependencies. The CNN-LSTM hybrid further enhances this by allowing local feature extraction before sequence learning, which is ideal for meteorological data with multiple interdependent variables (Gao et al., 2019). This hybrid architecture leverages the ability of CNNs to extract spatial features from sequences and LSTM networks to model temporal dependencies. When applied to sequence input data, it allows the model to learn complex patterns in the sequence of weather features and radiation over time (Shi et al., 2015).

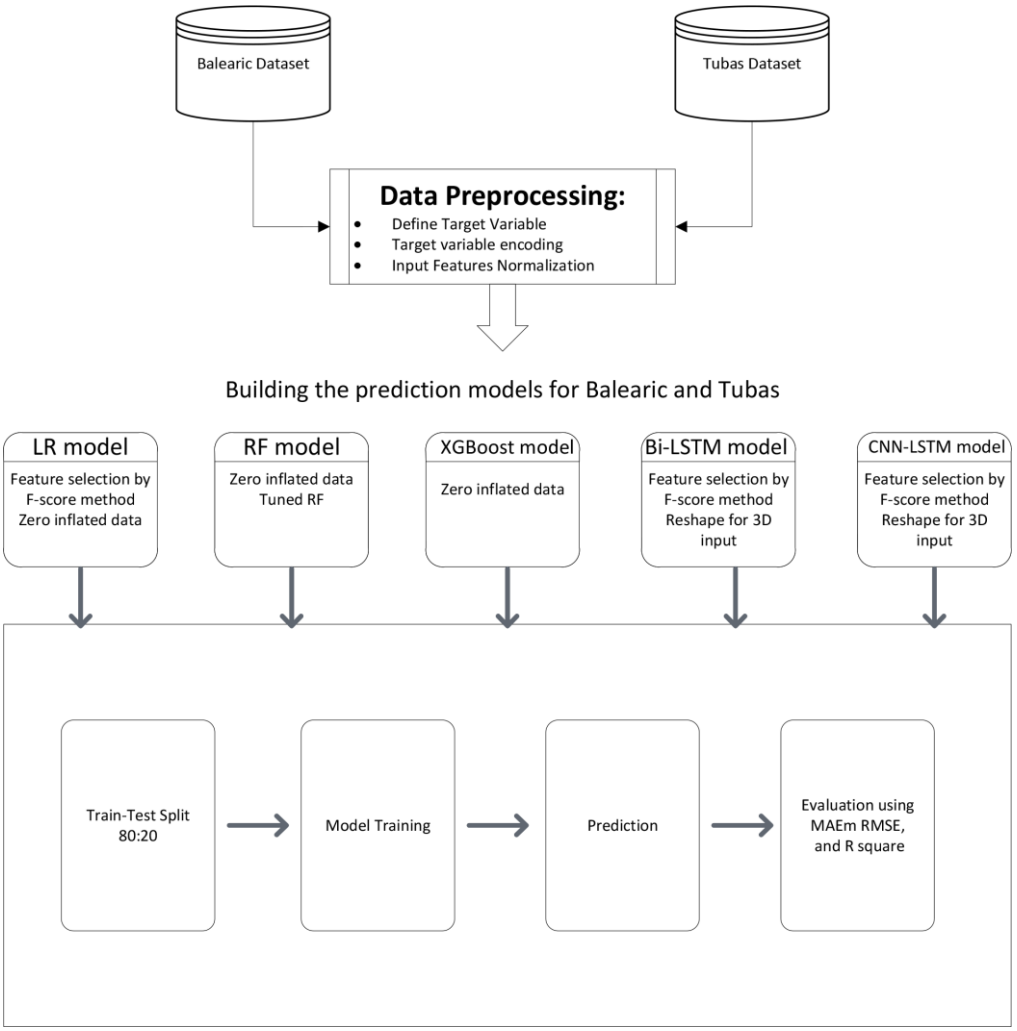


Figure 3-22 Modeling Workflow for Solar Power Forecasting

• Bi-LSTM: An advanced RNN model that processes the sequence data in both forward and backward directions. This presents a better understanding, which is especially relevant for datasets where predicting the target variable (solar output) depends on both prior and succeeding meteorological conditions (Graves & Schmidhuber, 2005).

Each model was trained independently on the Balearic and Tubas datasets, with preprocessing tailored to the needs of each algorithm.

3.7.2 Model Configuration

Table 3-10 highlights the parameters, hyper-tuning, settings, and the implementation method of each ML model employed for predicting the solar PV power under different meteorological conditions. Tree-based models like RF, LR, and XGBoost

were configured using standard parameters commonly used in the literature, while deep learning models like hybrid CNN-LSTM and Bi-LSTM were implemented using Keras API considering training parameters such as batch size, number of epochs, optimizer, and loss function to suit the temporal nature of the forecasting task.

Deep learning models used the MSE loss function and Adam optimizer for gradient descent. Time-series data were reshaped into 3D tensors [samples, timesteps, features] and normalized using ~~F~~Z-score scaling to improve training stability (Mellit et al., 2009).

Special consideration was given to zero-inflated values — instances where solar output equals zero, often during nighttime. These values were removed from traditional models (LR, RF, and XGBoost) to focus training on the generation period (Mellit & Kalogirou, 2008). Conversely, zero values were retained in sequence-based deep learning models (CNN-LSTM and Bi-LSTM) to benefit from full sequence context and enable the models to learn daily generation cycles (Yona et al., 2008).

Table 3-10 Configuration of Forecasting Models

Model	**Configuration Details.**	**Implemented Using**
Linear Regression	• Default settings • No hyperparameter tuning required	Scikit-learn (LinearRegression)
RF	• n_estimators = 300 • max_depth = 20 • min_samples_split = 2 • min_samples_leaf = 1	Scikit-learn (RandomForestRegressor)
Tuned RF	• n_estimators: [50, 100, 200] • max_depth: [5, 10, 20, None] • min_samples_split: [2, 5, 10]	GridSearchCV (3-fold CV)
XGBoost	• n_estimators = 500 • max_depth = 4 • learning_rate = 0.1 • Tree booster with regularization	XGBoost Python API (XGBRegressor)
CNN-LSTM	• Conv1D filters: 32, kernel size: 3 • MaxPooling layer • LSTM units: 50 • epochs = 20, batch size: 64 • Optimizer: Adam, loss: MSE • Input shape: [samples, timesteps, features]	Keras (TensorFlow backend)
Bi-LSTM	• Bidirectional LSTM with 100 units • epochs = 20, batch size: 64 • Optimizer: Adam, loss: MSE • Input shape: [samples, timesteps, features]	Keras (TensorFlow backend)

3.7.3 Model Performance Evaluation Criteria

To ensure a comprehensive evaluation of the predictive models, several statistical performance metrics were tested. These metrics provide quantitative insight into how closely the predicted solar power values match the actual observed values over the test set. As the target variable is continuous, regression-based evaluation criteria are most appropriate (Chai & Draxler, 2014).

The selected metrics, MAE, RMSE, and the Coefficient of Determination (R^2) offer unique perspectives on model performance. By analyzing multiple metrics, A comprehensive understanding of the strengths and weaknesses of the model can be obtained (Hyndman & Koehler, 2006)(Mellit & Kalogirou, 2008)

3.7.3.1 MAE

The MAE measures the average absolute difference between the predicted values and the actual target values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

y_i represents the actual observed value, $\hat{y}$ is the model's predicted value, and n is the number of observations. MAE is particularly useful for applications where instances are retained for tree-based models (RF and XGBoost) but not used for sequence-based models, where large deviations are not necessarily more harmful than smaller ones, such as during routine day-ahead solar predictions (Chicco et al., 2001).

