AI BASED SOLAR POWER FORECASTING

JIVITES DAMODAR
School of Artificial Intelligence
Amrita Vishwa Vidyapeetham
Coimbatore, India

RAMKUMAR K R
School of Artificial Intelligence
Amrita Vishwa Vidyapeetham
Coimbatore, India

SUKANTHAN
School of Artificial Intelligence
Amrita Vishwa Vidyapeetham
Coimbatore, India

MEHAN RANKA
School of Artificial Intelligence
Amrita Vishwa Vidyapeetham
Coimbatore, India

***Abstract*—With the growing demand for renewable energy, solar power has become an essential source of clean energy. However, predicting solar power output accurately is still a challenge due to changing weather conditions and environmental factors. In this project, we use machine learning—specifically, the Random forest algorithm to build a reliable model for forecasting solar power generation.We will be also comparing the results of other model such as XG boost and LSTM .Our model is trained using historical data that includes weather parameters like Ambient temperature,module temperature and solar irradiance, along with time-based features. After testing and evaluation, the model shows promising results in terms of accuracy and consistency. This work shows how AI techniques like Random Forest can play a key role in making solar energy more predictable and easier to integrate into power systems.**

***Index Terms*—Solar power prediction, Random Forest, machine learning, renewable energy, energy forecasting, ensemble learning.**

I. Introduction

In recent years, the need for clean and sustainable energy sources has increased rapidly due to environmental concerns and the depletion of fossil fuels. Among various renewable sources, solar energy stands out as one of the most promising and widely adopted options. It is abundant, free, and environmentally friendly. However, one of the main challenges in utilizing solar power effectively is its dependency on weather conditions, which makes it highly variable and difficult to predict.

Accurate forecasting of solar power generation is essential for optimizing energy usage, planning grid operations, and reducing the gap between energy supply and demand. Traditional methods for solar prediction often fall short due to their inability to capture complex patterns in weather-related data.

With the rise of Artificial Intelligence (AI) and Machine Learning (ML) techniques, it is now possible to develop more accurate and data-driven models for prediction tasks. this project, we focus on applying and comparing three machine learning approaches—Random Forest, XGBoost, and Long Short-Term Memory (LSTM)—for solar power prediction. Each model was trained using historical weather and solar data, including features such as temperature and solar irradiance. The LSTM model, although widely used for time-series data, produced lower accuracy in our experiments compared to the other two. Both Random Forest and XGBoost performed well, but Random forest slightly outperformed XG boost in terms of consistency.

Based on this analysis, Random forest was selected as the final model for implementation due to its balance of simplicity, speed, and performance. This paper presents our approach to collecting and preprocessing the dataset, training the model, evaluating its performance, and analyzing the results. The goal is to demonstrate how AI can contribute to the renewable energy sector by making solar power generation more predictable and reliable.

Identify applicable funding agency here. If none, delete this.

II. Related Work

Solar power forecasting is an important area of research because solar energy depends heavily on changing weather conditions like sunlight, clouds, and temperature. Many researchers have started using artificial intelligence (AI) and machine learning (ML) to improve the accuracy of solar power predictions.

[1] applied different machine learning methods like Deep Belief Networks (DBN), Support Vector Machine Regression (SVMR), and Random Forest (RF) to forecast solar power at the Buruthakanda Solar Park. These models were tested against a basic model called Smart Persistence (SP). The results showed that machine learning models performed better than SP. However, some models like DBN are complicated and need more time and resources to train.

[2] studied the use of Random Forest and Decision Tree algorithms to predict solar PV power output. They used a small dataset and showed that Random Forest gave more accurate results than the Decision Tree. But because their dataset was small, the model might not work well for bigger solar plants or longer time periods.

[3] gave a review of different machine learning techniques used in solar forecasting. It focused on comparing common models like Random Forest, Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbors (KNN). The paper explained how these models perform on short-term forecasting

and highlighted the importance of using good-quality weather and power data. However, this paper focused more on summarizing other research and did not include a real case study or hands-on results.

[4] proposed a short-term forecasting method using the Random Forest algorithm. It used weather forecast data and historical power generation to predict solar output for the next day. They built a simple interface to use the model and achieved about 93 percentage accuracy. Still, the method depends on real-time sensor and weather data, which may not be available everywhere.

In our project, we use the Random Forest algorithm with a publicly available dataset from Kaggle. This dataset contains solar power generation and sensor readings from two power plants in India over 34 days. Unlike other studies that use complex models or limited data, our approach is simple, practical, and uses open data. This makes it easier for students or beginners to understand and apply machine learning for solar forecasting.

