
DEPARTMENT OF CHEMICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS
CHENNAI-600 036

# DEEP LEARNING APPROACHES IN SUSTAINABLE ENERGY

## Analyzing Energy Generation Patterns and Forecasting of Solar Power Generation

A THESIS REPORT

Submitted by

**Aloy Banerjee**

**(Roll No- CH22M503)**

For the award of the degree

of

**MASTER OF TECHNOLOGY**

**IN INDUSTRIAL AI**

*Under the guidance of* **Dr. Apurba Das**

# THESIS CERTIFICATE

This is to certify that the thesis titled **DEEP LEARNING APPROACHES IN SUSTAINABLE ENERGY: Analyzing Energy Generation Patterns and Forecasting of Solar Power Generation**, submitted by **Aloy Banerjee (CH22M503),** to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Technology,** is a bonafide record of the research work, done by him under my supervision. The contents of this thesis, in whole or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

**DR. APURBA DAS**
Research Guide
Cognitive AI Head
TATA Consultancy Services

Place: Chennai
Date: 15th April 2024

# ACKNOWLEDGEMENTS

I want to express my deepest gratitude to my mentor, **Dr Apurba Das,** for his invaluable guidance, encouragement, and expert advice throughout my M.Tech project. His insights and expertise were helpful in the completion of this research, and his support has been a constant source of motivation.

I am also immensely grateful to my employer**, TATA Consultancy Services,** for sponsoring this M.Tech course. The opportunity to advance my education and skills through this sponsorship has been a pivotal step in my academic and professional growth. The support and understanding of my colleagues at **TATA Consultancy Services** have also been greatly appreciated, enabling me to balance my work and study commitments effectively.

This research would not have been possible without the professors of **IIT Madras** their suggestions and valuable feedback on this work. I sincerely thank all the professors who were directly/indirectly involved with the course.

Furthermore, I extend my heartfelt appreciation to my family and friends for their unwavering support and understanding throughout this journey. Their love and encouragement provided a constant source of strength during challenging moments.

# ABSTRACT

In the context of the accelerating global shift towards renewable energy sources, the accurate prediction of solar power generation has become paramount for optimizing energy distribution, enhancing grid stability, and facilitating efficient resource management. Our study introduces a LSTM-based XGBoost ensemble network, marking a significant innovation in the field of solar power forecasting. This novel approach, applied to data from two solar power plants in India as well as European Data (1986-2015) for 18 Countries leverages the synergistic capabilities of deep learning and ensemble modelling to dissect and understand the intricacies of solar energy generation.

To augment the robustness of our predictive framework, we integrated a Monte Carlo simulation & Quantile based approach, to navigate the inherent uncertainties in solar power generation. This addition not only enhances the model's resilience to variability but also significantly elevates its predictive precision.

The validation of our proposed approach's efficacy was conducted through a comprehensive evaluation of statistical metrics, including RMSE [68], Recall, and the $R^2$. The results from this rigorous validation process unequivocally demonstrate the superior forecasting performance of our ensemble network compared to pure LSTM and XGBoost approaches. For instance, for Plant 1 in Andhra Pradesh, our model achieved an RMSE of 1415.34 for a 1-day ahead forecast, which is a 49.50% improvement over the pure LSTM model that had an RMSE of 2802.86, and a 54.28% improvement over the pure XGBoost algorithm with an RMSE of 3095.7. With $R^2$ scores of 0.8643, and high recall rates of 0.91, our model significantly outperforms the baseline models, reflecting its capability to minimize false negatives effectively and provide highly accurate forecasts.

This paper not only highlights the potential of proposed networks in the realm of renewable energy prediction but also showcases the added value of incorporating Monte Carlo simulations & Quantile based approach to adeptly manage the uncertainties associated with solar power generation.



**KEYWORDS**: Sustainable Energy, Solar Power, Time Series Data, Ensemble Learning, Deep Learning, Energy Generation Patterns, Forecast Model

# ABBREVIATION

- **LSTM** - Long Short-Term Memory
- **ANN** - **A**rtificial Neural Network
- **RNN** - Recurrent Neural Networks
- **CNN** - Convolutional Neural Network
- **ARIMA** - Autoregressive Integrated Moving Average
- **SARIMA** - Seasonal Autoregressive Integrated Moving Average
- **GRU** - Gated Recurrent Unit
- **MLP** - Multi-Layer Perceptron
- **MSE** - Mean Squared Error
- **RMSE** - Root Mean Squared Error
- **MAE** - Mean Absolute Error
- **BPTT** - Back Propagation through Time
- **SDCV** - Standard Deviation Cross Validation
- **RMSESCV** - RMSE Score Cross Validation
- **RMSE(T)** - RMSE on Test Result Set
- **ML** – Machine Learning
- **DL** – Deep Learning

# LIST OF SYMBOLS

- **P(t):** Power output at time t
- **T_a (t):** Ambient temperature at time t
- **T_m (t):** Module temperature at time t
- **I(t):** Irradiation at time t
- **W_s (t):** Wind speed at time t
- **W_d (t):** Wind direction at time t
- **H(t):** Humidity at time t
- $\hat{\boldsymbol{P}}(\boldsymbol{t})$**:** Anticipated power output
- **L:** Loss function
- **R:** Recall
- $\boldsymbol{\theta}^{*}$**:** Optimal Parameter
- $\boldsymbol{S}(\boldsymbol{t})$**:** Solar power generation data at time $t$.
- $\boldsymbol{M}(\boldsymbol{t})$**:** Meteorological data at time $t$.
- $\boldsymbol{D}(\boldsymbol{t})$**:** Combined dataset at time $t$.
- $\boldsymbol{s_i}(\boldsymbol{t})$**:** The $i^{\text{th}}$ solar power generation data measurement at time $t$.
- $\boldsymbol{m_j}(\boldsymbol{t})$**:** The $j^{th}$ meteorological measurement at time $t$.
- $\boldsymbol{T}$**:** A sequence of timestamps.
- $\boldsymbol{D}$**:** Integrated dataset over all timestamps.
- $\boldsymbol{x_i}$**:** Feature vector of the $i^{th}$ sample.
- $\boldsymbol{y_i}$**:** Label of the $i^{th}$ sample.
- $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$**:** Prediction for the $i^{th}$ sample.
- $\boldsymbol{d}(\boldsymbol{x_i}, \boldsymbol{x_j})$**:** Distance between two points $x_i$ and $x_j$.
- $\boldsymbol{m_i}$**:** Imputed value for missing data point.
- $\boldsymbol{x'}$**:** Scaled value of feature $x$.
- $\boldsymbol{x}_{\mathbf{min}}, \boldsymbol{x}_{\mathbf{max}}$**:** Minimum and maximum values of feature $X$ in the dataset.
- $\boldsymbol{h}_{\mathbf{LNN}}^{(\boldsymbol{t})}, \boldsymbol{h}_{\mathbf{LSTM}}^{(\boldsymbol{t})}, \boldsymbol{h}_{\mathbf{GRU}}^{(\boldsymbol{t})}$**:** Hidden states at time step $t$ for RNN, LSTM, and GRU layers, respectively.

- $\boldsymbol{h_{\mathrm{RNN}}^{(t)}}$ **:** Hidden state of the RNN layer at time step $t$.
- $\boldsymbol{h_{\mathrm{LSTM}}^{(t)}, c_{\mathrm{LSTM}}^{(t)}}$: Hidden and cell states of the LSTM layer at time step $t$.
- $\boldsymbol{h_{\mathrm{GRU}}^{(t)}}$**:** Hidden state of the GRU layer at time step $t$.
- $\boldsymbol{W^{(l)}, b^{(l)}}$**:** Weights and biases of the $l^{th}$FCN layer.
- $\boldsymbol{a^{(l)}, h^{(l)}}$**:** Activation and output of the $l^{\text{th}}$ FCN layer.
- $\boldsymbol{y}$**:** Final predicted output.
- $\boldsymbol{m^{(l)}}$**:** Binary dropout mask for the $l^{\text{th}}$ layer.
- $\boldsymbol{Y}$**:** Set of predictions.
- $\boldsymbol{\mu, \sigma^2}$**:** Mean prediction and variance representing uncertainty.
- $\boldsymbol{L_q(y_i, \hat{y}_i)}$**:** Quantile loss function.
- $\boldsymbol{\hat{y}_x^q}$: Predicted quantile.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

In the current energy landscape, the shift toward renewable sources highlights the importance of precise solar power predictions. These predictions are crucial for sustainable energy efforts, playing a key role in energy management, environmental protection, and the sustainability of power systems. With a global emphasis on climate change mitigation, solar energy stands at the forefront of this transition, serving both as a marker of innovation and a practical solution to reduce carbon emissions for a sustainable future. Accurate solar power forecasts are essential, impacting energy storage management, especially in optimizing battery storage to ensure reliable electricity supply amidst solar power's variability.

Moreover, accurate forecasts enhance power system operations' safety and cost-effectiveness, allowing for proactive management to prevent disruptions and optimize resources, thereby lowering costs. They also enable the smooth integration of solar energy into existing grids, enriching energy diversity and promoting a sustainable energy mix. This precision in forecasting supports the micro-management of solar systems, reducing costs and improving efficiency. As we move towards a greener energy model, the significance of accurate solar power predictions becomes increasingly evident, a topic further explored in subsequent chapters through discussions on methodologies, technologies, and their implications.

## 1.2 The Crucial Role of Accurate Solar Power Predictions

Precise forecasts of solar power play a crucial role for various reasons. Initially, they enable the efficient management of electricity stored in batteries, ensuring a consistent power supply [1]. Secondly, these precise predictions contribute to enhanced safety measures and cost reduction in power system operations [2]. Thirdly, the use of accurate forecasting models facilitates the seamless integration of solar energy into existing power systems, resulting in economic and environmental advantages [3]. Moreover, precise predictions assist in the optimal management of solar power systems, leading to cost reduction and

improved energy efficiency [4]. In summary, accurate solar power predictions are instrumental in maintaining the stability and reliability of power generation, promoting the integration of renewable energy sources, and aiding in effective energy management [5-6].

Solar power offers promise in a society focused on sustainable energy. Solar energy is unpredictable and depends on the sun, unlike typical energy sources. This fluctuation makes grid integration and solar power maximization difficult. Accurate solar power estimates bridge the gap between sunlight's variability and the demand for a steady power source. Several researchers have used deep learning to forecast solar power. Solar energy output forecasting has been improved using ANN, RNN, CNN, LSTM, and ensemble learning models [7-9]. DL models like CNN-LSTM and DLSTM-RNN can forecast solar energy system power output, performance ratio, and soiling loss [11]. Comparing these models to real data shows their ability to forecast future trends. Feature selection and geographical analysis have also improved solar energy production prediction [10-12]. Figure 1.1 shows solar power production.

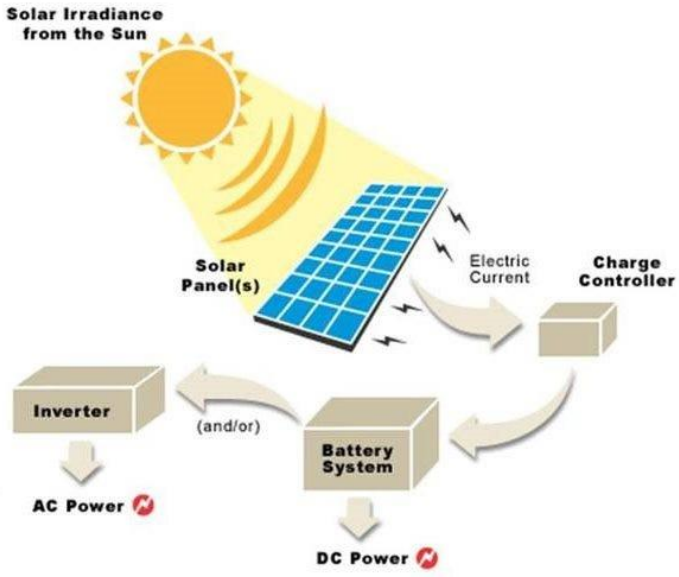


**Figure 1.1 Solar power generating**

### 1.2.1 Reliable models to support energy planning

Reliable models play a crucial role in supporting energy planning, offering valuable insights, predictions, and optimization strategies. These models leverage advanced computational techniques and data analytics to address the complexities of the energy

landscape [13]. These models inform energy policy and planning by predicting energy system evolution [14]. They help decision-makers achieve sustainable objectives by analyzing options and situations [15]. These models must be built using standardized methods to assure transparency and trustworthiness [16]. Balance model complexity to prevent untrustworthy black box models [17]. For simulation-based evaluations, correct and well-formatted data must be included. Energy planning may be incorporated into urban planning using accurate models to help stakeholders make decisions. Energy planning uses numerous credible models:

**Energy Demand Forecasting Models:**

- **Overview:** Predicts future energy consumption based on historical data, economic indicators, population growth, and technological trends.
- **Significance:** Helps utilities and policymakers anticipate future energy needs, facilitating optimal resource allocation and infrastructure development.

**Renewable Energy Resource Assessment Models:**

- **Overview:** Evaluate the potential of renewable energy sources (such as solar, wind, and hydropower) at specific locations.
- **Significance:** Guides the selection of suitable sites for renewable energy projects, optimizing energy production and minimizing environmental impact.

**Integrated Resource Planning (IRP) Models:**

- **Overview:** Assesses the optimal combination of energy resources, considering factors like costs, environmental impact, and reliability.
- **Significance:** Aids policymakers and utilities in long-term planning to achieve a balanced and sustainable energy mix.

**Grid Simulation Models:**

- **Overview:** Simulates the behavior of power grids under various conditions, considering factors like load variability, equipment failures, and renewable energy integration.

- **Significance:** Assists in designing resilient and efficient power grids, ensuring stability and reliability in the face of changing energy landscapes.

**Economic and Environmental Impact Models:**

- **Overview:** Analyzes the economic and environmental consequences of different energy scenarios and policies.
- **Significance:** Informs decision-makers about the trade-offs and benefits associated with different energy planning strategies, aiding in the development of sustainable policies.

**Energy System Optimization Models:**

- **Overview:** Utilizes mathematical optimization techniques to find the most cost-effective and efficient energy system configurations.
- **Significance:** Helps identify the optimal deployment of energy resources, considering technical, economic, and environmental constraints.

**Risk and Uncertainty Models:**

- **Overview:** Evaluate the uncertainties associated with energy planning decisions, considering factors like fuel prices, policy changes, and technological advancements.
- **Significance:** Enables decision-makers to assess and manage risks, ensuring robustness in the face of unforeseen challenges.

By integrating these reliable models into the energy planning process, policymakers, utilities, and stakeholders can make informed decisions that align with economic, environmental, and societal goals. The synergy of these models contributes to the development of resilient and sustainable energy systems.

## 1.3 Significance of Accurate Predictions

Accurate predictions hold significant importance across various domains, providing invaluable benefits and influencing decision-making processes. Here are some key areas where the significance of accurate predictions is evident:

**Grid Stability and Reliability:** Solar power's intermittent nature can disrupt grid stability if not properly managed. Precise solar power forecasts are essential for maintaining grid stability and dependability. Various articles suggest diverse techniques to enhance the precision of solar power prediction. Wang et al. provide a solar power prediction model using Mean-Shift clustering, SVM, and ResNet [18]. Guo and Ji propose a photovoltaic power forecasting technique that utilizes LCM and ASO to enhance the performance of the BPNN [19]. Jiao et al. provide a deep learning model named group solar irradiance neural network, which integrates a convolutional graph neural network and LSTM-RNN for predicting sun irradiance. Sabri and El Hassouni suggest a hybrid deep learning model that integrates a CNN and GRU for predicting solar power [21]. Mohamed and colleagues provide a strong hybrid machine-learning method that combines MLR and PCC for predicting solar power [22]. The approaches show enhanced precision and dependability in forecasting solar power, which helps enhance grid stability and reliability. Accurate projections let grid operators foresee variations in power production, allowing them to:

- **Dispatch other power sources:** They can proactively adjust the output of backup generators (e.g., natural gas) to compensate for dips in solar production, ensuring consistent power supply.
- **Curtailment minimization:** By anticipating high solar production periods, grid operators can minimize the need to "curtail" or reduce solar power generation, maximizing energy utilization.

**Environmental Advantages:** Solar power's integration into the grid relies heavily on accurate predictions to [23]:

- **Reduce reliance on fossil fuels:** By enabling better grid management and minimizing curtailment, accurate forecasts ensure greater utilization of clean solar energy, leading to reduced dependence on fossil fuels and their associated emissions.
- **Enhance renewable energy integration:** As the share of renewable energy sources in the grid increases, accurate predictions become even more critical for managing their collective variability and maintaining grid stability.

**Improved Efficiency and Planning:** Reliable forecasts empower various stakeholders to [24]:

- **Optimize energy storage:** Knowing anticipated solar production allows for efficient planning and utilization of energy storage solutions, enabling the capture and use of excess solar energy during peak production periods.
- **Enhance system planning:** Grid operators can leverage forecasts to plan for future grid infrastructure upgrades and expansion, ensuring the efficient integration of growing solar capacity.

Accurate solar power predictions are not just a desirable feature; they are essential for unlocking the full potential of solar energy. By enabling informed decision-making, ensuring grid stability, and promoting cost-effectiveness, accurate predictions pave the way for a cleaner, more sustainable energy future. As solar technology continues to evolve, advancements in forecasting methods will be crucial in maximizing the benefits of this abundant and renewable energy source.

## 1.4 Problem Formulation

Critical solar power forecast difficulties are addressed in the study. Solar power generation's time-series nature makes temporal dependence a major issue. The study uses the temporal dependency-capturing LSTM architecture along with the power of ensemble learning to overcome this problem and finally to augment the robustness of our predictive framework, we integrated a Monte Carlo simulation & Quantile based approach, a strategic move designed to navigate the inherent uncertainties in solar power generation. The efficient processing of the solar power plant dataset, which may include missing data and various factors, is another problem. To assure dataset quality and applicability for deep learning models, the study emphasizes thorough preprocessing, including missing values, normalizing input features, and scaling the target variable. The study also tackles overfitting in deep learning models, especially LSTMs based ensemble network, which may impair the generalization of fresh data. Training optimization using an Adam optimizer and early halting techniques reduces this risk. LSTMs based ensemble network performance is compared to standard forecasting approaches in the final study. This investigation is essential for demonstrating deep learning's benefits in solar power

prediction. The study addresses research goals and obstacles to improving solar power forecast methods considering the uncertainties in solar power generation. This will improve grid stability, energy distribution optimization, and renewable energy resource management.

## 1.5 Background study

Solar power, an element in the shift towards energy sources, is increasingly relying on solar photovoltaic (PV) technology. This growth highlights the importance of techniques for accurately forecasting solar energy generation moving away from traditional statistical models that depend on past weather data and sunlight measurements. While these traditional methods are useful, they often fail to capture the ever-changing patterns of power production during varying weather conditions. On the hand deep learning methods utilizing CNNs, RNNs and LSTMs excel at identifying time based and spatial connections within data. This leads to forecasts by analyzing patterns over time seasonal variations and utilizing various types of data such as satellite images and sensor data from PV systems. Despite the benefits of learning in improving solar energy predictions, challenges related to data quality, model transparency and scalability remain hurdles to overcome. Our objective is to enhance predictions of power output by employing state of the art learning techniques with diverse datasets. This effort aims to support the integration of energy into existing grids while advancing energy systems overall.

## 1.6 Motivation

Solar power generation as a sustainable, eco-friendly alternative to fossil fuels has grown due to the global demand for renewable energy. Solar energy can meet electricity needs, reduce greenhouse gas emissions, and combat climate change. The intermittent nature of solar power generation makes its integration into the electrical grid difficult. Advanced methods are needed to assess energy production patterns and accurately predict solar power output to solve problems and increase solar energy system efficiency and reliability. Simple statistical models or physical-based simulations may not adequately reflect the complex dynamics and variability of solar energy production. These systems may not be flexible or scalable enough to manage large-scale and real-time data from solar PV panels, weather predictions, and energy output records. Deep learning algorithms can identify complex

patterns and relationships in large datasets, making them promise in many domains. Deep learning models in solar power production may improve energy generation estimates by utilizing plenty of data from multiple sources. The motivation for this research is to explore the application of DL methods for analyzing energy generation patterns and forecasting solar power generation. By developing and evaluating deep learning models tailored to the specific characteristics of solar energy systems, this research aims to enhance the efficiency, reliability, and integration of solar power into the electrical grid. This study aims to enhance renewable energy technology and facilitate the transition to a more sustainable and resilient energy future.

## 1.7 Research Objectives

- **Developing an Ensemble Deep Learning Network:** Create and implement a novel Ensemble learning-based LSTM network designed for the prediction of solar power output, capitalizing on its advanced capability to understand and utilize temporal dependencies within time-series data.
- **Dataset Preprocessing:** Perform thorough preprocessing on the solar power plant dataset, including handling missing data, normalizing input features, and scaling the target variable to ensure data quality and model compatibility.
- **Hyperparameter Tuning:** Fine-tune hyperparameters of the Ensemble learning-based LSTM network to optimize its predictive accuracy. This involves experimenting with parameters such as learning rates, batch sizes, and network architecture to achieve the best performance.
- **Training Optimization:** Optimize the training process of the deep learning model by employing an Adam optimizer and implementing early stopping mechanisms to prevent overfitting and enhance the overall robustness of the predictive model.
- **Uncertainty handling:** Utilize Monte Carlo simulation & Quantile based approach to handle the uncertainty to improve the robustness of the predictive model.
- **Evaluation Metrics:** Evaluate the performance of the Ensemble learning-based LSTM network using key metrics such as MSE, RMSE, Recall, and $R^2$ values.

Assessing these metrics will provide insights into the accuracy and reliability of the predictions.

- **Comparative Analysis:** Conduct a comprehensive analysis comparing the Ensemble learning-based LSTM network approach with traditional forecasting methods to highlight the advantages and effectiveness of employing deep learning techniques in solar power prediction.

## 1.8 Organization of the thesis

Subsequent sections of the thesis are structured as follows:

In ***Chapter* 2,** related works in traditional models, ML & DL approaches, and forecasting solar power generation models are introduced in detail.

In ***Chapter* 3**, we described the problem statement and formulation which is expressed in detail.

In ***Chapter* 4**, we provide an explanation of the dataset used in this study.

In ***Chapter* 5**, we provide the proposed approach and the entire workflow of the system.

In ***Chapter* 6**, Experimental results presented in tabular format with plotted graphs for visualizing outcomes of proposed method.

In ***Chapter* 7**, the presented work is concluded along with the future scope of the present models to improve the performance.

In ***Chapter* 8**, this section contains additional content which describes the mathematical modelling, uncertainty handling and Transformer based model comparison with proposed approach.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Overview

Existing literature has delved into utilizing machine learning techniques, such as LSTM networks, for their proficiency in capturing the intricate patterns and temporal dependencies within solar power data. Researchers have explored meticulous preprocessing techniques to handle missing data, normalize input features, and scale target variables, enhancing the quality and suitability of datasets for accurate prediction models. Moreover, efforts have been made to mitigate the risk of overfitting in deep learning models by optimizing training processes. The accurate prediction of solar power output is critical in the dynamic landscape of renewable energy. Over the years, researchers have explored various methodologies to forecast solar energy generation, encompassing traditional time-series models and more advanced machine-learning techniques. This literature review provides a comprehensive overview of the existing research, emphasizing the limitations of conventional approaches and underscoring the potential benefits of deep learning in tackling the challenges posed by solar power generation's intermittent and dynamic nature.

## 2.2 Traditional Time-Series Models

Early efforts in solar power prediction often relied on traditional time-series models such as ARIMA and ETS. These models were instrumental in capturing trends and seasonality in solar power data. These models include AR, MA, ARMA, ARIMA, and Holt Winter's technique [25] [26]. However, with the rapid development of solar energy plants and the need for accurate predictions, researchers have explored more advanced techniques. One such technique is the use of ML and DL models, such as long LSTM networks [27] [28]. These models offer improved forecasting accuracy by considering factors like sun irradiation, temperature, and other random elements that affect photovoltaic power output. Additionally, data decomposition technology, like CEEMDAN, has been proposed to enhance prediction accuracy further [29]. Overall, the shift towards advanced models and techniques has led to more accurate solar power predictions than traditional time-series models.

