# Weather-Driven Solar Energy Forecasting: Advanced Predictive Analytics


**General Sir John Kotelawala Defense University**

**Faculty of Computing**

**Department of Computational Mathematics**


**BSc. Data Science and Business Analytics**

**Intake 39**

**PROJECT REPORT**

**GROUP PROJECT IN 3rd YEAR DATA SCIENCE**

| | Group Details | |
|---|---|---|
| **Group No. 02** | **Registration No** | **Name** |
| | **D/DBA/22/0002** | **HMRV Herath** |
| | **D/DBA/22/0007** | **RN Silva** |
| | **D/DBA/22/0017** | **TM Kahavidhana** |
| | **D/DBA/22/0028** | **KT Panditha** |
| **Project Details** | | |
| **Supervisor** | **Mrs. SMM Lakmali** | |

## Authors' Declaration

"I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. Also, I hereby grant to General Sir John Kotelawala Defence University the nonexclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books)."

Signature :

Student Name : HMRV Herath

Student Registration Number : D/DBA/22/0002

Date : 25/11/2024

Signature :

Student Name : RN Silva

Student Registration Number : D/DBA/22/0007

Date : 25/11/2024

Signature :

Student Name :TM Kahavidhana

Student Registration Number : D/DBA/22/0017

Date : 25/11/2024

Signature :

Student Name : KT Panditha

Student Registration Number : D/DBA/22/0028

Date : 25/11/2024

The above candidates have carried out the group project for the partial fulfillment of the BSc (Honors) in Data Science and Business Analytics Degree Thesis under my supervision.

Signature :

Supervisor Name : Mrs. SMM Lakmali

Date : 25/11/2024

# Acknowledgement

We would like to express our sincere gratitude to those who have been integral to the completion of our group project on Advanced Predictive Analytics on Weather-Driven Solar Energy Forecasting for the BSc (Honors) in Data Science and Business Analytics at the General Sir John Kotelawala Defence University. Our thanks to Mrs. SMM lakmali, our project supervisors for her invaluable guidance. The support from the Department of Computational Mathematics, Faculty of Computing and our fellow students has been instrumental. Special acknowledgement goes to Windforce PLC for providing the crucial dataset, enhancing the project's depth. Lastly, we appreciate all who helped and contributed to the success of this endeavor.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1: Introduction

## 1.1 Overview

The global transition to renewable energy sources is a crucial step toward mitigating climate change and addressing the increasing demand for sustainable energy. Among renewable sources, solar energy is one of the most widely available and environmentally friendly options. However, the amount of solar energy generated can fluctuate greatly depending on various weather factors, such as cloud cover, temperature, and time of day. This variability makes it challenging to predict energy output with precision, which in turn affects grid stability and the efficient management of energy resources.

In this context, accurate and reliable solar energy forecasting plays a critical role in minimizing fluctuations, optimizing energy storage, and ensuring grid stability. Predicting solar power output can help energy providers plan better, manage resources effectively, and make informed decisions regarding energy trading. For solar plant operators, a reliable forecasting model is also key to optimizing plant performance, minimizing downtime, and improving the overall efficiency of power generation.

This study addresses these challenges by designing, implementing, and evaluating a predictive model that integrates comprehensive weather and operational data to forecast solar power output effectively. The project specifically focuses on developing an accurate and efficient predictive model for short-term and long term solar power output forecasting at the Vydexa solar power plant in Sri Lanka, owned by WindForce Pvt Ltd. The need for such a model is from the company's identified requirement for a reliable method to anticipate energy production and address operational uncertainties. The model incorporates operational data from the plant and weather data obtained from an online resource to provide reliable predictions at 15-minute intervals and hourly for long term. The project focuses on this forecasting tool by which WindForce can gain actionable insights to improve its energy allocation, reduce operational uncertainty, and enhance grid stability while contributing to the country's growing demand for clean and sustainable energy.

## 1.2 Background and Motivation

Sri Lanka's energy sector is undergoing a transformation, as renewable energy is becoming more important and is increasingly being integrated into the national grid to reduce dependence on fossil fuels. WindForce PLC, the largest renewable energy developer in Sri Lanka, has played a pivotal role in this shift. Established in 2010, WindForce is a leader in renewable energy development and operates many solar, wind, and hydropower plants both in Sri Lanka and internationally.

WindForce's solar energy plants include 13 solar plants across Sri Lanka, Pakistan, Uganda, and Ukraine, collectively generating 265.17 GWh annually and preventing 188,220 metric tons of $CO_2$ emissions. The Vydexa Solar Power Plant in Vavuniya, is one of its key projects, and contributes 10 MW annually. While Vydexa is strategically located in a region with significant solar potential, the region's variable weather conditions - with irregular cloud cover, fluctuating solar irradiance, and rainfall patterns - can pose challenges to accurately forecasting power output.

WindForce identified the need for a robust and precise 15-minute (short term) and hourly (long term) power output forecasting model for Vydexa to address these challenges. Accurate predictions are vital for optimizing the plant's operational efficiency, reducing downtime, and enhancing grid reliability. This requirement forms the basis of our project and highlights its industrial relevance.

It is important to note that solar energy forecasting is still an evolving field with several unresolved challenges. Existing forecasting methods often have to struggle with the high variability of weather patterns and might lack consistency across different locations and geographic regions. Additionally, many forecasting models have issues with transparency, scalability, and performance reliability under extreme weather conditions. To address these challenges, we have the opportunity to utilize and apply machine learning to operational data from real-world solar plants to advance both industrial applications and academic research in renewable energy forecasting. By using real-world data from Vydexa, we aim to create a model that is tailored to the plant's specific conditions and contributes to better solutions in the renewable energy sector.

## 1.3 Research Problem

The research problem statement is how can the accurate and efficient prediction of short term and long term solar power output generation be achieved for a solar power plant while incorporating comprehensive weather data.

## 1.4 Solution

The solution proposed in this project is the development of a machine learning-based predictive model tailored to the specific needs of the Vydexa Solar Power Plant. This model leverages historical operational data from the plant and weather variables to forecast short-term solar power output at 15-minute intervals and long term solar power output at 1 hour.

## 1.5 Significance of the Study

The nature of solar energy poses challenges for power grid stability and resource allocation. Sudden changes in weather conditions, such as cloud cover or rainfall, can result in fluctuations in solar power output, making it difficult to balance supply and demand in the grid. By developing a

machine learning-based model tailored to the Vydexa Solar Power Plant, this project will provide WindForce with precise power output predictions. These forecasts will enable better scheduling of energy storage, more effective energy trading, and improved overall grid stability.

Through our predictive model, we aim to leverage advanced machine learning techniques like neural networks and ensemble methods, as identified in the systematic literature review, and by building upon it then create a customized, high-accuracy forecasting tool, focusing on variables like solar irradiation, temperature, relative humidity, and historical power generation data, which are crucial to reliable solar power predictions. Additionally, incorporating key weather variables into our model ensures its relevance to the unique weather and operational conditions of Vavuniya, while also including less frequently used but still critical variables like wind speed and cloud cover, the model aims to provide comprehensive and highly accurate predictions.

Accurate power output forecasting is essential for solar plant operators like WindForce. It enables better planning for plant maintenance, reduces energy losses, and ensures the optimal use of resources. With the practical and short term 15-minute forecasts, the Vydexa Solar Power Plant can fine-tune its operations, minimize downtime, and integrate its energy production more effectively with the grid. These operational improvements contribute to the economic and environmental sustainability of renewable energy systems in Sri Lanka.

This project not only addresses a specific industrial need but also contributes to the broader field of renewable energy forecasting. By integrating and exploring multiple machine learning methods, refining them for real-world conditions, the project proposes solutions to address the unique challenges of the Vydexa plant. Although the practical application of this study lies in its deployment at the Vydexa Solar Power Plant where the focus is primarily on this specific site, the methodology and model developed can serve as a blueprint for other solar power plants in similar tropical regions. The ability to adapt the model to other geographical and climatic zones could make this research a scalable solution for renewable energy forecasting on a larger scale.

This study serves as a practical example of how advanced technologies like machine learning can be applied to solve industry-specific problems and is significant because it not only addresses a critical operational need at the Vydexa Solar Power Plant but also contributes to advancing renewable energy forecasting techniques. The project combines modern data science technology with practical applications, paving the way for more efficient, reliable, and sustainable solar energy systems in Sri Lanka.

## 1.6 Research Aim

The aim of this project is to design, implement, evaluate, and deploy an accurate and efficient predictive model that integrates weather and operational data from the Vydexa solar power plant to forecast its short term (15 minute) and long term (1 hour) energy output.

## 1.7 Research Objectives

The objectives of this project are designed to address both the technical challenges of solar power forecasting and the practical needs of WindForce. The specific objectives are as follows:

1. To acquire and compile operational and weather data from the solar power plant .
2. To preprocess the data to ensure quality.
3. To explore and implement machine learning algorithms for solar energy prediction.
4. To evaluate, validate, and compare ML models for performance.
5. To optimize models to enhance forecasting accuracy.
6. To deploy the selected accurate model for practical implementation.

## 1.8 Structure of the Thesis

To present our research project in a logical and clear manner, this thesis is structured into eight chapters, with each contributing to a wholesome understanding of the research. The introduction chapter lays the foundation, highlighting the background, motivation, research problem, and significance of the study while setting the stage for the research objectives and aim.

The literature review elaborates on how we examined existing studies and methodologies, identifying gaps and challenges to provide a good foundation for the research. The methodology chapter meticulously details the approach we adopted, covering data acquisition, preprocessing, model development, and deployment, along with a clear project timeline. The tools and technologies chapter outlines the software, frameworks, and platforms utilized, emphasizing their role in achieving the research objectives. In the design and analysis chapter, the development of the predictive model is thoroughly explored, with insights into its construction and underlying processes.

The model evaluation and results chapter presents the findings, supported by evaluation metrics and comparisons, to validate the model's effectiveness. The model deployment chapter focuses on the practical implementation and operationalization of the predictive framework. Finally, the discussion and conclusion chapter synthesizes the findings, reflecting on their implications, limitations, and directions for future research. This thesis provides a comprehensive exploration of solar energy output prediction, delivering valuable insights for both academic and industrial applications.

## 1.9 Summary

This project focused on developing an accurate predictive model for short-term solar power output forecasting at the Vydexa Solar Power Plant, owned by WindForce Pvt Ltd in Sri Lanka. As solar energy becomes an increasingly important part of the global transition to renewable energy, its variable nature presents challenges for grid integration and energy management. To address these challenges, we, the project team, aimed to design a model that could provide precise

short term interval forecasts to optimize energy generation, improve operational efficiency, and enhance grid stability.

In summary, the project is to successfully develop a predictive model that addresses the specific forecasting needs of WindForce and contributes to the broader field of solar energy forecasting. By demonstrating the application of machine learning in this context, the groundwork for improving operational efficiency at the Vydexa plant is appropriately achieved. The project work not only can serve the practical needs of WindForce but also can offer insights that can be applied to renewable energy forecasting more broadly, supporting the global transition to more sustainable and reliable energy systems.

# Chapter 2: Literature Review

## 2.1 Overview

The integration of renewable energy, particularly solar power, into global energy systems necessitates precise forecasting models to handle its inherent variability. Recent advancements in weather-driven solar energy forecasting underscore the essential role of combining meteorological data with innovative computational methods to enhance forecast accuracy and reliability. This review synthesizes findings from various studies that explore the dynamic interplay between key meteorological variables and solar power output.

Significant challenges persist in the realm of solar forecasting, primarily due to the unpredictability of weather conditions and the complex nature of forecasting models. Issues such as the uncertainty in solar irradiance, temperature fluctuations, and cloud movements critically impact the accuracy of predictions. These elements necessitate sophisticated models that can adapt swiftly and accurately to changing environmental inputs. In response, researchers have increasingly turned to advanced machine learning techniques, including Neural Networks and Ensemble Methods, which offer substantial improvements over traditional models. These methods excel in processing large datasets and capturing the non-linear relationships inherent in meteorological factors influencing solar energy production.

Moreover, the review highlights the importance of continuous innovation in predictive technology and methodologies, urging ongoing refinement of models to incorporate real-time data and improve adaptability to diverse geographic and climatic conditions. The ultimate goal is to achieve a seamless integration of solar energy into power grids, supporting the transition towards more sustainable energy systems globally.

## 2.2 Problems Identified

The development of solar energy forecasting models has revealed several persistent challenges that significantly affect their precision and reliability. These issues stem from the variability in essential input data and the inherent complexities of forecasting models.

A primary challenge lies in the uncertainty of meteorological data, which has a critical impact on the accuracy of solar power predictions. Studies such as [1], [2], and [3] emphasize that fluctuations in solar irradiance, temperature, and cloud cover especially under dynamic conditions like partly cloudy or overcast skies frequently lead to significant discrepancies between predicted and actual energy output.

Another major issue is the complexity of predictive models themselves. Advanced techniques, including Extreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM), as

discussed in [4], [5], and [6], require extensive and high-quality data for effective training. However, these models often face difficulties due to discrepancies in data quality, particularly when applied to diverse geographic regions with varying climatic conditions. Furthermore, adapting these models to handle consistently unreliable real-world data remains a significant challenge.

The intermittent nature of solar power adds another layer of complexity to its integration into power grids. Accurate and timely forecasts, as highlighted by [7] and [8], are essential for informing power system operators and maintaining grid stability. The fluctuating nature of solar energy output necessitates highly dependable forecasting to support these operations.

Assessing the performance of forecasting models also presents considerable challenges. Studies such as [9], [10], and [11] discuss how inconsistencies in the use of evaluation metrics, like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), hinder comparative analysis across studies. These inconsistencies lead to varied interpretations of model performance and obstruct the establishment of best practices.

Collectively, these challenges underscore the need for continuous advancements in forecasting methodologies. As noted in [12] and [13], improvements in data collection and processing techniques and standardization of evaluation practices are critical to enhancing the reliability and practical application of solar energy forecasts.

## 2.3 Identification of Key Weather Variables

The literature review identifies consistent patterns in determining crucial weather variables that influence solar power generation and forecasting. Analysis of multiple studies highlights several key meteorological parameters as essential for accurate solar power prediction.

**General Solar and Meteorological Variables:** Solar irradiance and temperature are fundamental in solar energy studies. For example, [1] and [4] demonstrate that these parameters directly impact panel output efficiency and form the backbone of reliable forecasting models. The relationship between temperature and panel efficiency is particularly notable, as temperature coefficients and efficiency variations significantly affect overall system performance.

Wind characteristics, including speed and direction, also play a vital role in solar panel performance. Studies such as [14] and [7] underline how wind parameters influence panel cooling and operational efficiency. Proper consideration of these variables has been shown to markedly improve forecast accuracy.

Cloud cover and humidity represent another critical category of variables. As explored in [9] and [3], these factors greatly affect the solar irradiance received by panels. Incorporating cloud cover

and humidity data into models helps account for atmospheric interference with solar radiation transmission, thereby improving prediction precision.

**Advanced Meteorological Data:** More sophisticated parameters, such as precipitable water vapor and sea-level temperature, enhance forecasting accuracy. For instance, [2] illustrates how these detailed atmospheric conditions contribute to more precise predictions. Additionally, cumulative variables like total irradiation and rainfall, as utilized in [12], provide insights into the long-term environmental impacts on solar power generation.

**PV System-Specific Variables:** The physical characteristics and configuration of solar installations are significant determinants of power output. According to [5], factors such as panel efficiency, orientation, and system configuration are crucial. Long-term forecasting accuracy also depends on accounting for panel degradation and performance factors, as highlighted by [7], which addresses the natural decline in system efficiency over time.

**Historical Data and Technological Innovations:** Integrating historical generation data with current meteorological conditions significantly improves model accuracy. As noted in [10], historical performance data offers valuable context for prediction models. Furthermore, technological advancements in forecasting methods, such as hybrid and ensemble models discussed in [8], combine multiple machine learning techniques to enhance prediction reliability. The use of Numerical Weather Predictions (NWP), detailed in [6], marks a substantial advancement in integrating high-resolution weather data for solar power forecasting.

This comprehensive approach to identifying and integrating variables highlights the complex relationship between environmental conditions and solar power generation. The literature emphasizes that successful forecasting models must account for both basic meteorological parameters and system-specific characteristics, as well as historical performance data, to achieve optimal prediction accuracy.

## 2.4 Comparative Analysis of Methods and Technologies

The advancement of solar power forecasting methodologies in recent years is evident from a comprehensive examination of various studies, each employing unique technologies to enhance prediction accuracy and reliability. This section synthesizes insights from 16 research articles, highlighting the diversity and effectiveness of forecasting models and techniques.

**Statistical Methods:** Traditional statistical approaches, such as linear regression and autoregressive moving average models, remain foundational tools in solar forecasting. These methods provide baseline predictions and insights into temporal patterns, as demonstrated in [1]. However, their limitations in capturing complex variable interactions have led to the growing prominence of machine learning techniques.

**Machine Learning Techniques:** The integration of machine learning has revolutionized solar forecasting practices. For instance, [7] employs Deep Belief Networks, Support Vector Machines (SVM), and Random Forests to robustly handle complex variable interactions. Similarly, [14] demonstrates the superior performance of Support Vector Regression (SVR) in short-term forecasting scenarios.

**Deep Learning Techniques:** Advanced deep learning methods, particularly Long Short-Term Memory (LSTM) networks, have proven effective for capturing long-term dependencies in time-series data. Studies such as [4] highlight the utility of LSTM for both short-term and long-term forecasting, integrating techniques like Extreme Gradient Boosting (XGBoost) for enhanced performance. Furthermore, [3] explores modified LSTM architectures tailored specifically for medium and long-term photovoltaic power forecasting.

**Hybrid Models:** Combining algorithms to leverage their strengths has yielded promising results. For example, [11] integrates K-Nearest Neighbor (KNN) and SVM algorithms to balance computational efficiency with prediction accuracy, surpassing traditional models. Likewise, [10] demonstrates the effectiveness of combining Principal Component Analysis (PCA) with Artificial Neural Networks (ANNs) and Recurrent Neural Networks (RNNs) for improved feature selection.

**Ensemble Methods:** Ensemble approaches, as detailed in [8], utilize multiple machine learning models to mitigate individual model weaknesses, reducing variance and bias significantly. Similarly, [5] compares ensemble learning models such as Histogram Gradient Boosting and Light Gradient Boosted Machines, highlighting their superior predictive capabilities. The hybrid RNN-ANN architecture in [13], combined with specialized hyperparameter optimization, further exemplifies the potential of ensemble methods.

**Numerical Weather Predictions and Clustering Techniques:** Numerical Weather Predictions (NWP) and advanced clustering techniques enhance model adaptability. For instance, [12] categorizes days based on weather patterns to optimize model training, while [15] applies diverse methodologies to varying environmental conditions, improving forecasting accuracy.

**Advanced Data Processing:** Sophisticated data processing methods further enhance predictive accuracy. Studies such as [6] employ Gaussian Process Regression and Convolutional Neural Networks (CNNs) to transform irregular solar energy data into regular grids, improving geographic adaptability. Additionally, [16] integrates wavelet-based multiscale presentation techniques to refine forecasting models.

In conclusion, this comparative analysis underscores the significant strides made in solar power forecasting methodologies. Traditional statistical methods provide essential groundwork, while advanced machine learning, deep learning, and ensemble techniques offer substantial improvements in addressing the complexities of solar energy forecasting. The combined

application of these technologies enables precise, dynamic, and adaptable forecasting systems, which are crucial for the seamless integration of solar power into energy grids.

## 2.5 Assessment of Evaluation Metrics

The rigorous assessment of solar power forecasting models is essential for determining their efficacy and reliability. A variety of evaluation metrics, as outlined in several studies, provides a comprehensive understanding of model performance under different conditions and methodologies.

**Root Mean Square Error (RMSE) and Mean Absolute Error (MAE):** RMSE and MAE are the most commonly used metrics, offering insights into the average magnitude of prediction errors. RMSE, which is highly sensitive to large errors, is particularly crucial for models that need to minimize significant deviations, as illustrated in studies like [14] and [4]. Additionally, [10] reports precise MAE values of 0.0223 KWh and RMSE of 0.003 KWh, demonstrating exceptional accuracy. MAE, being less sensitive to outliers, provides a straightforward measure of average error magnitude and is widely used for its clarity in representing model accuracy.

**Mean Absolute Percentage Error (MAPE):** MAPE, utilized in studies such as [2], translates error into a percentage, making model accuracy easier to interpret. [16] emphasizes MAPE's value in comparing performance across systems with varying scales, as it normalizes errors relative to actual values, ensuring meaningful comparisons irrespective of dataset magnitude.