3.7.3.2 RMSE

RMSE is one of the most frequently used metrics in solar energy forecasting. It squares the error before averaging, thus penalizing larger errors more heavily (Mellit & Kalogirou, 2008) .

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

The squaring operation amplifies the impact of high-error predictions, making RMSE more sensitive to outliers compared to MAE. This property is particularly important in power systems, where overestimation or underestimation during peak hours can significantly affect grid stability and operational planning (Mellit & Kalogirou, 2008)

3.7.3.3 Coefficient of Determination (R^2 Score)

The R^2 score is a number between 0 and 1 that represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

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

Where $\bar{y}$ is the mean of the observed values. An R^2 value of 1 indicates a perfect fit, while values closer to 0 imply poor predictive capability. This measure is particularly useful for comparing models trained on the same dataset, as it provides a normalized assessment regardless of the magnitude of the predictions (J. Zhang et al., 2018).

While each metric individually provides a specific perspective, the three metrics were calculated across test samples to evaluate multiple models under different meteorological conditions.

3.8 Summary

This chapter presented the proposed methodology used in constructing the prediction models for solar PV power generated under different climate conditions. The next chapter will present the results of implementing the proposed models, along with a discussion of the obtained results, to obtain the model with the highest performance among the examined models utilizing different ML algorithms, and make a comparison between the effects of conditions varying in the efficiency of each model.

Chapter 4 Results

4.1 Introduction

This chapter presents the experimental results of predicting solar PV power generation under various climate conditions, specifically in the Balearic Islands and Tubas regions. The performance metrics presented in the previous chapter were used for evaluating the performance of the LR, RM, XGBoost, Bi-LSTM, and CNN-LSTM models.

4.2 Performance Evaluation Metrics Summary

Table 4-1 summarizes the metric performance of all models on both regional datasets, including R^2, MAE, and RMSE. The results are organized to reflect the structure of the study's methodology, comparing the performance of each forecasting model across the two regional datasets to observe the effect of climate variability.

Table 4-1 Performance Evaluation of Models for Solar Power Forecasting

Model	Region	RMSE	MAE	R^2
Linear Regression	Tubas	195.35	150.29	0.5202
	Balearic	416.23	344.86	0.6192
Random Forest	Tubas	142.25	91.97	0.7456
	Balearic	199.57	117.37	0.9125
Tuned RF	Tubas	141.72	93.55	0.7474
	Balearic	198.99	117.51	0.9130
XGBoost	Tubas	141.79	95.02	0.7472
	Balearic	178.89	110.25	0.9297
CNN-LSTM	Tubas	113.12	65.25	0.8266
	Balearic	122.41	72.77	0.9646
Bi-LSTM	Tubas	117.35	70.66	0.8134
	Balearic	98.84	62.88	0.9769

4.3 Comparative Analysis of Models

In this section, a comparative analysis of the performance of models based on their type.

4.3.1 Tubas Dataset Results

The Tubas dataset represents a climate with high variability in irradiance and temperature. This region faces a greater challenge for forecasting models due to the frequent fluctuations in meteorological features.

Linear Regression reports the weakest performance among all models, with an R^2 value of 0.52 and the highest RMSE (195.35). This is expected, as linear models are limited in capturing nonlinear relationships and temporal patterns in solar PV energy data (Murphy, 2012).

The tree-based models like (RF, Tuned RF, and XGBoost) showed significant improvements, all reaching R^2 values close to 0.75. These models benefit from the ability to model complex and non-linear interactions between input features. The tuning of hyperparameters in Random Forest (e.g., number of trees and tree depth) provided a slight performance improvement, with the best configurations found (n_estimators = 100, max_depth = 10, min_samples_split = 2)

The deep learning models Bi-LSTM and CNN-LSTM provided the best results. CNN-LSTM model achieved the highest R^2 (0.8266) and the lowest RMSE (113.12), indicating strong learning of temporal patterns, while Bi-LSTM followed closely with an R^2 of 0.8134**.**

For the MAE metric, which highlights how well a model performs across various conditions (Ahmed, 2023), the highest MAE was recorded by the Linear Regression model at 150.29, indicating poor performance and large deviations from the actual solar power output. Tree-based models, including RF, Tuned RF, and XGBoost, showed improved MAE values in the range of 91.97 to 95.02. The CNN-LSTM model and Bi-LSTM model achieved the lowest MAE at 65.25 and 70.66, respectively. These results demonstrate that deep learning models outperformed others in minimizing average

prediction errors and are more capable of reducing forecast errors in regions with high meteorological variability.

4.3.2 Balearic Islands Dataset Results

The Balearic Islands dataset is characterized by a Mediterranean climate with consistent sunlight and stable meteorological conditions.

Linear Regression, while still the weakest, performed better in Balearic than in Tubas, with an R^2 of 0.61. This improvement reflects the reduced variability in the Balearic data, which benefits simpler models.

The RF and Tuned RF models performed well, achieving R^2 values of 0.9125 and 0.9130, respectively. The slight improvement from tuning indicates that the default ensemble structure was already suitable for the data, with the best configurations found (n_estimators = 200, max_depth = 20, min_samples_split = 2)

XGBoost, known for its regularization and for preventing overfitting (Chen & Guestrin, 2016), outperformed both versions of Random Forest, with an R^2 of 0.93 and an RMSE of 178.89. This confirms XGBoost's ability to efficiently capture patterns in clean, structured datasets.

The best performance was again observed in deep learning models. Bi-LSTM achieved an R^2 of 0.9769 and the lowest RMSE (98.84), demonstrating excellent prediction. CNN-LSTM also performed strongly with an R^2 of 0.9646. These models successfully leveraged temporal dependencies and outperformed traditional models by a considerable margin.

Balearic Islands, and as R^2 and RMSE, all models performed better in terms of MAE, reflecting the relative ease of prediction in a consistent solar environment. The Linear Regression model also performed the weakest**,** but showed significant improvement compared to Tubas, with a lower MAE of 344.86. Tree-based models (RF, tuned RF, and XGBoost) showed important accuracy improvements, with MAE values of 117.37, 117.51, and 110.25, respectively. The Bi-LSTM model achieved the lowest MAE of 62.88, closely followed by CNN-LSTM at 72.77. This underscores their effectiveness in capturing both the short-term variability and long-term seasonal patterns inherent in solar energy generation.

In general, as observed in Table 4.1, models perform better in Balearic due to lower meteorological variability. Deep learning models consistently outperform tree-based and linear models.

4.4 Visual and Graphical Findings

This section introduces the graphical representations of the model performance results.

4.4.1 Residual Distribution

The residual error distributions provide key insights into forecasting consistency and the influence of climatic variability (Hyndman & Athanasopoulos, 2018). Figure 4-1 shows the residual across all models and both regions. In the Balearic dataset, most models demonstrate residuals that are sharply peaked and symmetrically centered on zero, particularly for high-performing models like Bi-LSTM, CNN-LSTM, and XGBoost. The residual plot for XGBoost in Balearic is tall and narrow, matching its strong performance ($R^2 = 0.9297$, MAE = 110.25), indicating a high concentration of low-error predictions. Random Forest and Tuned RF show similarly compact residual patterns in Balearic, further reinforcing their accuracy under stable weather conditions. Conversely, Linear Regression exhibits highly skewed and widely spread residuals in Balearic, reflecting its high RMSE (416.23) and weak adaptability to nonlinearity. In the Tubas dataset, where weather patterns are more volatile, residuals for all models widen significantly. The Bi-LSTM and CNN-LSTM models still maintain reasonably narrow distributions, but with increased variance, consistent with their R^2 scores of 0.8134 and 0.8266, respectively. The XGBoost residuals in Tubas, while broader than in Balearic, remain relatively symmetric and centralized, reflecting its resilience ($R^2 = 0.7472$, MAE = 95.02) despite challenging inputs. The residuals for RF and Linear Regression in Tubas are more dispersed with flatter peaks, indicating weaker predictive certainty. Overall, the residual plots confirm that forecasting accuracy and error compactness improve significantly under stable climatic conditions, and that models like XGBoost and deep learning architectures retain their robustness better than traditional linear or ensemble methods.