## 2.3 Machine Learning & Deep Learning Approaches

In response to the limitations of traditional models, researchers turned to machine learning & deep learning techniques to enhance the accuracy of solar power predictions. Regression models, support vector machines, and random forests were among the approaches explored. While these techniques exhibited improved performance compared to traditional models, challenges persisted in capturing the nuanced relationships within time-series data, especially when confronted with the non-linear and dynamic nature of solar power generation [30]. ML techniques have been widely used for predicting solar power generation. These techniques help to deal with the uncertainty caused by weather conditions and other factors affecting renewable energy production. Various ML models, such as NN, decision trees, and ensemble methods, have been employed for this purpose [31]. These models consider parameters like solar irradiance, temperature, and other factors that influence solar power generation [32]. Optimization algorithms can be used to enhance the accuracy of machine learning models for nonlinear prediction [33]. The accuracy of these models tends to decrease as the prediction horizon gets longer, but deep learning models like ConvLSTM have shown promising results [34]. Overall, ML and DL techniques offer a powerful tool for predicting solar power generation and can contribute to the efficient utilization of non-conventional energy sources [35-36]. Recognizing the potential disruptions caused by solar flares on Earth, the study analyzes various machine learning and deep learning approaches using the RHESSI dataset [37]. Several studies have addressed the crucial challenge of predicting driving energy for solar-powered Mars rovers [38-52].

## 2.4 Studies Related to Forecasting Solar Power Generation

Dwivedi et al., (2024) [53] studied that the worldwide production of renewable energy has seen a significant surge, mostly attributed to the establishment of large-scale renewable energy facilities. However, the challenge of overseeing renewable energy resources at these large sites is still significant owing to environmental factors that might result in reduced power generation, equipment breakdown, and degradation of asset lifespan. Hence, it is vital to identify surface imperfections on renewable energy installations to maintain operational effectiveness and productivity. Using a unique detection method, this work introduced a cost-effective surface monitoring system for renewable energy assets. High-quality photos of the assets are acquired and evaluated to identify solar panel and wind turbine blade deterioration. This study introduced the

Vision Transformer (ViT), a recent deep-learning model in computer vision that utilizes attention mechanisms to classify surface defects. Based on the findings, the suggested approach showcases its capacity to monitor and identify impairments in renewable energy assets, ensuring the effective and dependable functioning of renewable power plants.

Wang et al., (2024)[54] examined that the efficient functioning and organization of smart grids rely heavily on the management of operations and schedules. Forecasting modules for wind speed are crucial for efficiently overseeing wind power plants. Researchers have significantly contributed to the development of accurate forecasting models. However, predicting ideal performance with precision is still a difficult task. Data preparation methods are often used to process the unprocessed wind speed sequences. The trials demonstrated that the proposed wind forecasting framework outperforms prior benchmark models, offering a reliable solution for predicting wind speed and a powerful tool for managing power grid operations.

Chakraborty et al., (2023)[55] studied the challenges associated with the use of solar energy arise from its sporadic nature and dependence on climatic variables such as solar radiation, ambient temperature, precipitation, wind velocity, and other physical attributes such as dust accumulation. Therefore, it is crucial to calculate the size of photovoltaic power output for a certain geographic area. Machine learning models have become significant and are extensively used for forecasting the efficiency of solar power installations. Also introduced a new examination of how climatic factors affect the generation of solar PV electricity. It employs Ensemble ML (EML) models, including Bagging, Boosting, Stacking, and Voting, to conduct the analysis. The results provide a significantly elevated degree of predictive accuracy, around 96%, for Stacking and Voting EML models. The suggested study is comprehensive and can potentially be very beneficial in forecasting the efficiency of solar photovoltaic (PV) power plants on a big scale.

Nwokolo et al., (2023)[56] stated that the UN's Sustainable Development Goals (SDGs) have prompted several nations to use solar PV modules to augment the proportion of renewable energy in the global energy composition. Geographical and environmental conditions greatly influence the electrical performance of solar PV modules. Furthermore, the authors have created and tested 294 physical models based on six distinct solar PV power technologies. The primary difficulties in this study lie in creating a hybridized machine learning approach using the Gumbel probabilistic functional model. This involves a complex mathematical transformation process that necessitates

extensive knowledge of mathematical science. The primary objective is to create a precise and effective model for forecasting the prospective energy generation of solar photovoltaic systems. The results revealed that the intrinsic properties of CIGS thin-film modules make them very resistant to high temperatures, allowing them to retain 60.00–89.66% of their original module qualities for generating photovoltaic (PV) energy. This surpasses the performance of other technologies.

Zheng et al., (2023)[57] studied precise prediction of power generation from various renewable energy sources is crucial for scheduling the power output of RES. Previous research primarily emphasized the prediction of individual energy sources, neglecting the interconnection between diverse energy sources and hence failing to forecast power production for all energy sources continuously and accurately. This study presented a hybrid architecture to address the shortcomings in predicting power output from diverse renewable energy sources. "A CNN is created to extract the local correlations between multiple energy sources. The Attention-based Long Short-Term Memory (A-LSTM) network was created to recognize the nonlinear time-series features of weather conditions and individual energy. The findings indicate that the hybrid framework exhibits superior accuracy to more sophisticated models, such as artificial neural networks and decision trees."

Ozbek et al., (2022)[58] stated that photovoltaics is a highly efficient method of harnessing solar energy, also known as solar power production (SPP). The prediction of SPP is very significant as it helps to reduce the impact of unpredictable variations in the solar energy received and allows the operator to have prior access to data on solar power production. Accurate prediction of SPP is crucial for ensuring the provision of high-quality electricity to end-users. Recent research devised a deep learning method using the LSTM neural network. The comparison analysis reveals that the LSTM model performs superior to other models, as seen by its RMSE, MAE, and R values of 60.66 kWh, 30.47 kWh, and 0.9777, respectively.

Khan et al. (2022)[59] examined precise solar energy forecasting as crucial for facilitating the seamless integration of renewable energy into the operations of the current power system. Given the abundance of data at very detailed levels, there is a chance to use data-driven algorithms to enhance the accuracy of predicting solar energy output. This research introduces a more advanced and widely applicable stacked ensemble technique called DSE-XGB. It utilizes two deep learning

algorithms, ANN and LSTM, as basis models for forecasting solar energy. The use of extreme gradient boosting, a technique, is applied to combine the predictions generated by the base models, hence improving the accuracy of the solar PV output projection. The proposed model underwent evaluation utilizing four distinct solar power datasets to offer a thorough and complete assessment. The research used the shapely additive explanation framework to thoroughly understand the method's learning process. The efficacy of the suggested model was assessed by comparing the predicted outcomes with those of individual ANN, LSTM, and Bagging. The DSE-XGB approach demonstrated exceptional uniformity and steadiness in several case studies, irrespective of the factors involved. Additionally, it shows a significant improvement in the $R^2$ value, surpassing previous models by 10%–12%.

Agga et al., (2022)[60] stated that the escalating impact of climate change is compelling many countries to adopt renewable energy sources, with solar power emerging as a viable alternative to conventional energy sources. However, solar electricity production is heavily impacted by the weather, particularly the amount of sunlight received, which is very unpredictable and challenging to predict. As a result, this presents difficulties for power generation. Precise photovoltaic electricity forecasts may greatly enhance the efficiency of solar power plants. Ensuring the robustness of power plants' functioning is crucial for providing reliable energy to clients. The motivation for this study stems from the current acceptance and advancements in deep learning models and their effective use in the energy industry. The proposed model integrates two deep learning architectures, namely the LSTM and CNN. The CNN-LSTM architecture is superior to conventional ML and single DL models regarding prediction accuracy, precision, and stability.

Luo et al., (2021)[61] intended that solar energy serves as a viable complement to conventional energy sources. Precise prediction of PVPG is essential for the efficient operation of energy production, transmission, and distribution systems, guaranteeing the stability and dependability of power grids. This study presents a sophisticated deep-learning architecture designed to forecast photovoltaic power generation precisely. This study uses the LSTM network to address regression issues that include sequential data. Its objective is to address the limitations of current ML algorithms that rely only on large amounts of data, often leading to irrational predictions. Actual photovoltaic datasets are used to assess the practicality and efficiency of the models. A two-stage hybrid technique performs sensitivity analysis to select input feature variables. The data indicates that the proposed PC-LSTM model has greater predictive capability than the conventional LSTM

model. The PC-LSTM model outperforms typical machine learning and statistical techniques in PVPG forecasting, demonstrating superior accuracy.

Mellit et al., (2021)[62] discussed that Ultra-short-term photovoltaic power forecasting aids in real-time power grid management. Various weather conditions impact photovoltaic (PV) output, leading to fluctuations in energy pricing and affecting grid management. This study used optimum frequency-domain decomposition and DL to predict photovoltaic output shortly. Frequency domain analysis identifies the optimal frequency limits for the decomposition components. The suggested forecasting model reduces the MAPE by 52.97%, 64.07%, and 31.21% compared to the discrete wavelet transform, variational mode decomposition, and direct prediction models, respectively, based on real photovoltaic (PV) data gathered on rainy days. The CNN forecasting model surpasses the recurrent neural network and long-short-term memory model by decreasing the MAPE by 23.64% and 46.22%, respectively. It also boosts training efficiency by 85.63% and 87.68%, respectively. The study's results show that the model created has the potential to improve forecast accuracy and time efficiency.

Syu et al. (2021)[63] studied how environmental conservation has emerged as a crucial concern, and renewable energy stands as an optimal remedy for sustainable power production. Solar power production is a widely adopted renewable energy due to its affordability and little environmental impact. As a result, it has seen rapid expansion and significant investment from the industrial sector. An established solar business model exists, but various uncertainties impede its growth, particularly the scarcity of solar radiation. This study provided a hedging system to mitigate the low-radiation risk for solar investors. The methodology entails using Internet of Things (IoT) generated data and algorithms deployed at the edge to gauge solar radiation levels precisely. In addition, it provided several hedging options to mitigate this risk further. The investigation results indicate that the prediction models using edges obtain an $R^2$ value of 0.841 and a correlation coefficient of 0.917. The simulation results illustrate the effectiveness of the proposed hedging mechanism for investors who are buyers.

Zhen et al., (2020)[64] studied the growing penetration of solar photovoltaic (PV) in power systems has led to a heightened concern over the influence of unpredictable fluctuations in PV output on the reliable functioning of the power grid. To accurately predict solar PV power in the very near term, considering the impact of cloud movement, it is essential to gather surface

irradiance data based on observations of cloud patterns in the sky. Hence, introduces a hybrid mapping model that utilizes deep learning to properly establish the real-time correlation between sky images and surface irradiance. “The approach is specifically designed for solar PV power forecasting. The sky picture data is first preprocessed and then grouped using a combination of convolutional autoencoder feature extraction and the K-means clustering technique. Furthermore, a hybrid mapping model using deep learning techniques is developed for surface irradiance estimation. Ultimately, the simulation results are scrutinized and assessed using several deep-learning techniques, including CNN, LSTM, and ANN. The findings demonstrate that the model described in this research exhibits superior accuracy and is resilient across various weather situations.”

Li et al., (2020)[65] stated that the incorporation of photovoltaic (PV) electricity yields significant economic and environmental advantages. Though photovoltaic (PV) power generation is uncertain and irregular, its widespread adoption may make power grid planning and management challenging. PV power output forecasts are essential for providing high-quality electricity to customers and enhancing power system reliability. This research introduced a hybrid deep learning model using WPD and LSTM networks. Recently developed deep learning algorithms and their effective usage in the energy business inspired this study. A hybrid deep learning algorithm uses five-minute intervals to predict PV power one hour ahead. WPD initially breaks down the photovoltaic power series into subseries. After that, four autonomous LSTM networks are built for this subseries. Each LSTM network's predictions are rebuilt and weighted using linear weighting to produce the final forecast. Also compares individual LSTM, RNN, GRU, and MLP models. The proposed hybrid deep learning model performs well in accuracy and stability, as shown by MBE, MAPE, and RMSE values.

Sun et al., (2019)[66] stated that the unpredictability of cloud movement creates substantial ambiguity in short-term solar power prediction, which may impede the functioning of contemporary power systems. “This study presented a specific CNN called "SUNSET" that is designed to forecast the PV output for the next 15 minutes accurately, based on minute-averaged data. Over one year, the "baseline" model demonstrates a prediction skill of 16.3% in cloudy situations and 15.7% in all weather conditions, compared to a smart persistence forecast. Optimal input and output configurations are explored, and recommendations are provided. Both sky photos and PV output histories are deemed essential in input. In terms of results, training using PV output

data yields much better performance than utilizing the clear sky index (CSI). Prudent down-sampling may save training time up to 83% while maintaining accuracy. When dealing with lag term setups, choosing a history length that matches the prediction horizon is advisable. However, adopting a little shorter history resulted in a moderate increase of 0.5-0.9% in this scenario."

There is a wide range of authors who studied the forecasting solar power generation and give their findings as shown in table 2.1.

**Table 2.1 Comparison of reviewed techniques**

| Authors | Technique | Outcomes |
|---|---|---|
| **Dwivedi et al., (2024)[53]** | Deep learning | Based on the findings, the suggested approach showcases its capacity to monitor and identify impairments in renewable energy assets, ensuring the effective and dependable functioning of renewable power plants. |
| **Wang et al., (2024)[54]** | Data preparation | The results of the experiments have shown that the suggested wind forecasting framework surpasses previous benchmark comparative models. |
| **Chakraborty et al., (2023)[55]** | EML | The results provide a significantly elevated degree of predictive accuracy, around 96%, for Stacking and Voting EML models. |
| **Nwokolo et al., (2023)[56]** | Gumbel probabilistic functional | The results revealed that the intrinsic properties of CIGS thin-film modules make them very resistant to high temperatures and surpass the performance of other technologies. |
| **Zheng et al., (2023)[57]** | LSTM | The findings indicate that the hybrid framework exhibits superior accuracy to more sophisticated models, such as ANN and decision trees. |

| | | |
|---|---|---|
| **Ozbek et al., (2022)[58]** | LSTM | The comparison analysis reveals that the LSTM model performs superior to other models, as seen by its RMSE, MAE, and $R^2$ values of 60.66 kWh, 30.47 kWh, and 0.9777, respectively. |
| **Khan et al., (2022)[59]** | DSE-XGB | The DSE-XGB approach demonstrated exceptional uniformity and steadiness in several case studies, irrespective of the factors involved. |
| **Agga et al., (2022)[60]** | CNN-LSTM | The CNN-LSTM architecture is superior to conventional ML and single DL models regarding prediction accuracy, precision, and stability. |
| **Luo et al., (2021)[61]** | PC-LSTM | The PC-LSTM model outperforms typical machine learning and statistical techniques in PVPG forecasting, demonstrating superior accuracy. |
| **Mellit et al., (2021)[62]** | CNN | The CNN forecasting model outperforms the RNN and LSTM models by reducing the MAPE by 23.64% and 46.22%, respectively. |
| **Syu et al., (2021)[63]** | IoT | The investigation results indicate that the prediction models using edges obtain an $R^2$ value of 0.841 and a correlation coefficient of 0.917. |
| **Zhen et al., (2020)[64]** | Deep learning | The findings demonstrate that the model described in this research exhibits superior accuracy and is resilient across various weather situations. |
| **Li et al., (2020)[65]** | Hybrid deep learning | The proposed hybrid deep learning model performs well in accuracy and stability, as shown by MBE, MAPE, and RMSE values. |

| | | |
|---|---|---|
| **Sun et al., (2019)[66]** | CNN | The findings indicate that the PV output data training yields much better performance than the clear sky index (CSI). |

## 2.5 Challenges and Limitations

Despite the promise shown by deep learning approaches, challenges persist. The requirement for substantial amounts of data for effective training, potential overfitting, and the need for careful hyperparameter tuning are among the challenges that researchers grapple with. Additionally, interpretability and understanding of the internal workings of deep learning models remain areas of ongoing research.

The advantages offered by ensemble learning & deep learning, particularly Xgboost & LSTMs, include their capability to learn complex features and dependencies automatically, adapt to changing conditions, and handle non-linear relationships within the data additionally Monte Carlo simulation & Quantile based approaches finally handle the uncertainty to improve the robustness of the predictive model. This makes them inherently more adept at capturing solar energy generation's intermittent and unpredictable nature than traditional models and machine-learning techniques.

In conclusion, the literature reveals a paradigm shift in solar power prediction methods, with an increasing focus on leveraging ensemble learning & deep learning approaches, particularly Xgboost & LSTMs. While traditional time-series models and machine learning techniques laid the groundwork, solar power generation's dynamic and intermittent nature demands more sophisticated models capable of capturing complex temporal dependencies with high recall. The subsequent chapters of this research will delve into developing and evaluating a deep learning-based approach, contributing to the ongoing discourse on advancing solar power prediction methodologies.

# CHAPTER 3

# PROBLEM DEFINITION AND FORMULATION

Solar energy has several advantages, yet the initial cost of installing solar panels is substantial, making it unaffordable for some individuals. Regrettably, this is a drawback of solar panels; nevertheless, with the ongoing decrease in costs, the future seems promising. Solar panels are now expensive, but recent government initiatives and state-of-the-art technologies are reducing costs. **Deep learning has become a powerful tool in the sustainable energy sector**, providing creative ways to improve renewable energy sources' efficiency, dependability, and availability. Deep learning approaches in sustainable energy encompass a wide range of applications, including but not limited to:

- **Renewable Energy Forecasting:** Deep learning models can accurately predict renewable energy generation from sources such as solar, wind, and hydropower, enabling better integration into the grid and improved energy management.
- **Energy Efficiency Optimization:** Deep learning algorithms can optimize energy consumption in buildings, industries, and transportation systems by analyzing data from sensors, smart meters, and IoT devices, leading to reduced energy wastage and carbon emissions.
- **Grid Management and Stability:** Deep learning techniques aid in grid monitoring, fault detection, and real-time control, ensuring grid stability and reliability, especially in intermittent renewable energy sources.
- **Energy Storage and Management:** Deep learning algorithms optimize energy storage systems such as batteries and capacitors, improving their efficiency and lifespan while facilitating grid balancing and peak shaving.
- **Smart Grid Integration:** Deep learning enables the development of intelligent grid infrastructure capable of dynamically adapting to changing energy demands, integrating distributed energy resources, and supporting bidirectional energy flow.

In recent years, there has been a growing emphasis on renewable energy sources to mitigate environmental impacts and ensure sustainability in the energy sector. Solar power, in

particular, has gained significant attention due to its abundant availability and low environmental footprint. However, the intermittent nature of solar energy production poses challenges for its integration into the power grid, requiring accurate forecasting methods to ensure reliable energy supply. Additionally, understanding energy consumption patterns is crucial for optimizing energy usage and grid management.

***The primary challenges of this study are followed as:***

- **Solar Power Forecasting**: Solar power production is influenced by weather, time of day, and location. There's a critical need for advanced forecasting models that utilize deep learning to manage grid stability and optimize resources, given traditional models' limitations in handling the complex nature of solar energy production.
- **Alternative Energy Sources**: The reduction in local gas supplies, stringent environmental regulations, and rising energy demands necessitate exploring alternative fuels for power production.
- **Stability in Hybrid Renewable Energy Systems (HRES)**: HRES faces instability due to unpredictable weather, with existing models often lacking environmental adaptability. Upgrading weather prediction models is essential for managing these fluctuations.
- **Renewable Energy Research**: The surge in energy consumption, coupled with declining fossil fuel reserves, has prompted significant research into renewable energy solutions.
- **Efficiency Challenges**: Current renewable energy systems face efficiency issues due to a lack of consideration for input/output power balance, resulting in potential power loss.
- **Forecasting Challenges in Solar Power**: Accurate solar power forecasting is complicated by climate variability and weather conditions, underscoring the importance of understanding these influences for reliable energy predictions.

***Considering the problem mentioned above, we have streamlined our objective as below,***

- ❖ **Objective 1:** The principal aim of this investigation revolves around constructing a forecasting model leveraging data from two solar power plants situated in Andhra Pradesh and Maharashtra, each exposed to diverse weather patterns. Additionally, we

also have captured European Data spanned between 1986-2015 for 18 Countries. Refer to the dataset section for more information.

- **Objective 2:** Furthermore, this study seeks to enhance the recall capacity of the developed model. This enhancement is crucial for effectively forecasting solar power generation during the morning hours, coinciding with optimal Sun irradiance.

***To achieve these objectives, the research will undertake the following tasks:***

- **Task 1**: Analyze & pre-process the solar generation data.
- **Task 2**: Incorporate meteorological data along with power generation data.
- **Task 3**: Build a forecasting model targeting to improve efficiency and recall by focusing on handling the uncertainty to improve the robustness of the model.

Each task will be meticulously executed, with validation through statistical metrics such as RMSE, Recall and $R^2$ score, to establish a robust and accurate forecasting model for India's solar power generation.

## 3.1 Formulation of Problem Statement

This section rigorously analyzes exogenous factors responsible for forecasting solar power generation and tries to define the problem.

### 3.1.1 Variables

The variables used in the forecast are defined as follows: $P(t)$: Power output at time $t; T_a(t)$: Ambient temperature at time $t; T_m(t)$: Module temperature at time $t; I(t)$: Irradiation at time $t; W_s(t)$: Wind speed at time $t; W_d(t)$: Wind direction at time $t; H(t)$ : Humidity at time $t$

### 3.1.2 Recall

Along with Loss we also have to focus on calculating the recall of the model and our focus is to maximize the recall.

Recall can be formulated as,

$$\textbf{Recall(R) = TP / (TP + FN)}$$

Recall, often referred to as sensitivity, is the proportion of correct positive predictions to the total number of real positive occurrences, for as predicting solar power production in the morning when sunlight is available. In solar power forecasting, a high recall means that the model successfully captures a large proportion of the actual solar power generation events.

Now the question may arise why recall, as this is not a regression metric, so to answer the same we can say that we have divided solar power generation into two broad categories, firstly the generation which happens in the presence of high Solar irradiance mostly during morning 10 AM to 4 PM and other class belongs to solar power generation during 4 PM to morning 10 AM where mostly solar power generation in not effective.

### 3.1.3 Objective Function

Our goal is to understand & forecast the solar power generation during the duration of 10 AM to 4 PM.

The objective function for a forecasting problem is to minimize the error between the forecasted output and the actual output and maximize the recall. This can be represented as,

$$\mathbf{min} \sum_{t=1}^{T} \boldsymbol{L}\left(\boldsymbol{P}(\boldsymbol{t}), \hat{\boldsymbol{P}}(\boldsymbol{t})\right) - \boldsymbol{\lambda} \times (\mathbf{1} - \mathbf{R})$$

where $\hat{P}(t)$ is the anticipated power output. L is the loss function, which can be the mean absolute error or mean squared error, R is recall, and $\lambda$ acts as a balance between loss minimization and recall maximization.

### 3.1.4 Constraints

Below are a few constraints for the problem,

- **Physical constraints:** $P(t) \geq 0$, since power output cannot be negative.
- **System capacity constraints:** $P(t) \leq P_{\max}$ where $P_{\max}$ is the maximum output capacity of the solar installation.

- **Temporal constraints:** Solar power output must follow the natural daylight cycle, which can be incorporated into the model by including time-of-day and day-of-year as covariates.

### 3.1.5 Problem Formulation

Forecasting $P(t+1), P(t+2), \ldots, P(t+n)$ using historical data up to time $t$ as a time series problem.

### 3.1.6 Machine Learning Approach

Features $X(t)$ which includes $[P(t), T_a(t), T_m(t), I(t), W_s(t), W_d(t), H(t)]$ to predict $P(t+1)$. Create model $f$ with parameters $\theta$ predicts $\hat{P}(t+1) - f(X(t); \theta)$. Our objective is not only to minimize forecasting errors but also to ensure high recall in prediction. This dual objective ensures the model's proficiency in accurately forecasting power output while minimizing false negatives. The optimization goal is redefined as:

$$\mathbf{min}_{\boldsymbol{\theta}} \sum_{\boldsymbol{t-1}}^{\boldsymbol{T}} \boldsymbol{L}(\boldsymbol{P}(\boldsymbol{t+1}), \boldsymbol{f}(\boldsymbol{X}(\boldsymbol{t}); \boldsymbol{\theta})) - \boldsymbol{\lambda} \times (\mathbf{1} - \mathbf{R})$$

Finally, validate unseen test data for performance evaluation.

### 3.1.7 Optimization

The optimization involves finding the optimal parameters $\theta^*$ by minimizing the loss function L and focusing on maximizing the recall over the training data. Let**,** Input feature matrix $X$, target values vector $y$. Train model $f$ by minimizing loss $L$ with parameters.

$$\boldsymbol{\theta} : \boldsymbol{\theta}^* = \mathbf{arg\,min}_{\boldsymbol{\theta}} \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{L}\big(\boldsymbol{y_i}, \boldsymbol{f}(\boldsymbol{X_i}; \boldsymbol{\theta})\big) - \boldsymbol{\lambda} \times (\mathbf{1} - \mathbf{R})$$

where $N$ is the number of training samples & $\lambda$ is a regularization parameter that controls the trade-off between loss minimization and recall maximization. As part of the implementation approach, we must choose an appropriate model and features based on data characteristics and forecasting goals.