III. METHODOLOGY

We begin by collecting a dataset consisting of historical solar power output along with corresponding weather features such as temperature, humidity, wind speed, and solar irradiance. The data is then preprocessed by handling missing values, removing outliers, normalizing numerical features, and extracting time-based features such as hour of the day and day of the year. This preprocessing ensures the data is clean and ready for modeling. We use these features to train three different machine learning models: Random Forest, XGBoost, and LSTM. While Random Forest and XGBoost are tree-based ensemble models capable of handling structured data effectively, LSTM is a neural network architecture designed for time-series prediction. Each model is trained using a split of training and testing data, and performance is evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. After comparison, Random Forest is selected for final implementation due to its higher accuracy and stability across various test conditions

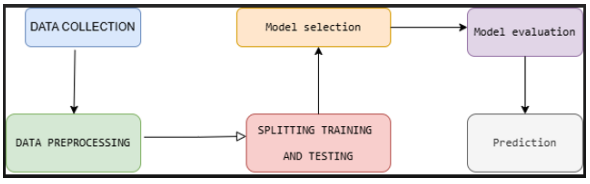

Fig. 1. Block diagram

A. DATASET OVERVIEW

This study employs a comprehensive dataset collected from two solar power plants in India over 34 days. The dataset consists of two primary data sources: one capturing power generation data at the inverter level and the other recording sensor readings at the plant level. These two data sources, when combined, provide a holistic view of solar power production, enabling accurate forecasting of future energy outputs. The power generation dataset records the real-time energy production from multiple inverters spread across the solar plant. Each inverter, connected to several solar panel arrays, generates direct current (DC) power, which is then converted into alternating current (AC) power. Observations are recorded at 15-minute intervals, providing detailed insights into the performance of the inverters. The key features in the power generation dataset include DATE TIME, which indicates the timestamp of each observation, PLANT ID, a unique identifier for the plant that remains constant across the file, SOURCE KEY, representing the unique ID of each inverter, DC POWER, which records the direct current power output generated by each inverter in kilowatts (kW), AC POWER, capturing the alternating current power output in kilowatts (kW), and TOTAL YIELD, reflecting the cumulative energy yield of the inverter up to each timestamp.

The sensor readings dataset, collected at the plant level, captures environmental and operational data through a single array of sensors placed optimally across the solar plant. These sensors provide critical contextual information about the factors influencing solar power generation. The features in the sensor readings dataset include DATE TIME, which corresponds to the timestamp of each observation and is recorded at the same 15-minute interval as the power generation data, PLANT ID, which remains constant for the entire file, and SENSOR PANEL ID, which is unique to the file as there is only one sensor panel for the plant.

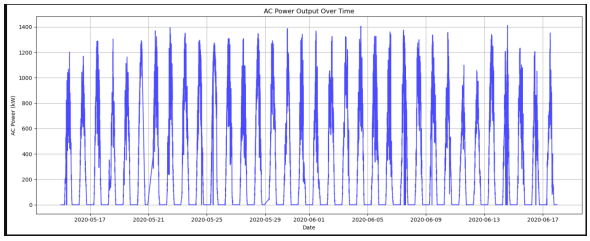

Fig. 2. AC VS TIME

Key environmental variables are captured, such as AMBIENT TEMPERATURE, which refers to the temperature of the surrounding air at the plant, MODULE TEMPERATURE, which measures the temperature of the solar panels' surface in degrees Celsius (°C), and IRRADIATION, which records the amount of solar irradiation received during each 15-minute interval, providing crucial insights into the energy input from the sun to the panels. By merging both the power generation and sensor data, this dataset enables a detailed analysis of solar energy production, accounting for both the mechanical performance of the inverters and the environmental conditions that affect their efficiency. The high-frequency nature of the data, with observations recorded every 15 minutes, allows for granular analysis of instantaneous and cumulative power output, making it an ideal foundation for developing machine learning models aimed at forecasting solar energy production with high accuracy

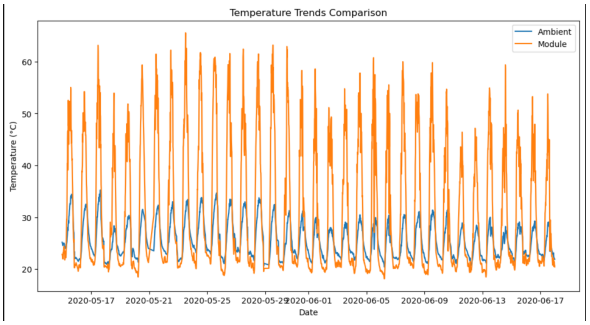

Fig. 3. temperature trend comparrision

B. DATA PREPROCESSING

Before training any predictive models, the raw dataset underwent several preprocessing steps to ensure data quality, consistency, and suitability for analysis. The dataset comprises two components: power generation data collected at the inverter level and sensor readings gathered at the plant level. These datasets were first individually cleaned and then merged on the common DATE TIME to create a unified dataset for analysis.