**Bias and R-value:** Bias, as discussed in [7], indicates whether a model consistently overestimates or underestimates outputs, which is critical for operational adjustments in power systems. Furthermore, [15] incorporates prediction accuracy metrics for directly comparing actual and predicted solar energy outputs. The R-value, measuring the correlation coefficient, is vital for assessing the linear relationship between observed and predicted values, as demonstrated in [5].

**Advanced Probabilistic Metrics:** The Continuous Ranked Probability Score (CRPS), employed in [13], evaluates probabilistic forecasts by measuring the difference between predicted and observed cumulative distributions. CRPS values reported range between 3.35% and 5.17% of rated power. Additionally, [3] introduces computation time as a metric, highlighting the practical importance of processing efficiency in operational scenarios.

**Comparative and Composite Metrics:** Several studies, such as [12], highlight comparative metrics, reporting accuracy improvements of 20.55% to 34.48% compared to traditional methods. [8] provides detailed RMSE ranges for various prediction horizons, with day-ahead forecasts ranging from 1.00% to 7.98% and week-ahead forecasts from 4.42% to 10.27%. Furthermore, [11] introduces specialized metrics, including Accuracy (AC), Sensitivity (Sn), and Specificity (Sf), offering a more holistic evaluation framework.

**Efficiency and Performance Metrics:** Study [15] highlights efficiency measurement metrics by comparing ANN and fuzzy logic models for solar irradiation prediction, along with comprehensive performance evaluations of photovoltaic systems. This multidimensional approach provides deeper insights into both model accuracy and system efficiency.

In conclusion, selecting evaluation metrics should align with the specific objectives of the forecasting model, emphasizing accuracy, reliability, and outlier management. The diversity of metrics across studies underscores the need for a multifaceted approach to thoroughly assess solar power forecasting models, ensuring they meet the standards required for practical deployment and operational efficiency.

## 2.6 Summary

The evolution of solar power forecasting has been significantly shaped by the integration and comparison of various machine-learning models and methodologies. Each study reviewed provides unique insights into the progression of forecasting techniques, emphasizing the need for enhanced prediction accuracy to support effective grid integration and energy management.

Studies such as [15] and [7] demonstrate the efficacy of Artificial Neural Networks (ANN) and Random Forest (RF) over traditional methods. These findings reveal that machine learning models adapt better to variable weather conditions and propose further improvements through technologies like sun-tracking and battery storage systems, which can enhance both the accuracy and efficiency of solar power systems.

The practical application of Support Vector Regression (SVR) for short-term forecasting is highlighted in [14]. This study underscores the importance of feature selection, showing that incorporating relevant features, such as recent power generation and cloud cover, significantly improves model accuracy. Conversely, it cautions against including less impactful weather parameters, which may lead to overfitting.

Medium and long-term forecasting challenges are addressed in [3], which discusses the limitations of relying on meteorological data from distant weather stations. It advocates for enhanced data collection methods using local data to improve the ability to capture seasonal variations, thereby increasing forecasting reliability over longer periods.

The advantages of ensemble methods are well-articulated in studies like [8] and [5]. These efforts highlight the ability of combined models to mitigate individual weaknesses, resulting in robust and reliable forecasting systems.

Emerging technologies have also contributed to advancements in solar forecasting. For example, [6] introduces the use of Convolutional Neural Networks (CNN) alongside Gaussian Process Regression, presenting a sophisticated approach for managing spatial data and improving

forecasting accuracy. Similarly, the incorporation of weather type clustering, as explored in [12], enhances prediction accuracy by categorizing weather patterns, offering a practical method for day-ahead forecasts.

Hybrid models, such as the KNN-SVM approach detailed in [11], illustrate the adaptability of combining machine learning techniques to meet specific forecasting needs, optimizing computational efficiency and prediction accuracy.

Collectively, these studies emphasize the transition towards data-driven, adaptive, and sophisticated machine-learning models in solar power forecasting. They advocate for continuous methodological evolution, integrating real-time data processing, advanced machine learning algorithms, and ensemble models to address the complex challenges of solar energy forecasting. This comprehensive analysis highlights both advancements in the field and the potential for future research to optimize and customize forecasting models for diverse regional and operational demands.

# Chapter 3: Methodology

In this project, Weather-Driven Solar Energy Forecasting  methodology played a crucial role in developing accurate models for predicting solar power output based on weather conditions. Since this project was conducted through the 1-year period, it is essential to have a defined methodology (structured workflow) to process and evaluate the data in a methodical manner. After discussions with supervisors, the methodology was divided into key stages, forming a streamlined workflow to ensure efficient execution.

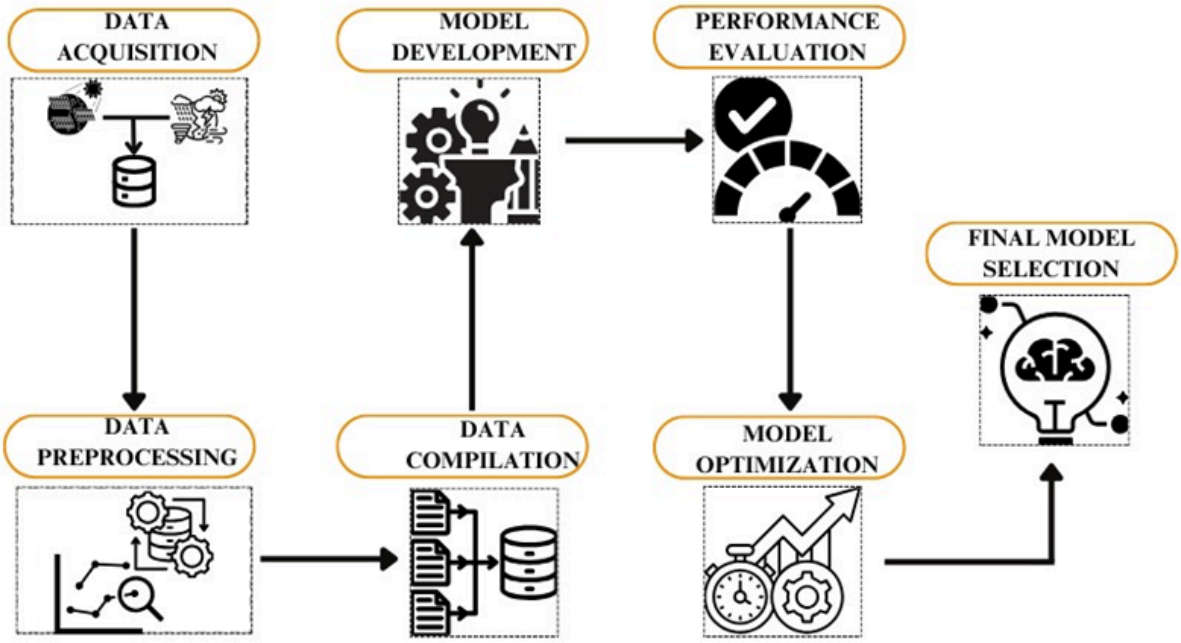



*Figure 3.1 : Methodology Diagram*

As shown in the diagram above, this methodology consists of 7 parts. It is described several parts below them are data acquisition, exploratory data analysis, data preprocessing, data compilation and model development, performance evaluation, model optimization as another one part then finally final model selection are explained below.

## 3.1 Data Acquisition

This study utilized secondary data collected from two distinct sources to develop and evaluate predictive models for solar power output. These datasets provided critical information on weather conditions and solar power generation, enabling comprehensive analysis and prediction. The data collection covered the period from January 1, 2023, to December 31, 2023 (12 A.M to 11:59 P.M), ensuring full temporal coverage for an entire year.

### Operational Data

The operational data was obtained from the Vydexa Solar Power Plant, located in Vavuniya. This dataset contained detailed measurements of electrical and solar parameters critical for model development. The variables included in the operational dataset are summarized in Table 3.1.

*Table 3.1: Operational variable set*

| Variable | Type | Description |
|---|---|---|
| **Time** | Categorical | The time of the day when the operational data was recorded. |
| **Current Phase A [A]** | Numeric | The electrical current in phase A (A). |
| **Current Phase B [A]** | Numeric | The electrical current in phase B (A). |
| **Current Phase C [A]** | Numeric | The electrical current in phase C (A). |
| **Total Active Power [MW]** | Numeric | The total active power in megawatts (MW). |
| **Irradiation** | Numeric | The amount of solar irradiation (W/m²). |

### Weather Data

The weather data was manually compiled from an online weather service specific to the Vavuniya region, sourced from Time and Date As (1995–2025). The weather data was collected at hourly intervals and covered the period from January to December 2023. This dataset provided crucial environmental parameters essential for solar power forecasting. The variables included in the weather dataset are listed in Table 3.2.

*Table 3.2: Weather variable set*

| Variable | Type | Description |
|---|---|---|
| **Date** | Categorical | The date when the weather data was recorded. |

| | | |
|---|---|---|
| **Time** | Categorical | The time of the day when the weather data was recorded. |
| **Temp** | Numeric | The temperature recorded at the given date and time (°C). |
| **Weather** | Categorical | The description of the weather conditions<br><br>(e.g., Sunny, Clear, Mild, Passing clouds, Scattered clouds, Partly sunny, Scattered clouds, Thundershowers). |
| **Wind** | Numeric | The wind speed recorded (km/h). |
| **Humidity** | Numeric | The percentage of humidity in the air. |
| **Barometer** | Numeric | The atmospheric pressure recorded (mbar). |

## 3.2 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) was conducted to gain valuable insights about distributions and interrelationships among various features by using statistical techniques and visualizations. By conducting the EDA phase our group was able to uncover the inherent patterns within the dataset.

**Objectives of EDA**

The primary goals of EDA included:

- Identifying significant missing and redundant values in the data.
- Exploring trends in weather variables and their potential correlations with solar power output.
- Evaluating variability in total active power and solar irradiance to uncover patterns.

**Descriptive Statistics**

Descriptive statistics is defined as a brief summary of what is known from within a particular dataset with regard to dependent or independent variables. These numerical indicators are important since they are mostly summarized, trend, variability, and distribution of data. For this study among various descriptive statistics included were count, mean, variance, standard deviation, minimum and maximum values. These already present the pattern of how spread and

ranged data values from one another; it will even help in determining the extent of variability and the identification of any possible outliers or extreme value, by calculating the statistics for primary important variables such as total active power, irradiation, temperature, wind speed, humidity, barometer. Therefore, it put a stronghold into further analysis for proper understanding of the dataset while preparing for model development.

**Histograms**

Histograms were used to identify distribution of each variable, such as total active power, irradiance, temperature, and all other weather parameters. These visual representations showed the internal tendencies, the variability, and the skewness of the data. Therefore, it put a stronghold into further analysis for proper understanding of the dataset while preparing for model development.

**Line plots**

Line plots were employed to visualize time-series data and identify temporal trends, seasonality, and fluctuations. Key variables, such as total active power, irradiation, temperature, and wind speed, were plotted over time. These plots helped pinpoint instabilities and gaps in the data, particularly in months with sparse or sporadic records.

**Box Plots**

Boxplots are an effective way to characterize distributions of important variables and identify potential outliers. By plotting the median, interquartile range (IQR), and whiskers for each variable, box plots clearly display the spread and skewness of the data. Any values sitting outside the whiskers are easily classified as outliers. This was essential to prepare the database for inputs in the models, as the critical point made data correction important.

**Correlation Heat Map**

Correlation heat map provides a comprehensive overview about relationships between different features. We compute the degree and direction of linear correlation between features by calculating correlation coefficients. A heat map format combined with a correlation matrix offers a thorough summary of pairwise relationships, highlighting possible interdependence.The correlation heat map provides quite an extensive picture of the interrelationships among all the features in the data set. Positive and negative dependencies were highlighted using gradient color representations. For example:

- A strong positive correlation was observed between irradiation and total active power.
- Weaker or more complex interactions were noted between variables such as temperature and humidity. These analyses aided in identifying the most significant variables for model development.

**Scatter Plot**

Scatter plots serve as an additional tool to examine and confirm the relationships between variables. By plotting pairs of features against one another, these visualizations demonstrated bivariate correlations and helped identify patterns, trends, or potential non-linear relationships. This visualization technique proved essential for gaining a more nuanced understanding of the dataset and validating the feature selection process.

**Average Relative Difference Equation**

An equation (1) was used to measure average relative difference to compare the active power with current phase A, current phase B, current phase C . Consequently, these parameters were deemed negligible and excluded from further analysis to maintain the focus on more impactful variables.

$$Formula = \frac{|Power - Current\ Phase|}{(Active\ Power + Current\ Phase)/2} \quad (1)$$

EDA was performed for the discovery of patterns, relationships, and anomalies in the data. In the analysis of distributions and the detection of outliers in the features of total active power, irradiance, and weather variables, techniques including descriptive statistics, histograms, line plots, box plots, scatter plots, and correlation heat maps were used. This phase gave a very good insight into variability, trends, and interdependencies of the data, thus laying a foundation for further analysis.

## 3.3 Data Preprocessing

Data preprocessing was a critical step to ensure data quality and compatibility with machine learning models. Key activities included:

- **Handling Missing and Inconsistent Data**

The preprocessing phase began with addressing the inconsistencies identified during the Exploratory Data Analysis (EDA). The operational data for January exhibited repetitive values, while February contained substantial missing data. These inconsistencies could not be resolved using standard imputation techniques. Consequently, under the guidance of the research supervisor, it was decided to exclude the data from January and February. The remaining dataset, covering the period from March to December 2023, was used for analysis and modeling. Any missing values within this period were addressed through imputation or removal, ensuring the dataset's reliability and consistency.

| Time | CURRENT PHASE A [A] | CURRENT PHASE B [A] | CURRENT PHASE C [A] | TOAL ACTIVE POWER [MW] | Irradiation |
|---|---|---|---|---|---|
| 1/1/2023 6:00 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:00 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:01 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:01 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:02 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:02 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:03 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:03 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:04 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:04 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:05 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:05 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:06 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:06 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:07 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:07 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |
| 1/1/2023 6:08 | 167.690247 | 165.706207 | 164.39296 | 9.425656 | 1083.101807 |

*Figure 3.2: Repetitive Data in January Operational Data*

| Time | CURRENT PHASE A [A] | CURRENT PHASE B [A] | CURRENT PHASE C [A] | TOAL ACTIVE POWER [MW] | Irradiation |
|---|---|---|---|---|---|
| 2/1/2023 7:01 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:02 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:03 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:04 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:05 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:06 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:07 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:08 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:09 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:10 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:11 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:12 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:13 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:14 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:15 | 0 | 0 | 0 | 0 | 0 |
| 2/1/2023 7:16 | 0 | 0 | 0 | 0 | 0 |

*Figure 3.3: Missing Data in February Operational Data*

- **Conversion of Categorical Variables**

Machine learning models require all input data to be in numerical format. To transform the categorical "Weather Condition" variable into a suitable numeric form, one-hot encoding was applied. This technique created separate binary columns for each category, assigning a value of "1" to indicate the presence of a specific category and "0" for its absence. For example, the "Cloud Cover" variable was converted into multiple binary columns, each representing a distinct weather condition. This transformation preserved the categorical nature of the variable while making it compatible with machine learning algorithms.

- **Frequency Alignment Using Interpolation**

To ensure compatibility between datasets, the weather data's hourly frequency needed to match the 1-minute frequency of the operational data. The original hourly weather data, collected from online sources, lacked the granularity required for effective analysis.To address this, missing data points were estimated for 1-minute intervals using an interpolation technique. Python's "time" method with forward and backward filling was employed for interpolation. This method generated intermediate values by analyzing trends within the existing hourly data, effectively converting the weather dataset into a 1-minute resolution. This alignment ensured consistency across all datasets and provided a solid foundation for further analysis and modeling.

- **Feature Scaling and Normalization**

To ensure uniformity across the dataset and provide optimal input for machine learning models, all numerical features were scaled using the MinMaxScaler. This technique normalized the data to a range between 0 and 1, reducing the influence of outliers and improving model performance. The normalization process helped maintain the integrity of the data while enhancing its suitability for predictive modeling.

By addressing these key aspects, handling missing and inconsistent data, converting categorical variables, and aligning data frequencies the preprocessing stage ensured the dataset was clean, consistent, and prepared for use in the machine learning models.

## 3.4 Data Compilation

Data compilation played a vital role in preparing the dataset for model development by integrating the relevant variables from both the operational and weather datasets. Under the guidance of the research supervisor, the exploratory data analysis (EDA) process included a thorough correlation analysis to identify the most relevant features for predictive modeling.

- **Feature Selection**

The correlation analysis revealed significant relationships between variables and the target output, leading to the selection of the most impactful features for further analysis. From the operational dataset, the variables total active power and irradiance were chosen due to their strong correlation with solar power output. Irrelevant features such as current phase A, B, and C were excluded, as they demonstrated negligible significance in predictive accuracy.

Similarly, from the weather dataset, the variables temperature, wind speed, humidity, and barometer were identified as essential features based on their correlation with solar power output. The weather condition variable, being categorical and less impactful, was also excluded from further analysis.

- **Dataset Integration**

The preprocessed weather data was then aligned with the refined operational dataset from the solar power plant. This process involved interpolating hourly weather data to match the 1-minute frequency of the operational data, ensuring consistency and compatibility across the datasets. The merged dataset contained operationally relevant parameters and key weather variables, forming a comprehensive and unified dataset.

- **Significance of Data Compilation**

By consolidating the preprocessed weather and operational datasets, the research established a robust foundation for model development. The integrated dataset provided a balanced combination of weather and solar power generation variables, enabling the creation of more accurate and insightful predictive models. This meticulous approach to data compilation not only improved the reliability of the dataset but also enhanced the potential for generating resource-relevant outcomes and supporting informed decision-making through meaningful insights.

## 3.5 Data Splitting into Training and Validation Sets

To evaluate the model's performance on unseen data, the dataset was manually divided into two parts:

- The training dataset included data from March 2023 to October 2023 and was used to train the model by identifying patterns and relationships within the data.
- The validation dataset contained data from November 2023 to December 2023. This dataset was excluded from the training process to ensure the model had no prior access to it, simulating real-world scenarios.

This splitting approach ensured that the validation dataset served as an independent set to evaluate the model's generalization capabilities, providing a robust and unbiased assessment of its performance.

## 3.6 Model development, performance evaluation and model optimization

After the completion of the feature engineering phase and the Exploratory Data Analysis (EDA), the next stage involved model building, performance evaluation, and model optimization. Since one of the critical components of this study was to develop accurate models for forecasting solar energy output based on weather-driven parameters.

The group used a methodical and iterative strategy to approach efficiency and accuracy because of the significance of this step and its potential complexity. During this stage we implemented and evaluated four machine learning algorithms for solar energy output prediction:

- Lasso Regression (Least Absolute Shrinkage and Selection Operator)
- Random Forest Regression
- Extreme Gradient Boosting (XGBoost)
- Long Short-Term Memory Neural Networks (LSTM)

These models were selected due to their effectiveness in handling nonlinear relationships, time-series data, and their capacity to produce reliable and accurate forecasts. Each model was trained on the prepared dataset, and its performance was evaluated using appropriate metrics such as Mean Squared Error (MSE), Coefficient of Determination (R-squared ), Mean Absolute Error(MAE), and Root Mean Square Error(RMSE) .The inclusion of Lasso Regression provided an additional perspective by leveraging its regularization capabilities to handle multicollinearity and feature selection, ensuring a simpler and interpretable model. The iterative process allowed for fine-tuning all models to achieve optimal results.

- **Least Absolute Shrinkage and Selection Operator(LASSO)**

The first model that the group built was the Lasso Regression (Least Absolute Shrinkage and Selection Operator) model. It was the least complex from all the models since it is a linear regression method that incorporates regularization to prevent overfitting and improve prediction accuracy. It automatically reduces irrelevant features by setting their coefficients to zero, which helps us to get a simplified and interpretable model. Therefore, we can use this to be particularly effective in handling datasets with a high number of predictors, where irrelevant features can degrade performance.

To build these both models(short term 15 minutes forecasting and long term 1 hour forecasting) by using LASSO Regression technique, our group had taken irradiation, wind, temp, humidity, and barometer as the input variables and total active power was taken as the target variable.

In the Data preprocessing process multiple CSV files were merged and cleaned also, numerical columns were imputed with the median; categorical columns were filled with zero and time-based shifting generated target variables for 15-minute and 60-minute forecasts also, lagged features (up to 5 steps) were added, and rows with missing values were removed. And here we used standard scaler for consistency.