For detailed mathemetical formulation refer Mathematical Modelling section.

# CHAPTER 4

# DATASETS

## 4.1 Overview

A fundamental step in the development of effective AI models involves the utilization of high-quality datasets. Considering the objective of our study we have to consider a supervised approach which leads to the use of labelled data. In our study we are planning to focus on doing forecasting of AC power generated from Solar power plants.

### 4.1.1 Dataset Description

The dataset is collected from two solar power plants situated in Andhra Pradesh and Maharashtra, each exposed to diverse weather patterns. Additionally, we also have captured European Data spanned between 1986-2015 for 18 Countries.

**Table 4.1 Dataset used in the study**

| Sampling Frequency | Dataset | No of Rows | Dataset Symbol |
|---|---|---|---|
| **15 min** | Plant1: Gandikotta, Andhra (2 months 22 power source) | 71808 | Plant 1 |
| | Plant2: Nasik, Maharashtra (2 months 22 power source) | 71808 | Plant 2 |
| | Plant1: Gandikotta, Andhra (24 months single power source) | 39168 | Plant 1 – 2 Years |
| **60 Min** | European Data (1986-2015) for 18 Countries – 24 records per Day – Sampling frequency 1 hours | 262969 | European Data |

Location of two solar power plants in India,


**Figure 4.4 Location of Solar Power Plant**

Countries data captured in European dataset,

**Table 4.2 Countries data available in European dataset used in the study**

| Country Name | |
|---|---|
| Belgium (BE) | Poland (PL) |
| Bulgaria (BG) | Portugal (PT) |
| Switzerland (CH) | Romania (RO) |
| Cyprus (CY) | Sweden (SE) |
| Czech Republic (CZ) | Slovenia (SI) |
| Germany (DE) | Slovakia (SK) |
| Spain (ES) | Estonia (EE) |
| Denmark (DK) | United Kingdom (UK) |

| | |
|---|---|
| Austria (AT) | Norway (NO) |

Here's an explanation of each parameter available in dataset of Indian Solar Power Plant:

| Column Name | Description |
|---|---|
| DATE_TIME | Timestamp indicating when the data was recorded. |
| PLANT_ID | Identifier for the specific solar power plant from which the data is collected. |
| SOURCE_KEY | Identifier for individual solar panels or inverters within the solar power plant. |
| DC_POWER | Direct current (DC) power generated by the solar panels, representing the electrical power output before conversion or inversion. |
| AC_POWER | Alternating current (AC) power generated by the solar panels, reflecting the electrical power output after conversion and inversion processes for feeding into the electrical grid. |
| DAILY_YIELD | Cumulative energy yield or power generated by the solar panels over a day, providing a measure of total energy output within a specific day. |
| TOTAL_YIELD | Cumulative amount of power generated by the plant up to that point in time. |
| AMBIENT_TEMPERATURE | Ambient temperature around the solar panels, impacting their efficiency and performance. |
| MODULE_TEMPERATURE | Temperature of the solar panels themselves, crucial because solar panel efficiency decreases as temperature rises. |
| IRRADIATION | Measurement of solar irradiance or sunlight intensity reaching the solar panels, providing information about available solar energy for conversion into electricity. |

Explanation of each parameter available in dataset of European Power Plant:

| Column Name | Description |
|---|---|
| Device Type | Type of device being monitored or managed |
| SN | Serial number of the device |
| Parent Device | Identifier of the parent device (if applicable) |
| Time | Timestamp indicating when the data was recorded |
| Production Compliance Country | Country where production compliance standards are applied |
| Input Mode | Mode of input for the device (e.g., manual, automatic) |
| Hardware Version | Version number of the hardware |
| Software Master Version | Version number of the software |
| DSP Version | Digital Signal Processor (DSP) version |
| Vice DSP Version | Version number of an additional DSP |
| FUSE Version | Version number of the FUSE component |
| DC Voltage PV1(V) | Direct Current (DC) voltage measured at PV1 |
| DC Voltage PV2(V) | DC voltage measured at PV2 |
| DC Current PV1(A) | DC current measured at PV1 |
| DC Current PV2(A) | DC current measured at PV2 |
| DC Power PV1(W) | DC power measured at PV1 |
| DC Power PV2(W) | DC power measured at PV2 |
| AC Voltage R/U/A(V) | Alternating Current (AC) voltage measured at phase R/U/A |
| AC Voltage S/V/B(V) | AC voltage measured at phase S/V/B |
| AC Voltage T/W/C(V) | AC voltage measured at phase T/W/C |
| AC Current R/U/A(A) | AC current measured at phase R/U/A |
| AC Current S/V/B(A) | AC current measured at phase S/V/B |
| AC Current T/W/C(A) | AC current measured at phase T/W/C |
| AC Output Frequency R(Hz) | Frequency of the AC output measured at phase R |
| Total AC Output Power (Active)(W) | Total active power output on the AC side |

| | |
|---|---|
| Cumulative Production (Active)(kWh) | Total cumulative production of active power |
| Daily Production (Active)(kWh) | Active power production within a day |
| Grid Status | Status of connection to the grid (e.g., connected, disconnected) |
| Total Grid Power(W) | Total power received from the grid |
| Leak Current(mA) | Leakage current measured |
| Total Consumption Power(W) | Total power consumption |
| Temperature- Inverter(℃) | Temperature of the inverter |
| Module temperature1(℃) | Temperature of a specific module |
| System Time | Time according to the system clock |
| Total Running Hour(h) | Total running time of the device |
| Countdown Time(s) | Countdown time for a specific event or operation |
| Bus Voltage(V) | Voltage on the bus |
| CT1-Current A(A) | Current measured by Current Transformer 1 |
| DC Component- Phase A/B/C | DC component measured for each phase |
| Back-up CPU Input Voltage Sample 1 | Voltage sampled by the backup CPU |
| Back-up CPU Input Current Sample 1 | Current sampled by the backup CPU |
| Insulation Impedance-Cathode to ground(KΩ) | Insulation impedance measured between cathode and ground |
| CT Power(W) | Power measured by Current Transformer |
| Inverter status | Status of the inverter |
| PV1 Insulation Impedance(KΩ) | Insulation impedance measured for PV1 |

| | |
|---|---|
| PV2 Insulation Impedance(KΩ) | Insulation impedance measured for PV2 |
| Temperature | Ambient temperature of the PV Panel |
| Irradiation | Solar Irradiance in watts/meter$^2$ |

# CHAPTER 5

# RESEARCH AND METHODOLOGY

## 5.1 Research Methodology

The heart of the study lies in implementing the XGBoost algorithm, renowned for its robustness in handling structured data and ensemble learning, and LSTM networks, an RNN specifically designed to capture temporal dependencies in sequential data. The model architecture is tailored to the unique characteristics of solar power generation time series, enabling it to learn and predict complex patterns over extended periods effectively.

Accurate solar energy forecasting is essential for enhancing the viability of solar power plants in the energy market and diminishing reliance on fossil fuels in socio-economic advancement. The primary goal of our effort is to precisely forecast solar energy.

### 5.1.1 Proposed framework

This work utilizes three distinct deep learning models, namely RNN, LSTM, and GRU along with one ensemble algorithm, namely Xgboost, and finally integrated with a Monte Carlo simulation & Quantile based approach [70] to point forecast the PV solar energy production specifically considering the high RMSE & Recall output. The objective of this research is to assess the significance of the proposed models, to determine the most effective algorithm for solar energy prediction that can achieve high recall. The procedure used in this study is shown in Figure 5.3. The procedure is divided into four primary phases: **data acquisition and merging, data preprocessing, model training and parameterization, and model testing and validation**. This part provides a comprehensive explanation of the execution of each processing and selection phase, along with crucial information on the structure and parameters of the created models.

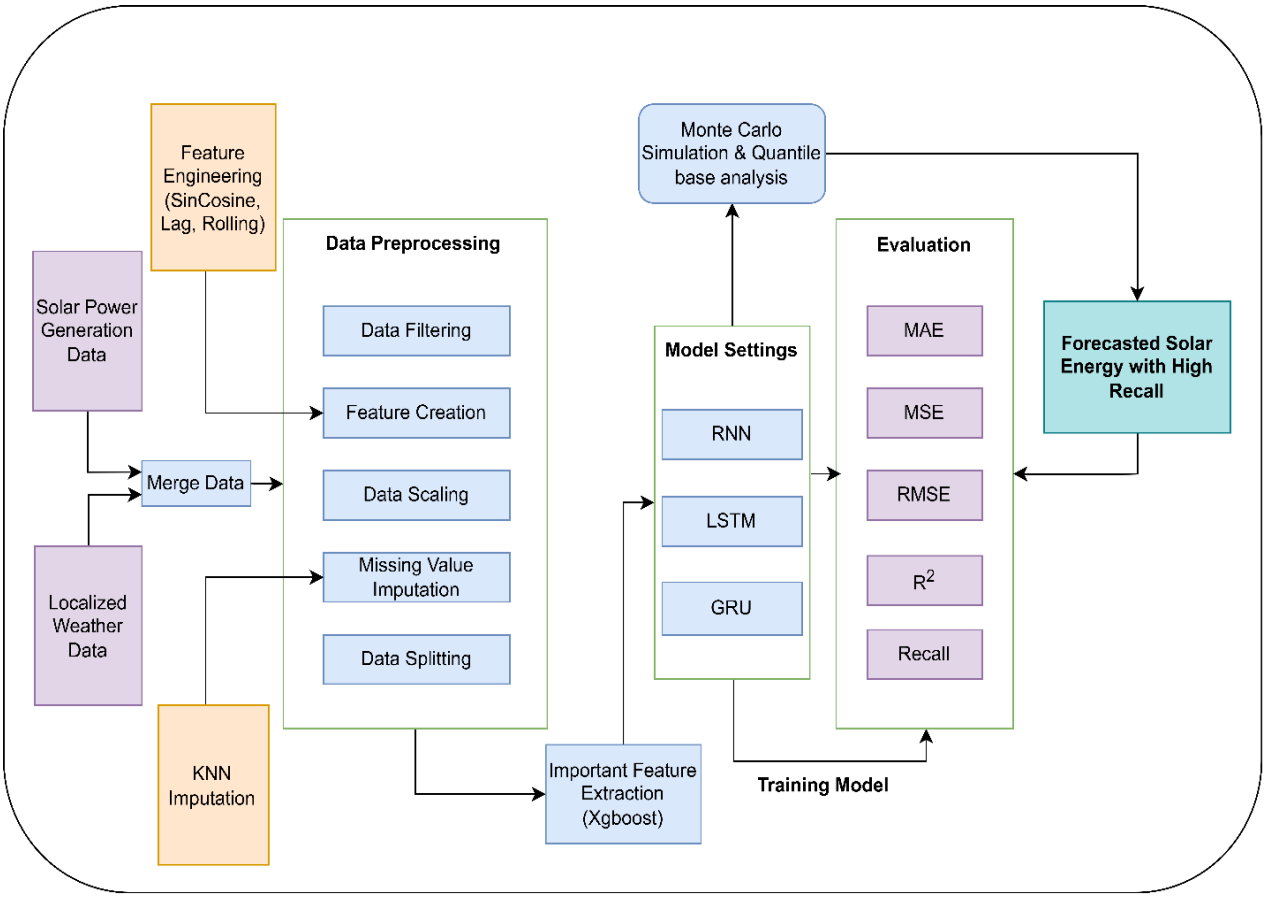


**Figure 5.3 Proposed Framework**

The proposed framework is described in the following steps:

### 5.1.1.1 Step 1: Power Generation & Weather Data

The dataset utilized in this study is detailed in Chapter 4, encompassing data from both Indian and European power plants. It combines power generation and weather information on a per-time-stamp basis, setting the stage for subsequent analysis phases.

### 5.1.1.2 Step 2: Pre-processing and Feature selection

Preprocessing and feature selection are essential in preparing data for machine learning, especially for solar power forecasting where data is highly variable due to weather conditions. These steps enhance the quality of data fed into the model, impacting its accuracy and reliability. The process starts by identifying and extracting four critical features from the dataset: solar power data and key weather parameters like temperature and irradiance.

We use Min-Max scaling to normalize these features between 0 and 1, balancing their influence on the model regardless of their original scales. This technique also minimizes the impact of outliers and noise, making the learning process more effective.

Given the time-series nature of weather data, recorded at intervals of either 15 or 60 minutes, we rely on historical data to forecast solar energy generation, enabling real-time predictions crucial for solar energy management. We adhere to the standard practice of splitting the dataset, allocating 80% for training and 20% for testing the models. The dataset has undergone meticulous preprocessing, which involves several steps:

- **Handling Missing Data:** Utilize the K-Nearest Neighbors (KNN) Imputation technique to address any gaps or incomplete data points. This approach estimates missing values based on the similarity of the available data, maintaining the integrity of the dataset.
- **Normalization of Input Features:** Normalization is a process of scaling numeric features to a standard range. It ensures that different features with varying scales contribute equally to the analysis.
- **Scaling of the Target Variable:** The target variable, which could be AC_POWER, DAILY_YIELD, Total AC Power Output, or another metric of solar power output, is scaled using Min-Max scaling to ensure it fits within a specific range, enhancing its suitability for modeling purposes.

Overall, this preprocessing is crucial for ensuring the quality and reliability of the dataset, making it ready for use in machine learning models or statistical analyses related to solar power generation.

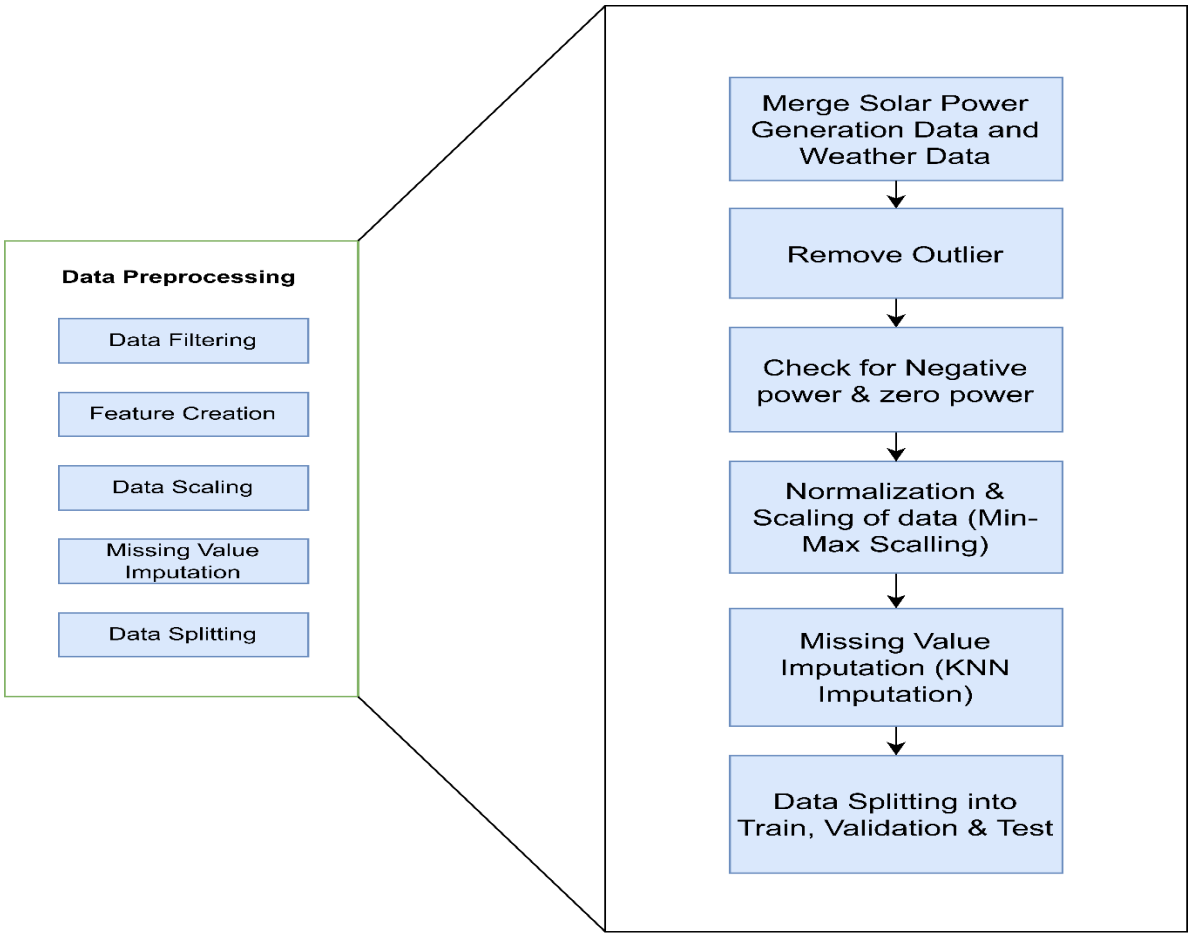


**Figure 5.3.1 Data Preprocessing**

### 5.1.1.3 Step 3: Training and models parametrizations

Each model must undergo parameterization throughout the training phase to provide accurate projections that consider the batch size. After configuring several parameters for our deep learning models throughout the training process and assessing the achieved outcomes, the Adam optimizer has shown superior performance. Consequently, it was chosen as the shared parameter for all the suggested deep-learning models. ADAM, a frequently used stochastic optimizer, has shown superior performance compared to other optimizers in empirical studies [32]. It enables models to adapt rapidly. Figure 5.3.2 illustrates the structure of a neural network, as seen below.

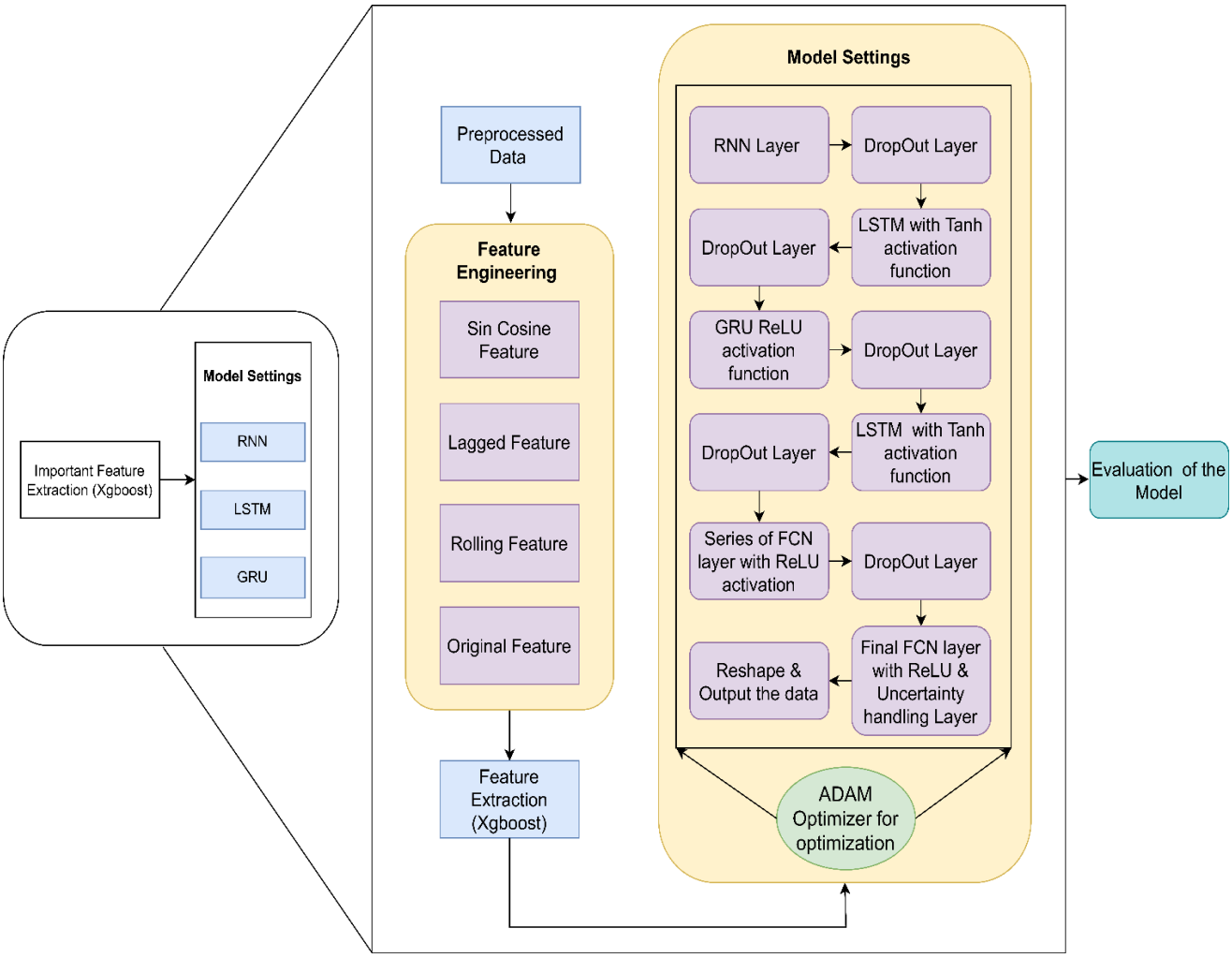


**Figure 5.3.2 Proposed model architecture for forecasting solar energy.**

As depicted in the network structure in Figure 5.3.2, we have a multi-layer neural network architecture consisting of:

1. An LSTM (Long Short-Term Memory) layer and tanh activation function, suitable for capturing long-term dependencies within the time-series data.
2. A GRU (Gated Recurrent Unit) layer and relu activation function, which helps the model to better capture patterns and sequences within the dataset without the vanishing gradient problem.
3. A series of fully connected layers with relevant activation function that process the data further after the recurrent layers.

4. A dropout layer applied after each LSTM, GRU, and fully connected layer to prevent overfitting by randomly setting a fraction of input units to 0 during training.
5. We also carefully introduced Monte Carlo simulation as well as Quantile based analysis, to minimize the inherent uncertainty in solar power generation.
6. Finally, a single neuron dense network to produce the forecasted output.

Feature engineering is another crucial aspect of the model development process, where various types of features are generated from the pre-processed data:

- **Sin Cosine Features**: These may involve transformations to capture cyclical patterns within the data, such as time of day or seasons.
- **Lagged Features**: Past values of the time-series are used as features to predict future values, providing the model with a sense of recent trends and patterns.
- **Rolling Features**: Statistics like rolling means or standard deviations are calculated over a window of past observations to capture the local behaviour of the time-series.
- **Original Features**: The core features from the dataset that are used directly in the model, such as the solar power generation data and weather conditions.

The architecture ensures that the model can capture both the short-term and long-term dependencies within the data, which is essential for accurate time-series forecasting. The use of a stacked combination of LSTM and GRU layers allows the model to benefit from the strengths of both types of recurrent units.

#### 5.1.1.4 Step 4: Performance metrics

To assess the efficacy and recall of the suggested deep learning models for predicting photovoltaic solar energy a collection of commonly used assessment measures is used. During this stage, we have implemented the most appropriate approach for the given context of deep learning and regression issues. To evaluate the models during testing and training, six performance measures have been utilized: MAE, MSE, RMSE, R squared ($R^2$), and Recall.

- **Mean absolute error (MAE)**

  Because the prediction error can be positive or negative, the average total value of the error is used to avoid the cancellation of positive and negative mistakes.

$$\boldsymbol{MAE} = \left(\frac{\boldsymbol{1}}{\boldsymbol{n}}\right)\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left|(\boldsymbol{x_i} - \widehat{\boldsymbol{x_i}})\right| \quad (3.1)$$

As can be seen in Eq. (3.1), where n is the number of observations, $xi$ represents the true value, while $\widehat{x_i})$ represent the prediction value.

- **Root mean square error (RMSE)**

We use the RMSE assessment model to estimate performance more precisely by determining the standard error based on the prediction findings outlined in Eq. (3.2):

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum(\boldsymbol{x_i} - \widehat{\boldsymbol{x_i}})^{\boldsymbol{2}}}{\boldsymbol{N}}} \quad (3.2)$$

Generally, RMSE is commonly used for evaluating the quality of predictions. As we can see in Eq. (3.2), where n is the number of the participating samples, $x_i$ is the real value while $\hat{x}_i$ is the predicted value.

