PLANT SCIENCE TODAY
Vol x(x): xx–xx
https://doi.org/10.14719/pst.8861

eISSN 2348-1900

RESEARCH ARTICLE

Daily solar power prediction using machine learning: A model-wise comparative study

P Khadeeja Faheema[1], M Radha[2*], G Vanitha[1], M Nirmala Devi[1], R Mahendiran[3] & S Vishnu Shankar[1]

[1]Department of Physical Science and Information Technology, Agricultural Engineering College and Research Institute, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India

[2]Department of Agricultural Economics, Anbil Dharmalingam Agricultural College and Research Institute, Trichirappalli 620 029, Tamil Nadu, India

[3]Department of Renewable Energy Engineering, Agricultural Engineering College and Research Institute, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India

*Correspondence email - radha@tnau.ac.in

Received: 14 April 2025; Accepted: 14 May 2025; Available online: Version 1.0: 13 June 2025

Cite this article: Khadeeja FP, Radha M, Vanitha G, Nirmala DM, Mahendiran R, Vishnu SS. Daily solar power prediction using machine learning: A model -wise comparative study. Plant Science Today (Early Access). https://doi.org/10.14719/pst.8861

Abstract

Solar energy produced by photovoltaic panels is a vital energy source that offers numerous benefits to both the environment and society. However, meteorological variables such as solar irradiation, weather patterns, precipitation and climate conditions present significant challenges to seamless energy integration into the power grid. Accurate forecasting is essential to maintain supply-demand balance, optimize energy storage and ensure grid stability. This study leverages machine learning (ML) techniques to predict solar power generation and address renewable energy integration challenges. Nine ML models were employed, including linear regression, auto regressive integrated moving average (ARIMA), artificial neural network (ANN), support vector machines (SVM), random forest (RF), decision tree, gradient boosting machine (GBM), light gradient boosting (LGBM) and extreme gradient boosting (XGBM). Inputs such as irradiance, humidity, minimum temperature, maximum temperature and surface pressure were used to train these models. The model performances were evaluated using metrics like root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The results highlighted ANN as the most effective model, achieving an RMSE of 274.84 kWh, MAE of 245.93 kWh and a MAPE of 5.26 %. This research contributes to the existing literature by addressing the relatively unexplored application of multiple machine learning models for predicting energy output from photovoltaic systems. A key novelty of this study is its ability to achieve accurate solar power forecasts using a limited dataset from a newly installed solar power plant, unlike many existing studies that rely on large volumes of data. Additionally, it explores the solar power production potential of Namakkal district in Tamil Nadu, India-a region with limited prior research.

Keywords: artificial intelligence; machine learning; renewable energy; solar power prediction

Introduction

Photovoltaic (PV) technology presents a viable alternative to fossil fuel-based energy sources, offering a sustainable solution to reduce greenhouse gas emissions (1). This is mainly attributed to the abundant solar resource, decreasing installation costs and ease of deployment of PV panels. One of the significant challenges in the solar industry is the unpredictable nature of solar energy production, which largely depends on weather conditions such as sunlight intensity, rainfall and cloud cover. This intermittent nature of solar power poses challenges for both power grids and end-users, potentially leading to disruptions in appliances and daily activities (2). Such variability can cause inconvenience, discouraging users from adopting this technology. Therefore, precise forecasting of PV power is essential for various applications, including energy management systems that facilitate the integration of solar energy with smart buildings, electric vehicles and energy storage systems (3, 4).

The accuracy of PV generation forecasting models depends on several factors, including the forecasting time horizon, data resolution, geographical location, meteorological variables and data quality. Based on the time horizon, forecasting can be classified broadly into different categories. Very short-term forecasting, which ranges from seconds to minutes, is primarily utilized for managing PV systems and microgrids (5). This timeframe is essential for real-time control operations such as inverter regulation, frequency balancing and immediate energy dispatch, all of which are critical for ensuring grid stability and avoiding sudden power fluctuations. Short-term forecasting, covering a period of 48 to 72 hr, plays a critical role in operational control and unit commitment (6). It supports decisions related to energy trading, load balancing and scheduling of conventional power plants, helping to optimize generation costs and resource use.