Lasso Regression was used for its ability to perform feature selection via regularization, reducing overfitting and noise. TimeSeriesSplit ensured temporal integrity during cross-validation, while GridSearchCV optimized the alpha parameter over a logarithmic scale. In here we set max iterations into a high value like 3000 for both models. This approach enhanced model robustness,

ensuring reliable and generalizable predictions for both short- and long-term forecasts. A next step trained model was tested on the prepared test dataset by using evaluation metrics like MSE, MAE, RMSE, and R-squared.

- **Random Forest Regression(RFR)**

Random Forest Regression was the second model that group was built; it was less complex compared to the other models that were built by group. This ensemble technique uses multiple decision trees to handle non-linear relationships and is robust against data noise. RF is well known for its adaptability, it can manage missing outliers, missing values, map nonlinear relationships and reduce the dimensions. In this study, RF was used to predict the solar power output .

To build these both models(short term 15 minutes forecasting and long term 1 hour forecasting) by using Random Forest Regression technique, our group had taken irradiation, wind, temp, humidity, and barometer as the input variables and total active power was taken as the target variable.

This model Utilizes multiple decision trees to improve predictive accuracy and control overfitting by averaging their outputs. For this model minimal feature scaling is required due to the non-parametric nature of decision trees. Our group trained this model on random subsets of data using bootstrapping, where multiple decision trees are grown and their predictions are averaged. Hyperparameter for both the models are set into n_estimators as 3000, max depth as 70, min_samples_split as 15 and random state as 42. A next step trained model was tested on the prepared test dataset by using evaluation metrics like MSE, MAE, RMSE, and R-squared.

- **Extreme Gradient Boosting(XGBoost)**

The third model developed for this project utilized Extreme Gradient Boosting (XGBoost), a robust and scalable machine learning algorithm. Compared to the other models—LASSO regression and Random Forest—XGBoost offered enhanced capability to capture nonlinear relationships, making it a more effective choice. XGBoost is renowned for its efficiency and flexibility, being well-suited for both regression and classification tasks due to its gradient boosting framework.

To achieve both short-term (15-minute) and long-term (1-hour) forecasting, two separate XGBoost models were built using the train_data and test_data datasets. The selected input variables included Irradiation, Wind Speed, Temperature, Humidity, and Barometric Pressure, with Total Active Power as the target variable.

**Data Preparation**

The data preparation process involved a custom function, create_lag_features, designed to generate lagged features for supervised learning. This function incorporated historical data to provide context for future predictions:

- **Short-term Forecasting**: Lag features were created for a 15-minute interval.
- **Long-term Forecasting**: Lag features were generated for a 120-minute interval.

Additionally, the target variable (Total Active Power) was shifted by the forecast horizon (15 minutes or 60 minutes) to align the inputs with the desired outputs. Rows with missing values resulting from this lagging and shifting process were removed to ensure data consistency. The final datasets were divided into:

- **Input Features**: x_train and x_test
- **Target Features**: y_train and y_test

**Model Architecture and Hyperparameters**

The XGBoost Regressor was configured with the following parameters to optimize its performance:

- **Objective**: reg:squarederror for regression tasks.
- **Number of Estimators (n_estimators)**: 50.
- **Learning Rate (learning_rate)**: 0.1 for balanced training.
- **Maximum Depth (max_depth)**: 8 to control overfitting.
- **Subsample and Column Subsampling (colsample_bytree)**: 0.8 for better generalization.

Both short-term and long-term models were trained using the same parameter configuration to maintain consistency.

**Training and Evaluation**

The models were trained on the prepared training datasets (x_train and y_train). Early stopping was implemented during training to avoid overfitting and ensure the models generalized well to unseen data.

After training, the models were tested on the test_data, and their performance was evaluated using the following metrics: MSE, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared.

- **Long Short Time Memory (LSTM)**

The fourth model developed was based on the Long Short-Term Memory (LSTM) neural network. Among the four models (LASSO regression, Random Forest, XGBoost, and LSTM), the LSTM model was the most complex, offering superior capability to capture nonlinear relationships in the data. Its strength lies in its ability to retain long-term dependencies, making it highly suitable for time-series forecasting and handling sequential data effectively.

To build the LSTM model for both short-term (15-minute) and long-term (1-hour) forecasting, two datasets were used: train_data and test_data. The selected input variables included Irradiation, Wind Speed, Temperature, Humidity, and Barometric Pressure, with Total Active Power as the target variable.

**Data Preprocessing**

The data preprocessing step was crucial for developing the LSTM model. A custom function, create_sequence, was used to transform raw time-series data into a supervised learning format. This function created input sequences using past values of the input variables over a specified time window and aligned them with the corresponding future target values.

- For 15-minute predictions, the sequence length was set to 15.
- For 1-hour predictions, the sequence length was set to 60.

To normalize the data, MinMaxScaler was applied to scale the input variables and target variable separately. The scalers were saved to ensure the same transformations were applied to the test data during evaluation.

**Model Architecture**

The architecture of the LSTM model consisted of:

- **Bidirectional LSTM Layer**: Processes input data in both forward and backward directions, capturing patterns effectively.
- **Dropout Layer**: Set at a 20% rate to prevent overfitting.
- **LSTM Layer**: Captures the complex patterns in sequential data.
- **Dense Layer**: Outputs the prediction for the target variable.

The model was compiled using the Adam optimizer and the Mean Squared Error (MSE) loss function to ensure reliable and accurate predictions.

**Training and Evaluation**

Both models (short-term and long-term) were trained using the following parameters:

- Dropout rate: 0.2
- LSTM units: 50
- Batch size: 32
- Epochs: 50

The trained models were evaluated using prepared test datasets, and performance metrics such as MSE, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared were used to assess their accuracy and generalization.

## 3.7 Final Model Selection

The final stage of the methodology consisted of model comparison and selection of the best performance based on evaluation metrics and practical considerations. Following the implementation and optimization of the four predictive models, LASSO Regression, Random Forest Regression, XGBoost, and Long Short-Term Memory-LSTM Neural Networks the group implemented a systematic evaluation to identify the most suitable model for solar energy forecasting.

Each model's performance was rigorously tested on the prepared test dataset, and their accuracy was evaluated using the following metrics:

**Mean Squared Error (MSE)**

The average squared differences between the predicted and actual values are measured by MSE. It is sensitive to notable variances because it gives greater weights to larger errors. This property is especially crucial in particular scenarios when significant forecasting errors are undesirable.MSE is calculated by equation (2).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \quad (2)$$

Here,$y_i$ stands for Actual value for the $i$ data point, $\widehat{y_i}$ stands for Predicted value for the $i$ data point and $n$ stands for the total number of data points.

**Coefficient of Determination (R-squared)**

The percentage of the dependent variable's variance that can be predicted from the independent variables is indicated by R-squared. It provides a goodness-of-fit metric for regression models, with values ranging from minus infinity to plus one. The model successfully captures the variation in the data when the value is near to 1. .R-squared is calculated by equation (3)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \overline{y_i})^2} \quad (3)$$

Here, $y_i$ stands for the Actual value for the $i$ data point, $\widehat{y_i}$ stands for the Predicted value for the $i$ data point, $\overline{y_i}$ stands for the mean of the actual values, numerator stands for the Residual sum of squares (unexplained variance) and denominator stands for the Total sum of squares (total variance).

**Root Mean Square Error(RMSE)**

The RMSE, or Root Mean Square Error, is a measure of the average magnitude of errors between predicted values and actual (measured) values. Its particular unequal sensitivity to large errors is due to squaring the differences before summing up, and thus, RMSE gives a more definite interpretation of the model's prediction accuracy, which is much more critical when larger deviations are to be observed.RMSE is calculated by equation (4).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2} \quad (4)$$

Here, $y_i$ represents the actual value for the $i$ data point, $\widehat{y_i}$ represents the predicted value for the $i$ data point, and $n$ is the total number of data points.

**Mean Absolute Error(MAE)**

MAE measures the average magnitude of errors between predicted and actual values. Contrary to RMSE, it does not square the difference making it less sensitive to outliers. MAE is easily interpretable as it directly represents average error in the same unit as that of the data. MAE is calculated by equation (5).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y_i}|^2 \quad (5)$$

Here, $y_i$ represents the actual value for the $i$ data point, $\widehat{y_i}$ represents the predicted value for the $i$ data point, and $n$ is the total number of data points.

## 3.8 Time Plan

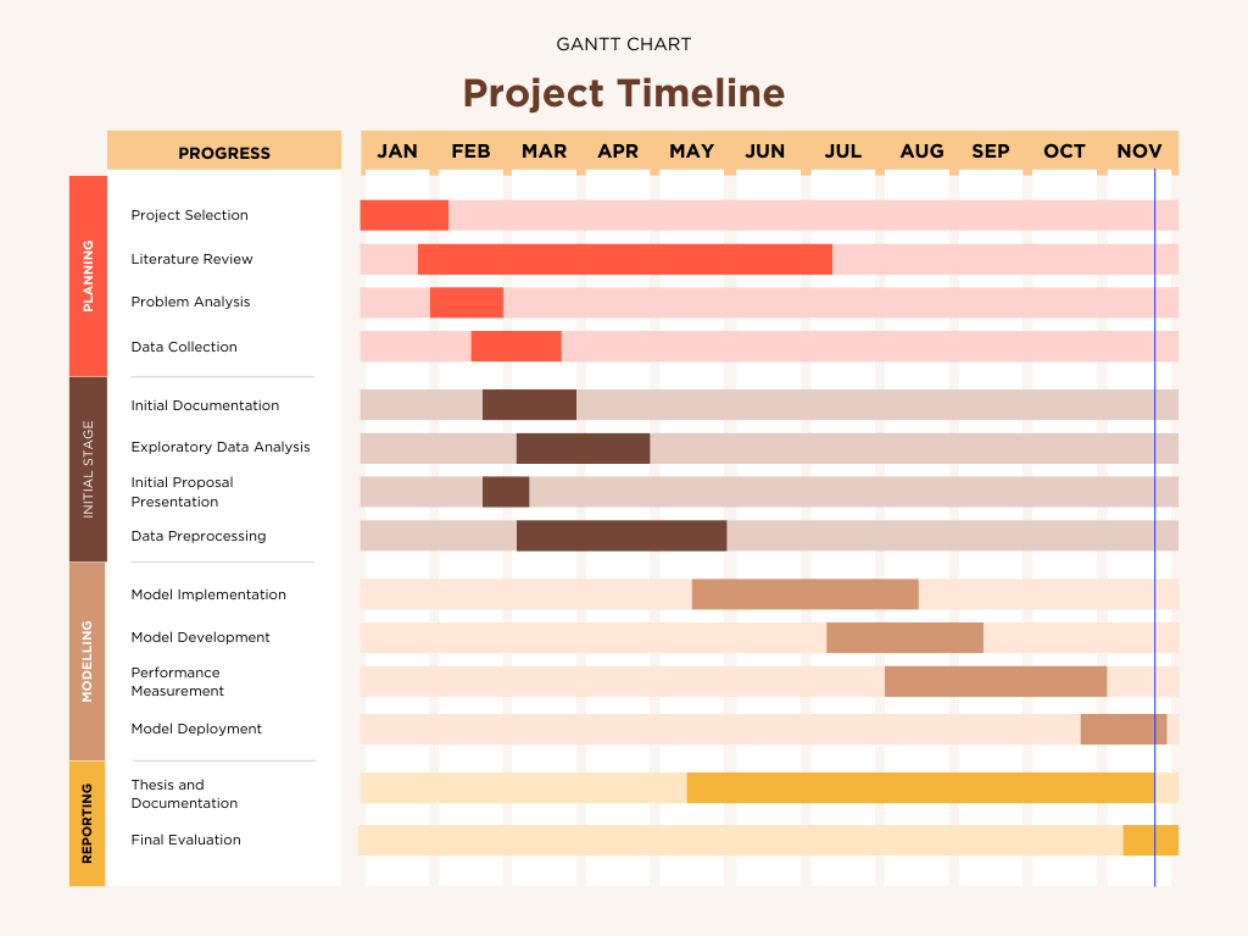



*Figure 3.4: Time Plan*

# Chapter 4: Tools and Technologies

This project used Google Colab and VS Code for collaborative machine learning, Python due to its versatility and support of libraries, Excel for data handling, and Power BI and R studio for finding trends through the visualization of data and also to analyze statistics in order to ensure collaboration is effective and that models are robust.

## 4.1 Microsoft Excel

Excel was used to organize, handle and explore the dataset, using its features such as filtering, sorting, and computations. These functionalities streamlined initial data preparation, ensuring readiness for advanced analysis and modeling.

## 4.2 Google Colab and VS Code

The main development and testing environment would be Google Colab and also here we used VS Code due to its network non-dependence. In here google colab and VS code both can collaborate in real time, smoothly integrate Python, and have free computational resources. Python would be the main language; it enables one to work through a typical workflow of data preparation to model evaluation. In here we used several libraries them are;

- NumPy-mathematical functions
- Pandas-data analysis and manipulation
- Matplotlib- data visualization
- SeaBorn- data visualization
- Keras- deep learning
- Tensorflow- machine learning
- SciPy- scientific computing

## 4.3 R Studio

R Studio was utilized for advanced statistical analysis, providing powerful tools for data analysis and also for some kind of data visualization. Its integration with R’s rich ecosystem of libraries facilitated efficient analysis, enhancing the overall workflow.

## 4.4 Power BI

Power BI has been used to visualize and analyze the data since it illustrates the relationships in the set of data much more clearly. Furthermore, it enables our group to transform difficult data into something actionable, hence being the key tool for understanding project findings.

## 4.5 Summary

The integration of tools like Google Colab, VS Code, Python, Excel, R studio, Power BI, along with powerful Python libraries, streamlined collaboration, data processing, and model development, contributed to the project's success in accurately forecasting solar power output.

# Chapter 5: Design and Analysis

## 5.1 Functional Requirements

- **Data Acquisition:** Data acquisition is a process of collection and combination of data from different sources. In this study we collected our data from two different sources. They are operational data from the solar power plant and collecting meteorological data from reliable sources. Exact and comprehensive data acquisition ensures that the predictive models are trained on relevant and high-quality inputs-a fact quite essential in generating reliable forecasts.
- **Data Preprocessing Software:** This implies data cleaning, data integration, data transformation, and data reduction of raw data for analysis with various software tools. To do this our group used python programming language and MS EXCEL. Data preprocessing ensures that the data is consistent and ready for modeling; it generally increases the efficiency and accuracy of machine learning algorithms.
- **Machine Learning Techniques:** In this study, advanced machine learning techniques have been developed in terms of predictive models that may comprehend and forecast the solar power output. Several of these are found quite useful in view of the complex and nonlinear relationship existing between different weather variables and energy production. This work implemented XGBoost, LSTM, Random Forest Regression, and the LASSO Regression models for accurate and reliable predictions.

## 5.2 Non-Functional Requirements

- **Prediction accuracy:** Prediction accuracy makes certain that the model provides forecasts that closely match actual solar power outputs. To get the correct forecasts is having a great impact in making informed operational decisions, such as grid management and energy storage. High accuracy minimizes forecasting errors, reducing the risk of operational inefficiencies.
- **Reliability:** Reliability can be described as how well the model performs consistently across different conditions, including weather and operational scenarios. A reliable model offers stakeholders dependable insight into the outputs of the system, engendering a sense of trust in the utility of the system itself.
- **Response Time:** The response time is the time the system takes to process the input data and provide the forecast. A small response time is essential in real-time applications, enabling operators to make decisions in time based on current and accurate forecasts. This makes the system effective for dynamic and fast-paced environments.

## 5.3 Software Requirements

Several software tools were utilized in this project to streamline the workflow and enhance efficiency. The primary software used includes Excel, Power BI, R Studio, Google Colab, and Visual Studio Code (VS Code).

**Google Colab** served as the primary Integrated Development Environment (IDE) for developing and implementing machine learning models. Its intuitive interface, free GPU availability, and seamless integration with Google Drive made it an ideal choice for training models and collaborating within the team.

For local development, **Visual Studio Code (VS Code)** was employed as an alternative IDE. Known for its lightweight, customizable features and support for Python, VS Code made local testing and debugging of code efficient before deployment on Google Colab.

**Python** was chosen as the programming language for its simplicity, versatility, and extensive library ecosystem. Libraries such as NumPy (mathematical functions), Pandas (data analysis and manipulation), Matplotlib and Seaborn (data visualization), Keras (deep learning), TensorFlow (machine learning), and SciPy (scientific computing) were used for data preprocessing, analysis, and model building.

**Excel** was utilized for data organization and preliminary analysis. Its capabilities in filtering, sorting, and formula-based computations made it a practical tool for cleaning and managing datasets.

**Power BI** enabled advanced visualization and exploration of data trends, providing valuable insights into hidden patterns and correlations. It helped identify inconsistencies, missing values, and relationships between variables.

**R Studio** was used for statistical analysis, complementing Python's data handling and machine learning capabilities.

## 5.4 Insights from Visualization Techniques

Visualization tools such as Power BI, Matplotlib, and Seaborn were instrumental in uncovering critical insights, as highlighted in the following figures:

- **Seasonality Identification:** Seasonal variations play a crucial role in solar power output. The analysis revealed recurring daily and monthly patterns, emphasizing the significance of seasonality in the data.

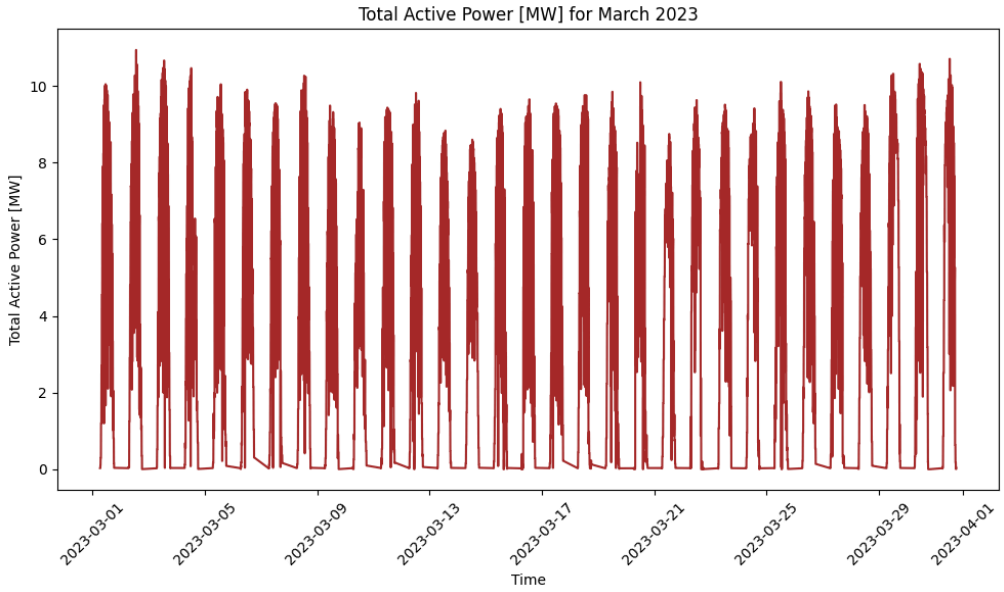



*Figure 5.1: Seasonality Identification Plot*

The Figure 5.1 shows the total active power output for March 2023, highlighting significant fluctuations that correspond with daylight hours. Such patterns are crucial for optimizing power management and predicting peak load times.

- **Comparison of Current Phases and Total Active Power:** Figure 5.2 illustrates the relationship between Current Phase A, B, C and Total Active Power over a specific period. Each line in the plot represents the trend of a current phase in comparison to the total active power.