- **$R^2$ score**

It is referred to as the coefficient of determination, which quantifies the proportion of variation in the dependent variable. R2 is used to assess the dispersed data points around a regression line that has been fitted.

$$\boldsymbol{R^2} = \boldsymbol{1} - \frac{\boldsymbol{SS_{res}}}{\boldsymbol{SS_{total}}} = \frac{\sum(\boldsymbol{y_i} - \widehat{\boldsymbol{y_i}})^{\boldsymbol{2}}}{\sum(\boldsymbol{y_i} - \boldsymbol{\mu})} \quad (3.3)$$

As shown in Equation (3.3), where $SS_{res}$ is the sum of squares of residuals, $SS_{total}$ is the total sum of the squares, $y_i$ is the true value, $\widehat{y_i}$ is the prediction value, and $\mu$ is the mean.

- **Mean square error (MSE)**

The average square sum of errors is used to avoid canceling positive and negative errors. The effects of a significant mistake value increase since the absolute error was squared. MSE is used to check how close estimates or predictions are to actual value using the Eq (3.4).

$$\boldsymbol{MSE} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{x_i} - \widehat{\boldsymbol{x_i}})^{\boldsymbol{2}} \quad (3.4)$$

- **Recall (R)**

Recall, often referred to as sensitivity, is the proportion of correct positive predictions to the total number of real positive occurrences, for as predicting solar power production in the morning when the sun's irradiance is present. In solar power forecasting, a high recall means that the model successfully captures a large proportion of the actual solar power generation events, following is the formulation Eq (3.5).

$$\mathbf{R = TP / (TP + FN)} \quad (3.5)$$

**True Positives (TP):** These are instances where the model correctly predicts solar power production, meaning it accurately forecasts that there will be solar power generation during the morning when the sun's irradiance is present mostly during morning 9AM to 4PM.

**False Negatives (FN):** These are instances where the model fails to predict solar power generation when it actually occurs. This means the model predicts no or low solar power output in the morning despite the presence of solar irradiance but low mostly during rest of the period in day.

In assessing our model, we prioritize achieving a high recall, acknowledging a potential compromise with the R2 metric.

### 5.1.2 Why Ensemble Deep Learning (Proposed Framework) Approach?

The architecture shown in Figure 5.3 introduces a novel hybrid predictive model for solar power output forecasting, combining Xgboost and LSTM networks with uncertainty handling layer. This model harnesses the distinct advantages of both approaches for a comprehensive analysis.

- **Feature Engineering**: The model incorporates seasonal patterns through sine and cosine time transformations, past data via lag features, and trends and volatility with rolling window statistics, capitalizing on solar power data's temporal nature.
- **KNN Imputation**: For missing data, K-Nearest Neighbours imputation is employed, preserving the data's time-series integrity by estimating missing values from similar time points.
- **Xgboost for Feature Selection**: Xgboost streamlines the dataset by ranking and selecting the most impactful features, reducing overfitting risks, and enhancing model generalizability.

- **Uncertainty Handling**: The model addresses prediction uncertainties through Monte Carlo simulation and quantile regression, balancing computational efficiency with comprehensive scenario forecasting.

By integrating Xgboost's feature selection and dimensionality reduction with LSTM's capacity for modelling temporal dependencies, coupled with sophisticated uncertainty handling, the model achieves precise forecasts with notable recall.

**Hybrid Network Advantage:** XGBoost filters noise and selects significant features, while LSTM network utilizes these features for accurate time-series predictions, offering a powerful two-stage modeling approach blending feature selection and temporal forecasting.

# CHAPTER 6

# EXPERIMENTAL RESULTS

This chapter provides a comprehensive examination of the experimental evaluation conducted to assess the predictive accuracy of various forecasting models for solar power generation at two distinct solar power plants in India and European Data spanned between 1986-2015 for 18 Countries. The primary focus is on enhancing model recall to improve early-hour power generation predictions, crucial for optimizing solar power utilization.

## 6.1 Modelling Stage

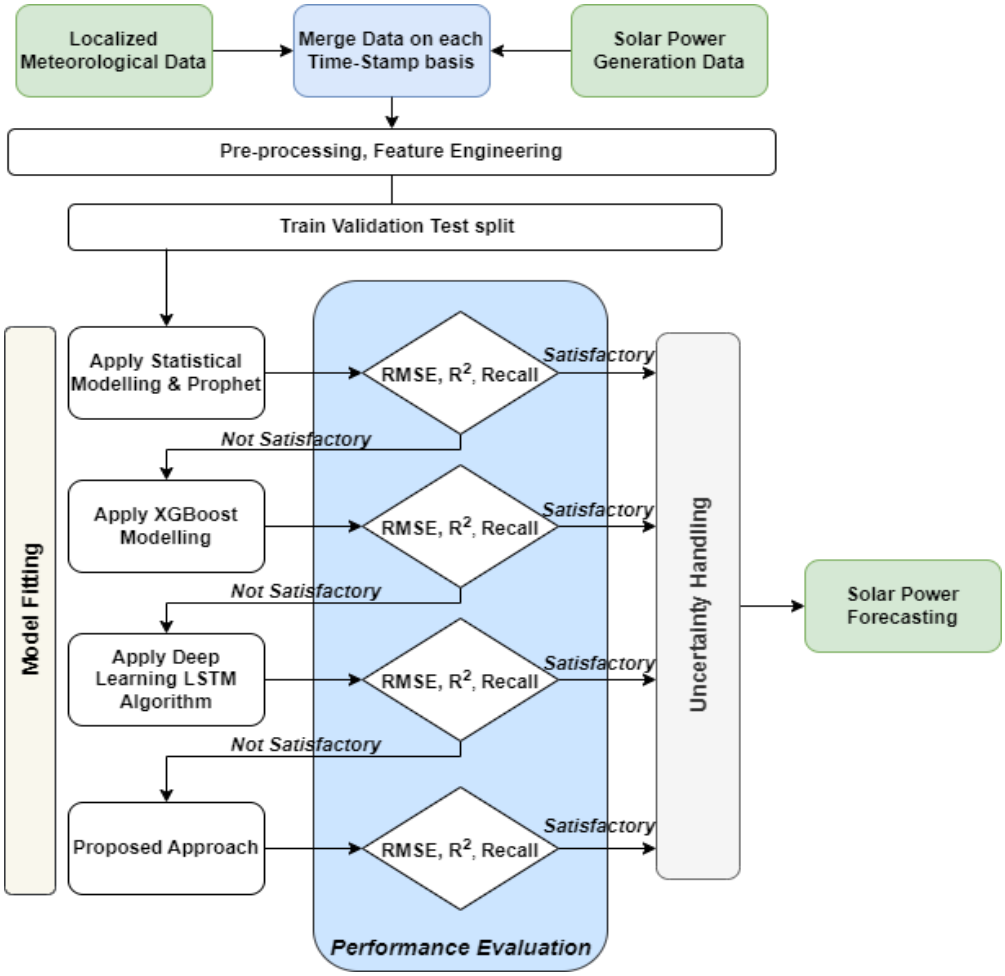


**Figure 6.0 Flow diagram followed for experimental results analysis**

The flowchart in figure 6.0 presents the methodological approach for forecasting solar power generation. After preprocessing and feature engineered data is split as per necessity and then

various models are applied sequentially, statistical modeling with Prophet, XGBoost modeling, and a deep learning LSTM algorithm and finally our proposed approach. Each model's performance is evaluated based on metrics such as RMSE, $R^2$, and recall. If the model's performance is satisfactory based on these metrics, it proceeds to uncertainty handling, which is an essential step in refining the solar power forecasting. The process culminates in performance evaluation, validating the effectiveness of the proposed approach in predicting solar power generation with a consideration for the inherent uncertainties in the data.

Our proposed hybrid model blends RNN, LSTM, and GRU layers with feature engineering and XGBoost selection to forecast solar energy trends. Overfitting is controlled with dropout layers, while ReLU and tanh activations capture nonlinearities. The model's accuracy is assessed using RMSE, $R^2$, and recall, prioritizing high recall for robust solar predictions.

## 6.2 Environment & Settings

The neural network and corresponding tuning were done using Python 3.9 and Tensorflow library in Google Colab. This platform provides GPU for a limited time, which needs to be enabled by changing the runtime for the Hardware accelerator to 'T4 GPU', which comes with 15GB memory and 12.7 GB system RAM for the free user.

Below are the Models tried to generate the experimental results,

**Table 6.1 List of Models**

| Model Name | Model Type | Model Symbol |
|---|---|---|
| **SARIMAX** | Statistical Modelling | M1 |
| **Prophet** | Global Forecasting Model | M2 |
| **XGBoost** | Tree based Boosting Algorithm | M3 |
| **LSTM** | Deep Learning based Algorithm | D1 |
| **Proposed Approach** | Combining Ensemble & Deep Learning Algorithm | D2 |
| **Transformer Based Approach** | Advance Deep Learning Algorithm | D3 |

## 6.3 Evaluation Framework and Initial Data Insights

The study prioritizes high recall in model evaluation, acknowledging potential compromises in other metrics like the $R^2$ score. This section introduces the initial data analysis, showcasing the variability in power generation across the two Indian solar power plants (Figures 6.1) and European Data spanned between 1986-2015 for 18 Countries (Figures 6.2).

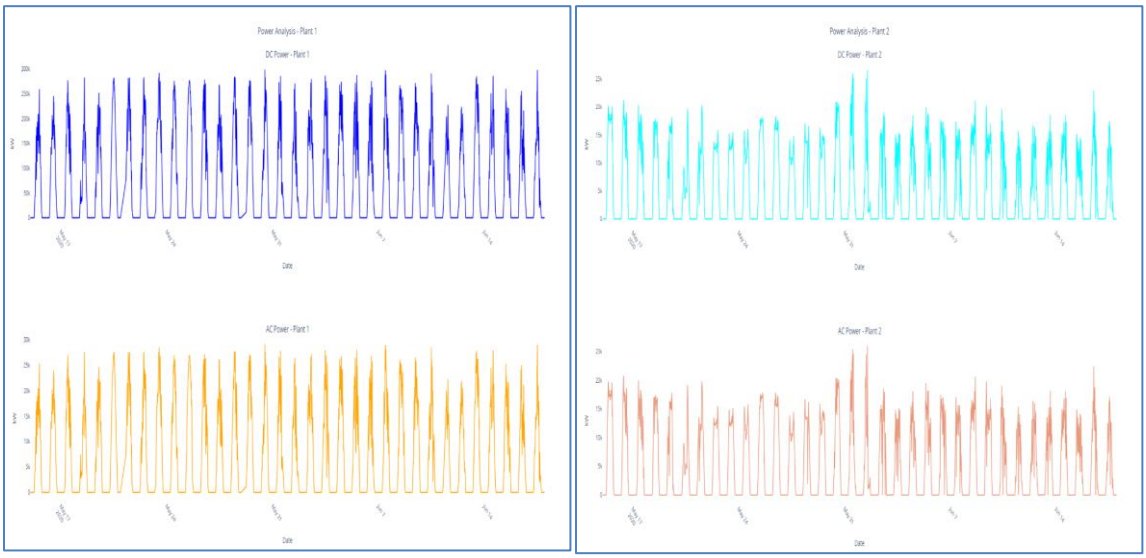

**Figure 6.1 Power analysis for Indian Plant 1 & Plant 2**

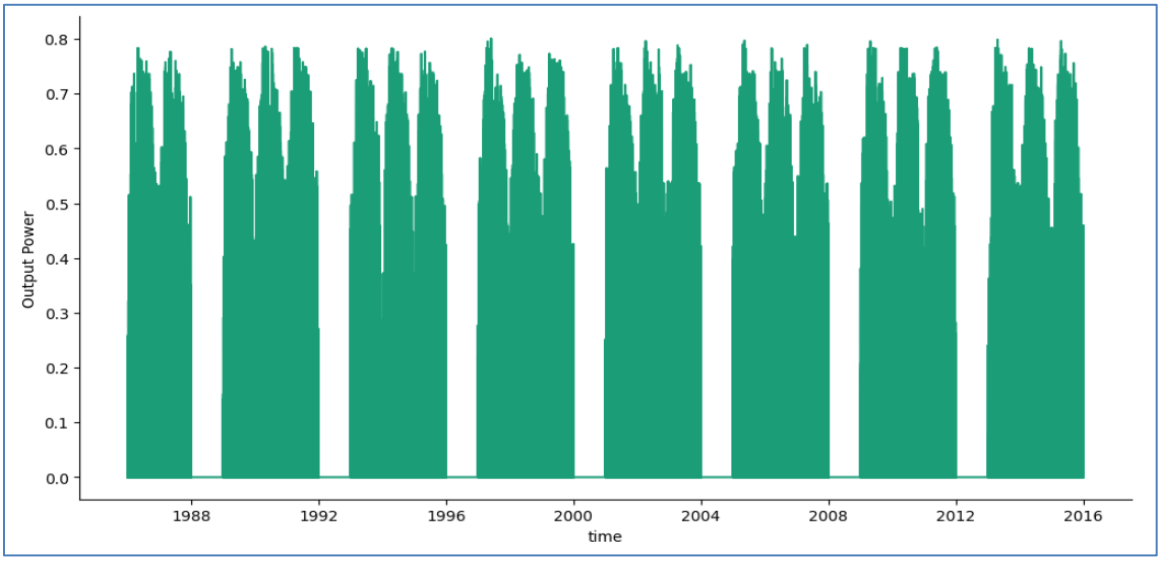


**Figure 6.2 Power analysis for European Country – Example plot for Belgium**

Detailed EDA for Indian Plants are available here,

## 6.4 Methodological Approach and Preliminary Results

### 6.4.1 Statistical Models and Ensemble Learning

Initial attempts employing Auto-Regressive and Facebook Prophet models yielded suboptimal results for the dataset of Indian plant hence as well as for European dataset. Table 6.2 contrasts the performance of SARIMAX and Prophet models for both Indian plants and European dataset, revealing the necessity for alternative approaches.

**Table 6.2 Result of SARIMAX and Prophet**

| Model Name | SARIMAX | | Prophet | |
|---|---|---|---|---|
| Plant Location | RMSE(T) | R Squared | RMSE(T) | R Squared |
| Plant1 | 4979.83 | 0.99 | 18425.75 | 0.8 |
| Plant2 | 70773.12 | -0.03065 | 17619761 | 0.35 |
| Plant1: 2 Years | 30459.58 | 0.2 | 21597545.32 | 0.33 |
| European Dataset | 0.91 | 0.16 | 0.89 | 0.30 |

After analyzing the data using PyCaret, we found that the tree based boosting algorithm like XGBoost holds promise for our study. By combining generation and meteorological data and then preparing the data, we achieved better outcomes, as detailed in Table 6.3. We refined our approach by tweaking various hyperparameters. Table 6.4 displays the adjusted hyperparameter values along with their respective results.

**Table 6.3 Performance of Xgboost Algorithm under different scenarios**

| Scenario | RMSE(T) | SDCV | RMSESCV |
|---|---|---|---|
| **Plant1: Gandikotta, Andhra (2 months 22 power source)** | | | |
| XGBoost - Aggregated Data Without Weather Conditions: 3-Day Test Period | 3117.78 | | |
| Combined_Power_Output_Predictions_without_Weather_data | 3028.36 | 2012.62 | 4176.16 |
| Combined_Power_Output_Predictions_with_Weather_data | 3193.60 | 879.66 | 3209.81 |
| Seperated_Power_Output_Predictions_without_Weather_data | 3218.29 | 1213.69 | 3750.51 |

| | | | |
|---|---|---|---|
| **Seperated_Generation_Power_Output_Predictions** | 3253.61 | 1060.89 | 4002.42 |
| **Aggregated generation incorporating sine cosine features** | 3603.74 | 886.91 | 3252.58 |
| **Aggregated generation incorporating cosine features** | 3427.65 | 832.21 | 3305.72 |
| **Aggregated generation incorporating cosine features up to 4 days** | 3407.01 | 889.33 | 3285.12 |
| **Aggregated generation incorporating Lag features** | 3193.60 | 879.66 | 3209.81 |
| **Aggregated generation-weather incorporating rolling features** | 3327.27 | 668.09 | 3225.68 |
| **Plant1: Gandikotta, Andhra (24 months single power source)** | | | |
| **XGBoost - Aggregated Data Without Weather Conditions: 3-Day Test Period** | 3775.82 | 1927.25 | 3713.54 |
| **Combined_Power_Output_Predictions_without_Weather_data** | 3530.29 | 1652.14 | 4071.56 |
| **Combined_Power_Output_Predictions_with_Weather_data** | 3381.44 | 1563.43 | 3946.95 |
| **Seperated_Power_Output_Predictions_without_Weather_data** | 3472.18 | 1024.47 | 3558.73 |
| **Seperated_Generation_Power_Output_Predictions** | 3655.71 | 1684.76 | 3582.22 |
| **Aggregated generation incorporating sine cosine features** | 3636.69 | 1075.93 | 3469.27 |
| **Aggregated generation incorporating cosine features** | 3624.52 | 1631.01 | 3380.48 |
| **Aggregated generation incorporating cosine features up to 4 days** | 3521.09 | 1456.79 | 3942.80 |
| **Aggregated generation incorporating Lag features** | 3681.20 | 1885.03 | 3649.56 |
| **Aggregated generation-weather incorporating rolling features** | 3792.54 | 2237.16 | 3356.81 |

**Table 6.4 Hyperparameter values**

| Param | Indian Dataset | European Dataset | Purpose |
|---|---|---|---|
| **booster** | gbtree | gbtree | Type of model to run |
| **colsample_bytree** | 1 | 2 | Subsample ratio of columns when constructing each tree |
| **eta** | 0.065 | 0.149 | Step size shrinkage - prevent overfitting |

| | | | |
|---|---|---|---|
| **max_depth** | 1 | 1 | Tree depth |
| **n_estimators** | 70 | 55 | Number of trees to be built or booster round for best result |
| **subsample** | 0.8 | 0.5 | Subsample ratio of training instance |

**Table 6.5 Xgboost: Best Result**

| **Model Name** | **XGBoost** | | |
|---|---|---|---|
| **Plant Location** | **RMSE** | **SDCV** | **RMSESCV** |
| **Plant 1** | 3095.67 | 733.11 | 2882.46 |
| **Plant 2** | Not Applied | | |
| **Plant 2 – 2 Years** | 3377.92 | 1124.09 | 3167.33 |
| **European Dataset** | 0.6452 | Not Applied | |

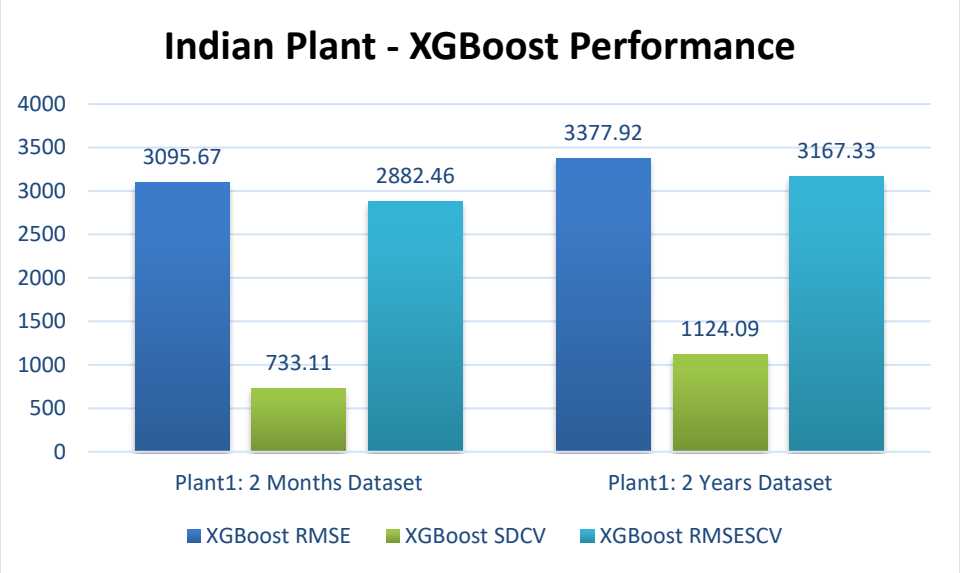


**Figure 6.3: Indian Plant – XGBoost Result Comparison**

After conducting XGBoost analysis on two different datasets on the same plant, one comprising a 2-month period and the other spanning 2 years of solar power generation data along with European dataset, several key findings have emerged.

**For the 2-month period dataset**:

- Utilizing XGBoost with aggregated data and excluding weather conditions for a 3-day test period yielded promising results.
- The Root Mean Square Error (RMSE) for the test set was recorded at 3095.67, indicating the model's predictive accuracy.
- Standard Deviation Cross Validation (SDCV) stood at 733.11, reflecting the consistency of the model's performance across different cross-validation folds.
- RMSE for Cross Validation (RMSESCV) was 2882.46, further affirming the robustness of the model.

**For the 2-year period dataset:**

- Analyzing a larger dataset spanning 2 years, focusing specifically on solar power generation, revealed slightly different outcomes.
- The XGBoost model achieved an RMSE of 3377.92 for the test set, indicating a slightly higher prediction error compared to the 6-month dataset.
- The Standard Deviation Cross Validation (SDCV) was recorded at 1124.09, suggesting a higher variability in the model's performance across cross-validation folds.
- RMSE for Cross Validation (RMSESCV) stood at 3167.33, showcasing the model's effectiveness in capturing the underlying patterns in the data over the 2-year period.

**For the European dataset:**

- The European solar power generation data presented a distinct scenario, considering predictions for a 10-time stamp ahead of forecast.
- A notable outcome was the impressively low Root Mean Square Error (RMSE) of 0.6452, a considerable improvement suggesting high precision in short-term forecasting.

The results reveal a significant boost in model efficacy after tuning various hyperparameters, underscoring the need for cautious optimization to avoid underfitting in specific scenarios.

Notably, the XGBoost algorithm outperforms the previously evaluated Prophet and SARIMA models in solar energy forecasting. This underscores XGBoost's capacity to manage diverse data types and forecasting periods, highlighting its adaptability and strength as a predictive tool in this field.

A detailed explanation is available here.

### 6.4.2 Deep Learning Approach

To seek better performance and higher recall value we progress our study into different deep-learning-based approaches specifically using the LSTM-based RNN model.

In the first approach we combined plant 1 and plant 2 generation data and tried to do intraday & day-ahead forecasting, but we observed that regional weather information was not captured which resulted from a failure in the day-ahead forecasting, so we moved to the second approach in which we first aggregate each plant's generation and weather data separately then move ahead for building LSTM network.

Below is the result comparison of 3 pure LSTM models applied to the dataset,

**Table 6.6 Pure LSTM Model Result – Indian Plant**

| Plant Location | Day Ahead | LSTM Model 1 | | LSTM Model 2 | | LSTM Model 3 | |
|---|---|---|---|---|---|---|---|
| | | RMSE(T) | Baseline | RMSE(T) | Baseline | RMSE(T) | Baseline |
| **Plant1** | 1 | 2871.60 | 3699.32 | 2778.47 | 3699.32 | 2832.88 | 3699.32 |
| **Plant2** | | 2375.08 | 3293.57 | 2376.99 | 3293.57 | 2383.42 | 3293.57 |
| **Plant1** | 7 | 2871.28 | 3699.32 | 2826.05 | 3699.32 | 2875.83 | 3699.32 |
| **Plant2** | | 2361.94 | 3293.57 | 2371.08 | 3293.57 | 2373.98 | 3293.57 |
| **Plant1: 2 Years** | 1 | Not Calculated | | | | | |
| **Plant1: 2 Years** | 7 | | | | | | |

We have come up with a combined model of Model 1 & Model 2 as we can see Model 1 works well on Plant 2 whereas Model 2 works well on Plant 1 which results as below,

**Table 6.7 Combined LSTM Model Result – Indian Plant**

| Plant Location | Day Ahead | Combined LSTM Model | |
|---|---|---|---|
| | | RMSE(T) | Baseline |
| **Plant 1** | 1 | 2802.86 | 3699.32 |
| **Plant 2** | | 2376.72 | 3293.57 |
| **Plant 1** | 7 | 2798.72 | 3699.32 |
| **Plant 2** | | 2368.05 | 3293.57 |
| **Plant 1: 2 Years** | 1 | 2850.12 | 3117.78 |
| **Plant 1: 2 Years** | 7 | 2715.45 | 3117.78 |

**Table 6.8 Combined LSTM Model Result – European Plant**

| Plant Location | Time Stamp Ahead | RMSE Score | $R^2$ |
|---|---|---|---|
| **European Solar Generation Data (1986-2015)** | Forecasting (10 Time Stamp Ahead) | 0.269 | 0.62 |
| | Forecasting (15 Time Stamp Ahead) | 0.296 | 0.59 |
| | Forecasting (20 Time Stamp Ahead) | 0.316 | 0.60 |
| | Forecasting (25 Time Stamp Ahead) | 0.346 | 0.59 |
| | Forecasting (30 Time Stamp Ahead) | 0.385 | 0.57 |
| | Forecasting (35 Time Stamp Ahead) | 0.426 | 0.52 |
| | Forecasting (40 Time Stamp Ahead) | 0.470 | 0.51 |
| | Forecasting (45 Time Stamp Ahead) | 0.516 | 0.49 |
| | Forecasting (50 Time Stamp Ahead) | 0.578 | 0.47 |

The findings from our analysis of the combined LSTM model suggest that this approach, integrating both Model 1 and Model 2, consistently outperforms or performs nearly as well as the standalone models across all prediction scenarios and time periods in our dataset. Specifically, the root means square error (RMSE) scores for the hybrid model consistently outshine those of the baseline models, indicating more precise forecasts when incorporating both generation and meteorological data. Additionally, our results indicate that this combined LSTM model surpasses the XGBoost model previously investigated.