Medium-term forecasting, which spans several days to a week, is important for maintenance scheduling, fuel

purchasing and energy storage planning. It provides utilities with the insight needed to plan operations over a slightly longer horizon, especially during variable weather conditions. Long-term forecasting extending from a few months to years, is vital for strategic planning, infrastructure development, investment decisions and policymaking. It supports capacity planning, grid expansion and evaluation of long-term renewable integration strategies under changing climatic and economic conditions (7). Many existing models are limited in their ability to incorporate real-time weather dynamics, which can hinder effective planning and compromise grid stability (8).

Statistical approaches like ARIMA, physical models such as solar cell modelling and the use of satellite images are also employed to predict solar module power generation (9–11). However, ML has demonstrated remarkable performance in forecasting due to its ability to handle complex data effectively. ML models employ advanced algorithms to analyse factors such as geographical location, weather conditions and solar panel efficiency (12). By integrating historical data with real-time weather information, these models provide accurate and reliable predictions of solar energy output. In recent years, the application of ML in solar forecasting has gained considerable attention, with numerous studies highlighting its potential to improve the accuracy and reliability of solar energy predictions.

Previous research utilized long short-term memory (LSTM), random forest, support vector machine (SVM) and ARIMA to predict energy generation and demand patterns (13). In one study, different regression techniques, such as least squares and SVM with multiple short-term functions (MSTF), were compared to develop prediction models (14). Meanwhile, some researchers employed the LGBM and K-nearest neighbors (KNN) for their work (15). Additionally, methods such as ANN, averaged perceptron, Bayes point machine, decision forest, decision jungle, LGBM, locally deep SVM, logistic regression, SVM and XGBM also used for prediction purpose (16).

Since solar radiation is the most influential variable affecting power generation, numerous studies have focused on forecasting solar radiation where measured data is available (17). Various ML and deep learning techniques have been applied for solar radiation and power generation forecasting, with ANN and SVM being the most commonly used methods (18). For time series problems, ANN like convolutional neural networks (CNNs), recurrent neural networks (RNNs), LSTM networks, gated recurrent units (GRU) and their bidirectional versions (Bi-GRU and Bi-LSTM) are widely adopted due to their high effectiveness (19). Additionally, a stacked autoencoder is often used as an efficient feature extraction technique (20).

This study stands out by applying a comprehensive comparison of nine machine learning models, including linear regression, ARIMA, ANN, SVR, RF, decision tree, GBM, LGBM and XGBM-to predict solar power generation. Unlike previous researches, which often relies on extensive datasets from established solar power systems, this work focuses on a newly installed solar power plant at Namakkal district in Tamil Nadu, India. By addressing the challenges of limited historical data and localized meteorological information, this study offers valuable insights into the practical application of ML models in emerging solar energy projects.

Materials and Methods

This section provides a comprehensive overview of the data sources, pre-processing techniques and the methodology employed to develop machine learning models-including linear regression, ARIMA, ANN, SVM, decision tree, random forest (RF), gradient boosting machine (GBM), LGBM and XGBM-for accurate daily solar power prediction. Fig. 1 illustrates the complete process flow for solar power estimation.

Data description

Data was obtained from a solar power plant situated in Namakkal, Tamil Nadu, India, with a total generation capacity of 1 MW across an area of 4 acres. The dataset comprises seven months of daily observations, covering 214 days from April 1, 2024 to October 31, 2024. To visualize the temporal variation in solar power generation over the study period, the time series plot of daily solar power output is presented in Fig. 2. Meteorological parameters, including irradiance, humidity, minimum temperature, maximum temperature and surface pressure, which were unavailable at the site, were sourced from the NASA website (https://power.larc.nasa.gov). To align the data with the study location, spatial resolution adjustments were made by selecting the nearest grid point corresponding to the target block in Namakkal district. Since the site data was in a raw format, pre-processing was necessary before using it for estimation. Key pre-processing steps involved handling missing values, removing duplicates and outliers.

Software and equipment

Our program is executed on Google Colab, allowing us to write and run Python code directly from the website. The processing is performed on a system equipped with an AMD Ryzen 5 5600H processor (6 cores, 12 threads), 8 GB of memory and hardware acceleration provided by Nvidia's GeForce GTX 1650.