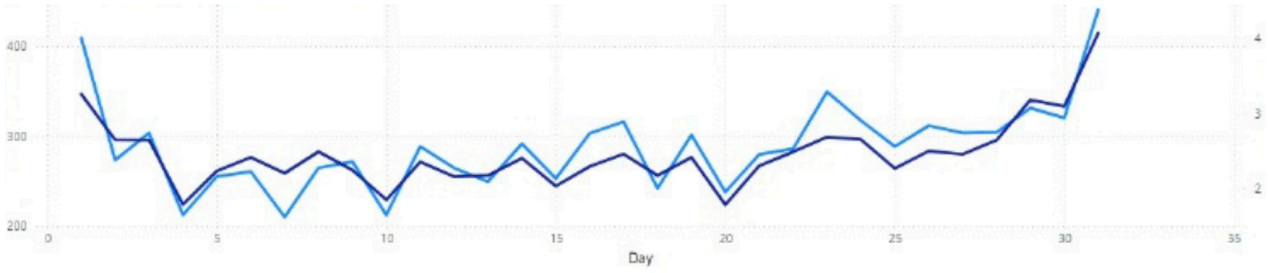
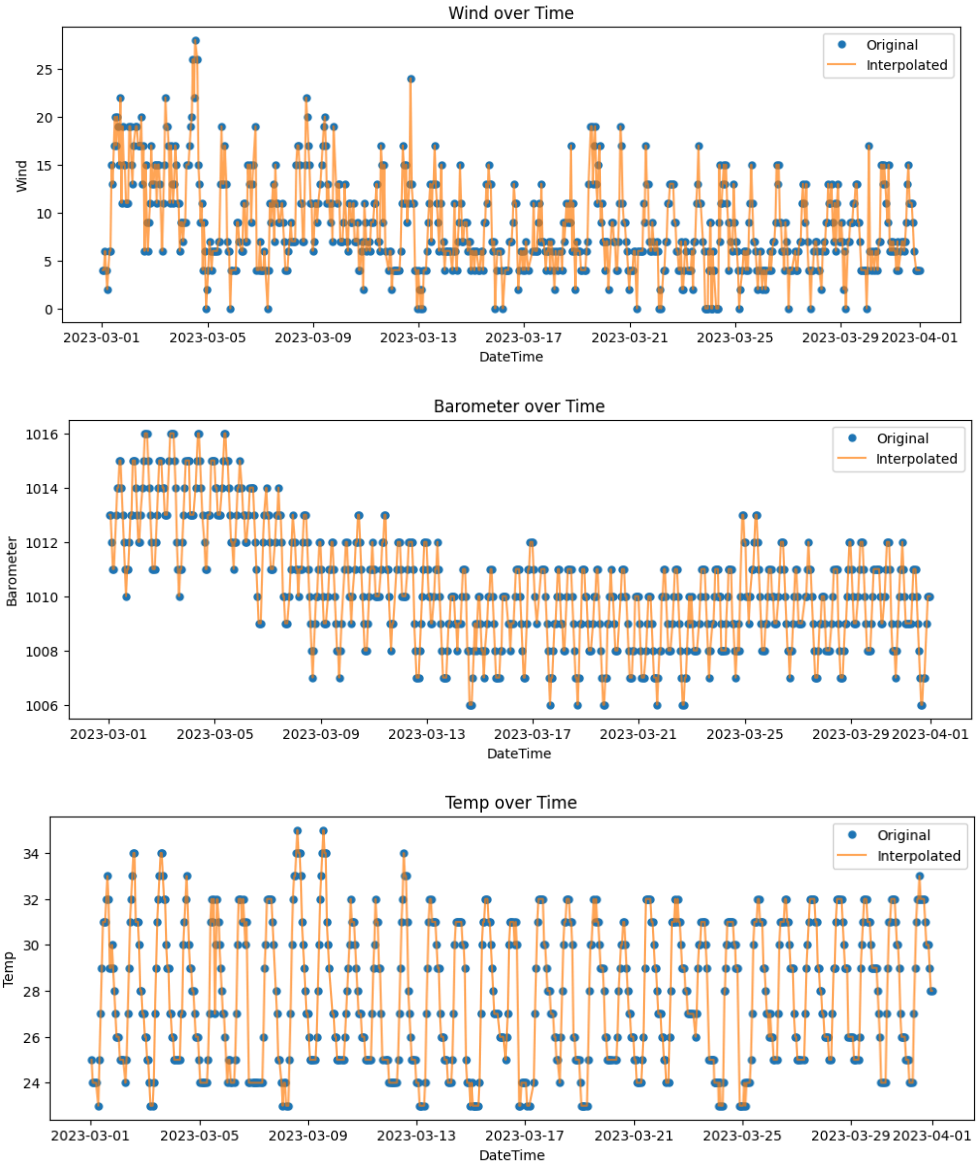



*Figure 5.2: Plot for Current Phase A,B,C compares to the Total Active Power*

As shown in Figure 5.2, Current Phase A, B, and C exhibit similar trends and patterns when compared to Total Active Power, indicating a strong correlation between these variables. This redundancy suggests that including all three current phases in the model would not add significant value to the predictive performance.

To streamline the dataset and enhance model efficiency, Total Active Power was selected as the representative variable for modeling, while Current Phase A, B, and C were excluded. This approach ensures that the model maintains its accuracy while avoiding unnecessary complexity.

- **Data Interpolation:** Data gaps in the dataset were addressed using advanced interpolation techniques, ensuring a continuous and accurate dataset for analysis and model training.



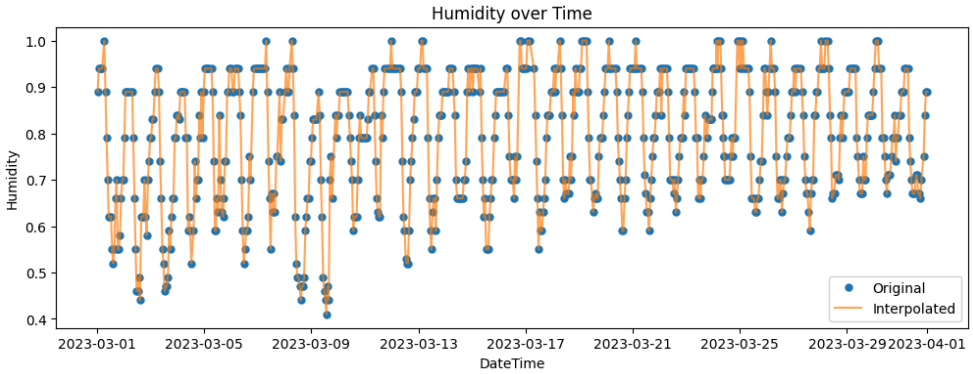



*Figure 5.3: Data Interpolation Plot*

The Figure 5.3 demonstrates the original data points with gaps and the interpolated values used to create a seamless dataset for training machine learning models.

- **Correlation Analysis**

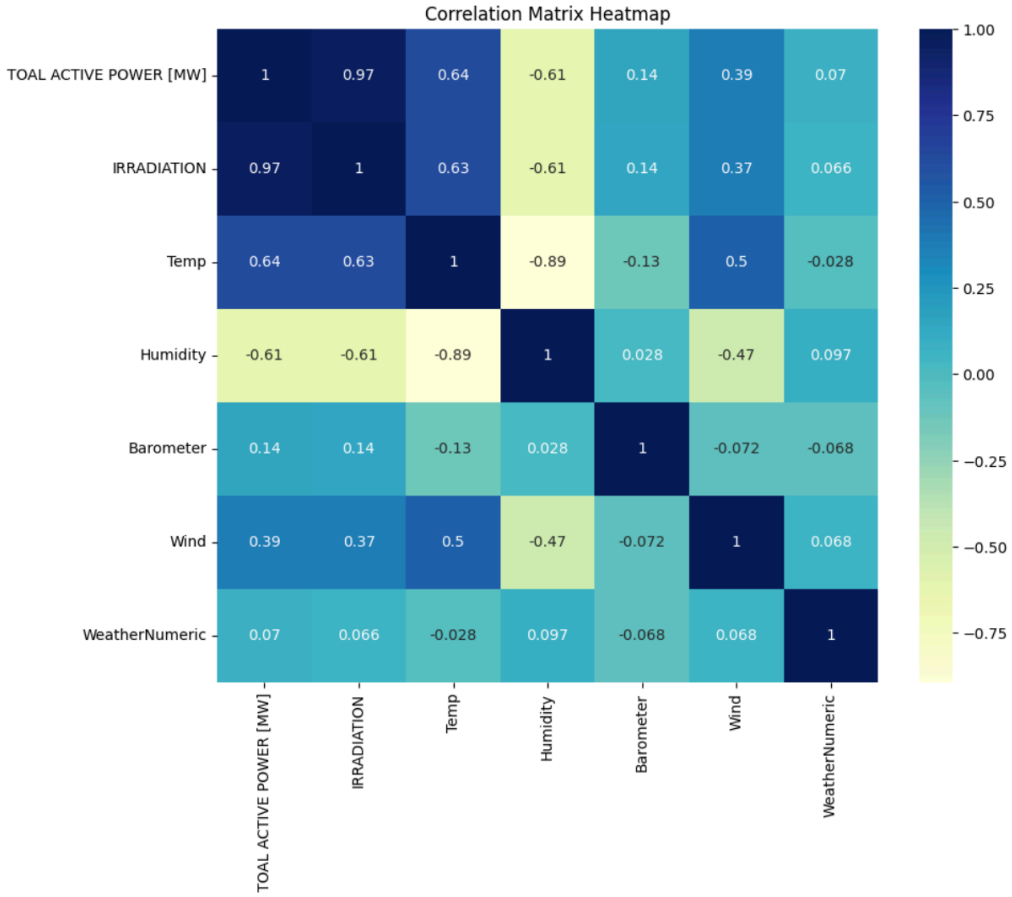



*Figure 5.4: Correlation Map for for Whole Data Analysis*

According to Figure 5.4, the correlation heatmap highlights significant relationships among the dataset variables. Irradiation shows a strong positive correlation with total active power (0.97), confirming its role as a primary predictor of solar power output. Similarly, temperature demonstrates a positive correlation (0.64) with total active power, emphasizing its influence on panel efficiency.

Wind speed exhibits a moderate positive correlation (0.39), indicating its partial impact on cooling and efficiency. On the other hand, humidity has a negative correlation (-0.61) with total active power, suggesting its adverse effect on solar performance. Barometric pressure shows an insignificant correlation (0.14), making it less relevant for modeling purposes.

These insights guided the selection of relevant variables for further predictive modeling.

- **Feature Importance Analysis:** Figure 5.5 illustrates the Feature Importance Plot derived using Lasso Regression, showcasing the significance of each feature in predicting solar power output. The coefficient values on the x-axis represent the contribution of each feature, with higher absolute values indicating greater importance.

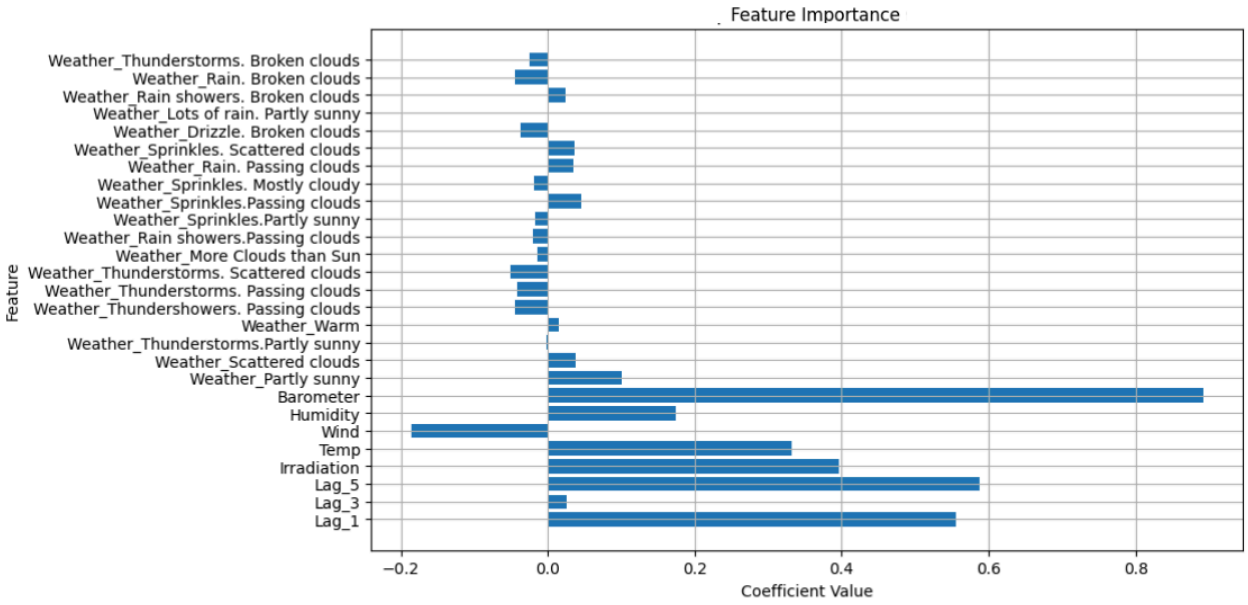



*Figure 5.5: Lasso Regression - Feature Importance*

Key variables, such as irradiation, temperature, wind speed, humidity, and barometric pressure, display significant coefficients, highlighting their strong influence on solar power prediction. In contrast, features like weather conditions and specific lag variables exhibit minimal or zero contributions, indicating their limited relevance to the model.

Based on the insights from this analysis, the following variables were selected for further modeling: Irradiation, Temperature, Wind Speed, Humidity, and Barometric Pressure.

The weather conditions were excluded due to their negligible importance, as confirmed by both the Correlation Map and this Lasso Feature Selection analysis. This refined feature selection ensures the inclusion of variables with meaningful predictive power, optimizing the model for accurate solar power forecasting.

## 5.4 Analysis of Methodology

In the analysis of methodology phase, a structured approach was followed for developing a data-driven solution on the prediction of solar power output. The identified steps in doing so are as follows:

1. Problem Definition: Defining the problem at hand regarding the prediction of Total Active Power using variables like weather variables and irradiation with high accuracy and reliability.
2. Data Collection: Collected operational data from the Vydexa Solar Power Plant and weather data from online sources to ensure completeness of the dataset.
3. Data Exploration and Cleaning: Researched structure and nature through the identification of missing values in the data set, and inconsistency of data.
4. Feature Selection: Identified key features that highly influence the solar power output, such as irradiation, temperature, wind speed, humidity and Barometer.
5. Model Selection, Training, and Evaluation: Carried out different machine learning models and tuned their parameters accordingly to have a better performance analysis by metrics like Mean Squared Error, $R^2$ score, Mean Absolute Error(MAE), Root Mean Squared Error(RMAE).

## 5.5 Summary

We used Excel, Power BI, R Studio, Google Colab, and VS Code. Our main IDE for model development would be Google Colab, while VS Code supported local coding. Python libraries NumPy and TensorFlow were used in data analysis and modeling. The visual insight from Power BI showed insignificant correlation with Total Active Power; repetition in the data, especially in January, missing data in February, and similar patterns across phases A, B, and C.

# Chapter 6: Model Evaluation & Results

## 6.1 Lasso Regression Model

Lasso Regression, or Least Absolute Shrinkage and Selection Operator, is a linear regression technique that incorporates L1L1-regularization to improve model generalization and accuracy. By adding a penalty proportional to the absolute values of the regression coefficients, Lasso forces some coefficients to shrink to exactly zero. This not only aids in preventing overfitting but also acts as a feature selection mechanism, retaining only the most influential predictors in the model. Given its strengths, Lasso Regression was employed to explore its predictive capacity for the solar power output dataset.

**Quantitative Metrics**

The performance of the Lasso Regression model was assessed using the following metrics:

The performance of the **Lasso Regression** model was evaluated for both short-term (15-minute) and long-term (1-hour) forecasting using the following metrics:

1. **Mean Absolute Error (MAE):** The model achieved an MAE of **2.088** for 15-minute forecasting and **2.574** for 1-hour forecasting, reflecting the average absolute difference between predicted and actual values. These values indicate reasonable predictive accuracy with slightly higher errors in the longer-term predictions.
2. **Root Mean Squared Error (RMSE):** The RMSE was calculated to be **2.592** for 15-minute forecasting and **2.961** for 1-hour forecasting. This metric penalizes larger errors more heavily and highlights areas where predictions deviate significantly from observed values.
3. **Mean Squared Error (MSE):** The model attained an MSE of **6.723** for 15-minute forecasting and **8.772** for 1-hour forecasting, indicating the squared differences between actual and predicted values. The higher MSE for the 1-hour forecast suggests greater prediction uncertainty over longer timeframes.
4. **R-Squared ($R^2$):** The $R^2$ value was **0.398** for 15-minute forecasting and **0.214** for 1-hour forecasting. These values show that the model explained approximately 39.8% of the variance in solar power output for short-term predictions and 21.4% for long-term predictions. While moderate, these results underline the need for further refinement, especially for long-term forecasts.

**Alpha (Regularization Strength):**
The alpha parameter was set to **0.028** for 15-minute forecasting and **0.032** for 1-hour forecasting. These values balance the complexity of the model while preventing overfitting.

**Visual Analysis for Lasso Model**

The performance of the Lasso regression model was evaluated using graphical visualizations for both 15-minute and 1-hour forecasts:

- **Full Time-Series Comparison**: The full time-series graphs for both 15-minute and 1-hour forecasts demonstrate that the Lasso model successfully captures the general trends and seasonality of solar power generation. However, discrepancies are visible in both cases during periods of high variability, with notable deviations at peaks and troughs. These deviations highlight the model's limitations in handling extreme fluctuations in solar power output.
- **Zoomed Analysis**: A detailed view of specific intervals reveals the model's ability to approximate daily patterns and sharp transitions. For the 15-minute forecast, the Lasso model captures the short-term variations reasonably well but struggles during abrupt changes, leading to mismatches in some areas. Similarly, for the 1-hour forecast, the model effectively tracks longer-term trends but occasionally lags behind or overestimates during rapid shifts, reflecting its challenges in managing dynamic changes over extended periods.

**Short-Term Forecasting (15 Minutes)**

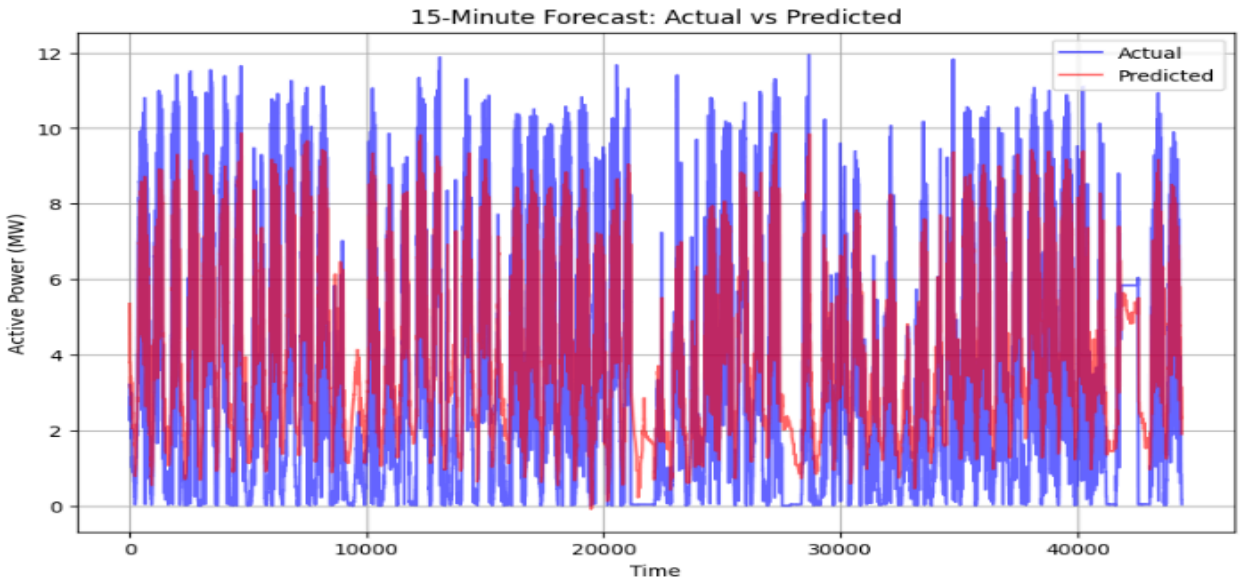



*Figure 6.1: Lasso Regression 15-Minute Comparison of Actual and Predicted Power over Time.*