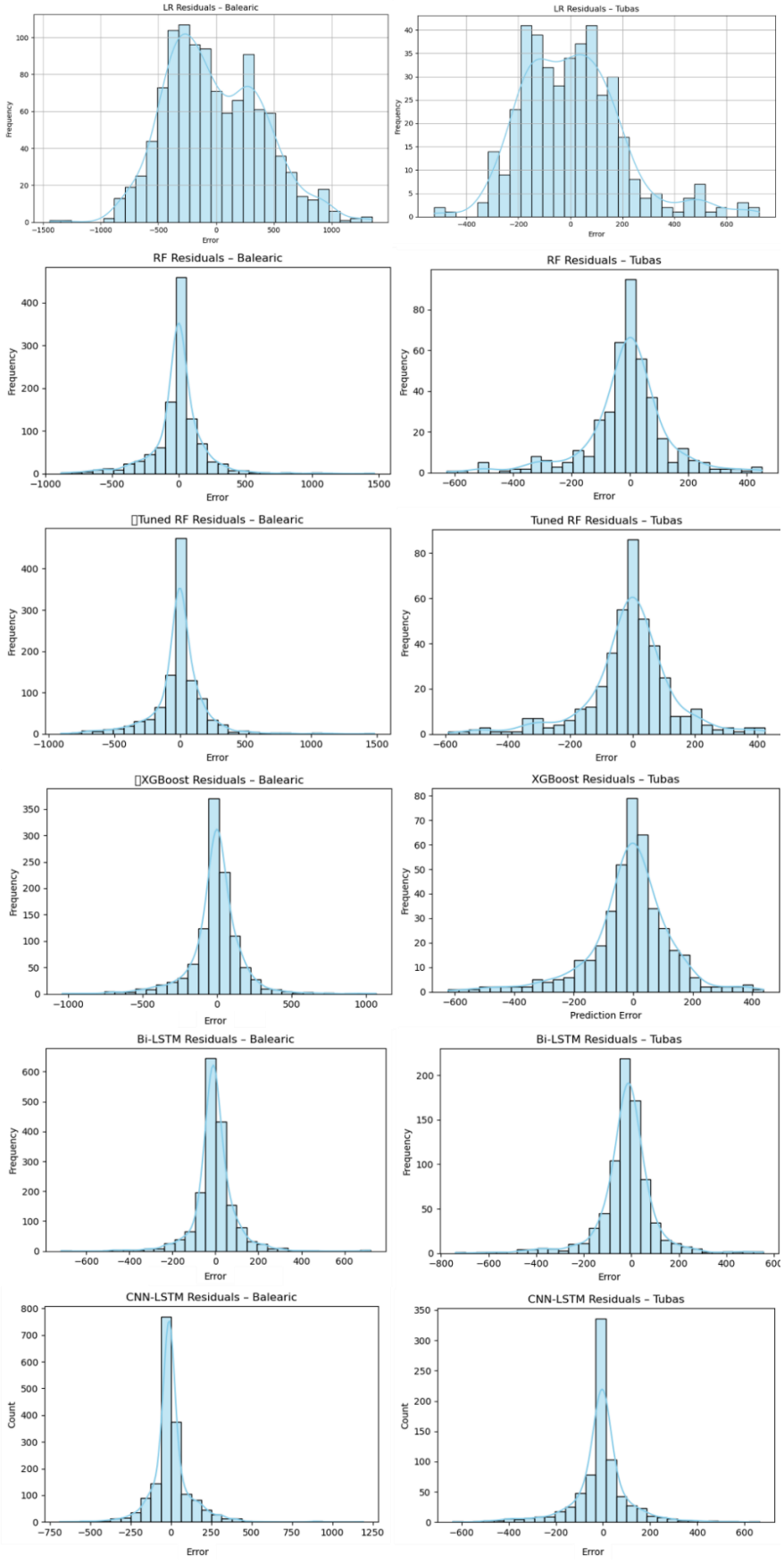


Figure 4-1 Residual across All Models and Both Regions

4.4.2 Feature importance Analysis

The feature importance plots for the Tubas and Balearic regions reveal substantial differences in how predictive models prioritize meteorological features (Ewald et al., 2024), highlighting the influence of local climatic variability on forecasting.

In the Balearic Islands dataset, Figure 4-2 show that solar radiation significantly dominates the model's decision-making process, with the F-score exceeding all other features. Minor contributions come from temperature, month, and wind speed. This suggests that solar power production in the Balearic Islands is primarily driven by radiation patterns, which remain relatively stable due to the region's climate. The minimal dependence on other features such as humidity, pressure, or time-based features reflects the consistent and predictable nature of solar exposure in this region.

In contrast, the Tubas dataset Figure 4-3 presents a more distributed feature importance profile. While solar radiation remains the most important predictor, other features such as temperature, humidity, and wind speed contribute significantly to the model. Temperature, in particular, ranks second with a relatively high F-score, indicating that it plays a crucial role in power generation variability, likely due to its impact on PV efficiency and cloud cover behavior. Furthermore, the presence of moderate humidity and wind speeds highlights the impact of instability in the climate in Tubas. The more balanced distribution of feature importance reflects the complexity and significant variability in this region, requiring the model to rely on a broader range of environmental indicators to obtain accurate predictions.