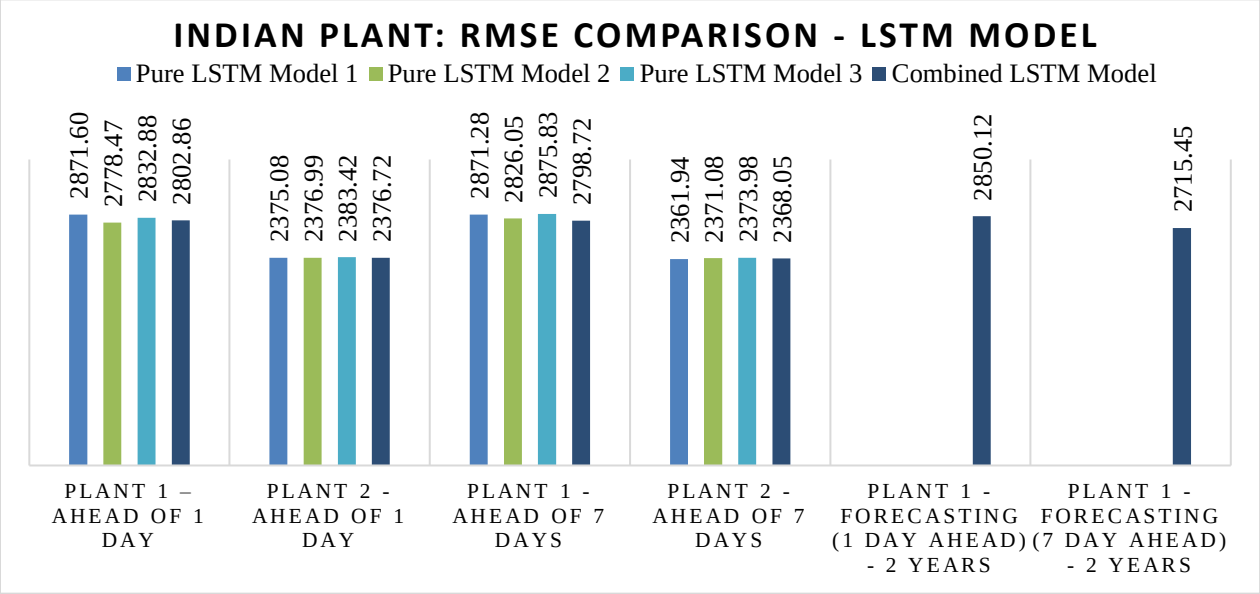


**Figure 6.4: Indian Plant – LSTM Result Comparison**

A detailed explanation is available here.

### 6.4.3 Hybrid Learning Approach (Proposed Approach)

All the models we've discussed up to now haven't done well in meeting our main goal of achieving a good Recall value. This leads us to delve deeper and look into the architecture we're suggesting in Figure 5.3. This architecture is based on ideas from **Agga et al. (2022)[60] and Luo et al. (2021)[61]**, who proposed a hybrid modeling approach in their research. We've also considered the analysis of uncertainty in solar power generation, introduced by **Lee, H.Y et al. (2019)[70],** to make our forecasting better.

Careful hyperparameter tuning yield below sets of hyperparameter for each dataset ,

**Table 6.9 Hybrid Model: Training Parameter**

| Dataset Location | Indian Dataset | | European Dataset |
|---|---|---|---|
| **Param** | **2 months duration** | **24 months duration** | **30 years duration** |
| No of Epoch | 300 | 300 | 200 |
| Optimizer | Adam | Adam | Adam |
| Loss | Mean Squared Error | Mean Squared Error | Mean Squared Error |
| Dropout Rate | 0.2 | 0.2 | 0.4 |
| Batch-Size | 1024 | 1024 | 5000 |
| Trainable Parameter | 424801 | 432599 | 642735 |

During our hyperparameter tuning process, we meticulously selected various parameters like activation functions, optimizers, and batch sizes. Our aim was to find the most effective combination that consistently produces the best results over a prolonged period of time. Below are the results of our simulations which we have performed on our Hybrid model,

**Simulation Criterion:**

- By Changing the **LSTM Unit**
- By Changing the **GRU Unit**
- By Changing the **Dropout Rate**
- By Changing the **Optimizer**
- By Changing the **Batch Size**
- By Changing the **Activation Function**

**Table 6.10 Combination tried on Indian Dataset**

| Scenario | Day Ahead | LSTM Units | GRU Units | Dropout Rate | Batch Size | Optimizer | LSTM Activation Layer | GRU Activation Layer | Activation FCN Layer |
|---|---|---|---|---|---|---|---|---|---|
| **Plant 1** | 1 | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU |
| | 7 | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU |
| | 1 | 15 | 15 | 0.2 | 1024 | Adam | Tanh | Tanh | ReLU |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 7 | 15 | 20 | 0.5 | 4096 | RMSprop | ReLU | Tanh | ReLU |
| | 1 | 15 | 20 | 0.3 | 1024 | RMSprop | Tanh | Tanh | Tanh |
| | 7 | 15 | 20 | 0.1 | 1024 | RMSprop | Tanh | ReLU | Sigmoid |
| | 1 | 10 | 25 | 0.5 | 4096 | SGD | Tanh | ReLU | ReLU |
| | 7 | 15 | 20 | 0.1 | 1024 | RMSprop | ReLU | Tanh | Sigmoid |
| | 1 | 20 | 15 | 0.1 | 2048 | SGD | Tanh | ReLU | ReLU |
| | 7 | 25 | 10 | 0.4 | 512 | Adam | ReLU | ReLU | Sigmoid |
| | 1 | 25 | 5 | 0.4 | 2048 | Adam | ReLU | Tanh | Tanh |
| | 7 | 20 | 10 | 0.3 | 512 | SGD | Tanh | Tanh | Sigmoid |
| | 1 | 25 | 5 | 0.4 | 2048 | Adam | ReLU | ReLU | Tanh |
| | 7 | 20 | 10 | 0.3 | 512 | SGD | Tanh | Tanh | ReLU |
| | 1 | 15 | 15 | 0.2 | 1024 | SGD | Tanh | ReLU | ReLU |
| | 7 | 20 | 15 | 0.2 | 2048 | Adam | ReLU | Tanh | Sigmoid |

After thoroughly confirming all combinations, we have found that the following results represent the most optimal combination for Indian Power Plants,

| Plant Location | Scenario | LSTM Units | GRU Units | Dropout Rate | Batch Size | Optimizer | LSTM Activation Layer | GRU Activation Layer | Activation FCN Layer | RMSE(T) | R Square | Avg. Recall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gandikota, Andhra | Plant 1 - Forecasting (1 Day Ahead) | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU | 1555.28 | 0.8067 | 0.8987 |
| Gandikota, Andhra | Plant 1 - Forecasting (7 Day Ahead) | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU | 1480.67 | 0.8053 | 0.864 |

**Figure 6.5: Indian Plant – Best combination of hyperparameter**

**Table 6.11 Combination tried on European Dataset**

| LSTM Units | GRU Units | Dropout Rate | Batch Size | Optimizer | LSTM Activation Layer | GRU Activation Layer | Activation FCN Layer |
|---|---|---|---|---|---|---|---|
| 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU |
| 15 | 15 | 0.2 | 1024 | Adam | Tanh | Tanh | ReLU |
| 15 | 20 | 0.3 | 1024 | RMSprop | Tanh | Tanh | Tanh |
| 10 | 25 | 0.5 | 4096 | SGD | Tanh | ReLU | ReLU |
| 20 | 20 | 0.4 | 2048 | Adam | Tanh | ReLU | Tanh |
| 25 | 15 | 0.2 | 512 | Adam | ReLU | Tanh | Tanh |
| 10 | 20 | 0.3 | 1024 | RMSprop | Tanh | ReLU | ReLU |
| 20 | 25 | 0.5 | 2048 | SGD | ReLU | ReLU | Tanh |
| 15 | 10 | 0.4 | 4096 | RMSprop | Tanh | ReLU | Tanh |

| Plant Location | Scenario | LSTM Units | GRU Units | Dropout Rate | Batch Size | Optimizer | LSTM Activation Layer | GRU Activation Layer | Activation FCN Layer | RMSE(T) | R Square | Avg. Recall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Different European Country | Forecasting (10 Time Stamp Ahead) | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU | 0.1794714 | 0.654 | 0.74896 |
| Different European Country | Forecasting (15 Time Stamp Ahead) | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU | 0.1958321 | 0.622 | 0.71785 |
| Different European Country | Forecasting (20 Time Stamp Ahead) | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU | 0.2149067 | 0.602 | 0.69234 |
| Different European Country | Forecasting (25 Time Stamp Ahead) | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU | 0.2370746 | 0.59 | 0.67543 |
| Different European Country | Forecasting (30 Time Stamp Ahead) | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU | 0.2627823 | 0.576 | 0.66321 |
| Different European Country | Forecasting (35 Time Stamp Ahead) | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU | 0.2925367 | 0.55 | 0.65232 |
| Different European Country | Forecasting (40 Time Stamp Ahead) | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU | 0.3278891 | 0.522 | 0.63754 |
| Different European Country | Forecasting (45 Time Stamp Ahead) | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU | 0.3704592 | 0.492 | 0.62345 |
| Different European Country | Forecasting (50 Time Stamp Ahead) | 15 | 15 | 0.2 | 1024 | Adam | Tanh | ReLU | ReLU | 0.4230585 | 0.48 | 0.60897 |

**Figure 6.6: European Plant – Best combination of hyperparameter**

Table 6.12 & 6.13 illustrate the detail results for both Indian power plant data as well as European power plant data after entire simulation,

**Table 6.12 Hybrid Model Result – Indian Plant**

| Plant Symbol | Day Ahead | RMSE(T) | $R^2$ | Avg. Recall |
|---|---|---|---|---|
| **Plant 1** | 1 | 1415.34 | 0.864 | 0.910 |
| **Plant 2** | 1 | 869.69 | 0.813 | 0.897 |
| **Plant 1** | 7 | 1363.38 | 0.856 | 0.930 |
| **Plant 2** | 7 | 868.39 | 0.816 | 0.878 |
| **Plant – 2 Years** | 1 | 1555.28 | 0.807 | 0.899 |
| **Plant – 2 Years** | 7 | 1480.67 | 0.805 | 0.864 |

Figure 6.7 illustrates the real solar power produced and predicted solar power for Indian Power Plant.

| | Real Solar Power Produced | Predicted Solar Power |
|---|---|---|
| **7** | 221,093.47 | 222,154.00 |
| **8** | 60,050.08 | 60,080.12 |
| **9** | 3,808,255.87 | 3,806,636.50 |
| **10** | 3,127,658.35 | 3,125,829.75 |
| **11** | 2,096,293.67 | 2,094,966.50 |
| **12** | 3,967,399.54 | 3,966,768.25 |
| **13** | 3,717,293.43 | 3,714,134.75 |
| **14** | 3,209,676.29 | 3,209,811.00 |
| **15** | 2,097,016.48 | 2,095,833.50 |
| **16** | 1,781,030.75 | 1,781,575.25 |
| **17** | 2,537,115.84 | 2,536,152.75 |

**Figure 6.7: Solar Power Predicted Values from Indian Plant**

**Table 6.13 Hybrid Model Result – European Plant**

| Plant Location | Scenario | RMSE(T) | $R^2$ | Avg. Recall |
|---|---|---|---|---|
| Different European Country | Forecasting (10 Time Stamp Ahead) | 0.179 | 0.654 | 0.749 |
| | Forecasting (15 Time Stamp Ahead) | 0.196 | 0.622 | 0.718 |
| | Forecasting (20 Time Stamp Ahead) | 0.215 | 0.602 | 0.692 |
| | Forecasting (25 Time Stamp Ahead) | 0.237 | 0.590 | 0.675 |
| | Forecasting (30 Time Stamp Ahead) | 0.263 | 0.576 | 0.663 |
| | Forecasting (35 Time Stamp Ahead) | 0.293 | 0.550 | 0.652 |
| | Forecasting (40 Time Stamp Ahead) | 0.328 | 0.522 | 0.638 |
| | Forecasting (45 Time Stamp Ahead) | 0.370 | 0.492 | 0.623 |
| | Forecasting (50 Time Stamp Ahead) | 0.423 | 0.480 | 0.609 |

In our forecasting process, we tweak the data to fit the timeframe we're interested in predicting – whether it's for the short or medium term. For example, when forecasting one day ahead, we move the AC_POWER or Total AC Power Output values forward by one day. This shift translates to 96 intervals for the Indian dataset and 24 intervals for the European dataset into the future. This adjustment helps our model accurately predict what will occur the following day. Conversely, for forecasting seven days ahead, we shift the AC_POWER or Total AC Power Output values forward by seven days. This provides us with a broader view of what might happen over the next week. By making these adjustments to the target series, we tailor our model to forecast either short-term changes or medium-term trends, depending on our specific forecasting requirements.

**For Indian Dataset:**

Data is gathered at 15-minute intervals, resulting in 96 equally spaced time intervals within a single day. To forecast one day ahead, 96 timestamps ahead predictions are required, while for a seven-day forecast, (96*7) = 672 timestamps ahead predictions are necessary. Following the application of this methodology to our dataset, the resultant plots depict the one-day and seven-day ahead predictions for both plants.

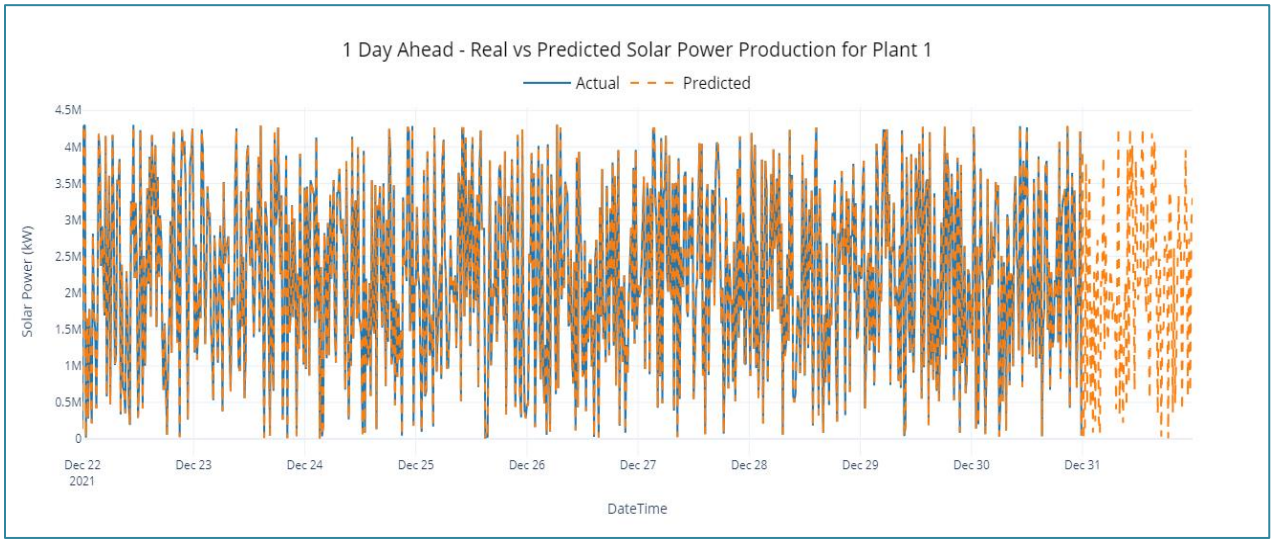


**Figure 6.8: Plant 1 - 1 Day ahead prediction**

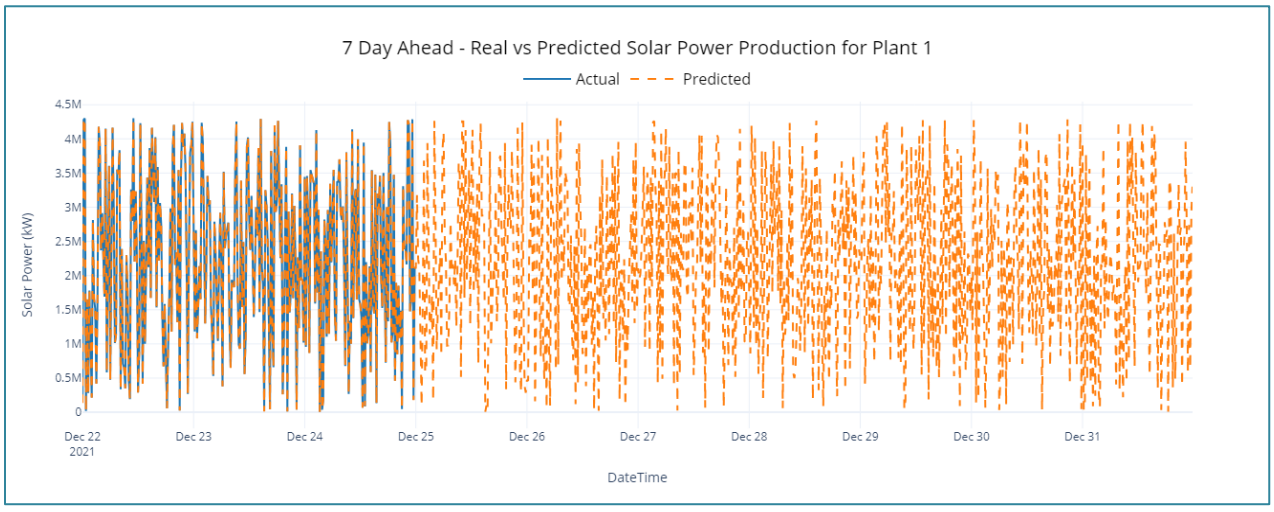


**Figure 6.9: Plant 1 - 7 Day ahead prediction**

**For European Dataset:**

Data is gathered at 60-minute intervals, resulting in 24 equally spaced time intervals within a single day. So here we have done forecasting for 10-50 timestamp ahead prediction.

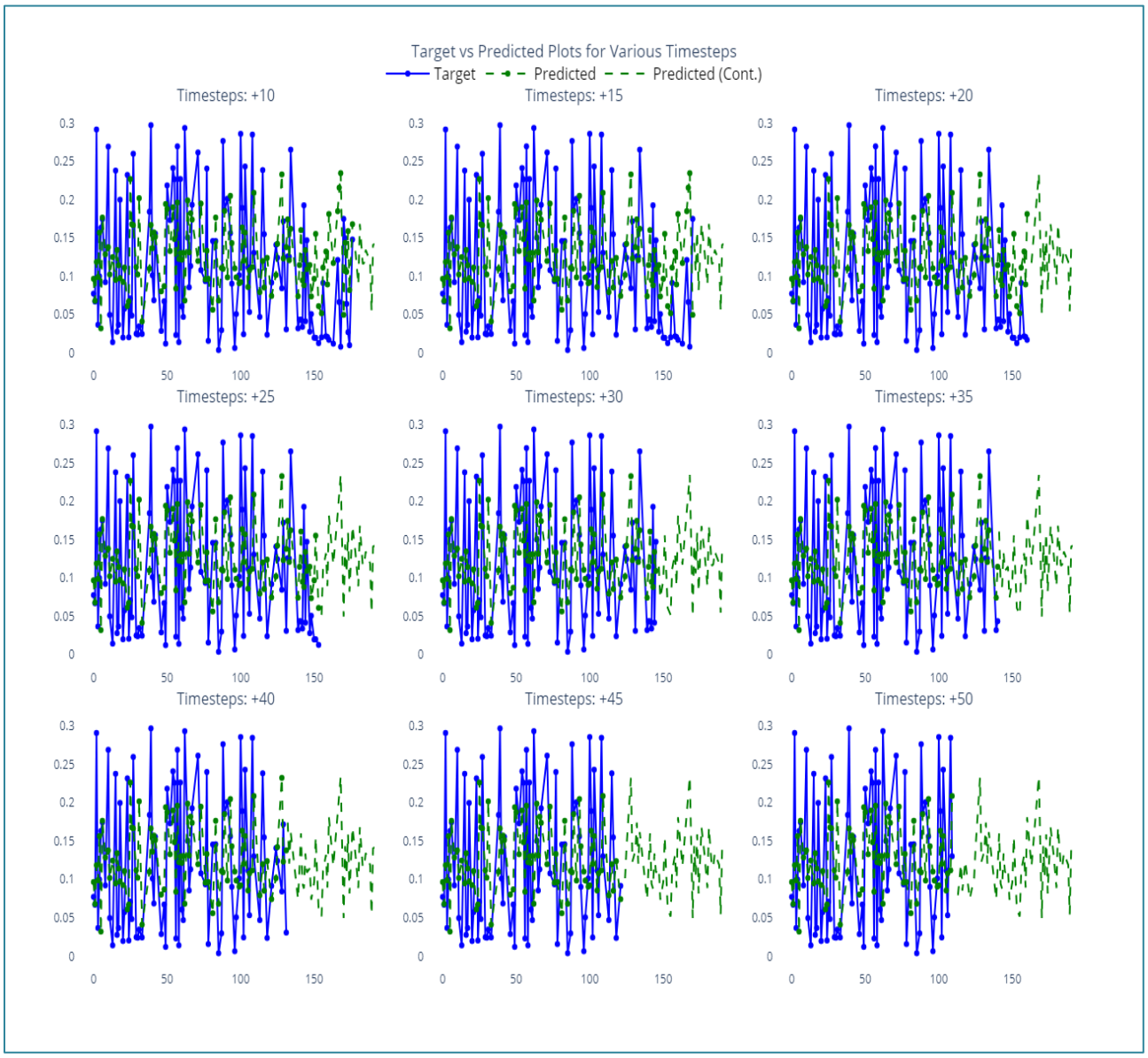


**Figure 6.10 Hybrid model performance on European data**

A detailed explanation is available here & here

Comparative Results of Different Learning method is shown below,

**Table 6.14 Model Performance Comparison on Different Dataset**

| Plant1: Gandi Kotta, Andhra Pradesh Plant2: Nasik, Maharashtra | | | Hybrid Model (Xgboost + RNN + LSTM + FCN) - Proposed Approach | | | Deep Learning Model (Pure LSTM) | | Pure Xgboost Algorithm | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Prediction Scenario** | **Location** | **RMSE Score** | **$R^2$** | **Recall** | **RMSE Score** | **$R^2$** | **RMSE Score** | **SDCV** | **RMSE CV** |
| **Plant1: (2 months 22 power source)** | Plant 1 (1 Day Ahead) | ***Gandi Kotta, Andhra*** | 1415.34 | 0.86 | 0.91 | 2802.86 | 0.81 | 3095.70 | 733.10 | 2882 |
| | Plant 1 (7 Day Ahead) | ***Gandi Kotta, Andhra*** | 1363.38 | 0.86 | 0.93 | 2798.72 | 0.83 | Not Applied | | |
| **Plant2: (2 months 22 power source)** | Plant 2 (1 Day Ahead) | ***Nasik, Maharashtra*** | 869.69 | 0.81 | 0.90 | 2376.72 | 0.78 | | | |
| | Plant 2 (7 Day Ahead) | ***Nasik, Maharashtra*** | 868.39 | 0.81 | 0.88 | 2368.05 | 0.75 | | | |
| **Plant1: (24 months single power source)** | Plant 1 (1 Day Ahead) | ***Gandi Kotta, Andhra*** | 1555.28 | 0.81 | 0.90 | 2850.12 | 0.80 | 3377.90 | 1124 | 3167 |
| | Plant 1 (7 Day Ahead) | ***Gandi Kotta, Andhra*** | 1480.67 | 0.81 | 0.86 | 2715.45 | 0.82 | Not Applied | | |
| **European Solar Generation Data (1986-2015)** | Forecasting 10 Time Stamp Ahead | ***18 Countries*** | 0.18 | 0.65 | 0.75 | 0.27 | 0.62 | 0.65 | Not Checked | |
| | Forecasting (15 Time Stamp Ahead) | ***18 Countries*** | 0.20 | 0.62 | 0.72 | 0.30 | 0.59 | Not Applied as initial result itself is not that good | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Forecasting (20 Time Stamp Ahead) | ***18 Countries*** | 0.21 | 0.60 | 0.69 | 0.32 | 0.60 | |
| | Forecasting (25 Time Stamp Ahead) | ***18 Countries*** | 0.24 | 0.59 | 0.68 | 0.35 | 0.59 | |
| | Forecasting (30 Time Stamp Ahead) | ***18 Countries*** | 0.26 | 0.58 | 0.66 | 0.39 | 0.57 | |
| | Forecasting (35 Time Stamp Ahead) | ***18 Countries*** | 0.29 | 0.55 | 0.65 | 0.43 | 0.52 | |
| | Forecasting (40 Time Stamp Ahead) | ***18 Countries*** | 0.33 | 0.52 | 0.64 | 0.47 | 0.51 | |
| | Forecasting (45 Time Stamp Ahead) | ***18 Countries*** | 0.37 | 0.49 | 0.62 | 0.52 | 0.49 | |
| | Forecasting (50 Time Stamp Ahead) | ***18 Countries*** | 0.42 | 0.48 | 0.60 | 0.58 | 0.47 | |