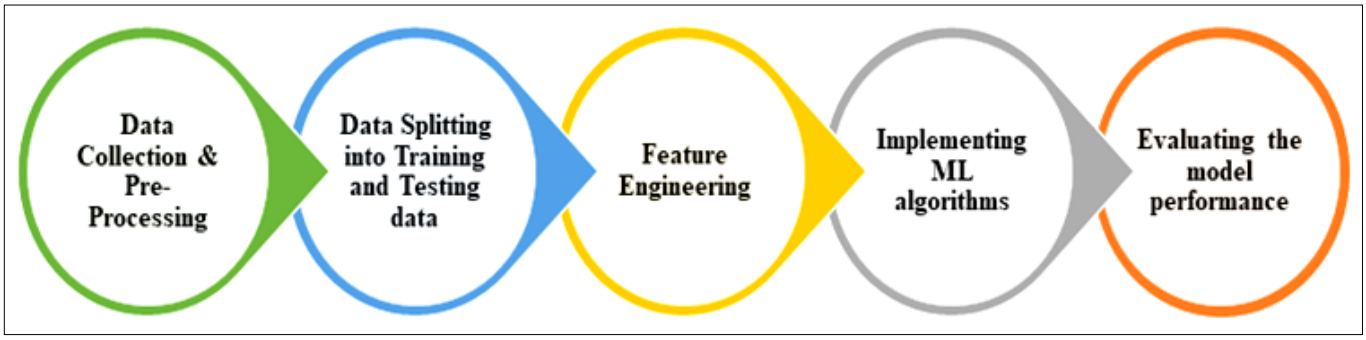


Fig. 1. Flowchart showing the methodology for solar power prediction.

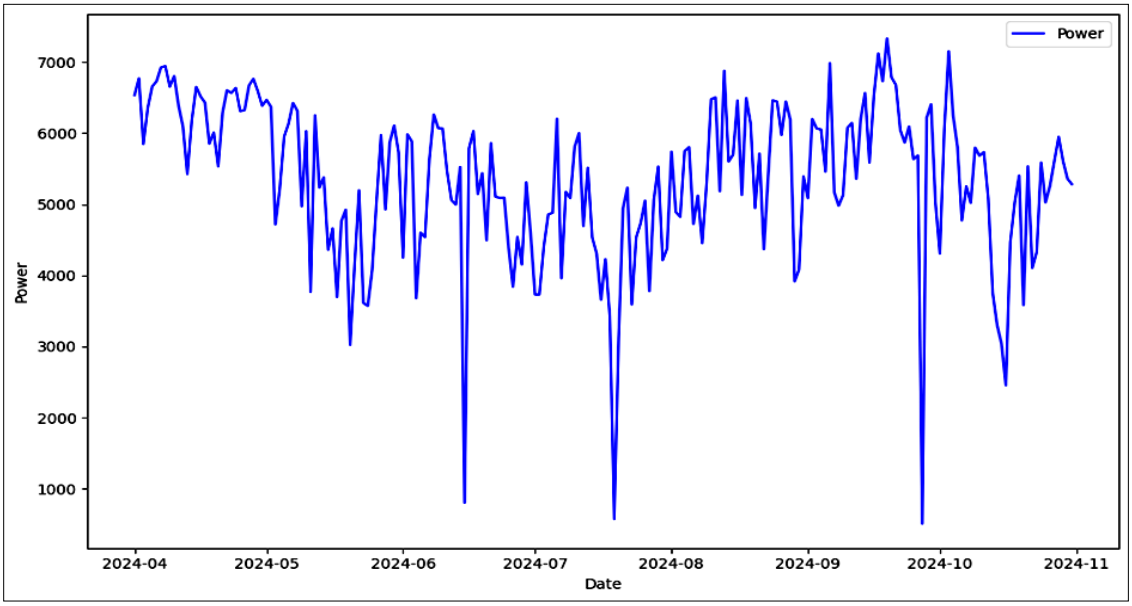


Fig. 2. Temporal variation in daily solar power output over the study period.

Data collection and pre-processing

The data collected from the site was initially in raw form, requiring pre-processing before it could be used for estimation. The key pre-processing tasks included removing duplicate values, interpolating missing data, detecting outliers and normalizing the dataset. Additionally, since it is a time series dataset, a stationarity test was conducted.

Removal of duplicates and interpolation for missing values

The collected meteorological data contained missing and duplicated values. As part of the pre-processing procedure, duplicate values were removed and the missing values were estimated using linear interpolation.

Detection of outliers

As part of the data pre-processing stage, outlier treatment was applied to enhance data quality and ensure the robustness of the prediction models. Due to the inherently volatile nature of solar power generation, the dataset contained several anomalous spikes and drops. Outliers were detected using the interquartile range (IQR) method, where values falling outside 1.5 times the IQR from the first and third quartiles were marked as anomalies. Instead of discarding these data points, interpolation was used to replace them. This approach helped preserve the temporal continuity of the time series while reducing the influence of extreme values on the model's learning process.