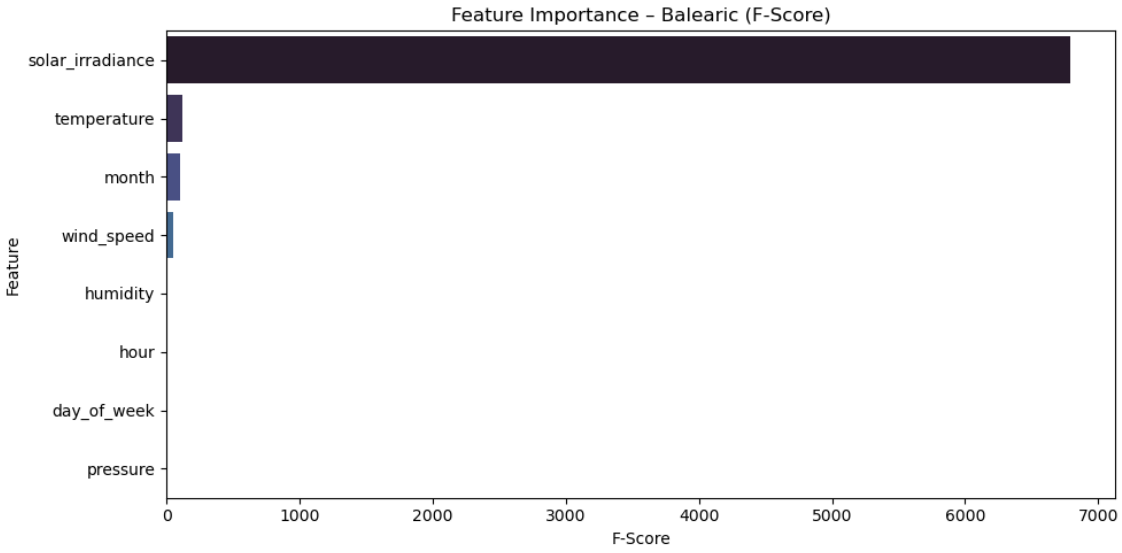


Figure 4-2 Feature Importance by F-Score – Balearic

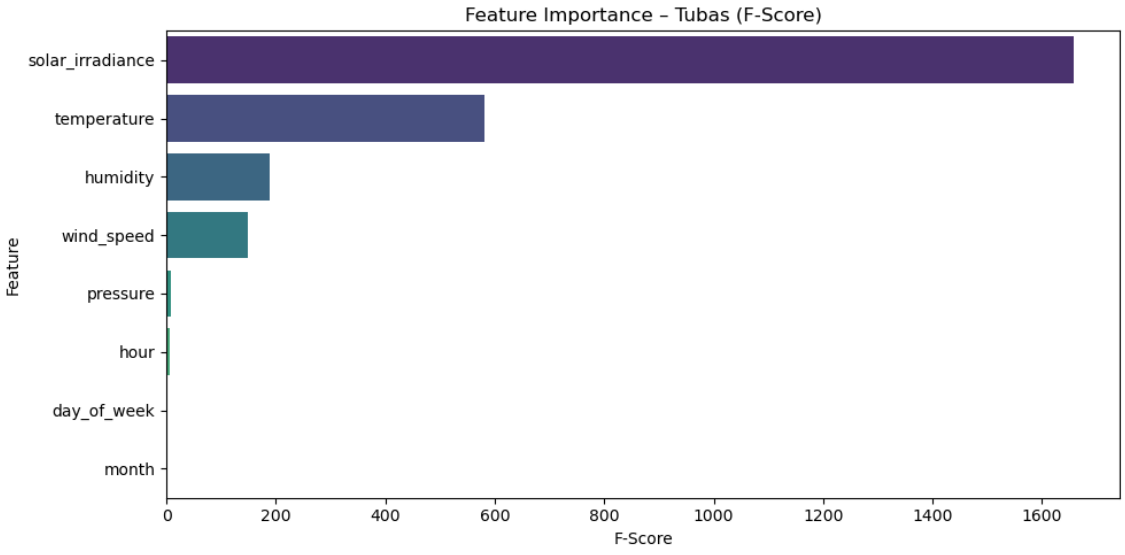


Figure 4-3 Feature Importance by F-Score – Tubas

4.4.3 Actual vs. Predicted Power Output

To complement the quantitative performance metrics, visualizations of actual versus predicted solar power output are presented for each model in both the Balearic Islands and Tubas. The predictions are visualized through line plots for sequence models (CNN-LSTM, Bi-LSTM) and scatter plots for regression-based models (LR, RF, XGBoost).

- **Linear Regression:**

In the Balearic region, the scatter plot for Linear Regression in Figure 4-4 and Figure 4-5 shows a wide dispersion of points around the regression diagonal at both low and high output values. This visual pattern correlates with its low R^2 of 0.62 and the highest RMSE among all models (416.23), revealing poor model generalization even in a relatively stable climate. In Tubas, although the plot demonstrates a slightly more compact alignment, substantial deviation from the ideal line remains evident.

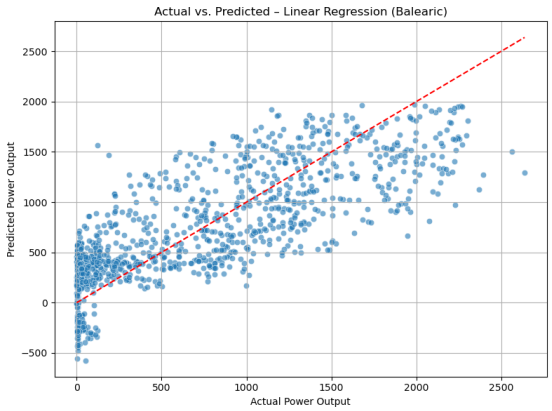


Figure 4-4 Predicted vs. Actual Solar Power Output Using Linear Regression – Balearic

Figure 4-5 Predicted vs. Actual Solar Power Output Using Linear Regression – Tubas

- **RF and Tuned RF**

The RF and Tuned RF models both demonstrated strong predictive capabilities in the Balearic dataset, as evidenced by closely clustered points around the diagonal in their scatter plots Figure 4-6. The tuned version reports little improvement without significantly altering performance in this consistent environment, as shown in Figure 4-7. In contrast, the performance in Tubas, the scatter plot in Figure 4-8, shows greater dispersion, particularly in mid-to-high output ranges. The Tuned model achieved an enhancement, with slightly denser predictions along the diagonal, Figure 4-9.

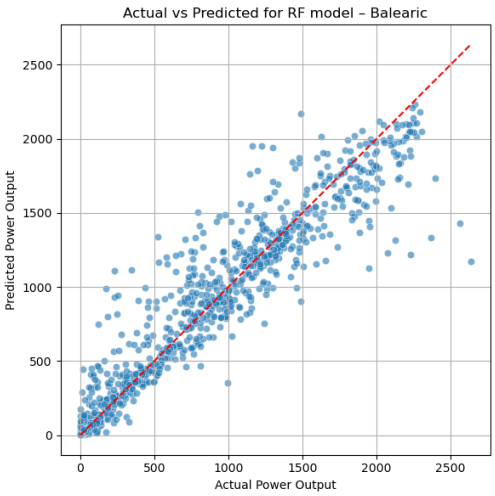


Figure 4-6 Predicted vs. Actual Solar Power Output Using RF – Balearic

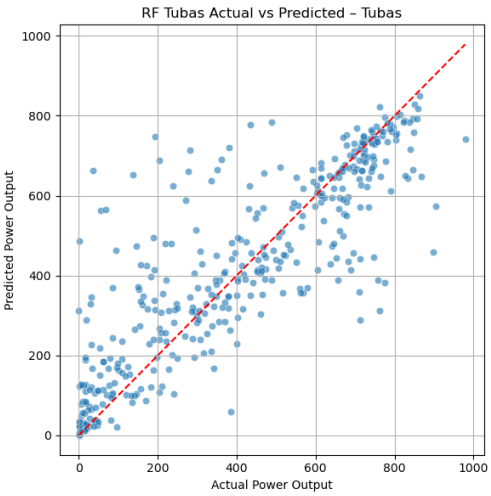


Figure 4-7 Predicted vs. Actual Solar Power Output Using RF – Tubas

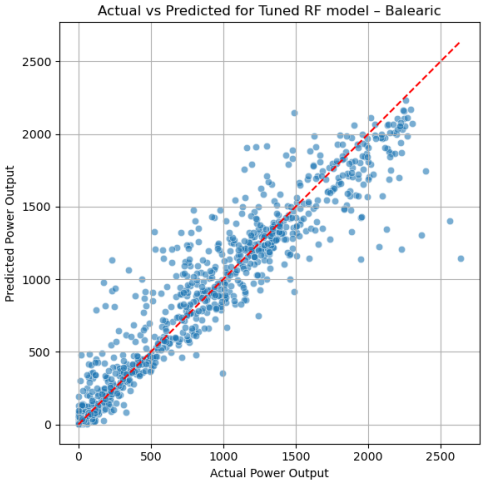


Figure 4-8 Predicted vs. Actual Solar Power Output Using Tuned RF – Balearic

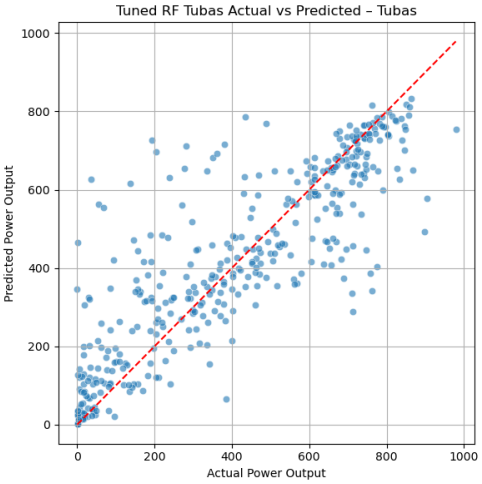


Figure 4-9 Predicted vs. Actual Solar Power Output Using Tuned RF – Tubas

Although Figures 4.6 to 4.9 illustrate the predicted vs. actual power output for RF and Tuned RF models in both regions, the difference in dispersion is not clearly visible. To provide a more objective comparison, the dispersion was quantified by computing the standard deviation of the prediction errors, defined as:

$$e_i = \hat{y}_i - y_i$$

Where $\hat{y}_i$ is the predicted solar power is output and y_i is the measured output.

as shown in Table 4-2, the computed standard deviations of the prediction errors

indicate very small differences between the RF and the tuned RF models for both the Balearic and Tubas datasets. This suggests that hyperparameter tuning resulted in only a limited reduction in error dispersion, which is consistent with the visual similarity observed in Figure 4-6 to Figure 4-9

Table 4-2 Standard deviation of prediction errors for RF and Tuned RF models in Balearic and Tubas datasets

Region	Model	Std. Deviation of Errors (MW)
Balearic	RF	199.41
Balearic	Tuned RF	198.80
Tubas	RF	141.60
Tubas	Tuned RF	141.25

- **XGBoost:**

The XGBoost model shows excellent scatter plot alignment in the Balearic dataset Figure 4-10, where predicted points closely follow the actual diagonal. This reflects its high R^2 (0.9297) and low MAE (110.25), confirming its strength in capturing structured, stable solar generation data. In Tubas, while the plot in Figure 4-11 shows a visible decrease in precision compared to Balearic, the predictions remain reasonably aligned.

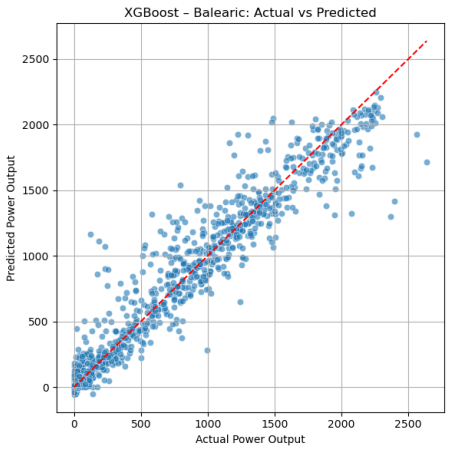


Figure 4-10 Predicted vs. Actual Solar Power Output Using XGBoost – Balearic

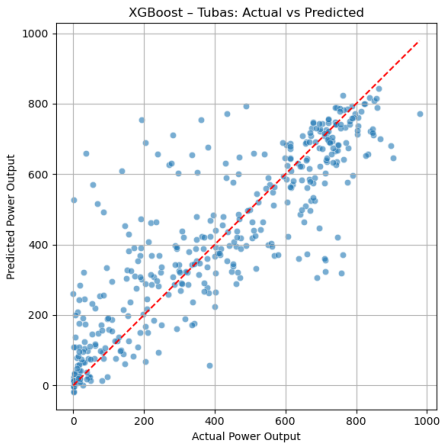


Figure 4-11 Predicted vs. Actual Solar Power Output Using XGBoost – Tubas

- **Bi-LSTM:**

The Bi-LSTM line plot for Balearic in Figure 4-12 reveals excellent predictive alignment with actual solar output over time. The model accurately tracks peaks and transitions, consistent with its superior R^2 of 0.9769 and the lowest RMSE (98.84) and MAE (62.88). In Tubas, although there is more variation between predicted and actual outputs due to irregular irradiance, the model still performs strongly Figure 4-13.

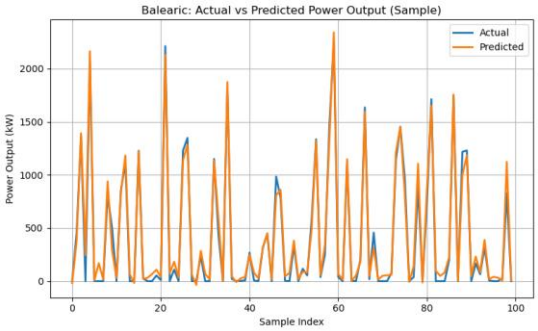


Figure 4-12 Time Series Comparison of Actual and Predicted Output Using Bi-LSTM – Balearic

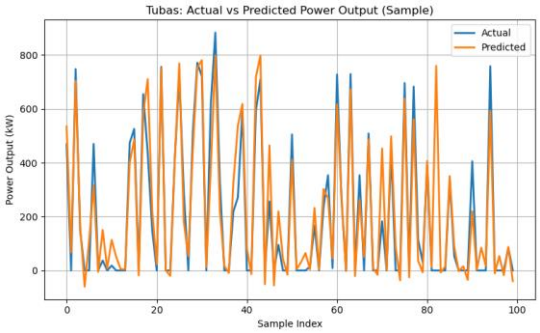


Figure 4-13 Time Series Comparison of Actual and Predicted Output Using Bi-LSTM – Tubas

- **CNN-LSTM**

The CNN-LSTM plot Figure 4-14 in Balearic demonstrates close adherence to actual values, successfully modeling both magnitude and timing of solar generation peaks. This is corroborated by its high R^2 of 0.9646 and low MAE of 72.77. In Tubas, the plot in Figure 4-15 shows slightly more pronounced deviations.

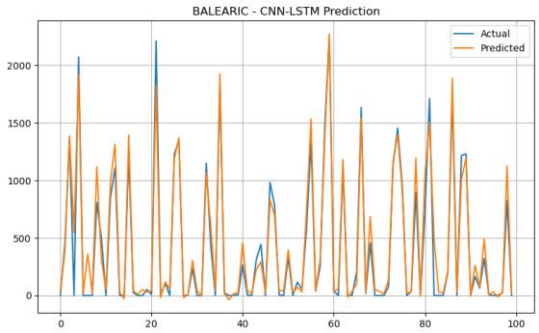


Figure 4-14 Time Series Comparison of Actual and Predicted Output Using CNN-LSTM – Balearic

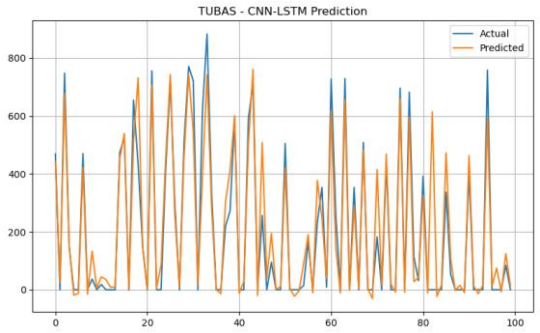


Figure 4-15 Time Series Comparison of Actual and Predicted Output Using CNN-LSTM – Tubas

4.5 Summary

This chapter presents the main findings and the experimental results of the constructed models in this study, the performance, and the accuracy achieved after implementing the solar PV power generation prediction models. The next chapter will discuss the results, present the conclusion inferred after conducting this study, the encountered obstacles and the mitigation methods, and the future works.