# Indian Solar Power Plant:

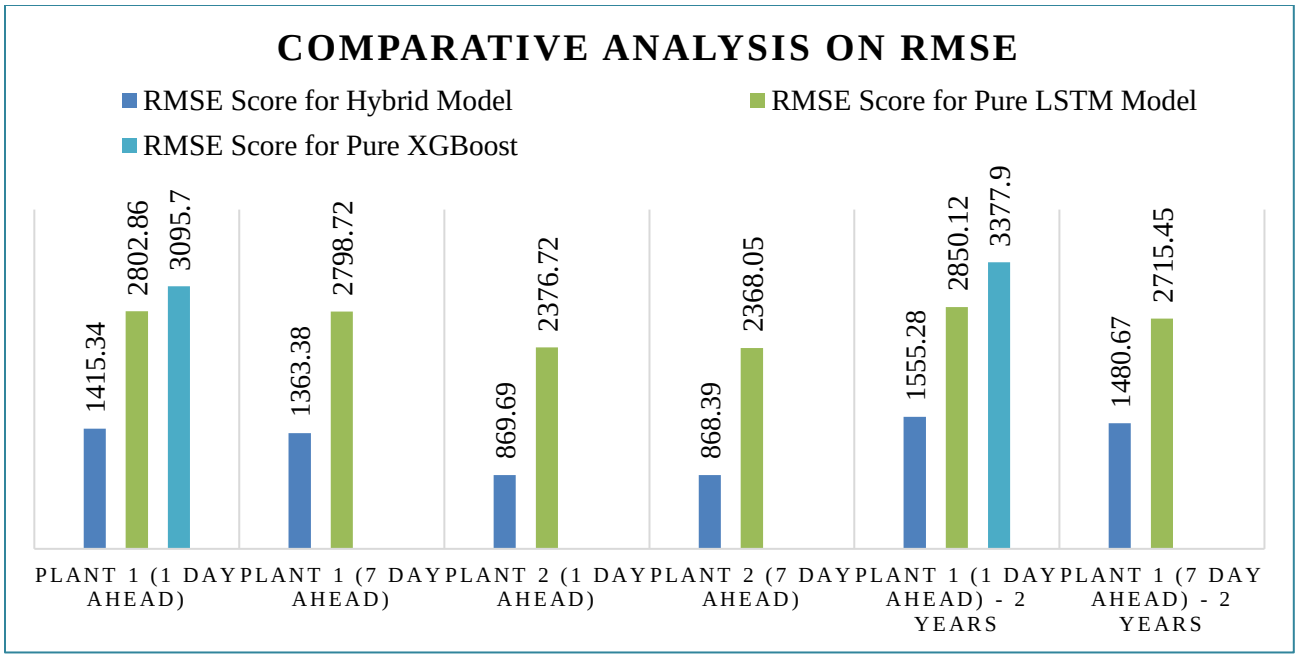


**Figure 6.11: RMSE Comparison for Indian Plant – blue bar indicating our proposed approach ensemble hybrid model result**

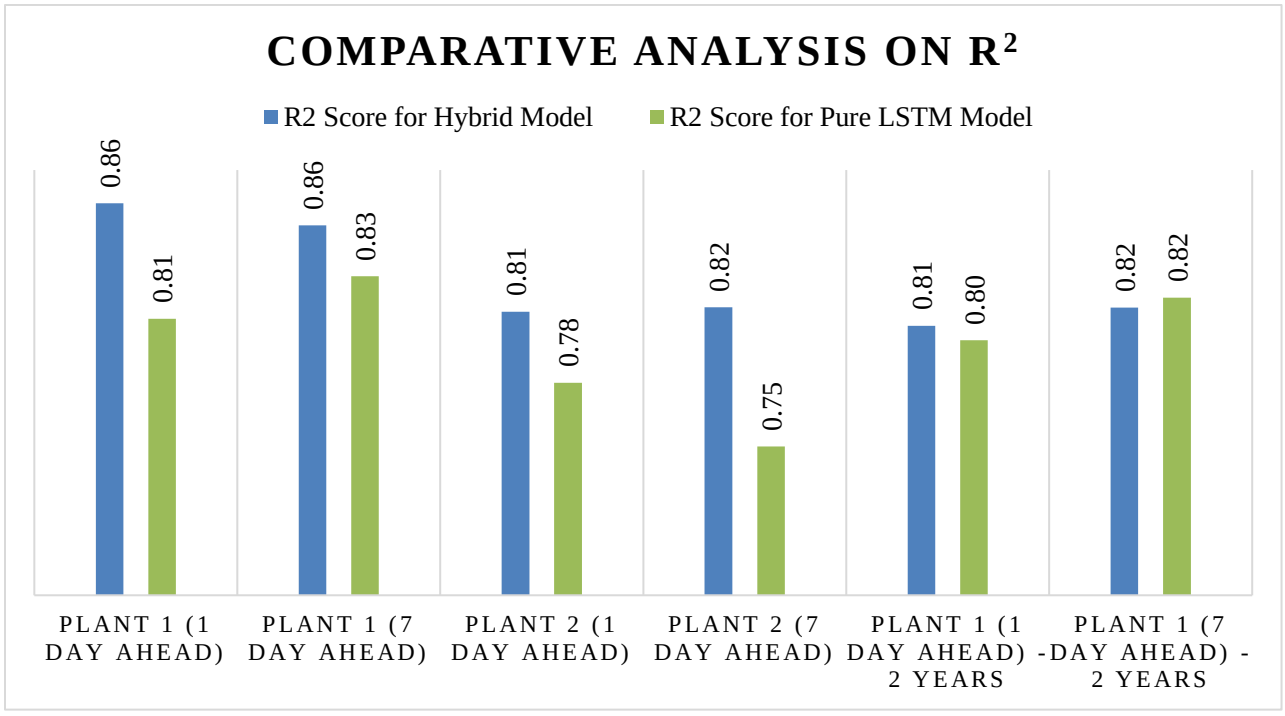


**Figure 6.12: R2 value for proposed approach ensemble hybrid model - Showing consistent high value**

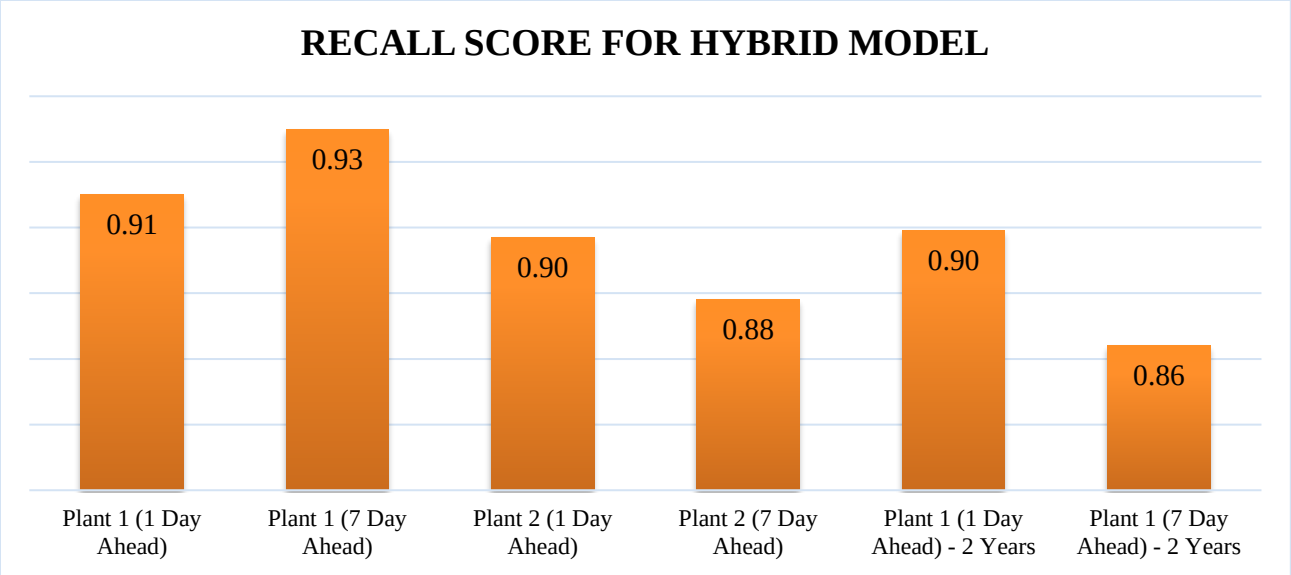


**Figure 6.13: Recall value for proposed approach ensemble hybrid model on Indian Data**

# European Solar Power Generation:

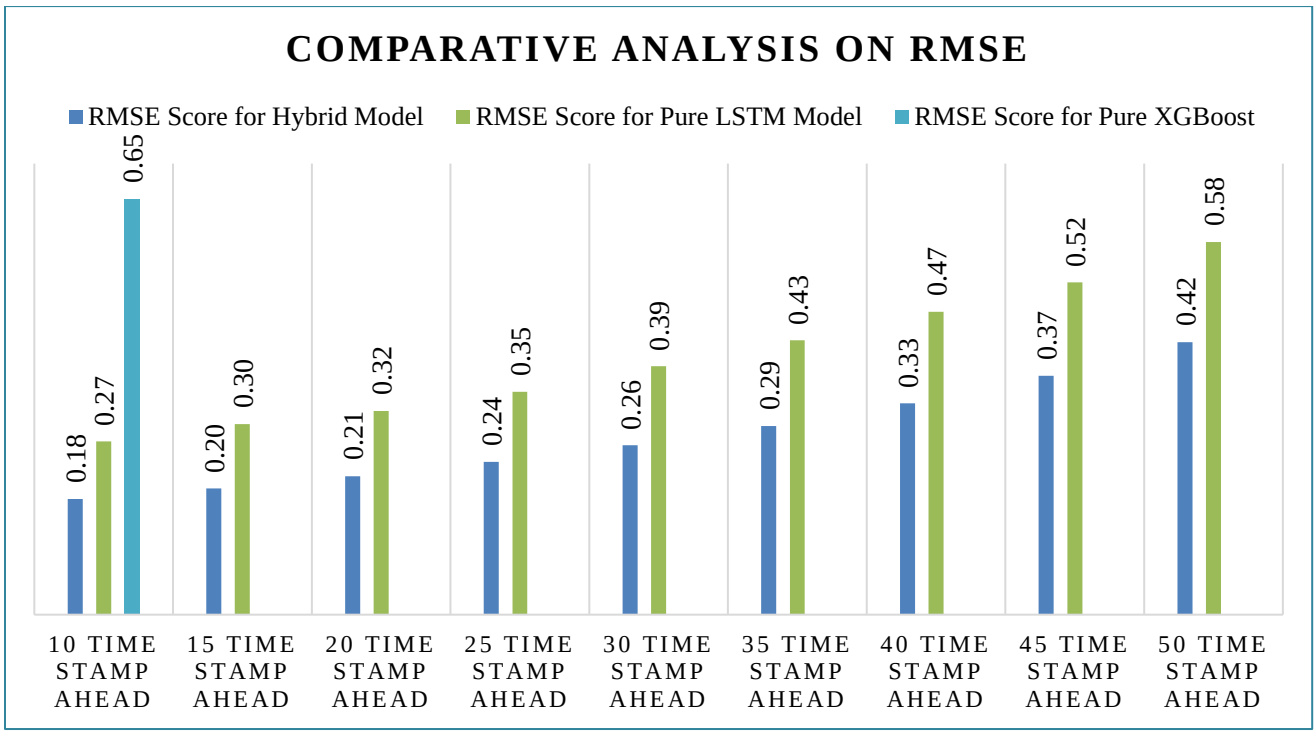


**Figure 6.14: RMSE Comparison for European Data – blue bar indicating our proposed approach ensemble hybrid model result**

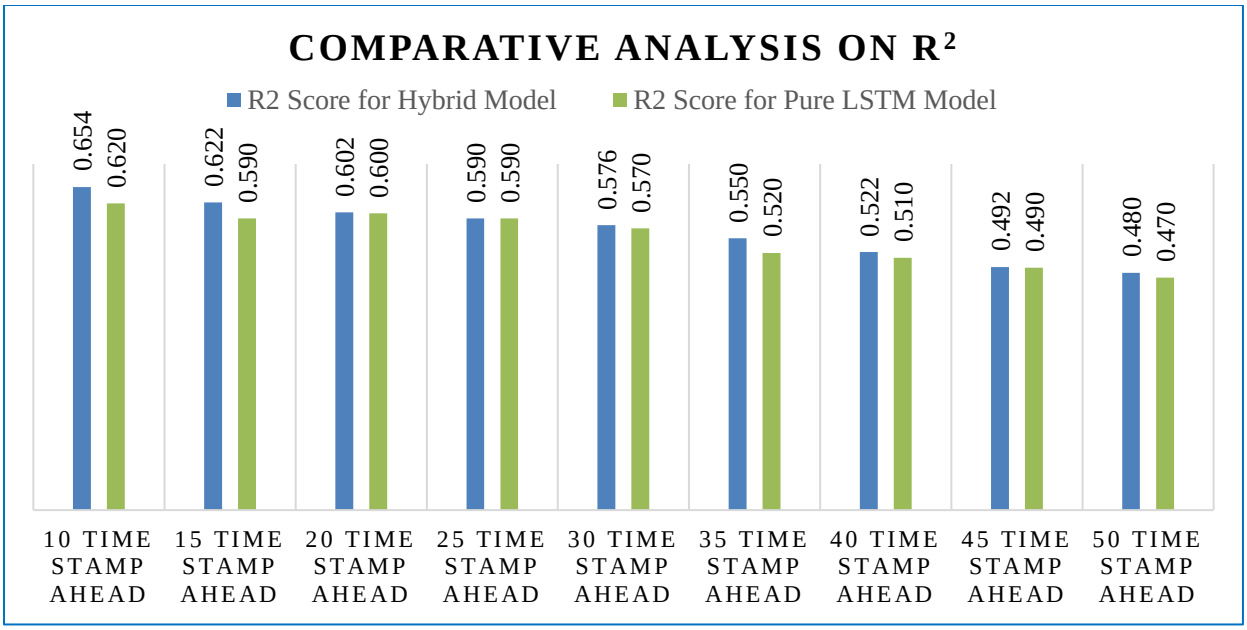


**Figure 6.15: R2 value for proposed approach ensemble hybrid model - Showing consistent high value**

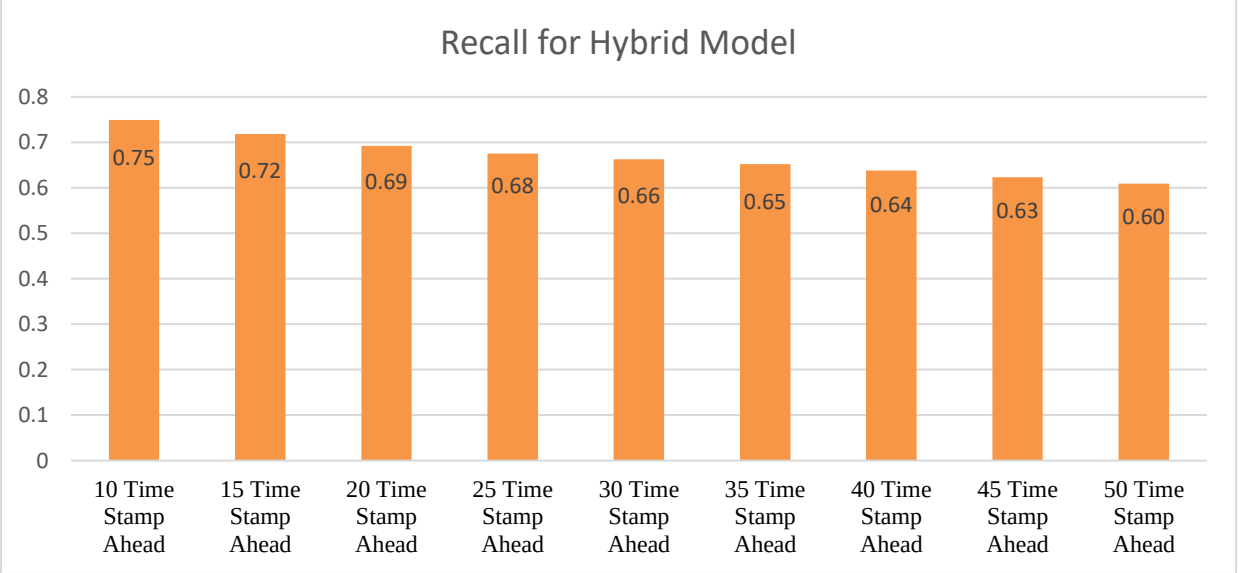


**Figure 6.16: Recall value for proposed approach ensemble hybrid model on European Data**

# CHAPTER 7

# CONCLUSION AND FUTURE SCOPE

This study adds to the current discussion on renewable energy prediction models by demonstrating the effectiveness of LSTM-based hybrid deep learning methods handling the uncertainties for predicting solar power. The results have important implications for improving grid stability and maximizing the use of renewable energy in the shift towards a sustainable energy environment. The research examines how well this technique captures intricate temporal patterns seen in solar power production. Exploring various statistical, ensemble, and deep learning methods leads to the following conclusion,

Based on the Chapter 6 result the following conclusions can be drawn:

1. The Hybrid Model, combining XGBoost with RNN and LSTM, displays superior predictive accuracy over the Pure LSTM Model and Pure XGBoost Algorithm across all datasets.
2. For Plant 1 (2 months data), the Hybrid Model's 1-day ahead prediction in Gandhi Kotta, Andhra Pradesh, achieves a notably lower RMSE of 1415.34 and a higher $R^2$ of 0.86, compared to other models.
3. The 7-day ahead forecast of the Hybrid Model for Plant 1 shows a consistent pattern of high accuracy with an RMSE of 1363.38 and an $R^2$ of 0.86, which is significantly better than the alternative models.
4. For Plant 2 (2 months data) in Nasik, Maharashtra, the Hybrid Model outperforms the others in the 1-day ahead forecast with an RMSE of 869.69 and an $R^2$ of 0.81.
5. The 7-day forecast for Plant 2 using the Hybrid Model also indicates superior performance with an RMSE of 868.39 and an $R^2$ of 0.82.
6. In the context of Plant 1's extended 24-month dataset, the Hybrid Model's 1-day ahead forecast scores an RMSE of 1555.28 and an $R^2$ of 0.81, revealing its effectiveness over longer periods.
7. The 7-day ahead forecast for the same 24-month dataset of Plant 1 gives an RMSE of 1480.67 and an $R^2$ of 0.79, indicating a marginal decrease in accuracy for longer forecast horizons.

8. The Recall metric for the Hybrid Model consistently shows its strength in predicting true positives across all scenarios, with high values close to 0.9, ensuring reliability in energy production forecasting.
9. The Deep Learning Pure LSTM Model generally exhibits higher RMSE scores and lower $R^2$ values across all scenarios, indicating less predictive reliability compared to the Hybrid Model.
10. The Pure XGBoost Algorithm is not employed for the 7-day forecasts and shows the highest RMSE and lowest $R^2$ in the 1-day ahead forecast for Plant 1, reflecting its lower standalone efficacy in short-term forecasts.
11. The European Solar Generation Data forecasting shows a decline in performance as the time step horizon extends for both models, but the Hybrid Model maintains a lead over the Pure LSTM, indicating better handling of complex temporal relationships.
12. For the European data, as the forecast horizon extends from 10 to 50-time stamps, the $R^2$ value for the Hybrid Model decreases from 0.65 to 0.48, showing the challenges in long-term forecasting.
13. The Pure LSTM Model's performance on the European dataset also deteriorates with increasing time steps, but it remains consistently less accurate than the Hybrid Model.
14. The application of the Pure XGBoost Algorithm for European data was limited due to its initial subpar performance, highlighting the necessity for more sophisticated approaches for long-term energy forecasting.
15. Overall, the Hybrid Model's integrated approach of machine learning and deep learning techniques across different geographies and timeframes demonstrates its robustness and potential for optimizing solar energy forecasting.

## 7.1 Limitation

- **Limited Data Scope**: The study may have been done by the availability and scope of the dataset, focusing primarily on two solar power plants within a specific geographical region as well as European dataset for 18 countries for 30 years. However, to make the model more useful globally, we need to gather data from other continents as well.

- **Global Forecasting Model & Advance Algorithm**: Utilize global forecasting model and integrate that model with proposed model to see if there is any improvement in performance happening. Also, exploration of attention mechanisms and transformer-based models in current problems.
- **Consideration of PV Panel Fault**: Considering PV panel faults in solar power prediction is essential as the efficiency and output of solar photovoltaic (PV) panels can be significantly affected by various faults, such as manufacturing defects, aging, environmental impacts, and mechanical damages. This particular aspect is not considered here in this study.

## 7.2 Future Directions

Future research directions should consider a multifaceted approach to enhance the robustness and applicability of the model. This entails investigating a wider array of ensemble and advanced deep learning architectures, such as attention mechanisms and transformer-based models, which may offer significant performance improvements. Incorporation of a broader set of features, potentially encompassing more diverse data types, could provide a more nuanced understanding of the underlying patterns and improve predictive accuracy. Moreover, adapting the model for real-time processing could greatly expand its practical applications.

# CHAPTER 8

# ADDITIONAL CONTENT

## 8.1 Mathematical Modelling

**Data Collection**:

Let's denote:

- $S(t)$ as the solar power generation data at time $t$.
- $M(t)$ as the meteorological data at time $t$.

**Integration of Data:**

Data from both sources is integrated on a timestamp basis, leading to a combined dataset

$$D(t) = \{S(t), M(t)\} \quad \ldots\ldots\ldots\ldots\ldots(1)$$

***Solar Power Generation Data**:*

$$S(t) = \{s_1(t), s_2(t), \ldots, s_n(t)\} \quad \ldots\ldots\ldots\ldots\ldots(2)$$

where $si(t)$ represents the $i^{th}$ solar power generation data measurement at time $t$, and $n$ is the number of solar power generation measurements.

***Meteorological Data**:*

$$M(t) = \{m_1(t), m_2(t), \ldots, m_k(t)\} \quad \ldots\ldots\ldots\ldots\ldots(3)$$

where $mj(t)$ represents the $j^{th}$ meteorological measurement at time $t$, and $k$ is the number of meteorological measurements.

From equation (1) we got,

$$\boldsymbol{D}(t) = \{S(t), M(t)\} = \{s_1(t), s_2(t), \ldots, s_n(t), m_1(t), m_2(t), \ldots, m_k(t)\} \ldots.(4)$$

For a sequence of timestamps $T = \{t1, t2, \ldots, tm\}$, the integrated dataset over all timestamps is given by:

$$D = U_{t \in T} D(t) \quad \ldots\ldots\ldots\ldots\ldots(5)$$

or more explicitly,

$$D = \{D(t_1), D(t_2), \ldots, D(t_m)\} \quad \ldots\ldots\ldots\ldots\ldots(6)$$

where ⋃ denotes the union of the datasets across all considered timestamps, essentially aggregating the data into a comprehensive dataset that includes both solar power generation and meteorological data for each timestamp.

This integrated dataset $D$ can then be utilized for further processing.

**Data Pre-processing:**

**Missing Value Imputation**: KNN imputation is utilized for handling missing values. For a missing data point at time t, we find k nearest neighbours based on a similarity measure (e.g., Euclidean distance) and impute the missing value by averaging these neighbours' values.