Normalization of data

The input data showed considerable variability among the different parameter values, making it difficult to model effectively. This wide range of values, represented by the minimum. (Min) and maximum. (Max) values in Table 1. To bring the data within a uniform range of 0 to 1. Min-Max normalization was applied using Eqn. 1.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad \text{(Eqn. 1)}$$

Here, X' denotes the normalized value of the data, X represents the input value and X_{min}, X_{max} refer to the minimum and maximum values of the dataset, respectively.

Stationarity tests

To ensure the reliability of the solar power prediction models, assessing the stationarity of the time series data. Stationarity indicates that the statistical properties of the series, such as mean and variance, remain constant over time. Two commonly used tests for checking stationarity are the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.

Augmented Dickey-Fuller (ADF) test

The ADF test is a widely used unit root test to check whether a time series is stationary. It extends the basic Dickey-Fuller test by including lagged differences to account for autocorrelation. The null hypothesis assumes the presence of a unit root, indicating non-stationarity, while the alternative hypothesis suggests stationarity.

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

The KPSS test is used to test the null hypothesis that a time series is stationary around a deterministic trend. Unlike the ADF test, the KPSS test assumes stationarity as the null hypothesis and checks for the presence of a unit root as the alternative hypothesis.

Data splitting into training and testing data

The next step in data pre-processing involves splitting the dataset into training and testing sets. Out of the 214 total observations, 80 % (171 observations) are allocated for

Table 1. Ranges of meteorological variables.

Unit	Irradiance (kWh/m²)	Humidity (%)	Minimum temperature (°C)	Maximum temperature (°C)	Surface pressure (kPa)	Power (kWh)
Mean	5.30	69.20	23.63	34.28	97.15	5405.86
Minimum value	1.13	37.17	20.87	27.36	96.63	2453.3
Maximum value	7.52	90.13	29.59	43.1	97.49	7333.3

training, while the remaining 20 % (43 observations) are used for testing. As the data follows a time series structure, a time-based splitting approach was used to preserve the temporal order and avoid data leakage.

Feature engineering

Feature engineering plays an important role in improving the performance of ML models for solar power prediction. This process involves extracting, transforming and selecting the most relevant features that influence solar power generation across different seasons. Several key techniques were implemented to refine the dataset and enhance model accuracy.

Creation of lag features

Lag features, representing past values of power generation, were introduced to help the model capture temporal patterns and improve predictive performance. To identify the most relevant lag intervals, autocorrelation function (ACF) and partial autocorrelation function (PACF) plots were analysed (Fig. 3). The ACF plot showed strong autocorrelation at the initial lags, especially up to lag 4, while the PACF plot indicated significant partial autocorrelations at lags 1, 2 and 3, with a sharp drop-off thereafter. Based on these observations, lagged values of solar power at 1, 2 and 3 time steps were selected and added as additional features. This data-driven lag selection enabled the model to effectively leverage recent historical power output and improve the accuracy of daily solar power prediction.

Interaction features

To enhance predictive power, interaction features were created by combining existing meteorological variables. The following new features were engineered:

Temperature difference (temp_diff)

The difference between daily maximum and minimum temperatures, which provides insights into daily temperature variations affecting solar power output.

Irradiance-humidity interaction (irr * humidity)

Captures the combined effect of solar radiation and humidity levels, which can impact solar panel efficiency. Higher humidity can reduce panel efficiency due to water vapor absorption, while low humidity combined with high irradiance can lead to overheating issues.

Irradiance-temperature difference interaction (irr * temp_diff)

Incorporates the relationship between solar irradiance and temperature fluctuations to improve model performance. Large temperature variations with high irradiance may indicate clear skies, which are favourable for solar energy generation, while smaller variations may suggest cloud cover, affecting output.

These interaction features are meaningful as they provide deeper insights into the environmental conditions affecting solar power generation. Such features have practical applications in optimizing solar energy systems, improving predictive models for solar farms and assisting in energy grid management by providing more accurate forecasts.