Initially, missing values were identified and handled appropriately. For time-series data, particularly at 15-minute intervals, missing entries were either forward-filled using previous values or interpolated based on surrounding observations to maintain temporal continuity. Duplicate records, if any, were removed to prevent data redundancy and bias in the model training phase.

The DATE TIME field was converted into a standard datetime format and further decomposed into relevant temporal features such as hour, day, and minutes to capture time-dependent trends in solar power generation.

Continuous numerical features such as DC POWER, AC POWER, TOTAL YIELD, AMBIENT TEMPERATURE, MODULE TEMPERATURE, and IRRADIATION were standardized or normalized to ensure that each feature contributed equally to the learning process, particularly for algorithms sensitive to feature scale.

Outlier detection techniques were also applied to identify and, if necessary, smooth extreme values in power and temperature measurements, which could otherwise distort model predictions. Additionally, to capture the performance of each inverter over time, rolling statistics such as moving averages and rolling standard deviations were computed for power outputs.

Finally, the dataset was resampled and aggregated where necessary to ensure alignment across all records, especially when merging sensor and inverter data. The resulting preprocessed dataset served as a clean, structured, and informative input for downstream exploratory analysis and machine learning model development.

To evaluate the model's ability to generalize to unseen data, we performed an 80-20 train-test split. This means that 80 percent of the preprocessed dataset was allocated for training the models, while the remaining 20 percent was reserved for testing. The train-test split was performed in a time-aware manner to ensure that the training data consisted of earlier timestamps and the test data contained later timestamps, preserving the temporal dependencies inherent in the time-series data.

This division allowed the models to learn from a substantial portion of the dataset while maintaining an independent test set that could be used to evaluate their predictive performance. The separation of data into training and testing sets also ensured that the models were not overfitting to the specific time intervals and were evaluated on their ability to forecast future solar power generation

C. MODEL SELECTION

Since the dataset contains a large amount of high-frequency time-series data, it was important to choose models that could not only handle this scale efficiently but also capture the non-linear relationships between environmental factors and power generation. Solar power output depends on variables like irradiation, temperature, and time of day, which don't follow a simple pattern. So, instead of basic models, we focused on more powerful machine learning approaches that are better suited for complex, real-world data.

We started by using Random Forest Regressor, which is well-known for its ability to handle large datasets and uncover non-linear patterns. It works by building multiple decision trees and combining their outputs, which helps in reducing overfitting and improving accuracy. It also gave us the added benefit of feature importance, helping us understand which variables had the most impact on power generation.

In addition to Random Forest, we also experimented with XGBoost, a boosting algorithm that builds trees sequentially and corrects the mistakes of previous ones. It's fast, scalable, and often gives better performance on structured datasets like ours. Since our dataset had both inverter-level and plant-level features, XGBoost did a good job of learning from the complex interactions among these features.

We considered using LSTM (Long Short-Term Memory) networks as well, given that our data is time-series in nature. LSTMs are specifically designed to learn patterns over time, and they would be a strong candidate for future improvements. However, due to the increased computational demands of training deep learning models on a large dataset, we prioritized tree-based methods for now, which delivered good accuracy and were easier to work with.

D. Random Forest

Random Forest is an ensemble learning algorithm that builds multiple decision trees using random subsets of data and features. For regression tasks, the final prediction is obtained by averaging the outputs of individual trees:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \tag{1}$$

where N is the number of trees and $f_i(x)$ is the output of the i^{th} tree. Random Forest handles large datasets efficiently, reduces overfitting, and provides insights into feature importance.

E. XGBoost

Extreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosting designed for high performance and scalability. It builds trees sequentially, each correcting the errors of its predecessor. The model minimizes a regularized objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

Here, l is the loss function (e.g., squared error), T is the number of leaves, w_j is the weight of leaf j, and γ, λ are regularization parameters. XGBoost improves both accuracy and generalization by incorporating second-order derivatives in its optimization.