Chapter 5 Discussion

This chapter presents a comprehensive discussion of the study's findings, highlighting the impact of climate variability on solar energy forecasting, and provides key conclusions and recommendations for future work.

5.1 Discussion

The results obtained from this study highlight the comparative effectiveness of various ML models in forecasting solar PV power generation under different climatic conditions, specifically the variable climate of Tubas and the more stable climate of the Balearic Islands.

Model Performance Overview

In both regions, a deep learning model outperformed traditional and ensemble learning models. Bi-LSTM and CNN-LSTM, both capable of capturing temporal dependencies, demonstrated better performance across all metrics (R^2, RMSE, MAE). These models proved to be significantly effective in the Balearic dataset, with Bi-LSTM achieving an R^2 of 0.9769 and the lowest MAE (62.88), indicating their strength in leveraging stable climate conditions.

Impact of Climatic Variability

In the Balearic Islands, where solar irradiance is consistently high, all models showed good performance. This was particularly evident for all models, even simpler models like Linear Regression, which benefited from the reduced complexity of the data, achieving a higher R^2 (0.62) in Balearic than in Tubas (0.52). Such improvements highlight that when external conditions are more uniform, even basic statistical models can approximate solar power trends to a reasonable extent.

On the contrary, Tubas presents a more complex environment for forecasting due to the variability in irradiance and other climatic factors. These fluctuations increase in the nonlinearity in the dataset. This is reflected in the elevated RMSE and MAE values across all models in Tubas, particularly for Linear Regression, which recorded the highest RMSE (195.35) and MAE (150.29), revealing its weak adaptability to unstable

conditions.

Tree-based ensemble methods such as RF and XGBoost performed better than linear models in both regions, but their performance in the Balearic was significantly more accurate than in Tubas. In the Balearic Islands, RF and XGBoost achieved R^2 values of 0.91 and 0.92, respectively, reflecting their ability to exploit consistent data. However, under the more fluctuating Tubas conditions, these scores reduced to 0.74 and 0.74 for RF and XGBoost models, respectively. Although these models are capable of modeling non-linear interactions, they are still limited in capturing the temporal dependencies inherent in fluctuating meteorological patterns.

The deep learning models (Bi-LSTM and CNN-LSTM) offered a significant advantage in robustness and accuracy, particularly under the high-variability conditions of Tubas. Unlike traditional models, these models are explicitly designed to capture time series trends. The Bi-LSTM leverages both past and future context during training, allowing it to interpret temporal fluctuations more effectively. Similarly, CNN-LSTM combines spatial pattern extraction with temporal modeling, enabling it to learn trends and sequences. These advantages are shown in the Tubas dataset, where CNN-LSTM achieved an R^2 of 0.82 and the lowest RMSE (113.12), outperforming all other approaches. The Bi-LSTM model followed with an R^2 of 0.81 and RMSE of 117.35, indicating high predictive accuracy despite the region's complex climate.

In the Balearic dataset, as noted earlier, Bi-LSTM demonstrated optimal prediction with $R^2 = 0.9769$. However, its key advantage is not only its superior performance under stable conditions, but also its ability to maintain high accuracy under fluctuations. This robustness enhances its practical advantage in realistic forecasting models, where weather conditions may deviate significantly from typical averages.

The main reason of this performance divergence lies in how models utilize meteorological inputs under varying climatic conditions. In the Balearic Islands, solar radiation dominates feature importance, underscoring the simplicity of prediction driven by the main factor. Meanwhile, Tubas exhibits a more distributed feature dependency profile. Temperature, humidity, and wind speed contribute significantly to prediction, illustrating the complex interactions of variables in influencing PV power generation. These factors introduce additional sources of uncertainty, which only models with advanced learning can handle effectively.

Overall, this comparison reveals that forecasting solar power under variable climate conditions demands models that can handle non-linear relationships and treat temporal data. The improved results of Bi-LSTM and CNN-LSTM confirm that deep learning models are equipped to adapt to the temporal complexity of the weather fluctuations. They maintain a consistent advantage in stable and fluctuating environments, but their advantage becomes significant when climatic variability increases, reinforcing their suitability for deployment in regions where the weather environment is not stable.

Model Generalization

The comparative analysis between tuned and normal models (e.g., RF and Tuned RF) indicates that while hyperparameter optimization provides marginal gains, the choice of model architecture plays a more significant role in performance enhancement. For example, tuning improved RF performance slightly in Balearic and Tubas, but CNN-LSTM and Bi-LSTM report a significant better performance.

These findings confirm that deep learning models, especially hybrid models, are essential for real-time PV forecasting systems operating in diverse climates.

Visualization Insights

Residual distribution analysis further validated quantitative results. In the Balearic dataset, the narrow and symmetrical residuals of CNN-LSTM, Bi-LSTM, and XGBoost indicate higher confidence and fewer outlier predictions. However, in Tubas, residuals were expanded across all models, reflecting the forecasting challenges posed by unstable weather patterns. Finally, the deep learning models maintained relatively compact residuals, reinforcing their robustness.

Feature Importance Insights

The analysis of feature importance revealed critical differences in features. In the Balearic Islands, solar radiation dominated the predictive features, aligning with the region’s climatic stability. In contrast, Tubas required models to account for a broader set of features, including temperature, humidity, and wind speed, underlining the complexity of its environmental dynamics.

5.2 Conclusion

This study has presented a comprehensive investigation into the application of ML models for forecasting solar PV power output across two distinct climatic regions: the Tubas district in Palestine and the Balearic Islands in Spain.

The experimental results demonstrated the outperformance of deep learning models, particularly the CNN-LSTM and Bi-LSTM models, in accurately predicting solar PV power generation. These models outperformed traditional algorithms in both stable (Balearic) and variable (Tubas) environments, achieving the highest R^2 scores and the lowest RMSE and MAE values. This underscores their ability to learn complex temporal dependencies and adapt to nonlinear patterns within the data.

Furthermore, the study highlighted the importance of feature selection and regional environmental characteristics in model performance. Solar irradiance was consistently the most influential feature, but in variable climates such as Tubas, additional meteorological parameters like temperature, humidity, and wind speed significantly contributed to forecasting accuracy. This result emphasizes the need for modeling that takes into account the nature of weather factors in forecasting renewable energy.

By integrating meteorological data with real PV production values, the study confirmed that ML based forecasting can support energy planning, optimize grid integration of solar power, and indirectly aid in CO_2 emissions reduction through optimizing the use of renewable energy.

5.3 Future Work

In the initial phases of conceptualizing this study, the focus was on building ML-based models for the environmental impact on renewable energy optimization and CO_2 emission reduction; however, after conducting a thorough literature review and investigating related works in this field, the study's direction shifted slightly while remaining within the same conceptual framework. The focus was redirected towards evaluating the efficiency of the ML approach in forecasting solar PV power generation across diverse climate conditions.

Although this study achieved promising results, several areas offer potential for further investigation and enhancement:

Applying ML models to more diverse climates, such as tropical or desert environments, each of these regions has unique climatic characteristics that can challenge the performance of ML models trained in temperate or Mediterranean climates.

Integrating additional factors may further improve forecasting accuracy, such as air quality and pollution, especially in urban or industrial areas where these factors affect solar radiation (X. Zhang et al., 2022).

Using data with longer periods will provide insight into the long-term impact of climate trends on solar energy generation and the effectiveness of forecasting models under climate change scenarios.

Combine the datasets from the Tubas and Balearic regions, then fit and evaluate models. Compare the results with models trained separately for each area in terms of accuracy and resource requirements. The goal is to make the system scalable, as having a separate model for each area is not scalable.