Random forest based feature importance

Random forest, a popular ensemble learning method, provides an effective approach to feature selection through its inherent ability to measure feature importance. This technique ranks input features based on their contribution to reducing impurity (e.g., Gini impurity or mean squared error) across all the trees in the ensemble (21). Features that result in larger reductions in impurity are considered more influential in predicting the target variable. For regression tasks such as solar power prediction, the importance of a feature X_j is typically computed as the total decrease in mean squared error (MSE) it contributes, averaged over all trees in the ensemble. Mathematically, the importance score $I(X_j)$ is given by Eqn. 2:

$$I(X_j) = \frac{1}{T}\sum_{t=1}^{T}\sum_{n \in N_t(X_j)} \Delta MSE_{n,t} \quad \text{(Eqn. 2)}$$

where T is the total number of trees, $N_t(X_j)$ is the set of all nodes in tree t that split on feature X_j and $MSE_{n,t}$ is the decrease in MSE caused by the split at node n in tree t . by implementing these feature engineering techniques, the dataset was transformed into a more informative and structured format, ensuring that the machine learning models receive the most relevant inputs for accurate solar power.

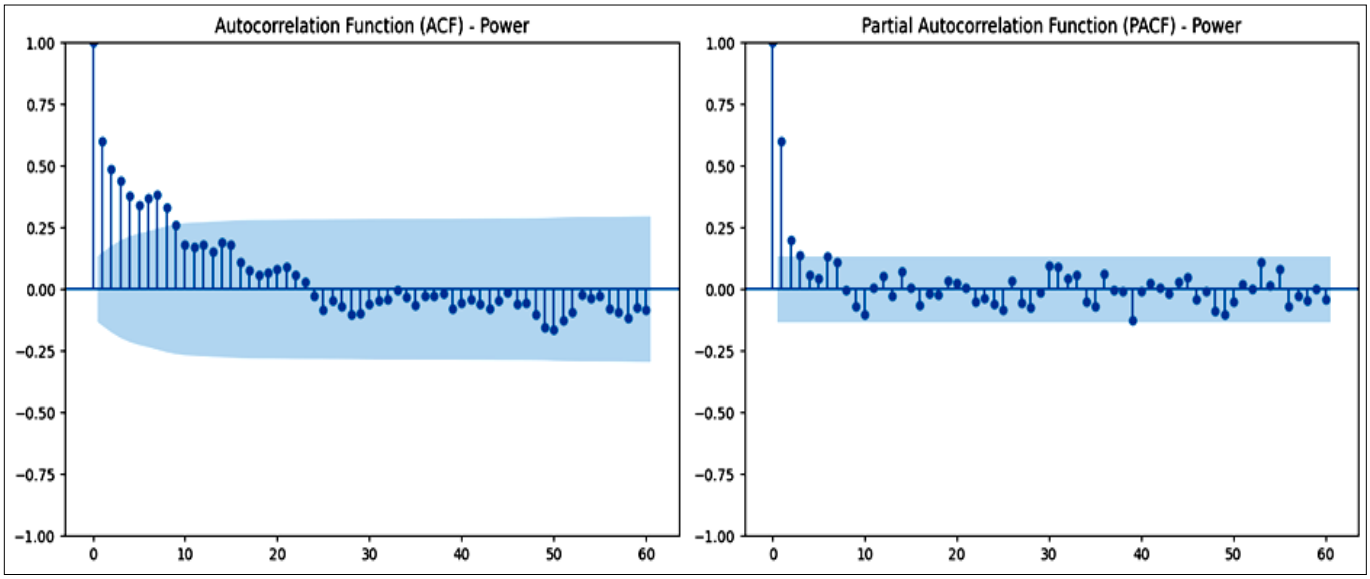


Fig. 3. ACF and PACF plots of solar power output.

Machine learning (ML) algorithm implementation

Model selection rationale

To ensure a comprehensive evaluation of forecasting performance, nine ML and statistical models were selected based on their algorithmic diversity, robustness and suitability for time series data with meteorological influences. The chosen models include linear regression, ARIMA, ANN, SVM, decision tree, random forest, GBM, LGBM and XGBM. This selection spans a wide range of model families-linear, tree-based, ensemble, neural networks and classical statistical forecasting-thereby enabling a balanced comparison between interpretable and high-performing models. Given the inherent non-linearity and volatility of solar power generation influenced by dynamic weather parameters, models capable of capturing complex patterns, such as ANN and tree-based ensembles, are particularly valuable. Statistical and linear models like ARIMA and linear regression serve as benchmarks to assess improvements offered by more advanced methods. Additionally, models like SVM and boosting algorithms bring the advantage of handling multi-dimensional and noisy data. Overall, the diverse model set allows for a robust analysis of prediction capabilities under varying assumptions, data structures and computational complexities.