F. Long Short-Term Memory (LSTM)

LSTM networks are a type of Recurrent Neural Network (RNN) specifically designed to capture long-term dependencies in sequential data. Each LSTM cell contains a cell state C_t and hidden state h_t, regulated by gates:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5}$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{6}$$
$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{7}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{8}$$
$$h_t = o_t \cdot \tanh(C_t) \tag{9}$$

LSTMs are particularly effective in solar power forecasting as they maintain memory across time steps, capturing both seasonal patterns and short-term fluctuations.

G. Evaluation Metrics

To evaluate the performance of the forecasting models, we used standard regression evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics are widely adopted in time-series forecasting tasks and provide insight into the accuracy and reliability of the models.

The Mean Absolute Error (MAE) measures the average magnitude of the errors between predicted and actual values, ignoring their direction. It is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{10}$$

The Mean Squared Error (MSE) is the average of the squared differences between the actual and predicted values. It penalizes larger errors more significantly and is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{11}$$

The Root Mean Squared Error (RMSE) is the square root of the MSE. RMSE retains the same unit as the predicted variable, making it more interpretable for practical applications:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{12}$$

These metrics collectively offer a comprehensive evaluation of model performance. Lower values across MAE, MSE.

IV. Result Analysis

In this study, the performance of three machine learning models—XGBoost, Random Forest, and Long Short-Term Memory (LSTM)—was evaluated for solar power forecasting. The models were assessed using three key evaluation metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) score. The following sections summarize the results of each model.

A. XGBoost

The XGBoost model demonstrated strong performance in forecasting solar power generation with the following results:

- Mean Squared Error (MSE): 1486.50
- Mean Absolute Error (MAE): 13.87
- R-squared (R^2) Score: 99.03%

The R^2 score of 99.03% indicates that the model explained a very high percentage of the variance in the solar power data. The relatively low MAE suggests that XGBoost performs well with small prediction errors on average. However, it is important to note that XGBoost was not subjected to feature scaling, and this might have slightly impacted its performance due to the unscaled features.

B. Random Forest

The Random Forest model also performed well, yielding the following results:

- Mean Squared Error (MSE): 1867.62
- Mean Absolute Error (MAE): 14.30
- R-squared (R^2) Score: 98.78%

The R^2 score of 98.78% demonstrates that Random Forest was able to capture most of the variance in the dataset.But random forest was consistent when compared to xg boost.As xg boost did not perform well in some peek values.

C. Long Short-Term Memory (LSTM)

The LSTM model, which is designed to capture temporal dependencies, showed the following results:

- R-squared (R^2) Score: 96.9%
- Mean Absolute Error (MAE): 31.41
- Mean Squared Error (MSE): 4168.9

This indicates that LSTM was able to make accurate point predictions. However, the lower R^2 score suggests that it did not capture as much of the variance in the data compared to XGBoost and Random Forest. The model still struggled to capture the broader trends in the data.

D. Comparison and Insights

- Random forest performed the best in terms of R^2 score (98.78%) and consistency indicating it explained the most variance in the solar power data. Despite not applying feature scaling, Random forest still demonstrated strong performance.

- Xg boost showed a R^2 score of 99%, performing similarly to Random forest in terms of variance explained.But at some peek values XG boost is not performing well so by considering the consistency we could say that Random forest have good accuracy and consistency.

- LSTM, despite being trained with scaled data, performed poorly in terms of R^2 score (97%). While the MAE and MSE values were exceptionally high compared with other two, indicating precise point predictions, the model did not capture the variance in the dataset as well as the tree-based models. This suggests that LSTM may not be ideal for this particular forecasting task due to its inability to capture broader trends in solar power generation.

In conclusion, while LSTM provided high precision in point predictions, it was outperformed by the tree-based models, particularly Random forest, in explaining the variance in the data. Random forest emerged as the best-performing model overall, providing the most balanced, accurate and consistent predictions for solar power forecasting.herefore, Random Forest emerged as the most effective model for solar power forecasting in this study, offering a practical and accurate approach for real-world renewable energy prediction systems.