Design and evaluate a mixture-of-experts approach for forecasting solar PV power generation.

References

Abadi, M., Agarwal, A., Barham, P., & others. (2016). TensorFlow: Large-scale machine learning on heterogeneous systems.

Abumohsen, M., Owda, A. Y., Owda, M., & Abumihsan, A. (2024). Hybrid machine learning model combining of CNN-LSTM-RF for time series forecasting of Solar Power Generation. E-Prime-Advances in Electrical Engineering, Electronics and Energy, 9, 100636.

Abumohsen, M., Owda, A. Y., Owda, M., Abumihsan, A., & Stergioulas, L. (2024). Forecasting Solar Power Generation Using Extreme Gradient Boosting: A Machine Learning Approach. 2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA), 505–510.

Acharya, N. (2024). Understanding Outlier Removal Using Interquartile Range (IQR). Medium.

Ackerman, S., Farchi, E., Raz, O., Zalmanovici, M., & Dube, P. (2020). Detection of data drift and outliers affecting machine learning model performance over time. ArXiv Preprint ArXiv:2012.09258.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. Organizational Research Methods, 16(2), 270–301.

Ahmed, M. W. (2023). Understanding mean absolute error (mae) in regression: A practical guide. Medium, Aug.

Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. Technologies, 9(3), 52.

AlSharabi, K., Bin Salamah, Y., Aljalal, M., Abdurraqeeb, A. M., & Alturki, F. A. (2025). Long-Term Forecasting of Solar Irradiation in Riyadh, Saudi Arabia, Using Machine Learning Techniques. Big Data and Cognitive Computing, 9(2), 21.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F. J., & Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. Solar Energy, 136, 78–111.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

Brownlee, J. (2020). Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery.

Chaaban, M. A. (2023). Irradiance and PV Performance Optimization. https://www.e-education.psu.edu/ae868/node/877

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?--Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 785–794.

Chicco, G., Napoli, R., & Piglione, F. (2001). Load pattern clustering for short-term load forecasting of anomalous days. 2001 IEEE Porto Power Tech Proceedings (Cat. No. 01EX502), 2, 6--pp.

Chicco, G., Napoli, R., & Piglione, F. (2014). Load pattern clustering for short-term load forecasting of anomalous days. Electric Power Systems Research, 73(2), 195–203.

Chuluunsaikhan, T., Nasridinov, A., Choi, W. S., Choi, D. Bin, Choi, S. H., & Kim, Y. M. (2021). Predicting the power output of solar panels based on weather and air pollution features using machine learning. Journal of Korea Multimedia Society, 24(2), 222–232.

de España (REE), R. E. (2025). ESIOS - The Electricity System Information Platform. https://www.esios.ree.es/

Deming, C., Dekkati, S., & Desamsetti, H. (2018). Exploratory data analysis and visualization for business analytics. Asian Journal of Applied Science and Engineering, 7(1), 93–100.

Diagne, M., David, M., Lauret, P., Boland, J., & Schmutz, N. (2013). Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renewable and Sustainable Energy Reviews, 27, 65–76.

Donoso, J., Behar, M., & Miranda, A. (2023). National Survey Report of PV Power Applications in Spain 2022. https://iea-pvps.org/wp-content/uploads/2023/08/National-Survey-Report-of-PV-Power-Applications-in-Spain-2022.pdf

Embarak, D. O., Embarak, K., & Karkal. (2018). Data analysis and visualization using python. Springer.

Essam, Y., Ahmed, A. N., Ramli, R., Chau, K.-W., Idris Ibrahim, M. S., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2022). Investigating photovoltaic solar power output forecasting using machine learning algorithms. Engineering Applications of Computational Fluid Mechanics, 16(1), 2002–2034.

Ewald, F. K., Bothmann, L., Wright, M. N., Bischl, B., Casalicchio, G., & König, G. (2024). A guide to feature importance methods for scientific inference. World Conference on Explainable Artificial Intelligence, 440–464.

Frederiksen, C. A. F., & Cai, Z. (2022). Novel machine learning approach for solar photovoltaic energy output forecast using extra-terrestrial solar irradiance. Applied Energy, 306, 118152.

Gao, M., Li, J., Hong, F., & Long, D. (2019). Short-term forecasting of power production in a large-scale photovoltaic plant based on LSTM. Applied Sciences, 9(15), 3192.

Gayathry, V., Kaliyaperumal, D., & Salkuti, S. R. (2024). Seasonal solar irradiance forecasting using artificial intelligence techniques with uncertainty analysis. Scientific Reports, 14(1), 17945.

Géron, A. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

Graphite Note. (2024). Understanding Target Variables in Machine Learning. https://graphite-note.com/understanding-target-variables-in-machine-learning/

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5–6), 602–610.

Hao, J., & Ho, T. K. (2019). Machine learning made easy: a review of scikit-learn package in python programming language. Journal of Educational and Behavioral Statistics, 44(3), 348–361.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts. https://otexts.com/fpp2/

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679–688.

in Data, O. W. (2023). Energy Production and Consumption. https://ourworldindata.org/energy-production-consumption

in Data, O. W. (2024). Share of Electricity Production from Renewables. https://ourworldindata.org/grapher/share-electricity-renewables

Institute for Health Metrics and Evaluation. (2021). Air Pollution – Research & Analysis. https://www.healthdata.org/research-analysis/health-risks-issues/air-pollution

International Energy Agency. (2023). Executive Summary – Renewables 2023. https://www.iea.org/reports/renewables-2023/executive-summary

International Energy Agency. (2024a). CO_2 Total Emissions by Region, 2000–2023. https://www.iea.org/data-and-statistics/charts/co2-total-emissions-by-region-2000-2023

International Energy Agency. (2024b). Solar PV. https://www.iea.org/energy-system/renewables/solar-pv

Kim, E., Akhtar, M. S., & Yang, O.-B. (2023). Designing solar power generation output forecasting methods using time series algorithms. Electric Power Systems Research, 216, 109073.

Kumar, R. D., Prakash, K., Sundari, P. A., & Sathya, S. (2023). A Hybrid Machine Learning Model for Solar Power Forecasting. E3S Web of Conferences, 387, 4003.

Magazzino, C., Mele, M., & Schneider, N. (2021). A machine learning approach on the relationship among solar and wind energy production, coal consumption, GDP, and CO2 emissions. Renewable Energy, 167, 99–115.

Mellit, A., & Kalogirou, S. A. (2008). Artificial intelligence techniques for photovoltaic applications: A review. Progress in Energy and Combustion Science, 34(5), 574–632.

Mellit, A., Kalogirou, S. A., Hontoria, L., & Shaari, S. (2009). Artificial intelligence techniques for sizing photovoltaic systems: A review. Renewable and Sustainable Energy Reviews, 13(2), 406–419.

Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

Nadeem, A., Hanif, M. F., Naveed, M. S., Hassan, M. T., Gul, M., Husnain, N., & Mi, J. (2024). AI-Driven precision in solar forecasting: Breakthroughs in machine learning and deep learning. AIMS Geosciences, 10(4), 684–734.

National Oceanic and Atmospheric Administration. (2023). Meteorological and Astronomical Seasons: Southern Hemisphere.

of Worldwide Energy Resources (POWER) Project, N. P. (2025). POWER Data Access Viewer. https://power.larc.nasa.gov/data-access-viewer/

Owda, M., Owda, A. Y., & Fasli, M. (2023). An Exploratory Data Analysis and Visualizations of Underprivileged Communities Diabetes Dataset for Public Good. 2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 581–585.