Linear regression

Linear regression is a machine learning technique designed to model the relationship between an independent variable (X) and a dependent variable (Y). It is widely applied for regression analysis and prediction tasks due to its simplicity and effectiveness. The fundamental equation (Eqn. 3) for linear regression is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon \qquad \text{(Eqn. 3)}$$

Here β_0 is the intercept, β1, β2.....β_n are regression co-efficient and € is the error term.

Autoregressive integrated moving average (ARIMA)

Autoregressive models predict future values based on past data points. The ARIMA model is an advanced version of the basic ARMA model, incorporating an integration (I) term to handle non-stationary data (22). It is a statistical model that relies solely on historical data to identify patterns and generate forecasts, making it a simple yet effective method for analysing time series data that exhibit stationarity. In the ARIMA model, the components AR, I and MA have distinct roles. The AR (auto-regressive) component captures the relationship between an observed value and its previous values (lags). The I (integrated) component ensures data stationarity by applying differencing, which involves replacing data points with the difference between the observed value and its lagged value (23). The MA (moving average) component models the relationship between an observed value and the residual errors from previous time steps. The model is represented using three parameters (p, d, q), where p denotes the number of lag observations, d represents the number of differencing operations applied for stationarity and q indicates the number of lagged error terms used. The general ARIMA model equation is Eqn. 4.

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t \qquad \text{(Eqn. 4)}$$

where y_t is the observed time series value at time t, ϕ(B) is the autoregressive operator, $(1 - B)^d$ is the differencing operator, θ(B) is the moving average operator and $€_t$ is the error term.

Artificial neural network (ANN)

An artificial neural network (ANN) is a ML technique inspired by the structure and function of biological neural networks (24). Since the late 1980s, ANNs have been applied to time series forecasting (25). They consist of interconnected information-processing units called neurons, also referred to as nodes or perceptrons, which serve as the computational units responsible for decision-making and data processing (26). A neural network (NN) is formed by linking multiple neurons through connections. Each neuron typically has one or more weighted input connections, an activation function and an output connection (Fig. 4).

An ANN is made up of three main layers: the input layer, hidden layer and output layer. In this structure, except for the neurons in the input layer, each node receives inputs that are multiplied by corresponding weights and then summed to calculate an activation value. This activation value is then passed through an activation function to generate the node's output. The resulting output is subsequently processed by the

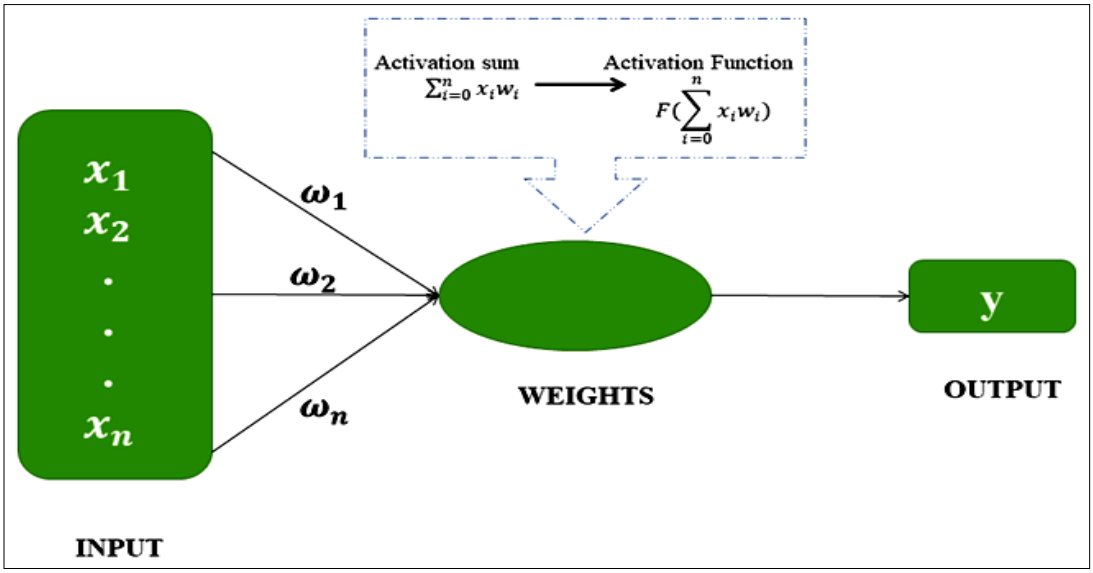


Fig. 4. Structure of a neuron.

following nodes through multiple layers until it reaches the output layer. This approach has demonstrated effectiveness in handling complex problems with numerous input variables.