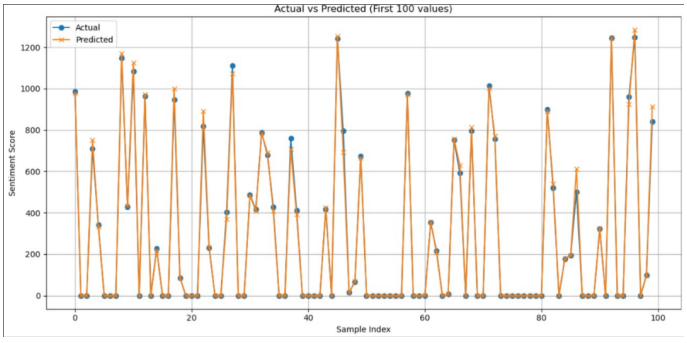

Fig. 4. actual vs predicted using Random forest

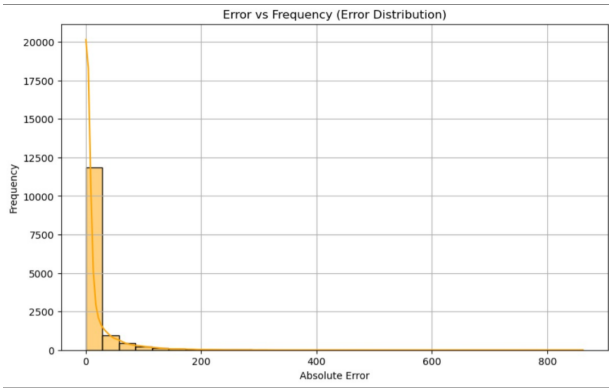

Fig. 5. error vs frequency

TABLE I
COMPARISON

Paper Title	Year & Citations	Model	Accuracy
Forecasting Solar Power Generation: An Evaluation of Regression Algorithms [5]	2023	Linear Regression Random forest	73.44% 93.81%
Prediction of Solar Power Generation Based on Random Forest Regressor Model [4]	2019	Random forest	93%
Machine Learning based Solar PV Power Prediction [2]	2023	Random forest	91%

V. CONCLUSION

This study explored the application of machine learning techniques—specifically XGBoost, Random Forest, and Long Short-Term Memory (LSTM)—for forecasting solar power generation based on inverter-level and sensor-level data collected from two solar power plants over 34 days. Through comprehensive preprocessing and careful feature engineering, the models were trained on a unified dataset incorporating both temporal and environmental features.

Among the models evaluated Random forest demonstrated superior performance, achieving the highest R^2 score and the lowest error metrics, indicating its ability to effectively capture the underlying patterns in the data. XG boost also performed well, but at some peek value it is not consistent. While LSTM showed promise in point-wise accuracy due to feature scaling, it struggled to capture the broader variance in the dataset, resulting in a comparatively lower R^2 score.

Overall, the results suggest that tree-based ensemble models are better suited for short-term solar power forecasting in this context.. These insights provide a solid foundation for developing data-driven energy forecasting tools to support smart grid operations and energy management.

VI. Future Scope

Although the current models have demonstrated robust performance, there is considerable potential for future enhancement. Integrating weather forecast data—such as humidity, wind speed, and cloud cover—could significantly improve model accuracy, especially for day-ahead or week-ahead forecasting.

Additionally, applying deep learning architectures like Convolutional Neural Networks (CNNs) or hybrid CNN-LSTM models could enable better spatial and temporal feature extraction from multivariate time-series data. Transfer learning techniques might also be explored to adapt models trained on one plant's data for use in other geographical locations with minimal retraining.

Moreover, incorporating real-time data streaming and online learning algorithms would allow for dynamic model updates, enhancing prediction reliability in rapidly changing environmental conditions. The development of lightweight, interpretable models could also be beneficial for integration into IoT devices and edge computing platforms used in modern solar energy systems.

With increasing global emphasis on renewable energy, refining these models and extending their capabilities holds great promise for optimizing energy generation, planning, and sustainable infrastructure development.

References

[1] P.A.G.M.Amarasinghe and S.K.Abeygunawardane, Application of machine learning algorithms for solar power forecasting in Sri Lanka,2018

[2] Ram Prakash Ponraj, Vijay Ravindran,Krishnakumar Chittibabu ,Akash T,Chandru A and Emmanuel Louis,"Machine Learning based Solar PV Power Prediction",2023

[3] Jwaone Gaboitaolelwe, Adamu Murtala Zungeru, Abid Yahya, Dasari Naga Vinod and Ayodeji Olalekan Salau." Machine Learning Based Solar Photovoltaic Power Forecasting: A Review and Comparison",2023

[4] Alexandra Khalyasmaa, Stanislav A. Eroshenko, Teja Piepur Chakravarthy, Venu Gopal Gasi, Sandeep Kumar Yadav Bollu, Raphaël Caire, Sai Kumar Reddy Atluri and Suresh Karrolla,"Prediction of Solar Power Generation Based on Random Forest Regressor Model",2019

[5] Arsh Dayal ,"Forecasting Solar Power Generation: An Evaluation of Regression Algorithms",2023