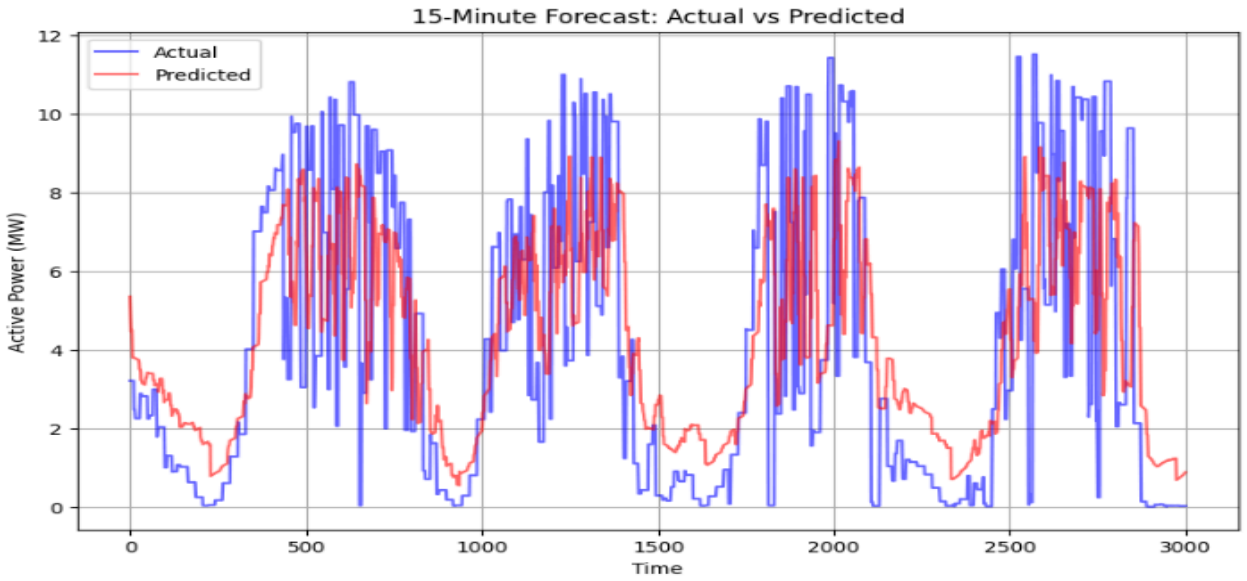



*Figure 6.2: Lasso Regression 15-Minute Prediction for the First 3000 Data Points.*

**Long-Term Forecasting (1 Hour)**

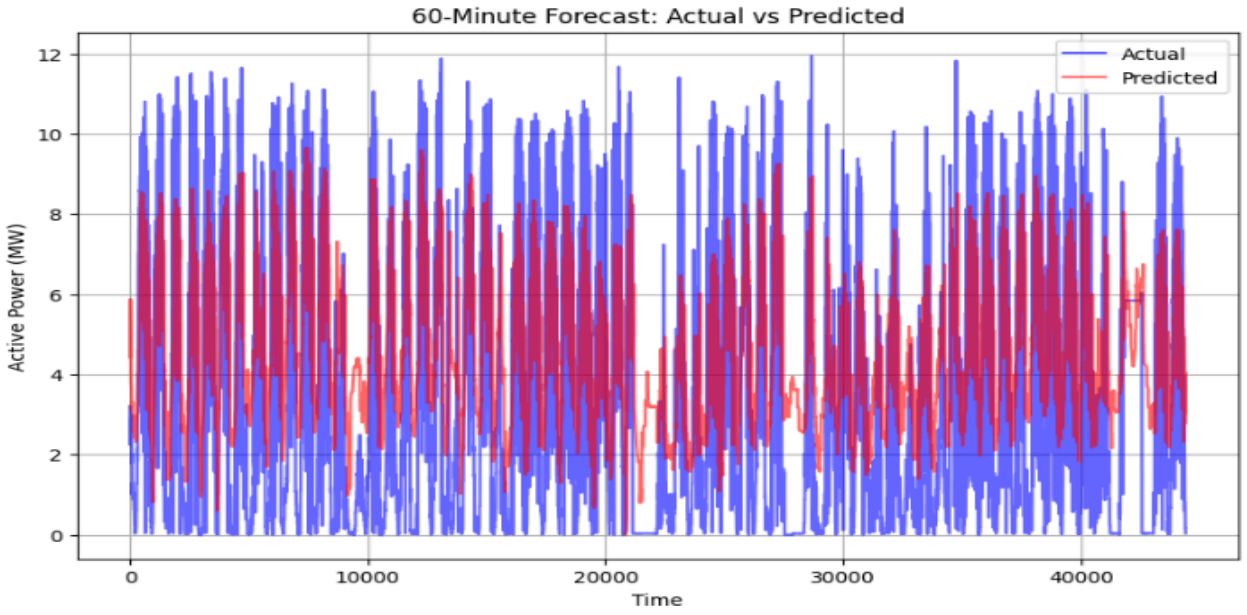



*Figure 6.3: Lasso Regression 1-Hour Comparison of Actual and Predicted Power over Time.*

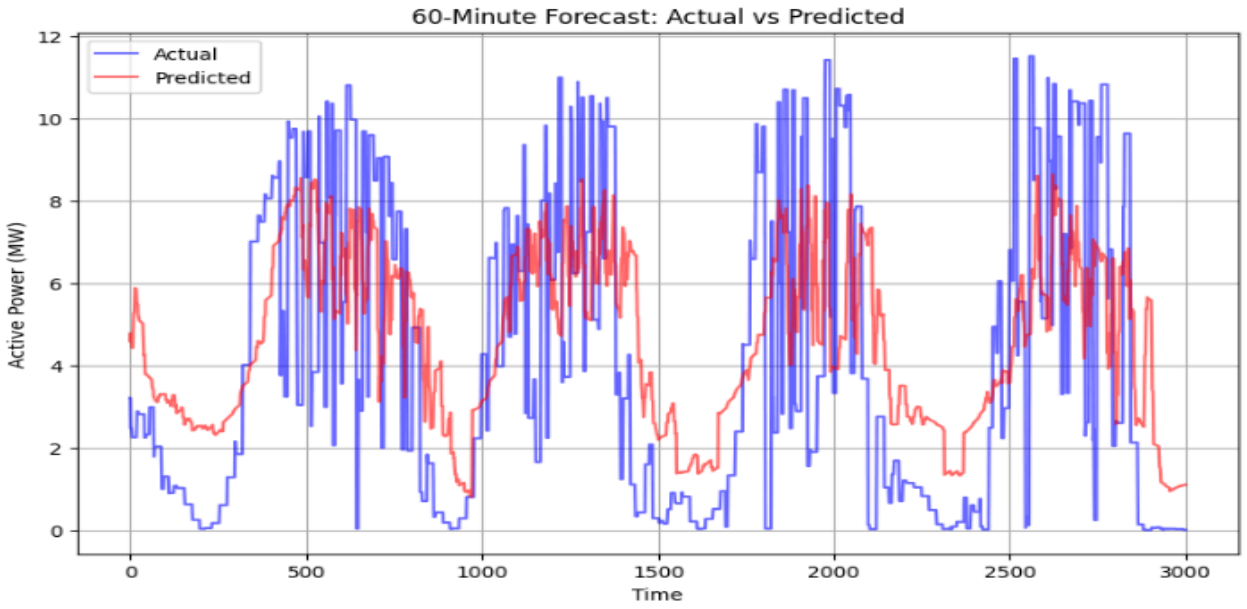



*Figure 6.4: Lasso Regression 1-Hour Prediction for the First 3000 Data Points.*

## 6.2 Random Forest Regression Model

The Random Forest Regression model was evaluated to assess its ability to forecast the total active power output of a solar power plant based on weather conditions. The evaluation focused on quantitative metrics, visual comparisons, and an analysis of the model's strengths and limitations.

**Quantitative Metrics**

1. **Mean Squared Error (MSE)**: The model reported an MSE value of **3.034** for 1 hour forecasting and **3.031** for 15 minute forecast, reflecting the average squared difference between the actual and predicted solar power outputs. While this error indicates room for improvement in predictive accuracy, it is acceptable for modeling non-linear time-series relationships.
2. **Coefficient of Determination (R-squared)**: An R-squared value of **0.657** for 1 hour forecasting and **0.657** for 15 minute forecast was obtained, demonstrating that the model explains approximately 65.70% of the variance in the target variable. This value suggests that the Random Forest model captures significant patterns in the data but may be limited in handling certain complex dependencies.

### Visual Analysis

To complement the numerical evaluation, the model's performance was analyzed through graphical representations:

1. **Full Time-Series Comparison**: The actual versus predicted power output graph shows a general alignment, with the model successfully capturing the overall trend and seasonality in the solar power generation. However, discrepancies are evident during certain peak and trough intervals, where the model appears to under- or over-predict.
2. **Zoomed Analysis of Initial Data Points**: A closer examination of the initial 1000 data points highlights the model's ability to approximate sharp transitions and daily patterns. Nevertheless, some deviation during abrupt changes indicates that the model may face challenges in precisely modeling rapid fluctuations in solar power output.

### Short-Term Forecasting (15 Minutes)

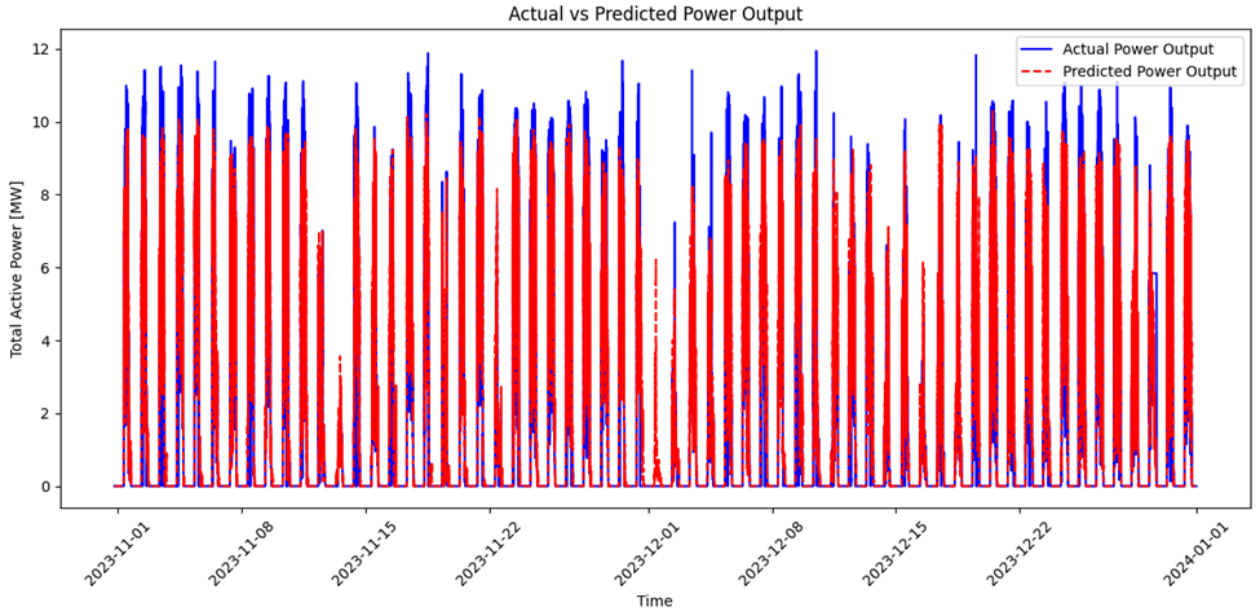



*Figure 6.5: Random Forest 15-Minute Comparison of Actual and Predicted Power over Time.*

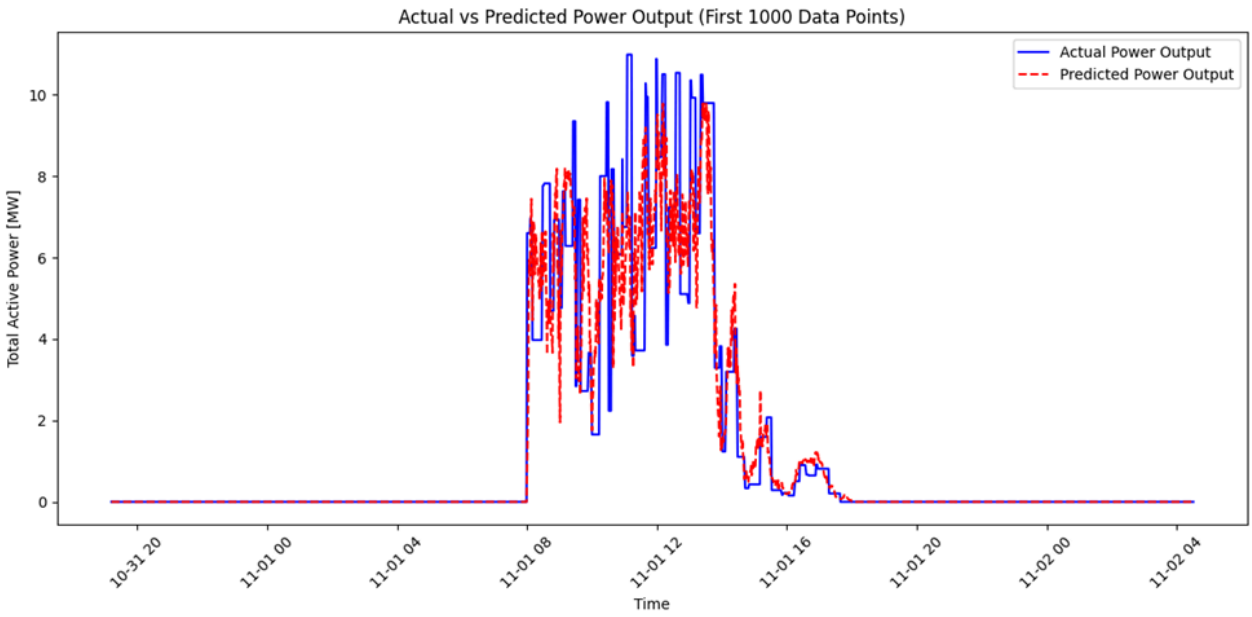



*Figure 6.6: Lasso Regression 15-Minute Prediction for the First 1000 Data Points.*

## Long-Term Forecasting (1 Hour)

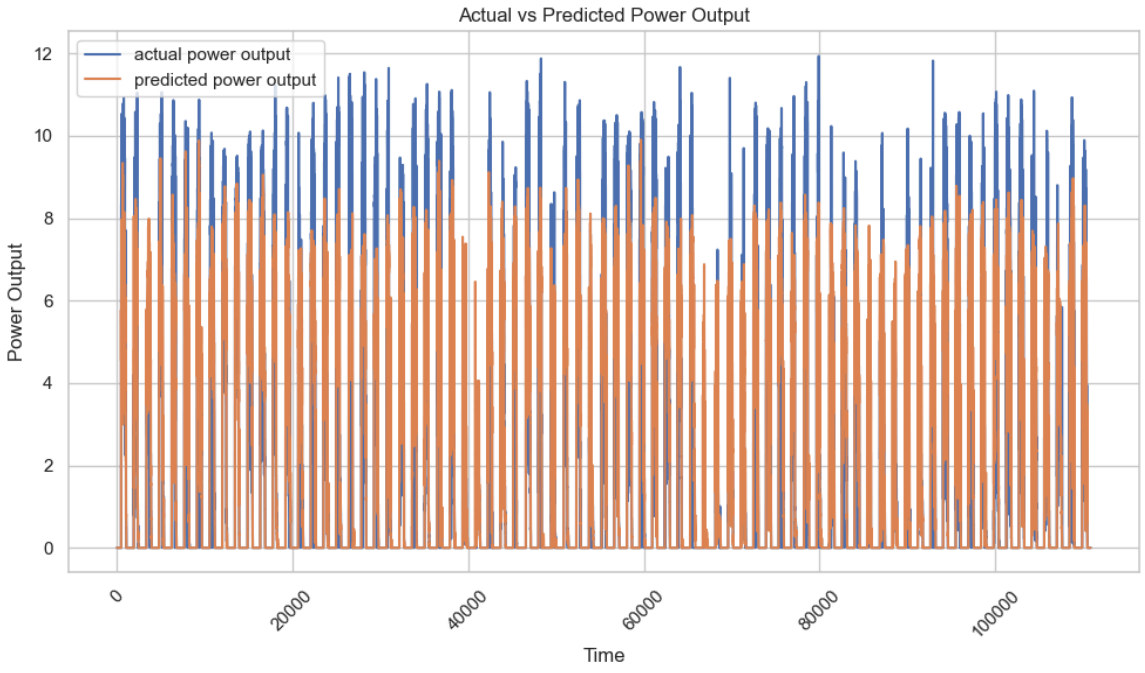



*Figure 6.7: Random Forest 1 hour Comparison of Actual and Predicted Power over Time.*

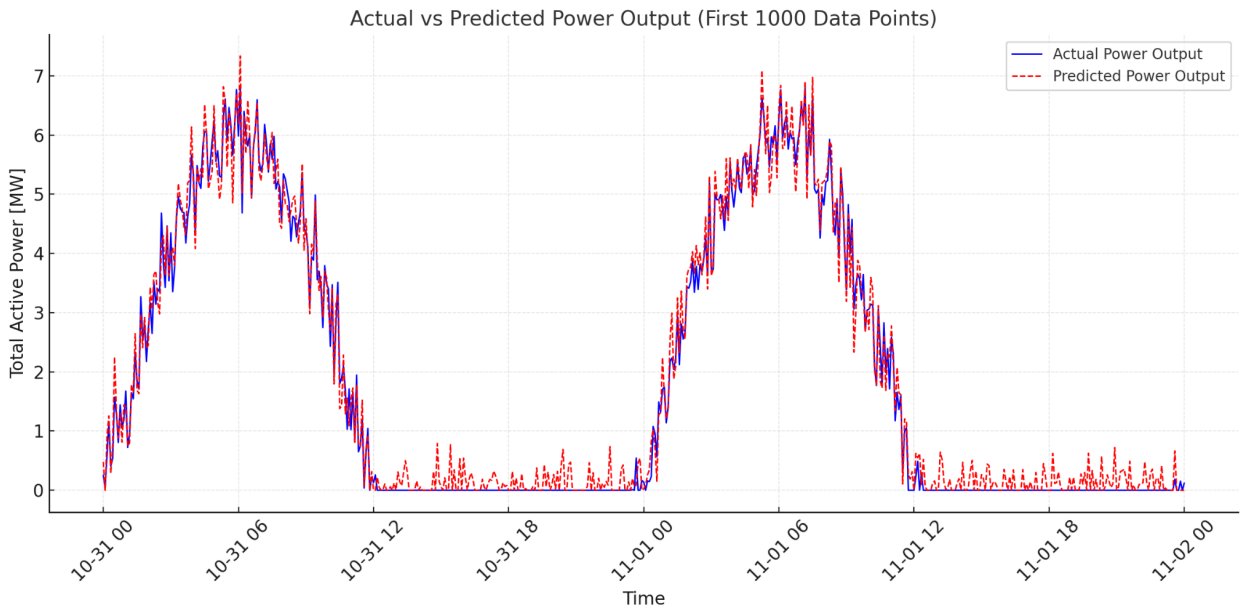



*Figure 6.8: Lasso Regression 1 hour Prediction for the First 1000 Data Points.*

**Model Robustness**

The Random Forest model incorporates several hyperparameter configurations to enhance its robustness and adaptability to the dataset:

- **Number of Trees (n_estimators)**: A high value of 3000 trees improves stability and reduces variance in predictions by averaging results across a large ensemble.
- **Tree Depth (max_depth)**: Limiting the maximum depth to 70 prevents overfitting while allowing sufficient model complexity to capture non-linear relationships.
- **Minimum Samples Split (min_samples_split)**: This parameter, set to 15, ensures that each tree node splits only if a minimum number of samples is present, enhancing the model's resilience to noise.

## 6.3 Extreme Gradient Boosting

The XGBoost (Extreme Gradient Boosting) model was evaluated to assess its effectiveness in predicting the total active power output of a solar power plant using weather features. The evaluation considered both quantitative performance metrics and visual analysis to determine the model's accuracy and reliability.

**Quantitative Metrics**

1. **Mean Squared Error (MSE):** For the **15-minute forecast**, the MSE was **2.608**, reflecting a moderate average squared difference between predicted and actual values. For the **1-hour forecast**, the MSE was **3.828**, indicating a slight increase in prediction error over longer intervals.
2. **Root Mean Squared Error (RMSE):** The **15-minute forecast** achieved an RMSE of **1.615**, demonstrating that the model effectively minimized large deviations in predictions over short-term intervals. For the **1-hour forecast**, the RMSE was **1.956**, showing reasonable performance but with slightly higher penalties for larger errors in long-term predictions.
3. **Mean Absolute Error (MAE):** The **15-minute forecast** achieved an MAE of **0.794**, highlighting the model's ability to maintain a low average error between predicted and actual values in short-term predictions. For the **1-hour forecast**, the MAE was **1.135**, indicating a slightly higher average deviation over extended forecasting periods.
4. **R-squared ($R^2$):** The **15-minute forecast** had an $R^2$ value of **0.705**, demonstrating that the model explained 70.5% of the variance in solar power output, indicating strong short-term predictive power. The **1-hour forecast** achieved an $R^2$ value of **0.56**, showing that the model captured 56% of the variance in long-term forecasts, though there is room for improvement.