Steps as follows,

1. **Distance Calculation**: Compute the distance between *xi* and all other points in the dataset with non-missing values for the same feature. The distance can be calculated using various metrics, such as the Euclidean distance for continuous variables. For two points $xi$ and $xj$ in an *n*-dimensional space, the Euclidean distance is given by:

$$d(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sqrt{\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} (\boldsymbol{x_{ik}} - \boldsymbol{x_{jk}})^{\boldsymbol{2}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots(7)$$

2. **Find *k* Nearest Neighbors**: Select the *k* closest points (neighbors) based on the distance metric.
3. **Imputation**: Replace the missing value with the average (for continuous variables) or the mode (for categorical variables) of the *k* nearest neighbors. If the missing feature is continuous, the imputed value *mi* is calculated as:

$$m_i = \frac{1}{k}\sum_{j=1}^{k} x_j \quad \ldots\ldots\ldots\ldots\ldots\ldots(8)$$

where $x_j$ are the feature values of the *k* nearest neighbors.

**Min Max Scaling:** The scaling of a feature x is done as follows:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \ldots\ldots\ldots\ldots\ldots\ldots(9)$$

where:

- $x_{min}$ is the minimum value of feature X in the dataset.

- $x_{max}$ is the maximum value of feature X in the dataset.
- $x'$ is the scaled value, falling in the range [0, 1].

**Feature Engineering:**

We have performed detailed feature engineering in the data which mostly consists of cyclic, lag, rolling features along with original features.

**Cyclic Features via Sine-Cosine Transformations:** Cyclic features help in modelling the natural cycles, like hours of the day, days of the week, or months in a year, in a way that captures the continuity at the cycle endpoints. For a given time variable *t*, which could represent any cyclic measure (e.g., hour, day), we map it onto a unit circle using sine and cosine transformations to preserve its cyclic nature.

Given:

- *t*: Time variable (e.g., hour of the day).
- *T*: The total period of the cycle (e.g. *T*=24 for hours in a day).

The cyclic features are computed as:

- **Sine component**: $sin(\frac{2\pi t}{T})$ ……*(10)*
- **Cosine component**: $cos(\frac{2\pi t}{T})$……*(11)*

This transformation ensures that values at the end and beginning of the cycle are close to each other, maintaining the cyclic continuity.

**Lag Features:** Lag features are created to include information from previous time steps, which can be critical for time-series forecasting. For a given series $X = \{x_1, x_2, \dots, x_n\}$ and a lag of $\tau$ time steps, the lag feature $X_{\text{lag\_}\tau}$ at time *t* is defined as:

$$X_{\text{lag\_}\tau}(t) = x_{t-\tau} \text{……*(12)*}$$

For instance, with $\tau = 1$, $X_{\text{lag}_\tau}(t)$ $r$epresents the value of *x* from the previous time step.

**Rolling Features:** Rolling features capture the moving aggregate statistics (e.g., mean, sum) over a window of *w* periods to incorporate trends and patterns over time. For a series $X = \{x_1, x_2, \dots, x_n\}$ and a window size *w*, the rolling mean and rolling sum at time *t* can be defined as follows:

- **Rolling Mean**: $\text{RollingMean}_w\ (t) = \frac{1}{w}\sum_{i=tw+1}^{t} x_i$ ……*(13)*

- **Rolling Sum**: $\text{RollingSum}_w w(t) = \sum_{i=t-w+1}^{t} x_i$ ……*(14)*

Other rolling features can include standard deviation, variance, min, and max, each providing different insights into the past behavior of the series over the window.

**Integration in Feature Engineering:**

The comprehensive feature set after engineering may look like this for a dataset $D$ with initial features $X$ and a time variable $t$ :

1 **Cyclic Features**: For each cyclic time variable, compute its sine and cosine components.

2 **Lag Features**: Determine the relevant lags $\tau$ and compute the lag features for the selected variables.

3 **Rolling Features**: Choose a window size $w$ and calculate the rolling statistics for the selected variables.
The enhanced dataset $D'$ now contains:

- **Original features**: $X = \{x_1, x_2, \ldots, x_n\}$
- **Cyclic features**: $\left\{\sin\left(\frac{2\pi t}{T}\right), \cos\left(\frac{2\pi t}{T}\right)\right\}$ for each cyclic time variable
- **Lag features**: $\{X_{\text{lag\_}\tau}(t)\}$ for selected $\tau$ and variables
- **Rolling features**: $\{\text{RollingMean}_w(t), \text{RollingSum}_w(t)\}$ for selected window sizes and variables

These mathematical formulations of feature engineering techniques enhance the dataset, making it more suitable for complex modelling techniques by capturing essential patterns, trends, and dependencies in the data.

**XGBoost Feature Extraction:**

**Objective Function:**
XGBoost optimizes an objective function that consists of a loss function $L$ and a regularization term. For a set of training data $\{(x_i, y_i)\}_{i=1}^{n}$ where $x_i$ is the feature vector of the $i$-th sample and $y_i$ is its label, and a prediction $\hat{y}_i$ for the $i$-th sample, the objective function $Obj$ is given by:

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad \ldots\ldots\ldots\ldots\ldots(15)$$

- $L(y_i, \hat{y}_i)$ is the loss function that measures the difference between the predicted value $\hat{y}_i$ written as $\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i)$ and the actual label $y_i$.

- $\Omega(f_k)$ is the regularization term that penalizes the complexity of the model. $\Omega(f) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$, Here, $f_k$ represents the $k$-th decision tree in the ensemble, and $K$ is the total number of trees.
- The regularization term typically includes parameters that control the tree's depth and the minimum loss reduction required to make a further partition on a leaf node of the tree.

**Feature Importance:**

XGBoost provides several methods to calculate feature importance, which quantitatively evaluates how each feature contributes to the predictive power of the model. The most common methods include:

1 **Weight (Frequency)**: This is the simplest form of feature importance given by XGBoost, which counts how many times a feature is used to split the data across all trees. If a feature is used often to make decisions, it is considered more important.

2 **Gain**: The gain of a feature is calculated as the average gain of the feature across all its occurrences in the trees of the model. If a feature $f$ is used to split a node, resulting in two child nodes with values $\hat{y}_L$ and $\hat{y}_R$ respectively, and the split results in a gain $G$, then the gain can be quantified in terms of reduction in loss, given by the formula:

$$G = \frac{1}{2}\left[\frac{(\sum_{i-L} g_i)^2}{\sum_{i(L} h_i+\lambda} + \frac{(\sum_{i\in R} g_i)^2}{\sum_{i\in R} h_i+\lambda} - \frac{(\sum_{iCN} g_i)^2}{\sum_{iCN} h_i+\lambda}\right] - \gamma \quad ...(16)$$

where:

- $L$ and $R$ denote the sets of instances in the left and right child nodes, respectively.
- $N$ denotes the set of instances in the parent node.
- $g_i$ and $h_i$ are the first and second order gradients of the loss function with respect to the prediction for instance $i$, respectively.
- $\gamma$ and $\lambda$ are regularization parameters.

3 **Cover**: The cover metric measures the average number of samples affected by the splits on a given feature across all trees. It reflects the relative quantity of observations concerned by a feature.
Mathematically, for a given feature $f$, its importance based on gain can be computed as:

$$\text{Importance } (f) = \frac{\sum \text{ Gain of splits using } f}{\text{Number of splits using } f} \ldots\ldots\ldots\ldots\ldots(17)$$

Where the gain from a split on feature $f$ is improvement in performance (e.g., reduction in Gini impurity or entropy for classification, variance reduction for regression) from that split.

The formula of $f$ provides an aggregate measure of a feature's overall importance across all trees in the model, while the formula for $G$ details the specific gain from a single split involving feature f in one tree.

**Feature Selection Process:**

After training the XGBoost model, the feature importances are assessed using one of the methods described above. Features are then ranked based on their importance scores. The selection process involves retaining features with importance scores above a certain threshold or selecting a top N set of features. This process helps in identifying and keeping the most relevant features that contribute significantly to the model's predictive performance, thereby enhancing model simplicity, interpretability, and often performance on unseen data.

**Deep Learning Network with Uncertainty handling Layer:**

Let $x^{(t)}$ be the input vector at time step $t$, and let $h_{\text{LNN}}^{(t)}$, $h_{\text{LSTM}}^{(t)}$, and $h_{\text{GRU}}^{(t)}$ be the hidden states of the RNN, LSTM, and GRU layers at time step $t$, respectively.

- **Recurrent Neural Network Layer (RNN):**

$$h_{\text{RNN}}^{(t)} = \text{RNN}\left(x^{(t)}, h_{\text{RNN}}^{(t-1)}; \theta_{\text{RNN}}\right) \ldots (17)$$

Here, $h_{RNN}^{(t-1)}$ is the hidden state from the previous time step, and $\theta_{\text{RNN}}$ represents the RNN parameters.

- **Long Short-Term Memory Layer (LSTM):**

$$h_{\text{LSTM}}^{(t)} = \text{LSTM}\left(h_{\text{RNN}}^{(t)}, h_{\text{LSTM}}^{(t-1)}, c_{\text{LSTM}}^{(t-1)}; \theta_{\text{LSTM}}\right) \ldots (18)$$

The LSTM layer takes the output of the RNN layer $h_{\text{RNN}}^{(t)}$ along with its own previous hidden state $h_{\text{LSTM}}^{(t-1)}$ and previous cell state $c_{\text{LSTM}}^{(t-1)}$ $\cdot\, \theta_{\text{LSTM}}$ are the parameters of the LSTM.

- **Gated Recurrent Unit Layer (GRU):**

$$h_{\text{GRU}}^{(t)} = \text{GRU}\left(h_{\text{LSTM}}^{(t)}, h_{\text{GRU}}^{(t-1)}; \theta_{\text{GRU}}\right) \ldots (19)$$

The GRU layer receives the output from the LSTM layer $h_{\text{LSTM}}^{(t)}$ and its previous hidden state $h_{\text{GRU}}^{(t-1)}$ $\cdot\, \theta_{\text{GRU}}$ denotes the GRU parameters.
The output of the GRU layer $h_{\text{GRU}}^{(t)}$ is then passed through fully connected layers

(FCN):

- **Fully Connected Neural Layer (FCN):**

$$a^{(l)} = \text{ReLU}\left(W^{(l)} h_{\text{GrU}}^{(t)} + b^{(l)}\right) \quad \ldots(20)$$
$$h^{(l)} = \text{Dropout}\left(a^{(l)}, p\right)$$

Where:

- $W^{(l)}$ and $b^{(l)}$ are the weights and biases of the I-th FCN layer.
- $a^{(l)}$ is the activation of the l-th FCN layer.
- $h^{(l)}$ is the output of the I-th FCN layer after applying dropout with probability $p$.
- ReLU is used as the activation function to introduce non-linearity.

The output of the final FCN layer $h^{(l)}$ is used for the prediction:

- **Output Layer:**

$$\hat{y} = \text{Final Activation}\left(W^{\text{out}} h^{(l)} + b^{\text{out}}\right) \ldots(21)$$

Here $W^{\text{out}}$ and $b^{\text{ont}}$ are the weights and biases of the output layer, and Final Activation can be a SoftMax or linear activation function depending on the problem.

**Uncertainty Layer:**

❖ **Monte Carlo Dropout for Uncertainty Estimation:**

In a neural network, dropouts are typically used during training to prevent overfitting. However, Monte Carlo Dropout uses dropout during prediction as well to provide an estimate of the model's uncertainty.

❖ **Training Phase with Dropout:** During training, for each forward pass and for each dropout layer $l$, a binary dropout mask $m^{(l)}$ is sampled from a Bernoulli distribution with probability $p$ (the dropout rate):

$$m^{(l)} \sim \text{Bernoulli}\ (p)$$

The output of each dropout layer $h^{(l)}$ is then:

$$h^{(l)} = m^{(l)} * a^{(l)}$$

Where $a^{(l)}$ is the activation from the previous layer.
Prediction Phase with Monte Carlo Dropout:
During prediction, the same dropout procedure is applied, and multiple forward passes ( $M$ times) are made through the network to obtain a set of predictions:

$$\hat{Y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_M\}$$

The mean of these predictions gives the expected value of the output, and the variance gives an estimate of the uncertainty:

$$\mu = \frac{1}{M}\sum_{i=1}^{M} \hat{y}_i$$
$$\sigma^2 = \frac{1}{M}\sum_{i=1}^{M} (\hat{y}_i - \mu)^2 \quad \text{...(22)}$$

Here, $\mu$ is the mean prediction and $\sigma^2$ is the variance representing the uncertainty.

- **Quantile Regression for Forecast Uncertainty Analysis:**

Quantile regression differs from ordinary least squares regression in that it estimates the conditional quantiles of the response variable, providing a more complete view of the possible outcomes.

**Quantile Loss Function:** The loss function for quantile regression for a given quantile $q$ (where $0 < q < 1$ ) is:

$$\mathcal{L}_q(y_i, \hat{y}_i) = q \cdot \max(y_i - \hat{y}_i, 0) + (1 - q) \cdot \max(\hat{y}_i - y_i, 0)$$

Where $y_i$ is the true value and $\hat{y}_i$ is the predicted value.

**Quantile Forecasting:** The model is trained to minimize the quantile loss, and the output is the predicted quantile $\hat{y}_i^{\mu}$ of the target distribution:

$$\hat{y}_i^q = \text{argmin}_{\hat{y}_i}\ \mathcal{L}_q(y_i, \hat{y}_i)$$

By computing predictions for multiple quantiles (e.g., $q = 0.1,0.4,0.5,0.6,0.9$ ), we obtain a distribution of predicted values that capture the range possible outcomes.

❖ **Combined Approach for Uncertainty Handling:**

The integrated approach combines the uncertainty estimates from Monte Carlo dropout with the quantile predictions from Quantile Regression.

- **Combining Predictions and Uncertainties:**
  For each quantile $q$, you perform Monte Carlo predictions to obtain $\hat{Y}^q = \{\hat{y}_1^q, \hat{y}_2^q, \ldots, \hat{y}_M^q\}$. Then, you calculate the mean and variance for each quantile set of predictions:

$$\mu^q = \frac{1}{M}\sum_{i=1}^{M} \hat{y}_i^q$$

$$(\sigma^q)^2 = \frac{1}{M}\sum_{i=1}^{M} \left(\hat{y}_i^q - \mu^q\right)^2$$

- **Interpretation of Combined Uncertainty:**
  The mean predictions $\mu^q$ across different quantiles provide a range of forecasts representing different confidence levels. The associated variances $(\sigma^q)^2$ provide an estimate of the uncertainty for each quantile forecast.

## 8.2 Forecast Uncertainty Analysis Modelling

Our approach to forecast uncertainty analysis integrates two powerful methodologies: Monte Carlo Simulation and Quantile-Based Forecasting[70]. Monte Carlo Dropout, a practical technique leveraging dropout layers, provides ease of implementation and computational efficiency during both training and prediction phases. However, its tendency towards optimistic uncertainties may underestimate true uncertainty. Complementing this, our Quantile-Based Forecast Uncertainty Analysis employs Quantile Regression to generate diverse scenarios, capturing a wide range of confidence levels. By focusing on key percentiles and deriving insightful metrics like mean and variance, we attain a comprehensive understanding of the forecast distribution. This integrated approach equips us with valuable insights to make informed decisions amidst uncertainty, blending computational efficiency with comprehensive uncertainty assessment.

| Plant ID | Real Solar Power Produced | Predicted Solar Power | DateTime | Source ID | 60th_mean | 50th_var | Q40to60 | Q10to90 | upper_limit | lower_limit |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.093130e+08 | 2.083390e+08 | 2020-06-11 13:30:00 | 1BY6WEcLGh8j5v7 | 2.166726e+08 | 2.517499e+14 | 2.083390e+08 | 2.083390e+08 | 2.083390e+08 | 2.083390e+08 |
| 1 | 2.008957e+07 | 1.832925e+07 | 2020-06-11 13:30:00 | 1IF53ai7Xc0U56Y | 1.906242e+07 | 1.948575e+12 | 1.832925e+07 | 1.832925e+07 | 1.832925e+07 | 1.832925e+07 |
| 1 | 7.337972e+06 | 6.765022e+06 | 2020-06-11 13:30:00 | 3PZuoBAID5Wc2HD | 7.035623e+06 | 2.654400e+11 | 6.765022e+06 | 6.765022e+06 | 6.765022e+06 | 6.765022e+06 |
| 1 | 6.423608e+06 | 6.732382e+06 | 2020-06-11 13:30:00 | 7JYdWkrLSPkdwr4 | 7.001677e+06 | 2.628848e+11 | 6.732382e+06 | 6.732382e+06 | 6.732382e+06 | 6.732382e+06 |
| 1 | 2.017522e+07 | 1.863017e+07 | 2020-06-11 13:30:00 | McdE0feGgRqW7Ca | 1.937538e+07 | 2.013084e+12 | 1.863017e+07 | 1.863017e+07 | 1.863017e+07 | 1.863017e+07 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2 | 2.827201e+08 | 2.820992e+08 | 2020-06-09 12:30:00 | wCURE6d3bPkepu2 | 2.933832e+08 | 4.615638e+14 | 2.820992e+08 | 2.820992e+08 | 2.820992e+08 | 2.820992e+08 |
| 2 | 6.388164e+06 | 6.673086e+06 | 2020-06-09 12:30:00 | z9Y9gH1T5YWrNuG | 6.940009e+06 | 2.582744e+11 | 6.673086e+06 | 6.673086e+06 | 6.673086e+06 | 6.673086e+06 |
| 2 | 1.215303e+09 | 1.214478e+09 | 2020-06-09 12:30:00 | zBIq5rxdHJRwDNY | 1.263057e+09 | 8.554743e+15 | 1.214478e+09 | 1.214478e+09 | 1.214478e+09 | 1.214478e+09 |
| 2 | 2.247736e+09 | 2.245427e+09 | 2020-06-09 12:30:00 | zVJPv84UY57bAof | 2.335245e+09 | 2.924328e+16 | 2.245427e+09 | 2.245427e+09 | 2.245427e+09 | 2.245427e+09 |
| 2 | 7.273453e+06 | 6.788574e+06 | 2020-06-09 12:45:00 | YxYtjZvoooNbGkE | 7.060117e+06 | 2.672915e+11 | 6.788574e+06 | 6.788574e+06 | 6.788574e+06 | 6.788574e+06 |

**Figure 8.1: Uncertainty Measurement in Solar Power Generation for Plant 1, 2 Years Data**

In conclusion we can say that while the uncertainty inherent in solar power generation presents a notable challenge for accurate prediction due to the complex interplay of weather, environmental factors, and materials, our analysis indicates that this uncertainty can be effectively quantified and may be lower than traditionally expected. Addressing this reduced uncertainty is still paramount for reliable forecasting and efficient grid management. Through the exploration of various techniques, including Monte Carlo Simulation, Bayesian Neural Networks, Gaussian Processes, Ensemble Methods, and Deep Ensemble, we find that Monte Carlo Simulation, and specifically Monte Carlo Dropout, stands out as a practical approach owing to its straightforward implementation and computational efficiency. However, it is vital to remain cautious of the possibility of optimistic uncertainties that may underestimate the true variability. To mitigate this, we incorporate a Quantile-Based Forecast Uncertainty Analysis utilizing Quantile Regression. This method not only facilitates the generation of diverse scenarios but also imparts a more nuanced appreciation of the forecast distribution, suggesting that the confidence in the predictions might be well-founded. By amalgamating these methodologies, we construct a comprehensive and resilient framework for uncertainty assessment, which fortifies decision-making amidst the dynamic and often unpredictable solar power generation landscape.

## 8.3 Forecast using Transformer based Modelling

As part of hyperparameter tuning we have tried multiple combination as below,

1. Varying the node of LSTM & GRU unit
2. Change the nearest neighbour for KNN imputation
3. Varying the dropout rate

By doing the same we have come up with below result,

**Table 8.1 Transformer Model Performance Comparison by Hyperparameter Tuning**

| Hyperparameter Tuning of Transformer Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Scenario** | **Max Epochs** | **Dropout Rate** | **KNN Neighbour** | **Unit of Neuron** | **Day Ahead** | **RMSE** | **R2** |
| **Plant 1** | Varying Dropout & Unit of Neuron | 2000 | 0.95 | 10 | 16 | 1 | 1379.92 | 0.87 |
| **Plant 1** | Varying Dropout & Unit of Neuron | 2500 | 0.95 | 10 | 16 | 1 | 1427.77 | 0.85 |
| **Plant 1** | Varying Dropout & Unit of Neuron | 2000 | 0.98 | 10 | 16 | 7 | 1462.73 | 0.85 |
| **Plant 1** | Varying Dropout & Unit of Neuron | 2000 | 0.99 | 10 | 16 | 7 | 1457.51 | 0.85 |
| **Plant 1** | Varying Dropout & Unit of Neuron | 2500 | 0.95 | 10 | 16 | 7 | 1435.11 | 0.85 |
| **Plant 2** | Varying Dropout & Unit of Neuron | 2000 | 0.95 | 10 | 16 | 1 | 1191.23 | 0.81 |
| **Plant 2** | Varying Dropout & Unit of Neuron | 2500 | 0.95 | 10 | 16 | 1 | 1193.42 | 0.81 |
| **Plant 2** | Varying Dropout & | 2000 | 0.6 | 10 | 16 | 7 | 1186.34 | 0.82 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Unit of Neuron | | | | | | | |
| **Plant 2** | Varying Dropout & Unit of Neuron | 2500 | 0.7 | 10 | 16 | 7 | 1198.05 | 0.81 |
| **Plant 1** | Varying KNN Neighbour | 2000 | 0.95 | 25 | 16 | 1 | 1388.87 | 0.86 |
| **Plant 1** | Varying KNN Neighbour | 2500 | 0.95 | 20 | 16 | 1 | 1375.40 | 0.87 |
| **Plant 1** | Varying KNN Neighbour | 2000 | 0.95 | 50 | 16 | 7 | 1384.28 | 0.86 |
| **Plant 1** | Varying KNN Neighbour | 2500 | 0.95 | 30 | 16 | 7 | 1381.04 | 0.86 |
| **Plant 1** | Varying KNN Neighbour | 2500 | 0.95 | 45 | 16 | 7 | 1381.99 | 0.86 |
| **Plant 2** | Varying KNN Neighbour | 2000 | 0.95 | 40 | 16 | 1 | 1183.90 | 0.82 |
| **Plant 2** | Varying KNN Neighbour | 2500 | 0.95 | 40 | 16 | 1 | 1186.23 | 0.81 |
| **Plant 2** | Varying KNN Neighbour | 2000 | 0.6 | 10 | 16 | 7 | 1186.34 | 0.82 |
| **Plant 2** | Varying KNN Neighbour | 2500 | 0.7 | 10 | 16 | 7 | 1186.92 | 0.81 |

Comparative study of the transformer model with all other applied model for Indian Dataset,

**Table 8.2 Model Performance Comparison**

| **Dataset Name** | **Day Ahead** | **Models** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Proposed Approach** | | **Transformer Based Model** | | **Deep Learning Model (Pure LSTM)** | | **XGBoost** | |
| | | **RMSE** | **R2** | **RMSE** | **R2** | **RMSE** | **R2** | **RMSE** | **R2** |
| **Plant 1** | 1 | 1415.34 | 0.86 | 1379.92 | 0.87 | 2802.86 | 0.81 | 3377.92 | 0.65 |
| | 7 | 1363.38 | 0.86 | 1381.04 | 0.86 | 2798.72 | 0.83 | NA | |
| **Plant 2** | 1 | 869.69 | 0.81 | 1183.90 | 0.82 | 2376.72 | 0.78 | NA | |
| | 7 | 868.39 | 0.82 | 1186.34 | 0.82 | 2368.05 | 0.75 | | |