Palestine Investment Fund. (2025). PIF Launches “Tubas” Solar Power Plant in Tubas Governorate with a Production Capacity of 8 MW. https://www.pif.ps/items/view/71677

Perera, M., De Hoog, J., Bandara, K., Senanayake, D., & Halgamuge, S. (2024). Day-ahead regional solar power forecasting with hierarchical temporal convolutional neural networks using historical power generation and weather data. Applied Energy, 361, 122971.

Rahmad Ramadhan, L., & Anne Mudya, Y. (2024). A Comparative Study of Z-Score and Min-Max Normalization for Rainfall Classification in Pekanbaru. Journal of Data Science, 2024(04), 1–8.

Saini, M. K., Saroha, S., & others. (2023). Solar Power Forecasting using Machine Learning Approaches: A Review. 2023 9th IEEE India International Conference on Power Electronics (IICPE), 1–6.

Shah, A., Viswanath, V., Gandhi, K., & Patil, N. M. (2024). Predicting Solar Energy Generation with Machine Learning based on AQI and Weather Features. ArXiv Preprint ArXiv:2408.12476.

Shaker, L. M., Al-Amiery, A. A., Hanoon, M. M., Al-Azzawi, W. K., & Kadhum, A. A. H. (2024). Examining the influence of thermal effects on solar cells: a comprehensive review. Sustainable Energy Research, 11(1), 6.

Sharma, A., & Zhang, L. (2024). Solar Irradiance Forecasting for Grid-Integrated Photovoltaic Systems: A Case Study. Forecasting, 4(1), 66–84. https://doi.org/10.3390/forecast4010005

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in Neural Information Processing Systems, 28.

Simeunović, J., Schubnel, B., Alet, P.-J., & Carrillo, R. E. (2021). Spatio-temporal graph neural networks for multi-site PV power forecasting. IEEE Transactions on Sustainable Energy, 13(2), 1210–1220.

Solar Energy Industries Association. (2023). Climate Change. https://seia.org/initiatives/climate-change

Subramanian, E., Karthik, M. M., Krishna, G. P., Prasath, D. V., & Kumar, V. S. (2023). Solar power prediction using Machine learning. ArXiv Preprint ArXiv:2303.07875.

U.S. Energy Information Administration. (2023). Solar Energy and the Environment. https://www.eia.gov/energyexplained/solar/solar-energy-and-the-environment.php

Urraca, R., Huld, T., Gracia-Amillo, A., Martinez-de-Pison, F. J., Kaspar, F., & Sanz-Garcia, A. (2018). Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data. Solar Energy, 164, 339–354.

Urraca, R., Martinez-de-Pison, E., Sanz-Garcia, A., Antonanzas, J., & Antonanzas-Torres, F. (2017). Estimation methods for global solar radiation: Case study evaluation of five different approaches in central Spain. Renewable and Sustainable Energy Reviews, 77, 1098–1113.

Utopus Insights. (2023). Benefits of Power Forecasting. https://www.utopusinsights.com/benefits-of-power-forecasting

VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc."

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. Renewable Energy, 105, 569–582.

Yang, D., & Kleissl, J. (2013). Solar Energy Forecasting and Variability. Renewable and Sustainable Energy Reviews, 21, 103–113. https://doi.org/10.1016/j.rser.2012.12.046

Yona, A., Senjyu, T., Saber, A. Y., & Funabashi, T. (2008). Application of neural networks to 24-hour-ahead PV power forecasting. IEEE Power Engineering Society General Meeting.

Yu, Y., Cao, J., & Zhu, J. (2019). An LSTM short-term solar irradiance forecasting under complicated weather conditions. IEEE Access, 7, 145651–145666.

Zhang, J., Florita, A. R., Hodge, B.-M., Lu, S., & Hamann, H. F. (2018). A probabilistic approach to solar power forecasting. IEEE Transactions on Sustainable Energy, 9(3), 1680–1690.

Zhang, X., Zhang, M., Cui, Y., & He, Y. (2022). Estimation of Daily Ground-Received Global Solar Radiation Using Air Pollutant Data. 10(April), 1–13. https://doi.org/10.3389/fpubh.2022.860107

Zhou, H., Liu, Q., Yan, K., & Du, Y. (2021). Deep Learning Enhanced Solar Energy Forecasting with AI-Driven IoT. Wireless Communications and Mobile Computing, 2021(1), 9249387.

Zulkifly, Z., Baharin, K. A., & Gan, C. K. I. M. (2021). Improved machine learning model selection techniques for solar energy forecasting applications. International Journal of Renewable Energy Research (IJRER), 11(1), 308–319.

منهج مقارن لتوقع الطاقة الشمسية باستخدام التعلم الآلي في ظل ظروف مناخية متنوعة

عبد الرؤوف محمد شاكر شرباتي

لجنة الإشراف:

الدكتور مجدي عودة

الدكتورة هند سوسة

الدكتور محمد جبرائيل

الملخص

ت ه ه ال راسة على مقارنة فاءة ن اذج ال عل الآلي (ML) في ال إن اج ال اقة ال ة ال ه وضو ئ ة (PV) في وف م اخ ة م لفة خلال الف ة م أغ 2022 إلى ي ل 2023. ت ج ع ال انات م م قع جغ اف : م ي ة اس في فل ، ال ي ت ب قل ات م اخ ة ع الة، وج ر ال ار في إس ان ا ذات ال اخ ال ق . وت ال انات ب انات أرصاد ج ة م الإشعاع ال ي، ودرجة ال ارة، وال ة، وال غ ، وس عة ال اح م قاع ة ب انات و الة ناسا، الإضافة إلى ب انات إن اج ال اقة ال ة م ش ة ه اء اس وال ة ال ه ائ ة الإس ان ة Red Eléctrica de España .

ش ل م ه ة ال راسة ت ال انات وت ل لها واس افها وت ها لل جة. ت اس ام ع ة ن اذج لل عل الآلي، م ها الان ار ال ي، الغا ة الع ائ ة(Random Forest) ، ال ع ال ر ي ال ي (XGBoost) ، ال اك ة لة ال ث ائ ة الات اه(Bi-LSTM) ، وال ذج اله ال ع ب ال ات الع ة الال اف ة وال اك ة لة ال .(CNN-LSTM) ت ت ق أداء ال اذج اس ام م ش ات ال أ ال عي ال ر (RMSE) ، وم س ال أ ال ل (MAE) ، ومعام ال ي .(R^2) وق حقق ن اذج ال عل الع ، وخاصة Bi-LSTM و CNN-LSTMاله ، أف أداء في لا ال ق . ا ت أن اس اذ أك ت اً تقل ات ال ق ، إلا أن ال اذج الع قة أ ه ت ث اتاً في الأداء.

أ هـ ت الـ ائج أن الإشعاع الـ ي ان العامـ الأكـ تأثـ اً في الـ في لا الـ قع ، بـ ا ساه درجة الـ ارة والـ ة وسـ عة الـ اح ملـ في الـ ا ذات الـ ق الـ قل . وتـ لـ الـ راسة إلى أن نـ اذج الـعلـ العـ هي الأذ لـ قع الـ اقة الـ ة في بـ ات مـ اخـ ة مـ عة. وتُعـ هـ ه الـ ائج مهـ ة لـعـ تـ امـ الـ اقة الـ ة وتـ جـ ه الاسـ ارات في الـ ة الـ ة.

الـ لـ ات الـ فـ احـ ة: الـ إنـ اج الـ اقة الـ ة، الـعلـ الآلي، الـعلـ العـ ، الـ وف الـ اخـ ة الـ غـ ة.