**Visual Analysis**

The performance of the XGBoost model was further analyzed through visual comparisons of actual versus predicted power outputs:

1. **Overall Time-Series Trends**: The full dataset comparison reveals that the model accurately captures daily and seasonal patterns in solar power generation. The alignment between the actual and predicted curves is particularly evident during stable weather conditions. However, slight deviations during periods of abrupt change suggest areas for further refinement.
2. **Zoomed Examination of Initial Data Points**: A focused view of the first 3000 data points provides insights into the model’s ability to track short-term fluctuations. The predicted values closely follow the observed peaks and troughs, highlighting the model’s capacity to respond to rapid variations. However, minor underestimations during sharp peaks underscore the need for improved handling of extreme cases.

## Short-Term Forecasting (15 Minutes)

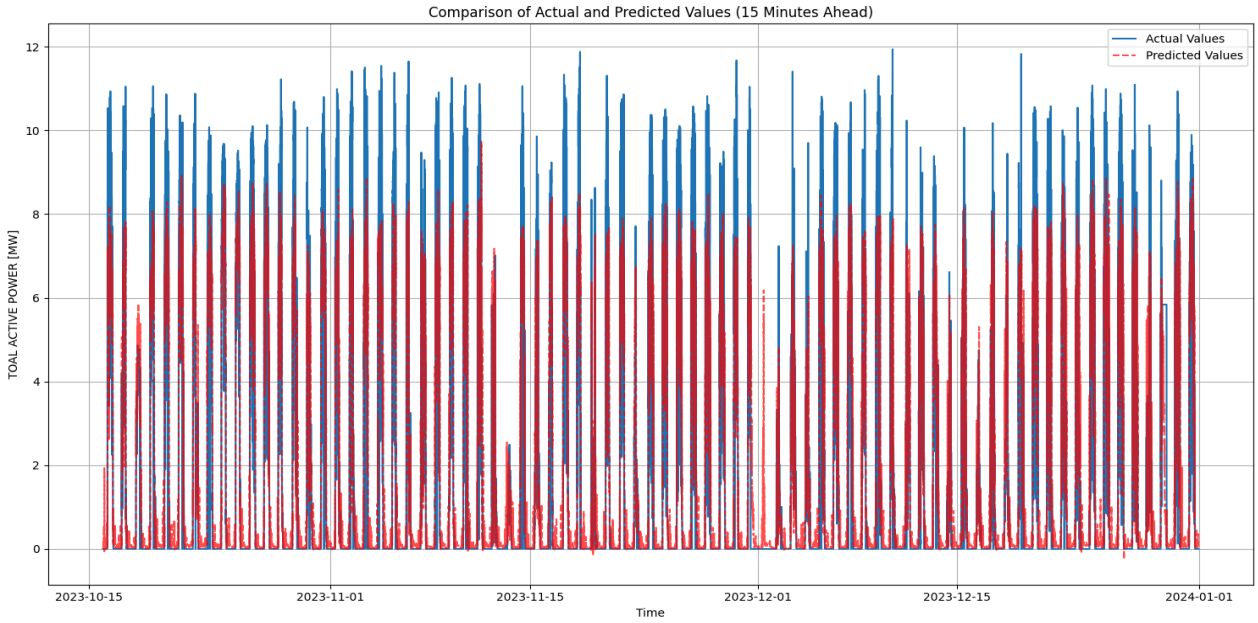



*Figure 6.9: XGBoost 15-Minute Comparison of Actual and Predicted Power over Time*

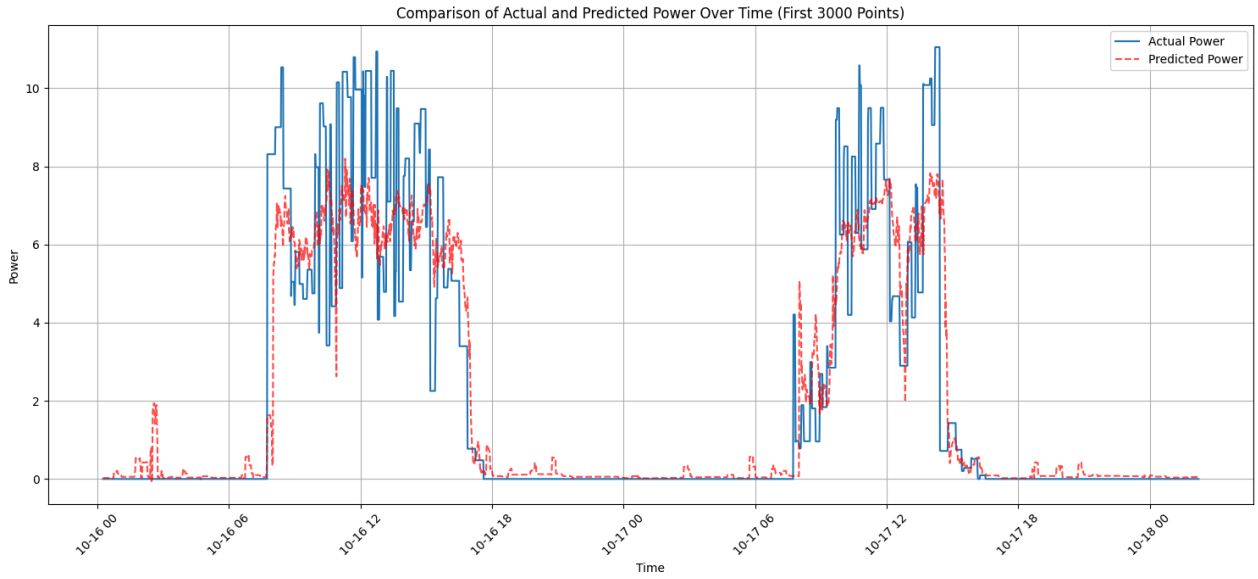



*Figure 6.10: XGBoost 15-Minute Prediction for the First 3000 Data Points.*

## Long-Term Forecasting (1 Hour)

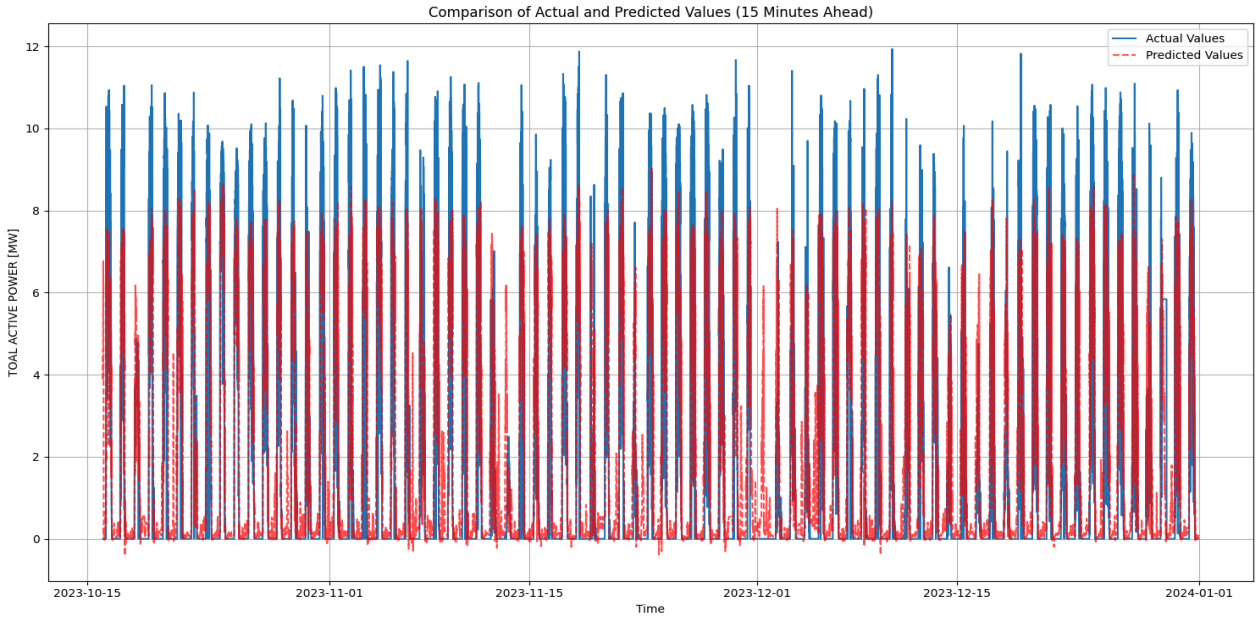



*Figure 6.11: XGBoost 1-Hour Comparison of Actual and Predicted Power Over Time.*

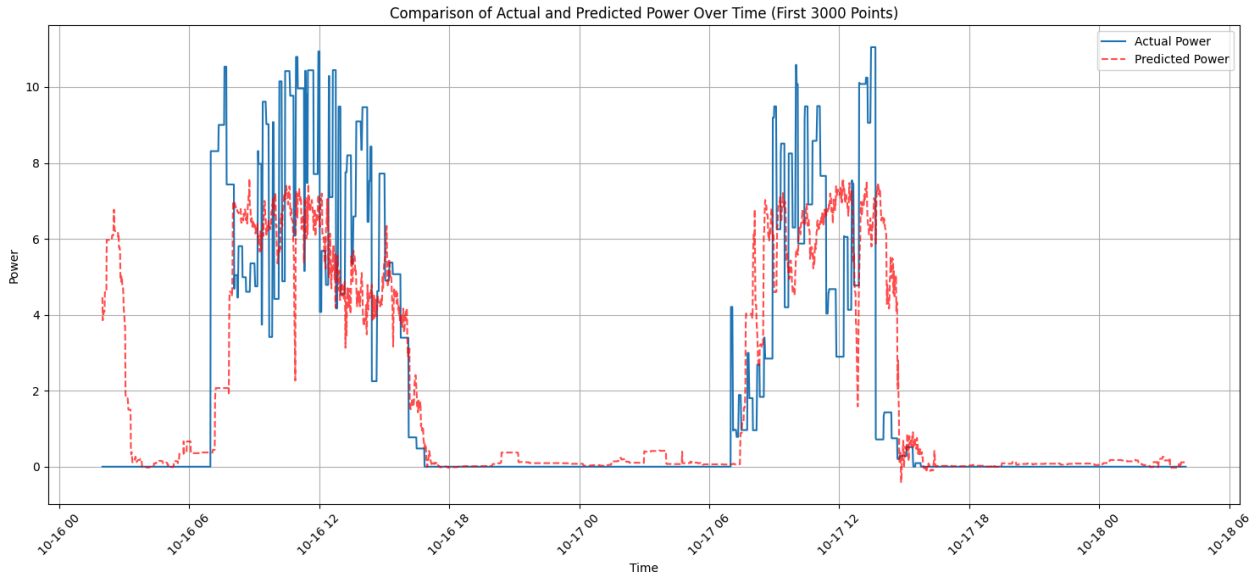



*Figure 6.12: XGBoost  1-Hour Prediction for the First 3000 Data Points.*

**Model Robustness**

The XGBoost model's architecture incorporates features and hyperparameter tuning aimed at enhancing robustness and performance:

- **Number of Boosting Rounds (n_estimators)**: Set to 50, this parameter balances training time and model complexity, ensuring consistent improvements in prediction accuracy without overfitting.
- **Tree Depth (max_depth)**: A value of 8 was chosen to prevent overfitting while retaining sufficient granularity to capture intricate patterns in the data.
- **Learning Rate**: By setting the learning rate to 0.1, the model achieves a balance between convergence speed and generalization, ensuring stability during training.

The inclusion of early stopping mechanisms further contributes to the model's resilience by halting training when no substantial improvement is observed, thereby reducing the risk of overfitting.

## 6.4 Long Short-Term Memory (LSTM )

The performance of the proposed Long Short-Term Memory (LSTM) model was evaluated using key metrics and visual analyses to assess its predictive accuracy and generalization capability for solar power output forecasting.

**Quantitative Metrics**

1. **Mean Squared Error (MSE)**: The model achieved an MSE value of **1.746** for 1 hour forecasting and **1.169** for 15 minute forecast, indicating a low average squared difference between the predicted and actual power output values. This highlights the model's ability to produce forecasts with minimal deviation, thereby demonstrating precision in capturing the underlying patterns in the time-series data.
2. **Coefficient of Determination (R-squared)**: An R-squared value of **0.802** for 1 hour forecasting and **0.811** for 15 minute forecast, was obtained, suggesting that approximately 81.1% of the variance in the 15 minute short term prediction and it's actual solar power output and is explained by the model. This high R-squared value signifies strong goodness of fit and reinforces the suitability of the LSTM model for forecasting tasks in this domain.
3. **Root Mean Squared Error (RMSE):** 15-Minute Forecasting, The RMSE of **1.292** penalizes larger deviations, showcasing the model's robustness for dynamic short-term variations. 1-Hour Forecasting, The RMSE of **1.321** indicates reliable long-term performance, though slightly less precise compared to the short-term forecast.
4. **Mean Absolute Error (MAE):** 15-Minute Forecasting, With an MAE of **0.578**, the model demonstrated consistent accuracy in capturing short-term variations. 1-Hour Forecasting,

An MAE of **0.643** suggests slightly higher errors for longer forecasts, yet still within an acceptable range for time-series tasks.

### Visual Analysis

To further substantiate the quantitative findings, a comparison of actual versus predicted power outputs was conducted through graphical visualization[M3] :

1. **Overall Time-Series Fit**: The plot displaying actual and predicted values across the entire test period shows a strong overlap between the two series. The LSTM model successfully captures both the periodic fluctuations and the dynamic variations in solar power output over time, particularly during periods of high activity.
2. **Zoomed View for Critical Intervals**: A closer inspection of critical intervals reveals the model's capability to track sudden changes in power generation with reasonable accuracy. While minor deviations are visible during peaks and troughs, the overall alignment suggests that the model effectively learns temporal dependencies.

### Short-Term Forecasting (15 Minutes)

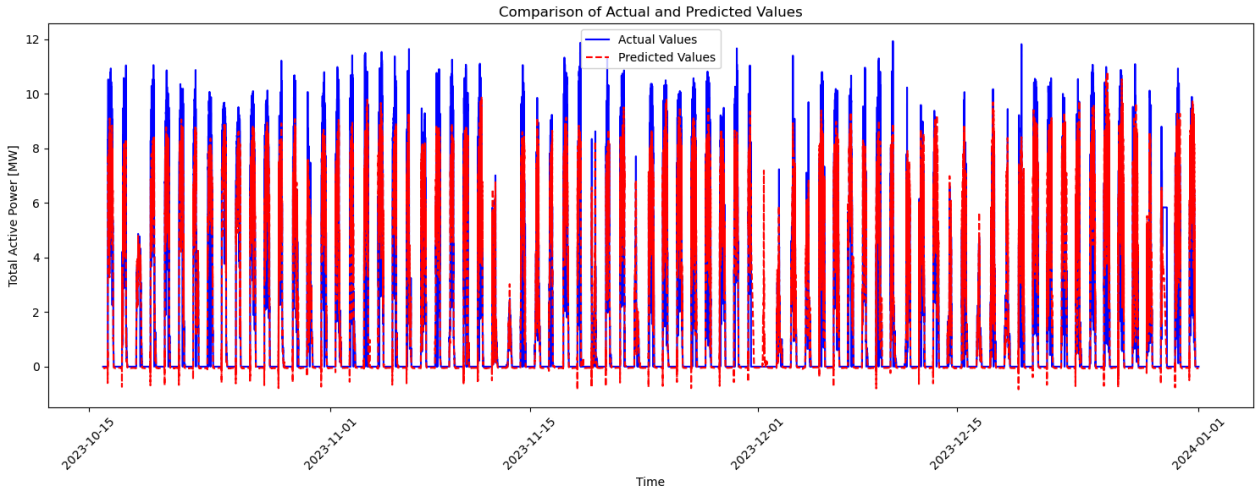



*Figure 6.13 : LSTM 15-Minute Comparison of Actual and Predicted Power Over Time.*

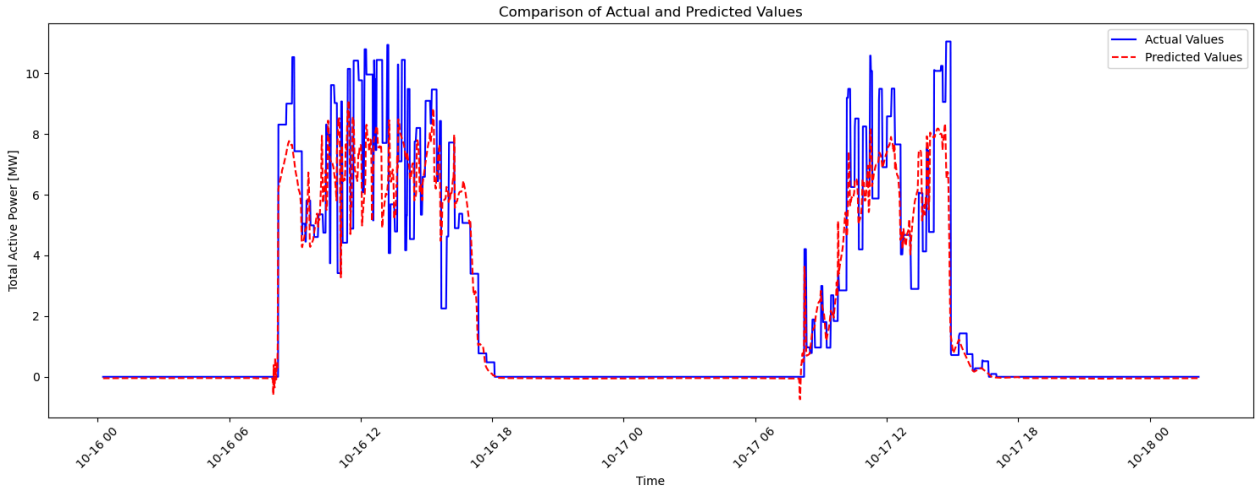



*Figure 6.14: LSTM 15-Minute Prediction for the First 3000 Data Points.*

## Long-Term Forecasting (1 Hour)

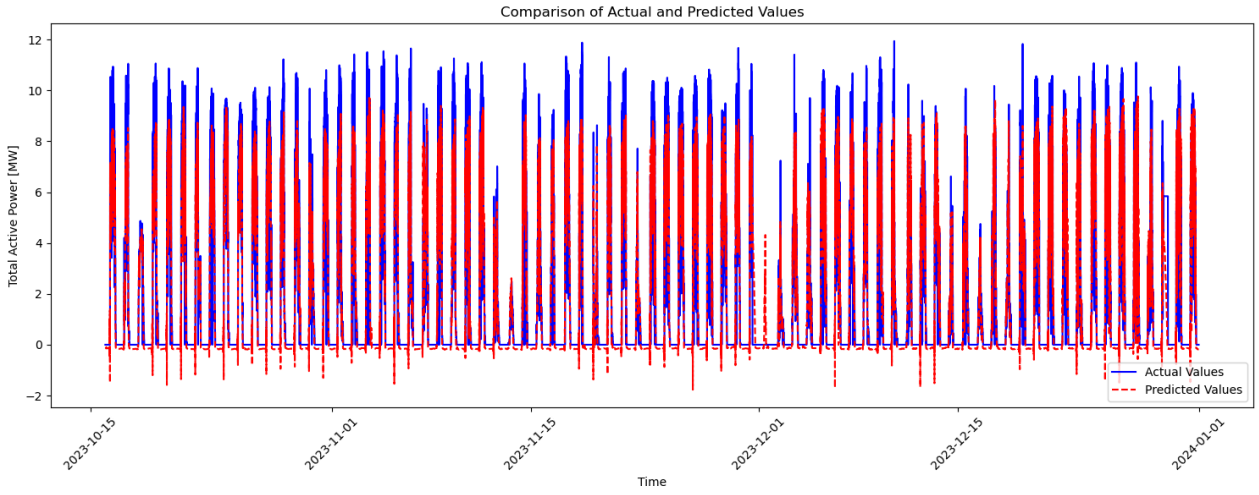


*Figure 6.15: LSTM 1-Hour Comparison of Actual and Predicted Power Over Time.*

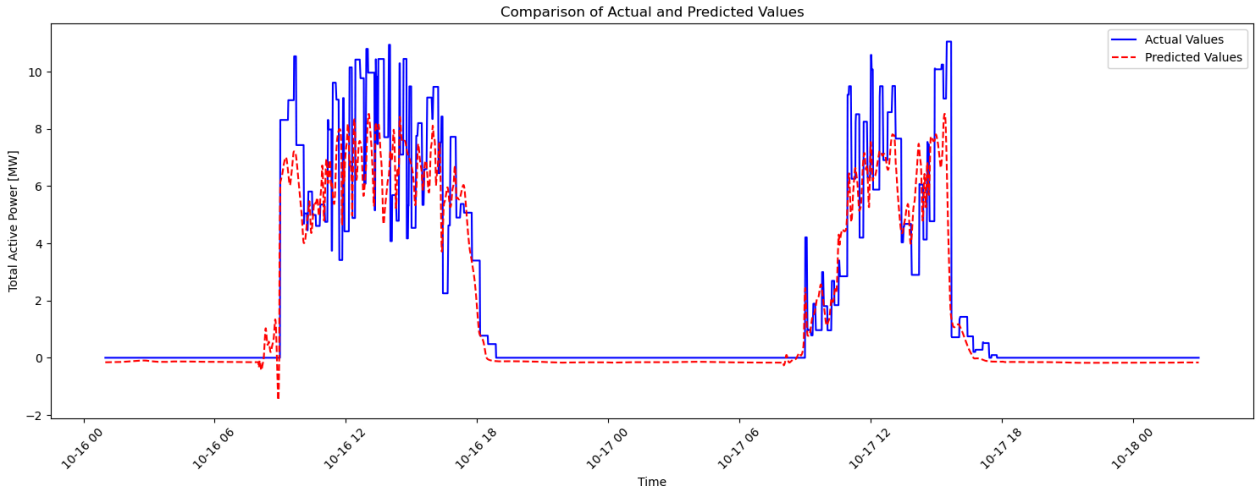



*Figure 6.16: LSTM 1-Hour Prediction for the First 3000 Data Points.*

**Model Robustness**

Several measures were incorporated to enhance the model's robustness and reduce the risk of overfitting:

- **Dropout Regularization**: A dropout rate of 20% was used during training to randomly deactivate a portion of neurons, ensuring that the model generalized well to unseen data.
- **Early Stopping**: This technique was implemented to terminate training once the model performance on validation data ceased improving, thereby preventing overfitting to the training set.