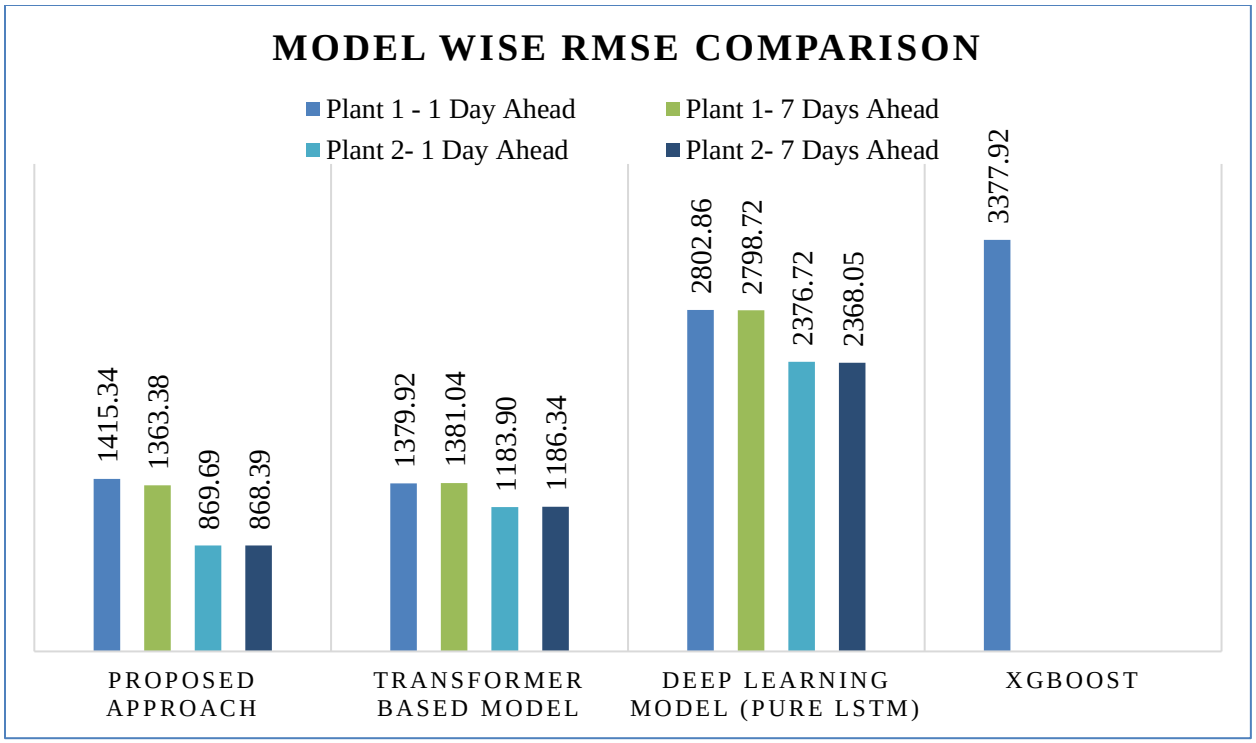


**Figure 8.2: RMSE score comparison for different models on Power Generation**

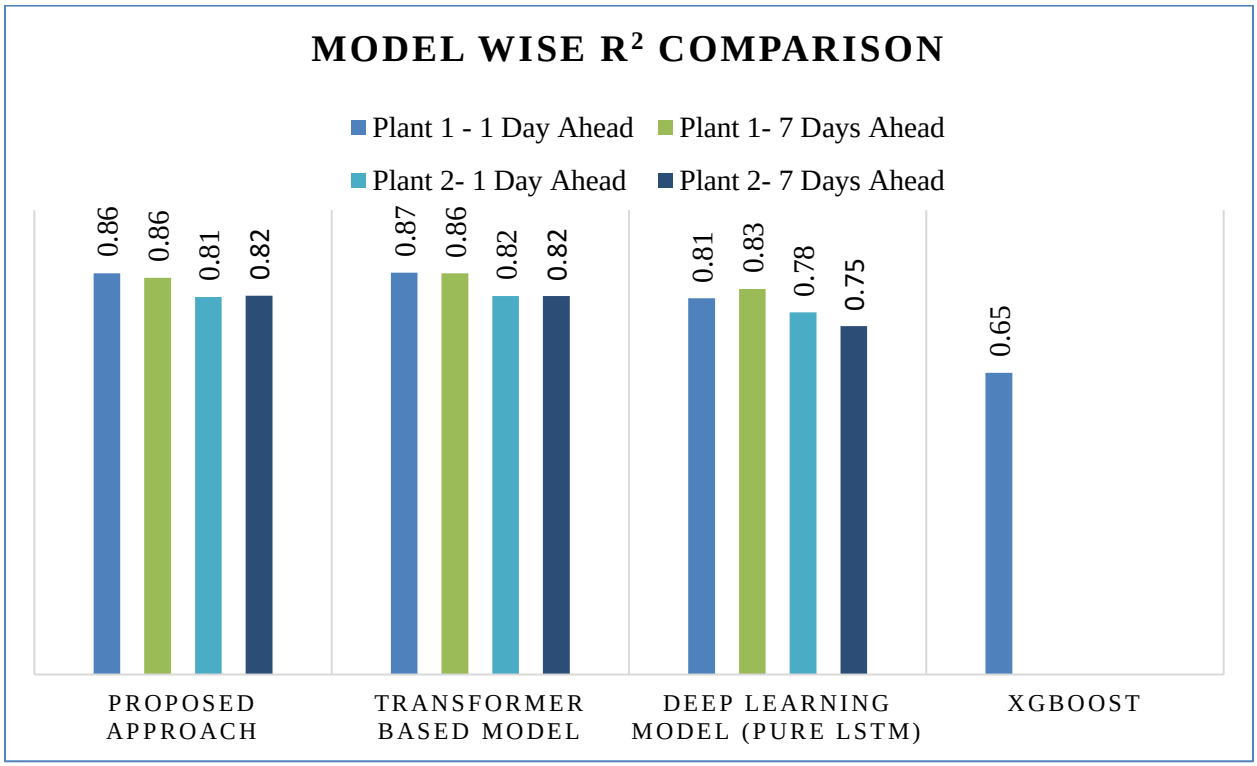


**Figure 8.3: R2 score comparison for different models on Power Generation**

## Conclusion:

The graphical analysis of the Root Mean Square Error (RMSE) and R-squared ($R^2$) values from multiple forecasting models reveals distinct performance characteristics for Plant 1 and Plant 2, predicting 1-day ahead and 7-days ahead electricity generation. The Proposed Approach generally exhibits superior predictive accuracy with lower RMSE values, especially for Plant 2, and maintains relatively high $R^2$ values, indicating a robust model performance. Conversely, the Transformer Model, while consistent across both plants, does not achieve the lower RMSE levels of the Proposed Approach, despite demonstrating competitive $R^2$ values. Additionally, it is worth noting that the Proposed Approach exhibits a notable advantage in terms of computational efficiency and training speed compared to the Transformer Model. Pure LSTM Model underperforms in comparison, with higher RMSE values and moderate $R^2$ values. Notably, XGBoost demonstrates the highest RMSE for Plant 1 with a 1-day ahead for the and exhibits a significant discrepancy in performance with an $R^2$ of 0.65, suggesting less reliability for the data set and forecasting horizon considered. These comparative insights are pivotal for model selection in predictive maintenance and operational planning within the energy sector, where the balance between error minimization and the explanation of variance is critical.

# REFERENCES


[1]. Kundjanasith, Thonglek., Kohei, Ichikawa., Keichi, Takahashi., Chawanat, Nakasan., Kazufumi, Yuasa., Tadatoshi, Babasaki., Hajimu, Iida. (2023). Toward Predictive Modeling of Solar Power Generation for Multiple Power Plants. IEICE Transactions on Communications, doi: 10.1587/transcom.2022ebt0003

[2]. "A Novel Forecasting Model for Solar Power Generation by a Deep Learning Framework With Data Preprocessing and Postprocessing." IEEE Transactions on Industry Applications, undefined (2023). doi: 10.1109/tia.2022.3212999

[3]. Quoc-Thang, Phan., Yuan-Kang, Wu., Quoc, Dung, Phan., Hsin-Yen, Lo. (2023). A Novel Forecasting Model for Solar Power Generation by a Deep Learning Framework With Data Preprocessing and Postprocessing. IEEE Transactions on Industry Applications, doi: 10.1109/TIA.2022.3212999

[4]. Hamad, Alharkan., Shabana, Habib., Muhammad, Islam. (2023). Solar Power Prediction Using Dual Stream CNN-LSTM Architecture. Sensors, doi: 10.3390/s23020945

[5]. E., Subramanian., M., M., Karthik., G., Krishna., Dharam, Prasath., V., Kumar. (2023). Solar Power Prediction Using Machine Learning. arXiv.org, doi: 10.48550/arXiv.2303.07875

[6]. Mustafa, Yasin, Erten., Hüseyin, Aydilek. (2022). Solar Power Prediction using Regression Models. Uluslararası mühendislik araştırma ve geliştirme dergisi, doi: 10.29137/umagd.1100957

[7]. Irfan, Jamil., Hong, Lucheng., Muhammad, Aurangzaib., Rehan, R., Jamil., Hossam, R., Kotb., Abdulaziz, Alkuhayli., Kareem, M., AboRas. (2023). Predictive evaluation of solar energy variables for a large-scale solar power plant based on triple deep learning forecast models. alexandria engineering journal, doi: 10.1016/j.aej.2023.06.023

[8]. Lisang, Liu., Kaiqi, Guo., Cheng-Hao, Ke., Dongwei, He. (2023). A Photovoltaic Power Prediction Approach Based on Data Decomposition and Stacked Deep Learning Model. Electronics, doi: 10.3390/electronics12132764

[9]. Jamshid, Aghaei. (2023). Geographic information system-based prediction of solar power plant production using deep neural networks. Iet Renewable Power Generation, doi: 10.1049/rpg2.12781

[10]. I., V., Gontovaya. (2023). Power generation forecast for a solar plant with a deep-learning method. doi: 10.5194/egusphere-egu23-2107

[11]. "Futuristic deep learning algorithms for long-term solar power prediction." undefined (2023). doi: 10.21203/rs.3.rs-2830639/v1

[12]. Hasan, Alkahtani., Theyazn, H., H., Aldhyani., Saleh, Nagi, Alsubari. (2023). Application of Artificial Intelligence Model Solar Radiation Prediction for Renewable Energy Systems. Sustainability, doi: 10.3390/su15086973

[13]. Taco, Niet., Nastaran, Arianpoo., Kamaria, Kuling., Andy, Wright. (2022). Increasing the reliability of energy system scenarios with integrated modelling: a review. Environmental Research Letters, doi: 10.1088/1748-9326/ac5cf5

[14]. Michael, O., Dioha., Magnus, C., Abraham-Dukuma., Natalia, Bogado., Francis, N., Okpaleke. (2020). Supporting climate policy with effective energy modelling: A perspective on the varying technical capacity of South Africa, China, Germany and the United States. Energy research and social science, doi: 10.1016/J.ERSS.2020.101759

[15]. Tanveer, Ahmad., Dongdong, Zhang., Wahab, Ali, Shah. (2020). Efficient Energy Planning With Decomposition-Based Evolutionary Neural Networks. IEEE Access, doi: 10.1109/ACCESS.2020.3010782

[16]. Edmund, Widl., Giorgio, Agugiaro., Jan, Peters-Anders. (2021). Linking Semantic 3D City Models with Domain-Specific Simulation Tools for the Planning and Validation of Energy Applications at District Level. Sustainability, doi: 10.3390/SU13168782

[17]. Shimeng, Hao., Tianzhen, Hong. (2021). The Application of Urban Building Energy Modeling in Urban Planning. doi: 10.1007/978-3-030-71819-0_3

[18]. Xiaoyang, Wang., Siming, Chen., Yun, Lin, Sun., Si, Chen. (2023). Short-term Power Prediction Method of Photovoltaic Based on Output Clustering in Smart Grid. Advances in Engineering Technology Research, doi: 10.56028/aetr.5.1.477.2023

[19]. Mingkang, Guo., Wenxuan, Ji. (2023). Research on Photovoltaic Power Prediction Method for Power Grid Safety. doi: 10.1109/CISCE58541.2023.10142818

[20]. Xuan, Jiao., Xingshuo, Li., Dingyi, Lin., Weidong, Xiao. (2022). A Graph Neural Network Based Deep Learning Predictor for Spatio-Temporal Group Solar Irradiance Forecasting. IEEE Transactions on Industrial Informatics, doi: 10.1109/tii.2021.3133289

[21]. N., M., Sabri., Mohammed, El, Hassouni. (2022). Accurate photovoltaic power prediction models based on deep convolutional neural networks and gated recurrent units. Energy Sources Part A-recovery Utilization and Environmental Effects, doi: 10.1080/15567036.2022.2097751

[22]. Muamar, Mohamed., Farhad, E., Mahmood., Mehmmood, Abdulla, Abd., Ambrish, Chandra., Bhim, Singh. (2022). Dynamic Forecasting of Solar Energy Microgrid Systems Using Feature Engineering. IEEE Transactions on Industry Applications, doi: 10.1109/TIA.2022.3199182

[23]. Nwaigwe, K. N., Mutabilwa, P., & Dintwa, E. (2019). An overview of solar power (PV systems) integration into electricity grids. *Materials Science for Energy Technologies*, *2*(3), 629-633.

[24]. Inman, R. H., Pedro, H. T., & Coimbra, C. F. (2013). Solar forecasting methods for renewable energy integration. *Progress in energy and combustion science*, *39*(6), 535-576.

[25]. Weisong, Wang. (2023). Prediction of photovoltaic power generation based on LSTM and transfer learning digital twin. Journal of physics, doi: 10.1088/1742-6596/2467/1/012015

[26]. Qianqian, Li., Dongping, Zhang., Ke, Sheng, Yan. (2023). A Solar Irradiance Forecasting Framework Based on the CEE-WGAN-LSTM Model. Sensors, doi: 10.3390/s23052799

[27]. Sakshi, Shukla., Sarita, Sheoran., Sumanta, Pasari. (2022). Prediction of Solar Energy using Time Series Methods. doi: 10.1109/ICACRS55517.2022.10028997

[28]. Mustafa, Yasin, Erten., Hüseyin, Aydilek. (2022). Solar Power Prediction using Regression Models. Uluslararası mühendislik araştırma ve geliştirme dergisi, doi: 10.29137/umagd.1100957

[29]. Thummuluru, Kavitha., S., Hemalatha. (2023). Forecasting the Solar Power with Differentiation of Data Figures using Neural Networks. doi: 10.1109/ICCCI56745.2023.10128530

[30]. Said, Benkirane., Azidine, Guezzaz., Mourade, Azrour., Abderrahim, Beni-Hssane. (2023). A Novel Machine Learning Approach for Solar Radiation Estimation. Sustainability, doi: 10.3390/su151310609

[31]. Rachna., Ajay, K., Singh. (2023). Prediction of Photovoltaic Power Generation using Machine Learning - A Review. doi: 10.1109/InCACCT57535.2023.10141769

[32]. Shashikant, Kaushaley., Binod, Shaw., Jyoti, Ranjan, Nayak. (2023). Optimized Machine Learning-Based Forecasting Model for Solar Power Generation by Using Crow Search Algorithm and Seagull Optimization Algorithm. Arabian journal for science and engineering, doi: 10.1007/s13369-023-07822-9

[33]. Chibuzor, N, Obiora., Ali, N., Hasan., Ahmed, A., Ali. (2023). Predicting Solar Irradiance at Several Time Horizons Using Machine Learning Algorithms. Sustainability, doi: 10.3390/su15118927

[34]. E., Subramanian., M., M., Karthik., G., Krishna., Dharam, Prasath., V., Kumar. (2023). Solar Power Prediction Using Machine Learning. arXiv.org, doi: 10.48550/arXiv.2303.07875

[35]. K, Venu., Somasekhar, Jayaram., Kadian, Renu. (2023). Solar Radiation Prediction using Machine Learning Model. doi: 10.1109/ICSCDS56580.2023.10104904

[36]. K.N., Sangeetha., Suganthi, P. (2023). Integrating Machine Learning Algorithms for Predicting Solar Power Generation. E3S web of conferences, doi: 10.1051/e3sconf/202338701004

[37]. Moulshree, Anjum, M., Goel, A. (2022). Solar Flare Prediction using Machine Learning Algorithms on RHESSI Dataset. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 1-5). https://doi.org/10.1109/ICSCDS53736.2022.9760755

[38]. Higa, S., Iwashita, Y., Otsu, K., Ono, M., Lamarre, O., Didier, A., & Hoffmann, M. (2019). Vision-Based Estimation of Driving Energy for Planetary Rovers Using Deep Learning and Terramechanics. IEEE Robotics and Automation Letters, 4(4), 4297-4304. https://doi.org/10.1109/LRA.2019.2928765

[39]. Kartini, U. T., Choiroh, U. N., & Kartini, U. T. (2021). Very Short Term Photovoltaic Power Generation Station Forecasting Based On Meteorology Using Hybrid model Decomposition-Deep Neural Network. In 2021 Fourth International Conference on Vocational Education and Electrical Engineering (ICVEE) (pp. 1-6). https://doi.org/10.1109/ICVEE54186.2021.9649698

[40]. Silva, J. F., Silva, F. A., Kazmierkowski, M. P., & Kazmierkowski, M. (2021). Computational Intelligence for Modeling, Control, Optimization, Forecasting and Diagnostics in Photovoltaic Applications [Book News]. IEEE Industrial Electronics Magazine, 15(3), 79-80. https://doi.org/10.1109/MIE.2021.3071208

[41]. Üstün, Ä., Üneş, F., Mert, Ä., Karakuş, C., & Mert, I. (2020). A comparative study of estimating solar radiation using machine learning approaches: DL, SMGRT, and ANFIS. Energy Sources Part A: Recovery, Utilization, and Environmental Effects, 42(19), 2367-2384. https://doi.org/10.1080/15567036.2020.1781301

[42]. Moulshree, Anjum, M., Goel, A. (2022). Solar Flare Prediction using Machine Learning Algorithms on RHESSI Dataset. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 1-5). https://doi.org/10.1109/ICSCDS53736.2022.9760755

[43]. Zech, M., & von Bremen, L. (2021). Understanding the relationship between clouds and surface downward radiation forecast errors with Unsupervised Deep Learning. European Meteorological Society Annual Meeting. https://doi.org/10.5194/ems2021-471

[44]. Zaman, M., Saha, S., Eini, R., Abdelwahed, S. (2021). A Deep Learning Model for Forecasting Photovoltaic Energy with Uncertainties. In 2021 IEEE Green Energy and Smart Systems Conference (IGESSC) (pp. 1-6). https://doi.org/10.1109/IGESSC53124.2021.9618681

[45]. Kothona, D., Panapakidis, I. P., & Christoforidis, G. C. (2021). An Hour-Ahead Photovoltaic Power Forecasting Based on LSTM Model. In 2021 IEEE Madrid PowerTech (pp. 1-6). https://doi.org/10.1109/PT46648.2021.9494841

[46]. Phan, Q. T., Wu, Y. K., & Phan, Q. D. (2022). An Approach Using Transformer-based Model for Short-term PV generation forecasting. International Conference on Applied System Innovation. https://doi.org/10.1109/ICASI55125.2022.9774491

[47]. Sedai, A. K., Dhakal, R., Gautam, S., Dhamala, A., Bilbao, A., Wang, Q., ... Pol, S. (2023). Performance Analysis of Statistical, Machine Learning and Deep Learning Models in Long-Term Forecasting of Solar Power Production. Forecasting, 5(1), 153-171. https://doi.org/10.3390/forecast5010014

[48]. Cano-Martínez, J., Peñalvo-López, E., León-Martínez, V., & Valencia-Salazar, I. (2023). Dynamic energy prices for residential users based on Deep Learning prediction models of

consumption and renewable generation. The Renewable Energies and Power Quality Journal (RE&PQJ), 1(21), 226. https://doi.org/10.24084/repqj21.226

[49]. Tanha, S. N., Mim, S. A., Roy, P., & Razzaque, M. A. (2021). Prediction of Energy Harvesting in Solar Powered Small Cells Networks. In 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI) (pp. 1-6). https://doi.org/10.1109/STI53101.2021.9732578

[50]. Gorantla, K. R., & Roy, A. (2023). Generalizable Solar Irradiation Prediction using Large Transformer Models with Sky Imagery. In 2023 18th International Conference on Machine Vision and Applications (MVA). https://doi.org/10.23919/MVA57639.2023.10216081

[51]. Hari, N. G., & Jisha, G. (2022). Solar Irradiance Prediction using Deep Learning-Based Approaches. In 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) (pp. 1-6). https://doi.org/10.1109/CSDE56538.2022.10089282

[52]. Sahu, H., Rao, W., Troisi, A., Ma, J. (2018). Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. Advanced Energy Materials, 8(1), 1701032. https://doi.org/10.1002/aenm.201801032

[53]. Dwivedi, Divyanshi, K. Victor Sam Moses Babu, Pradeep Kumar Yemula, Pratyush Chakraborty, and Mayukha Pal. "Identification of surface defects on solar PV panels and wind turbine blades using attention-based deep learning model." Engineering Applications of Artificial Intelligence 131 (2024): 107836.

[54]. Wang, Jianzhou, Xinsong Niu, Lifang Zhang, Zhenkun Liu, and Xiaojia Huang. "A wind speed forecasting system for the construction of a smart grid with two-stage data processing based on improved ELM and deep learning strategies." Expert Systems with Applications 241 (2024): 122487.

[55]. Chakraborty, Debojyoti, Jayeeta Mondal, Hrishav Bakul Barua, and Ankur Bhattacharjee. "Computational solar energy–Ensemble learning methods for prediction of solar power generation based on meteorological parameters in Eastern India." Renewable Energy Focus 44 (2023): 277-294.

[56]. Nwokolo, Samuel Chukwujindu, Anthony Umunnakwe Obiwulu, and Julie C. Ogbulezie. "Machine learning and analytical model hybridization to assess the impact of climate change on solar PV energy production." Physics and Chemistry of the Earth, Parts A/B/C 130 (2023): 103389.

[57]. Zheng, Jianqin, Jian Du, Bohong Wang, Jiří Jaromír Klemeš, Qi Liao, and Yongtu Liang. "A hybrid framework for forecasting power generation of multiple renewable energy sources." Renewable and Sustainable Energy Reviews 172 (2023): 113046.

[58]. Ozbek, Arif, Alper Yildirim, and Mehmet Bilgili. "Deep learning approach for one-hour ahead forecasting of energy production in a solar-PV plant." Energy Sources, Part A: Recovery, Utilization, and Environmental Effects 44, no. 4 (2022): 10465-10480.

[59]. Khan, Waqas, Shalika Walker, and Wim Zeiler. "Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach." Energy 240 (2022): 122812.

[60]. Agga, Ali, Ahmed Abbou, Moussa Labbadi, Yassine El Houm, and Imane Hammou Ou Ali. "CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production." Electric Power Systems Research 208 (2022): 107908.

[61]. Luo, Xing, Dongxiao Zhang, and Xu Zhu. "Deep learning based forecasting of photovoltaic power generation by incorporating domain knowledge." Energy 225 (2021): 120240.

[62]. Mellit, A., A. Massi Pavan, and V. Lughi. "Deep learning neural networks for short-term photovoltaic power forecasting." Renewable Energy 172 (2021): 276-288.

[63]. Syu, Jia-Hao, Mu-En Wu, Gautam Srivastava, Chi-Fang Chao, and Jerry Chun-Wei Lin. "An IoT-based hedge system for solar power generation." IEEE Internet of Things Journal 8, no. 13 (2021): 10347-10355.

[64]. Zhen, Zhao, Jiaming Liu, Zhanyao Zhang, Fei Wang, Hua Chai, Yili Yu, Xiaoxing Lu, Tieqiang Wang, and Yuzhang Lin. "Deep learning based surface irradiance mapping model for solar PV power forecasting using sky image." IEEE Transactions on Industry Applications 56, no. 4 (2020): 3385-3396.

[65]. Li, Pengtao, Kaile Zhou, Xinhui Lu, and Shanlin Yang. "A hybrid deep learning model for short-term PV power forecasting." Applied Energy 259 (2020): 114216.

[66]. Sun, Yuchi, Vignesh Venugopal, and Adam R. Brandt. "Short-term solar power forecast with deep learning: Exploring optimal input and output configuration." Solar Energy 188 (2019): 730-741.

[67]. Bharat Girdhani, Meena Agrawal (2023).Comparison and Statistical analysis of various machine learning techniques for daily prediction of solar GHI representing India's overall solar radiation: https://doi.org/10.21203/rs.3.rs-2801060/v1

[68]. Lu, L., Sali, A., Noordin, N., Ismail, A., & Hashim, F. (2023). Prediction of Peatlands Forest Fires in Malaysia Using Machine Learning. Forests, 14(7), 1472.

[69]. Sit, Muhammed & Demiray, Bekir & Xiang, Zhongrun & Ewing, Gregory & Sermet, Yusuf & Demir, Ibrahim. (2020). A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources. 10.31223/osf.io/xs36g.

[70]. Lee, H.Y.; Kim, N.W.; Lee, J.G.; Lee, B.T. Uncertainty-Aware Forecast Interval for Hourly PV Power Output. IET Renew. Power Gener. 2019, 13, 2656–2664.