## 6.5 Comparative Analysis

The comparative analysis aims to identify the best-performing model among the four machine learning approaches—Lasso Regression, Random Forest, XGBoost, and Long Short-Term Memory (LSTM)—based on their ability to accurately predict solar power output at both 15-minute and 1-hour intervals. By evaluating each model using key performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared ($R^2$), this study determines which model demonstrates superior predictive accuracy and reliability. This assessment facilitates the selection of the most effective model for solar energy forecasting.

*Table 6.1 Comparison of Quantitative Metrics for 15 minute predictions*

| Matric | Lasso Regression | Random Forest | XG Boost | LSTM |
|---|---|---|---|---|
| **MSE** | 6.723 | 3.031 | 2.650 | 1.169 |
| **RMSE** | 2.592 | 1.741 | 1.627 | 1.292 |
| **MAE** | 2.088 | 0.851 | 0.818 | 0.578 |
| **R Square** | 0.398 | 0.657 | 0.700 | 0.811 |

The table provides a comparison of four machine learning models—Lasso Regression, Random Forest, XGBoost, and Long Short-Term Memory (LSTM) for 15-minute interval predictions, evaluated using four metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared ($R^2$). Among the models, LSTM exhibited the best performance with the lowest MSE (1.169), RMSE (1.292), and MAE (0.578), alongside the highest R2R^2R2 value of 0.811, indicating superior predictive accuracy and variance explanation. XGBoost achieved the second-best results, with an MSE of 2.650, RMSE of 1.627, MAE of 0.818, and an $R^2$ of 0.700. Random Forest performed moderately well with an MSE of 3.031 and an $R^2$ of 0.657, while Lasso Regression showed the weakest performance, recording the highest MSE (6.723), RMSE (2.592), and MAE (2.088) with the lowest $R^2$ value of 0.398.

*Table 6.2 Comparison of Quantitative Metrics for 1 hour predictions*

| Metric | Lasso Regression | Random Forest | XGBoost | LSTM |
|---|---|---|---|---|
| **MSE** | 8.772 | 3.034 | 3.882 | 1.746 |
| **RMSE** | 2.961 | 1.741 | 1.970 | 1.321 |
| **MAE** | 2.574 | 0.851 | 1.112 | 0.643 |
| **R Square** | 0.214 | 0.657 | 0.561 | 0.802 |

The second table presents a similar comparison of the same four models for 1-hour interval predictions. LSTM once again demonstrated the best overall performance with the lowest MSE (1.746), RMSE (1.321), and MAE (0.643), as well as the highest $R^2$ value of 0.802. Random

Forest maintained strong results, achieving an MSE of 3.034, RMSE of 1.741, MAE of 0.851, and an $R^2$ of 0.657. XGBoost followed with an MSE of 3.882, RMSE of 1.970, MAE of 1.112, and an $R^2$ of 0.561. Lasso Regression displayed the weakest performance in this case as well, with the highest MSE (8.772), RMSE (2.961), MAE (2.574), and the lowest $R^2$ value of 0.214.

These results further highlight the effectiveness of the LSTM model for both short- and long-term predictions.

# Chapter 7: Model Deployment

The Final SolarFluxPredict application was deployed using Streamlit and a Long Short-Term Memory (LSTM) model, leveraging the latter's capabilities for time-series forecasting. This deployment was designed to provide real-time solar power predictions through an interactive interface, enabling dynamic user interaction.

The deployment process began with the selection of Streamlit as the framework due to its ability to transform Python scripts into interactive web applications with minimal code. This open-source framework allowed seamless integration of machine learning models with data visualization tools, making it an ideal choice for the project.

The application was deployed locally, running on a localhost address (e.g., http://localhost:8514) accessible via a standard web browser. This setup facilitated ease of testing and debugging during the development phase. The use of Streamlit ensured that the deployment process was lightweight and efficient, requiring no extensive front-end development expertise.

To support the deployment, various tools and technologies were utilized:

- **Anaconda Distribution:** This platform enabled the creation of an isolated Python environment containing all the necessary dependencies.
- **Visual Studio Code (VS Code):** VS Code served as the primary code editor, streamlining script editing, debugging, and project management tasks.
- **Streamlit Framework:** The framework played a critical role in deploying the LSTM model as a web application, allowing users to interact with prediction results dynamically.

The LSTM model was designed to predict solar power output for two selectable timeframes:

1. **15-minute predictions**, which are suitable for short-term operational adjustments.
2. **1-hour predictions**, which support broader energy management strategies.

Interactive widgets such as radio buttons, datetime inputs, and file path inputs were incorporated into the application, enabling users to tailor their predictions dynamically. These features enhanced the overall usability and adaptability of the tool.

### Features of the SolarFluxPredict Application

The SolarFluxPredict application is an interactive and robust tool designed to predict solar power output dynamically. The application’s features ensure user-friendly interaction, high accuracy in forecasting, and a seamless workflow. Below is a detailed explanation of its features, supported by relevant figures.

The primary interface of the application, shown in Figure 1, allows users to interact with the tool through a clean and organized design. It integrates a welcoming description of the application's capabilities along with a visually engaging header image that contextualizes its purpose.

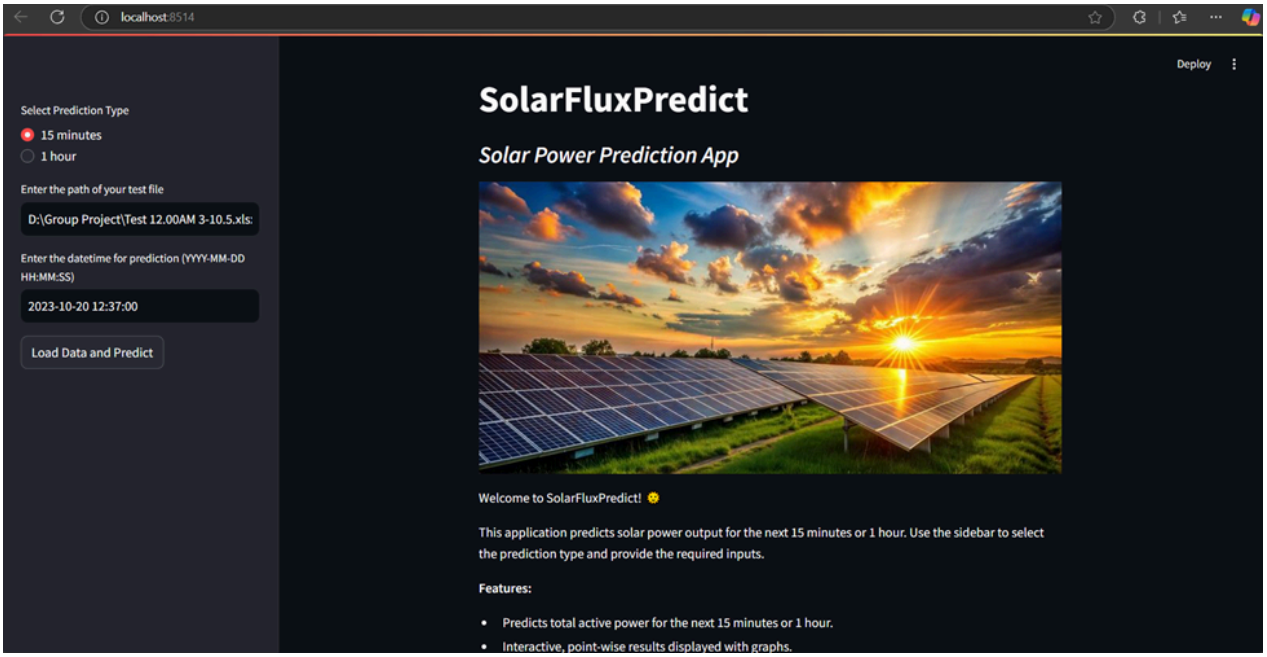



*Figure 7.1: SolarFluxPredict Application Interface*

1. **Dynamic Data Prediction**

The application employs pre-trained Long Short-Term Memory (LSTM) models to predict solar power output based on historical data. Users can toggle between two selectable prediction timeframes:

- **15-minute predictions**: Suitable for immediate, short-term operational adjustments.
- **1-hour predictions**: Useful for broader energy management and planning.
  This selection is facilitated through interactive radio buttons, as illustrated in Figure 2.

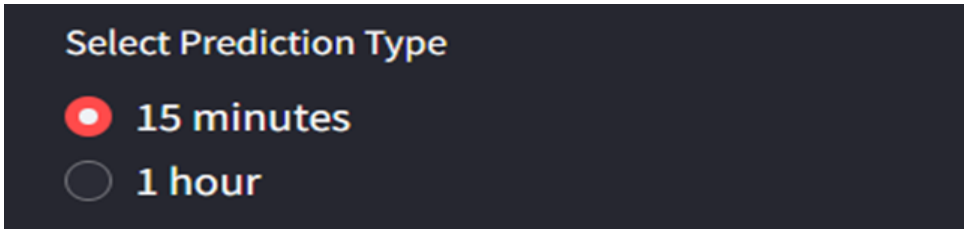



*Figure 7.2: Prediction Type Selection*

2. **Interactive Widgets**

The user interface provides several intuitive widgets to enhance usability. These widgets, shown in Figure 3, include:

- **File path input**: Users can specify the path to their test data file, enabling the application to process external datasets seamlessly.
- **Datetime input**: Users can select the specific time for prediction, allowing for flexibility in analyzing past or future solar power trends.

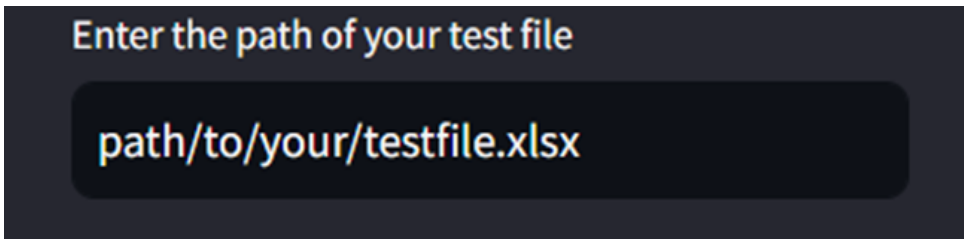



*Figure 8.3: File Path Input Widget*

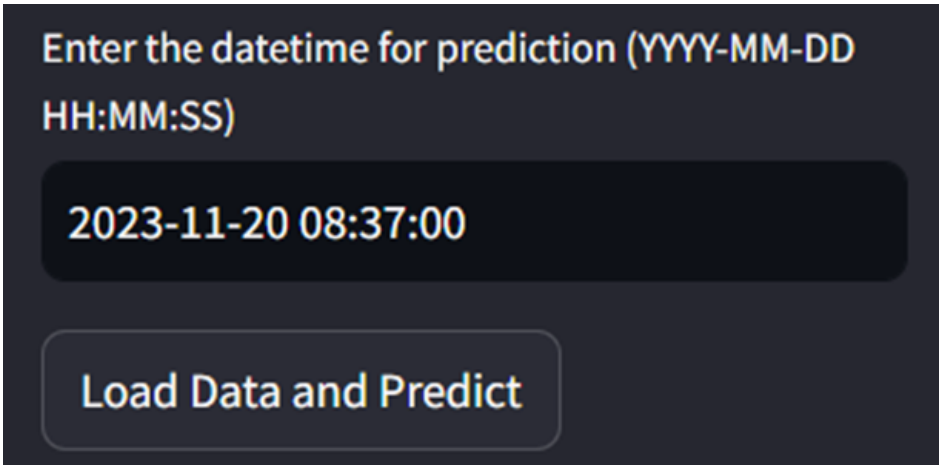



*Figure 7.4: Datetime Selection Widget*

**3. Real-Time Visualization**

The application leverages Plotly to generate dynamic and point-wise plots, showcasing prediction results over time. These visualizations allow users to zoom, pan, and explore trends interactively, enhancing their understanding of solar power variations. For instance:

- Figure 5 displays the 15-minute prediction results, highlighting the power output variations with respect to time.


*Figure 7.5: 15-Minute Prediction Plot*

- Figure 6 showcases the 1-hour prediction, providing insights into longer-term power trends.

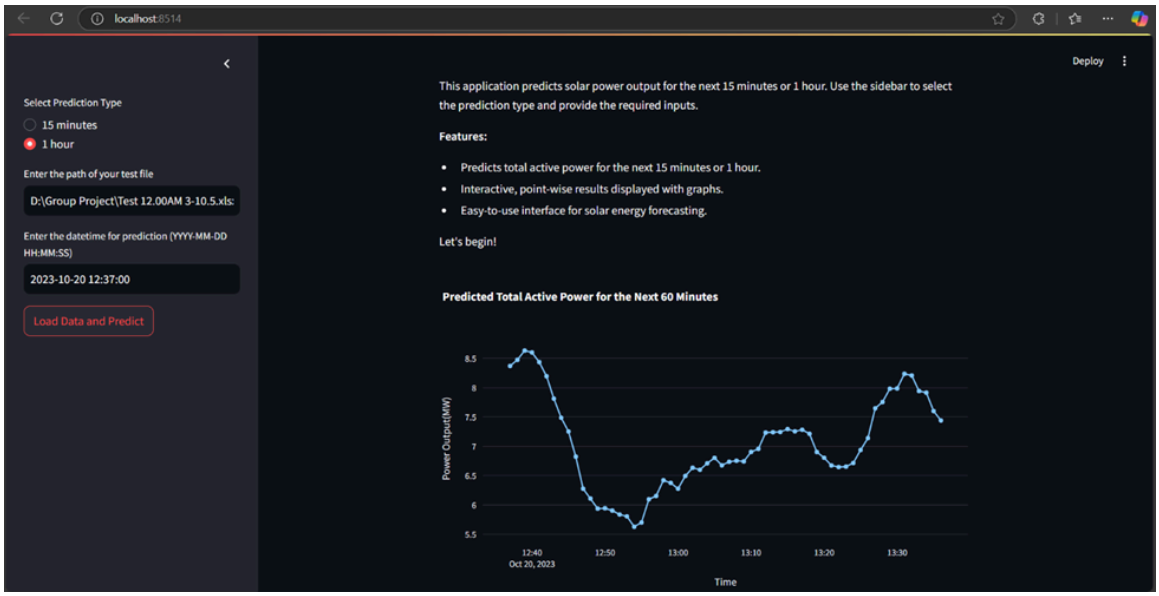



*Figure 7.6: 1-Hour Prediction Plot*

**4. Scalability and Accessibility**

The application runs locally on a **localhost address** (e.g., http://localhost:8514), making it easily accessible through any standard web browser. This design ensures scalability and adaptability to different environments, allowing for future enhancements without significant reconfiguration.

**5. Integration of Machine Learning Models**

The backend incorporates pre-trained LSTM models along with feature and target scalers to ensure the predictions maintain accuracy and align with the original data's scale. This seamless integration ensures reliability in the generated forecasts.

**6. Efficient Workflow**

The application provides immediate results upon data submission, making it suitable for real-time decision-making. Its lightweight design ensures it can operate effectively without requiring advanced technical expertise

# Chapter 8: Discussion & Conclusion

The evaluation phase of the developed system was previously discussed. This chapter depicts the end of the entire research project. This chapter addresses system recommendations and future enhancements that can be made to develop the system further for use in many scenarios.

## 8.1 Project Conclusion

The “Weather-Driven Solar Energy Forecasting: Advanced Predictive Analytics” project aimed to develop a predictive model for short-term solar power output using weather data, addressing inefficiencies in existing forecasting methods. This project, commissioned by WindForce PLC, sought to improve operational decision-making and enhance energy management through accurate and reliable predictions.

The models were developed and evaluated: Random Forest, XGBoost, LSTM (Long Short-Term Memory), and Lasso Regression. The Random Forest model demonstrated strong performance in handling non-linear relationships with an R-squared value of 0.7370 and a Mean Squared Error (MSE) of 2.1586. The XGBoost model performed better, with an R-squared value of 0.7006 and an MSE of 1.6279, showcasing its strength in managing feature importance and iterative learning.

The LSTM model, designed specifically for sequential and time-series data, achieved the highest R-squared value of 0.8138 and the lowest MSE of 0.0116, highlighting its capability to capture temporal dependencies and dynamic patterns. On the other hand, the Lasso Regression model, with its feature selection capabilities, achieved an R-squared value of 0.2768 and an MSE of 2.8418. While it was effective for reducing model complexity and interpreting feature importance, it performed modestly compared to the other models.

Based on the evaluation metrics, the LSTM model was selected as the best-performing model for this project. Its ability to capture complex temporal patterns, low error rate, and superior predictive accuracy made it the most suitable model for short-term solar power forecasting.

## 8.2 Limitations

During the course of this project, we faced several issues and challenges, and each required careful consideration and problem-solving. These challenges included the aspects of data collection, model development, validation, and practical integration.

**Data Collection and Preprocessing Challenges:** The first challenge we encountered in the early stages of the project was when missing data from January and February 2023 were identified. Due to gaps in the weather data for these months, it was necessary to exclude this data from the analysis, which reduced the available training data for the models. Additionally, there were difficulties in obtaining weather data from Sri Lanka's government meteorological department, which led to the decision to source weather data from an online platform. While this solution was effective, the challenge remained in ensuring the quality and consistency of the weather data over time.

**Data Integration and Synchronization:** A significant challenge that was encountered came from the need to integrate operational data from the Vydexa Solar Power Plant, which was recorded at one-minute intervals, with the hourly weather data into a comprehensive dataset. This difference in time intervals meant that the data had to be interpolated and then integrated to ensure that the operational and weather datasets could be effectively combined. This was an important step we undertook to ensure that both data streams aligned properly and that the models could be trained on accurate, synchronized datasets.

**Model Selection:** Selecting the most appropriate machine learning models posed another challenge. After the Exploratory Data Analysis, we as the project team initially experimented with a range of algorithms, but the final selection of models to compare - LSTM, XGBoost, Random Forest, and Lasso Regression - was made based on insights from the literature review conducted as part of the project process. Fine-tuning the models also required significant effort, including selecting the right hyperparameters and performing feature engineering to improve the models' predictive capabilities.

**Evaluation of models:** While the models performed reasonably well, certain models did not meet the desired accuracy thresholds in the initial evaluations. This required further adjustments and refinements in the models, including the application of feature engineering techniques to identify the most significant weather variables affecting power output. The limitations observed which can be considered for the models used can be given as below:

- In the LSTM Model, Minor prediction errors were observed during abrupt transitions in power generation, which may indicate challenges in modeling highly non-linear behaviors.

- The LSTM model's reliance on pre-normalized data suggests that preprocessing steps could impact real-time forecasting accuracy when raw weather data inputs are used.
- The relatively high MSE and moderate R-squared values in the Random Forest model suggest that the model struggles with highly dynamic variations and certain intricate dependencies in the data.
- The computational cost of the Random Forest model for training and predicting with 3000 trees is substantial, which could limit its real-time applicability. Unlike sequential models like LSTMs, Random Forest may not optimally leverage time-series dependencies inherent in the dataset.
- Despite Random Forest model's strong overall performance, the model exhibits minor inaccuracies during extreme variations in power output. This limitation could stem from the inherent complexity of sudden weather changes, which are challenging to predict.
- The computational cost associated with boosting algorithms, while lower than some ensemble methods, remains a consideration for real-time applications.

**Practical Integration:** Although the models have shown promise, the final challenge is deploying the selected and optimized predictive model for practical use at the Vydexa Solar Power Plant. The integration of the model into a real-time system or dashboard is an ongoing task, and we are currently exploring solutions to create a user-friendly interface for plant operators. This will allow them to access the forecasts, make informed decisions about energy production, and adjust operations accordingly.

## 8.3 Future Works and Recommendations

While the project successfully developed a short-term forecasting model, several areas for future improvement and exploration remain:

1. **Model Generalization**: Expanding the training dataset to include weather and solar power data from multiple solar plants across different geographic locations can improve the model's generalizability and robustness under diverse conditions.
2. **Feature Engineering**: Exploring additional meteorological variables, such as wind speed and cloud cover, or employing advanced time-series decomposition techniques could further refine the model's predictive capabilities.
3. **Deployment and Testing**: Collaborating with WindForce PLC to deploy the model on-site and conducting real-world performance evaluations will help identify practical challenges and inform iterative improvements.

4. **Hybrid Model Development**: Future work could explore combining the strengths of the Random Forest, XGBoost, and LSTM models into a hybrid ensemble model. This approach may further improve prediction accuracy by leveraging the complementary advantages of these algorithms.
5. **Incorporating Sustainability Goals**: Aligning the model's application with Sustainable Development Goals (SDGs), such as promoting renewable energy adoption and improving energy efficiency, can further enhance its societal impact.
6. **Incorporating Seasonal and Geographic Variability**: Expanding the model to account for seasonal changes and geographic differences in weather patterns will ensure its applicability across diverse locations, making it a more versatile tool for solar power forecasting.
7. **Integration with Energy Storage Systems**: The predictive outputs from the model can be integrated with energy storage systems to optimize the charging and discharging cycles. This integration will help in addressing intermittency issues and improving grid stability.

In conclusion, the "Solar Flux Predict" project has established a solid foundation for weather-driven solar power forecasting and demonstrated its potential for operational optimization. By implementing the recommended enhancements and focusing on real-world applicability, the model can evolve into a powerful decision-support tool for renewable energy stakeholders.

## 8.4 Research Dissemination

As part of this research, the findings and methodologies were shared and presented at the **International Research Conference of SLTC 2024**. The paper, titled *"A Systematic Literature Review of Weather-Driven Solar Energy Forecasting: Advanced Predictive Analytics"*, highlights the study's contributions and its relevance in advancing knowledge in solar energy forecasting.

Key contributions of the publication include:

- Reviewing state-of-the-art predictive models for solar energy influenced by weather conditions.
- Analyzing methodologies incorporating machine learning techniques to enhance forecasting accuracy.
- Discussing the application of these models in real-world scenarios, emphasizing their potential to optimize energy production.

This publication underscores the impact and significance of this study within the academic community.

The full paper is accessible at [https://irc.sltc.ac.lk/conference-proceedings/].

# REFERENCES


[1] D. Solanki, U. Upadhyay, S. Patel, R. Chauhan, and S. Desai, "Solar Energy Prediction using Meteorological Variables," in *2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)*, Bhubaneswar, India: IEEE, Jul. 2018, pp. 16–19. doi: 10.1109/ICRIEECE44171.2018.9009175.

[2] Y.-J. Zhong and Y.-K. Wu, "Short-Term Solar Power Forecasts Considering Various Weather Variables," in *2020 International Symposium on Computer, Consumer and Control (IS3C)*, Taichung City, Taiwan: IEEE, Nov. 2020, pp. 432–435. doi: 10.1109/IS3C50286.2020.00117.

[3] N. Son and M. Jung, "Analysis of Meteorological Factor Multivariate Models for Medium- and Long-Term Photovoltaic Solar Power Forecasting Using Long Short-Term Memory," *Applied Sciences*, vol. 11, no. 1, p. 316, Dec. 2020, doi: 10.3390/app11010316.

[4] O. Bamisile, C. J. Ejiyi, E. Osei-Mensah, I. A. Chikwendu, J. Li, and Q. Huang, "Long-Term Prediction of Solar Radiation Using XGboost, LSTM, and Machine Learning Algorithms," in *2022 4th Asia Energy and Electrical Engineering Symposium (AEEES)*, Chengdu, China: IEEE, Mar. 2022, pp. 214–218. doi: 10.1109/AEEES54426.2022.9759719.

[5] D. Chakraborty, J. Mondal, H. B. Barua, and A. Bhattacharjee, "Computational solar energy – Ensemble learning methods for prediction of solar power generation based on meteorological parameters in Eastern India," *Renewable Energy Focus*, vol. 44, pp. 277–294, Mar. 2023, doi: 10.1016/j.ref.2023.01.006.

[6] K. Chen, Z. He, K. Chen, J. Hu, and J. He, "Solar energy forecasting with numerical weather predictions on a grid and convolutional networks," in *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*, Beijing: IEEE, Nov. 2017, pp. 1–5. doi: 10.1109/EI2.2017.8245549.

[7] P. A. G. M. Amarasinghe and S. K. Abeygunawardane, "Application of Machine Learning Algorithms for Solar Power Forecasting in Sri Lanka," in *2018 2nd International Conference On Electrical Engineering (EECon)*, Colombo: IEEE, Sep. 2018, pp. 87–92. doi: 10.1109/EECon.2018.8541017.

[8] Jiahui Guo *et al.*, "An ensemble solar power output forecasting model through statistical learning of historical weather dataset," in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, Boston, MA, USA: IEEE, Jul. 2016, pp. 1–5. doi: 10.1109/PESGM.2016.7741059.

[9] M. V. Khaire, A. G. Thosar, and V. N. Pande, “Prediction of Solar Power Generation Using NWP and Machine Learning,” in *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, Ravet IN, India: IEEE, Aug. 2023, pp. 1–6. doi: 10.1109/ASIANCON58793.2023.10270030.

[10] S. Diop, P. S. Traore, and M. Lamine Ndiaye, “Power and Solar Energy Predictions Based on Neural Networks and Principal Component Analysis with Meteorological Parameters of Two Different Cities: Case of Diass and Taïba Ndiaye,” in *2022 IEEE International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)*, Tunis, Tunisia: IEEE, Oct. 2022, pp. 1–6. doi: 10.1109/CISTEM55808.2022.10044065.

[11] N. Saxena *et al.*, “Hybrid KNN-SVM machine learning approach for solar power forecasting,” *Environmental Challenges*, vol. 14, p. 100838, Jan. 2024, doi: 10.1016/j.envc.2024.100838.

[12] M. Rana, I. Koprinska, and V. G. Agelidis, “Solar power forecasting using weather type clustering and ensembles of neural networks,” in *2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, Canada: IEEE, Jul. 2016, pp. 4962–4969. doi: 10.1109/IJCNN.2016.7727853.

[13] J. M. Filipe, R. J. Bessa, J. Sumaili, R. Tome, and J. N. Sousa, “A hybrid short-term solar power forecasting tool,” in *2015 18th International Conference on Intelligent System Application to Power Systems (ISAP)*, Porto, Portugal: IEEE, Sep. 2015, pp. 1–6. doi: 10.1109/ISAP.2015.7325543.

[14] A. Alfadda, R. Adhikari, M. Kuzlu, and S. Rahman, “Hour-ahead solar PV power forecasting using SVR based approach,” in *2017 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, Washington, DC, USA: IEEE, Apr. 2017, pp. 1–5. doi: 10.1109/ISGT.2017.8086020.

[15] J. K. Pandey, “Prediction for Solar Energy Different Climatic Conditions to Harvest Maximum Energy,” in *2023 International Conference on Networking and Communications (ICNWC)*, Chennai, India: IEEE, Apr. 2023, pp. 1–6. doi: 10.1109/ICNWC57852.2023.10127523.

[16] F. Harrou, F. Kadri, and Y. Sun, “Forecasting of Photovoltaic Solar Power Production Using LSTM Approach,” in *Advanced Statistical Modeling, Forecasting, and Fault Detection in Renewable Energy Systems*, F. Harrou and Y. Sun, Eds., IntechOpen, 2020. doi: 10.5772/intechopen.91248.

# APPENDICES

Research Publication – International Research Conference of SLTC 2024

## A Systematic Literature Review of Weather-Driven Solar Energy Forecasting: Advanced Predictive Analytics


HMRV Herath
*Department of Computational Mathematics*
*Faculty of Computing*
*General Sir John Kotelawala Defence University*
Ratmalana, Sri Lanka
39-dba-0002@kdu.ac.lk

RN Silva
*Department of Computational Mathematics*
*Faculty of Computing*
*General Sir John Kotelawala Defence University*
Ratmalana, Sri Lanka
39-dba-0007@kdu.ac.lk

PM Yakupitiyage
*Department of Electrical, Electronics and Telecommunication Engineering*
*General Sir John Kotelawala Defence University*
Ratmalana, Sri Lanka
pmyakupitiyage@gmail.com

TM Kahavidhana
*Department of Computational Mathematics*
*Faculty of Computing*
*General Sir John Kotelawala Defence University*
Ratmalana, Sri Lanka
39-dba-0017@kdu.ac.lk

KT Panditha
*Department of Computational Mathematics*
*Faculty of Computing*
*General Sir John Kotelawala Defence University*
Ratmalana, Sri Lanka
39-dba-0028@kdu.ac.lk

SMM Lakmali
*Department of Computational Mathematics*
*Faculty of Computing*
*General Sir John Kotelawala Defence University*
Ratmalana, Sri Lanka
lakmalismm@kdu.ac.lk



***Abstract*— The rapid adaptation of renewable energy resources such as solar energy to meet the growing global demand for energy necessitates the accurate and reliable forecasting of energy output to ensure grid stability and efficient resource allocation. In light of this, the topic was chosen to conduct a Systematic Literature Review (SLR) with the aim of gathering knowledge on predictive analytics methodologies, techniques, and approaches to weather-driven solar energy forecasting. The review objectives of ascertaining appropriate predictive analytics techniques significantly used variables, and best evaluation metrics for the same were met by the SLR conducted according to guidelines as proposed by B.Kitchenham [1]. The initial selection of 35 related studies was then narrowed to 15 for further detailed review. The results of the review indicate that Neural Networks (NN), Linear Regression, and Support Vector Machine models are the most used technologies, whereas the most significant variables considered in the studies were solar irradiation, temperature, historical power generation data, and relative humidity. The studies also emphasize the use of evaluation metrics such as RMSE and MAE for validating model accuracy. These findings provide valuable insights into predictive analytics in weather-driven solar energy forecasting and offer recommendations for best-suited approaches such as hybrid predictive models to implement in enhancing the accuracy and reliability of weather-driven solar energy forecasts.**



***Keywords*— *Solar energy, weather, forecast, neural networks, machine learning***


### I. INTRODUCTION

Increasing adaptation of renewable power sources like solar energy into the global energy mix has emphasized the need for precise and reliable forecasting methods[2]. At a time like this with an energy shortage, solar energy is one of the most abundant and most environmentally friendly sources. However, its fundamentally irregular nature poses a huge challenge to grid stability and energy planning[3]. As the demand for sustainable energy grows, the timely endeavour of optimizing solar energy output to meet this demand by utilizing predictive analysis by leveraging weather variables has emerged as a promising solution to enhance the accuracy of solar energy output forecasting[4] whilst maintaining grid stability and efficiency. Despite the many recent advances in solar energy forecasting, there are still a number of unresolved challenges to be tackled in this discipline. Most studies point to a remarkable gap that exists in the design of robust model comparison studies, as ones which would review forecasting accuracy show inconsistency of predictive models across geographic and climatic conditions. Current methods often face challenges such as high variability in weather patterns, regional climatic differences, and the influence of unexpected weather changes on solar irradiance, which complicates accurate energy forecasting. Due to the volatile nature of weather patterns and their geographic variability, a single model rarely achieves consistent accuracy worldwide, and this absence of a single, universally optimal model for different geographic and operational conditions limits forecasting accuracy. This therefore represents a missing link into the practical knowledge needed for industry users to determine which forecasting methods are most appropriate for their local conditions and operation. While machine learning methods promise good solutions, transparency and performance reliabilities of those methods raise a number of questions, particularly on conditions of extreme weather variables that might affect grid stability.

Solar energy forecasting entails utilizing a range of statistical and machine-learning methods to anticipate future solar

Resources and Code Work

https://kduac-my.sharepoint.com/personal/39-dba-0007_kdu_ac_lk/_layouts/15/onedrive.aspx?ga=1&id=%2Fpersonal%2F39%2Ddba%2D0007%5Fkdu%5Fac%5Flk%2FDocuments%2FGPDS%2FCodes%20for%20models%2Ezip&parent=%2Fpersonal%2F39%2Ddba%2D0007%5Fkdu%5Fac%5Flk%2FDocuments%2FGPDS

Raw weather Data

| Temp | Weather | Wind | Humidity | Barometer | DateTime |
|---|---|---|---|---|---|
| 25 | Clear. | 4 | 0.89 | 1013 | 3/1/2023 0:40 |
| 24 | Clear. | 4 | 0.94 | 1013 | 3/1/2023 1:40 |
| 24 | Clear. | 6 | 0.94 | 1012 | 3/1/2023 2:40 |
| 24 | Passing clouds. | 4 | 0.94 | 1011 | 3/1/2023 3:40 |
| 24 | Clear. | 2 | 0.94 | 1011 | 3/1/2023 4:40 |
| 23 | Passing clouds. | 6 | 1 | 1013 | 3/1/2023 6:40 |
| 25 | Scattered clouds. | 6 | 0.89 | 1014 | 3/1/2023 7:40 |
| 27 | Passing clouds. | 15 | 0.79 | 1014 | 3/1/2023 8:40 |
| 29 | Passing clouds. | 13 | 0.7 | 1015 | 3/1/2023 9:40 |
| 31 | Passing clouds. | 17 | 0.62 | 1015 | 3/1/2023 10:40 |
| 31 | Passing clouds. | 20 | 0.62 | 1014 | 3/1/2023 11:40 |
| 31 | Passing clouds. | 17 | 0.62 | 1013 | 3/1/2023 12:40 |
| 32 | Passing clouds. | 20 | 0.55 | 1012 | 3/1/2023 13:40 |
| 33 | Scattered clouds. | 19 | 0.52 | 1011 | 3/1/2023 14:40 |
| 32 | Scattered clouds. | 15 | 0.55 | 1010 | 3/1/2023 15:40 |
| 29 | Scattered clouds. | 22 | 0.66 | 1011 | 3/1/2023 16:40 |
| 29 | Passing clouds. | 11 | 0.7 | 1011 | 3/1/2023 17:40 |
| 30 | Passing clouds. | 19 | 0.55 | 1012 | 3/1/2023 18:40 |
| 29 | Thundershowers. Passing clouds. | 15 | 0.55 | 1013 | 3/1/2023 19:40 |
| 28 | Passing clouds. | 15 | 0.58 | 1013 | 3/1/2023 20:40 |
| 27 | Passing clouds. | 11 | 0.66 | 1015 | 3/1/2023 21:40 |
| 26 | Passing clouds. | 11 | 0.7 | 1015 | 3/1/2023 22:40 |

Interpolated Weather Data

| DateTime | Temp | Weather | Wind | Humidity | Barometer |
|---|---|---|---|---|---|
| 2023-03-01 00:40:00 | 25 | Clear. | 4 | 0.89 | 1013 |
| 2023-03-01 00:41:00 | 24.98333333 | Clear. | 4 | 0.890833333 | 1013 |
| 2023-03-01 00:42:00 | 24.96666667 | Clear. | 4 | 0.891666667 | 1013 |
| 2023-03-01 00:43:00 | 24.95 | Clear. | 4 | 0.8925 | 1013 |
| 2023-03-01 00:44:00 | 24.93333333 | Clear. | 4 | 0.893333333 | 1013 |
| 2023-03-01 00:45:00 | 24.91666667 | Clear. | 4 | 0.894166667 | 1013 |
| 2023-03-01 00:46:00 | 24.9 | Clear. | 4 | 0.895 | 1013 |
| 2023-03-01 00:47:00 | 24.88333333 | Clear. | 4 | 0.895833333 | 1013 |
| 2023-03-01 00:48:00 | 24.86666667 | Clear. | 4 | 0.896666667 | 1013 |
| 2023-03-01 00:49:00 | 24.85 | Clear. | 4 | 0.8975 | 1013 |
| 2023-03-01 00:50:00 | 24.83333333 | Clear. | 4 | 0.898333333 | 1013 |
| 2023-03-01 00:51:00 | 24.81666667 | Clear. | 4 | 0.899166667 | 1013 |
| 2023-03-01 00:52:00 | 24.8 | Clear. | 4 | 0.9 | 1013 |
| 2023-03-01 00:53:00 | 24.78333333 | Clear. | 4 | 0.900833333 | 1013 |
| 2023-03-01 00:54:00 | 24.76666667 | Clear. | 4 | 0.901666667 | 1013 |
| 2023-03-01 00:55:00 | 24.75 | Clear. | 4 | 0.9025 | 1013 |
| 2023-03-01 00:56:00 | 24.73333333 | Clear. | 4 | 0.903333333 | 1013 |
| 2023-03-01 00:57:00 | 24.71666667 | Clear. | 4 | 0.904166667 | 1013 |
| 2023-03-01 00:58:00 | 24.7 | Clear. | 4 | 0.905 | 1013 |
| 2023-03-01 00:59:00 | 24.68333333 | Clear. | 4 | 0.905833333 | 1013 |
| 2023-03-01 01:00:00 | 24.66666667 | Clear. | 4 | 0.906666667 | 1013 |
| 2023-03-01 01:01:00 | 24.65 | Clear. | 4 | 0.9075 | 1013 |

Merged and Cleaned Data

| Time | TOAL ACTIVE POWER [MW] | Irradiation | Temp | Wind | Humidity | Barometer |
|---|---|---|---|---|---|---|
| 2023-10-16 00:00:00 | 0 | 0 | 25 | 6 | 0.94 | 1011.666667 |
| 2023-10-16 00:01:00 | 0 | 0 | 25 | 5.95 | 0.94 | 1011.65 |
| 2023-10-16 00:02:00 | 0 | 0 | 25 | 5.9 | 0.94 | 1011.633333 |
| 2023-10-16 00:03:00 | 0 | 0 | 25 | 5.85 | 0.94 | 1011.616667 |
| 2023-10-16 00:04:00 | 0 | 0 | 25 | 5.8 | 0.94 | 1011.6 |
| 2023-10-16 00:05:00 | 0 | 0 | 25 | 5.75 | 0.94 | 1011.583333 |
| 2023-10-16 00:06:00 | 0 | 0 | 25 | 5.7 | 0.94 | 1011.566667 |
| 2023-10-16 00:07:00 | 0 | 0 | 25 | 5.65 | 0.94 | 1011.55 |
| 2023-10-16 00:08:00 | 0 | 0 | 25 | 5.6 | 0.94 | 1011.533333 |
| 2023-10-16 00:09:00 | 0 | 0 | 25 | 5.55 | 0.94 | 1011.516667 |
| 2023-10-16 00:10:00 | 0 | 0 | 25 | 5.5 | 0.94 | 1011.5 |
| 2023-10-16 00:11:00 | 0 | 0 | 25 | 5.45 | 0.94 | 1011.483333 |
| 2023-10-16 00:12:00 | 0 | 0 | 25 | 5.4 | 0.94 | 1011.466667 |
| 2023-10-16 00:13:00 | 0 | 0 | 25 | 5.35 | 0.94 | 1011.45 |
| 2023-10-16 00:14:00 | 0 | 0 | 25 | 5.3 | 0.94 | 1011.433333 |
| 2023-10-16 00:15:00 | 0 | 0 | 25 | 5.25 | 0.94 | 1011.416667 |
| 2023-10-16 00:16:00 | 0 | 0 | 25 | 5.2 | 0.94 | 1011.4 |
| 2023-10-16 00:17:00 | 0 | 0 | 25 | 5.15 | 0.94 | 1011.383333 |
| 2023-10-16 00:18:00 | 0 | 0 | 25 | 5.1 | 0.94 | 1011.366667 |
| 2023-10-16 00:19:00 | 0 | 0 | 25 | 5.05 | 0.94 | 1011.35 |
| 2023-10-16 00:20:00 | 0 | 0 | 25 | 5 | 0.94 | 1011.333333 |
| 2023-10-16 00:21:00 | 0 | 0 | 25 | 4.95 | 0.94 | 1011.316667 |