Support vector machine (SVM)

Vapnik introduced the concept of SVM, a kernel-based machine learning technique designed for both classification and regression tasks (27). SVR applies the principles of SVM to regression problems. SVM utilizes kernel functions and the kernel trick to transform low-dimensional data into a higher-dimensional feature space. In this transformed space, linear solutions correspond to non-linear solutions in the original lower -dimensional space. This makes SVM an effective choice for addressing various naturally non-linear problems. A kernel function, denoted as k(x, x') , is a symmetric function that measures the similarity between observations based on their feature values. SVMs utilize various kernel functions, broadly categorized into linear and non-linear types. Consequently, SVMs can be classified as L-SVMs, which employ linear kernel functions and K-SVMs, which use non-linear kernels. The radial basis function (RBF) is the most commonly used and recommended kernel due to its adaptability in optimization, robustness and high efficiency (28). Examples of kernel functions used in SVMs are linear, polynomial, radial basis and sigmoid which can be mathematically presented as follows (Eqn. 5 - Eqn. 8):

Linear $k(x,x') = x^T x'$ (Eqn. 5)

Polynomial $k(x,x') = (\gamma x^T x' + c)^d$ (Eqn. 6)

Radial basis $k(x,x') = \exp(-\gamma \|x - x'\|^2)$ (Eqn. 7)

Sigmoid $k(x,x') = \tanh(\gamma x^T x' + c)$ (Eqn. 8)

The prediction calculation of an SVM can be mathematically expressed as in Eqn. 9.

$$y = \sum_{i=1}^{N} (a_i k(x_i, x') + b) \quad \text{(Eqn. 9)}$$

In this context, y represents the predicted output value, b is the bias parameter, a_i denotes the coefficient and x_i is the input vector from the training set at the i^{th} instance, with a total of n training vectors.

Decision tree

The decision tree is a tree-based machine learning technique used for classification and regression. It divides data into smaller, more manageable subsets by selecting features that offer the most information gain. It does this through simple 'if-then' rules, allowing it to uncover complex relationships among features such as location, temperature and time. This makes it particularly effective for tasks like predicting crime categories based on contextual variables (29). Information gain, derived from entropy, serves as the primary criterion for selecting the optimal attribute in decision trees (30). The formula for information gain and entropy are as follows (Eqn. 10 & 11):

Information Gain = E (S) (Eqn. 10)

$$E(S) = -\sum_{i-1}^{c} p_i \log_2 p_i \quad \text{(Eqn. 11)}$$

The entropy before data separation is represented as E (S), where S_i denotes the subset of data resulting from the separation based on specific attributes, n is the number of subsets and S is the size of the initial dataset. The probability of class i within the dataset is denoted by P_i . Although decision trees can be prone to overfitting when dealing with complex data, they are widely used due to their computational efficiency and easy interpretability. In this study, the decision tree model is applied to predict solar power generation, leveraging its ability to handle non-linear and interpretable data.

Random forest

The random forest regressor is an ensemble-based machine learning algorithm that combines the predictions of multiple decision trees to improve accuracy and robustness. Random forest is especially effective for handling non-linear patterns and interactions among predictors, making it well-suited for predicting solar energy output, which is influenced by complex and dynamic weather conditions. The model operates by constructing multiple decision trees during training, each on a randomly sampled subset of the data (bootstrapping) and averaging their outputs to generate a final prediction. The mathematical expression for the random forest prediction $\hat{y}$ for an input feature vector x is given in Eqn. 12:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad \text{(Eqn. 12)}$$

Where T is the number of trees in the ensemble and $f_t(x)$ represents the prediction of the t^{th} regression tree for input x . This ensemble approach reduces variance and improves generalization compared to individual decision trees. Random selection of features at each split ensures diversity among trees, which enhances the robustness of the model. Overall, random forest served as a reliable and interpretable model in capturing the non-linear dependencies between weather variables and solar power output (29).

Gradient boosting machine

The gradient boosting regressor (GBR) is a powerful algorithm for regression tasks, particularly effective in modeling complex and non-linear data, such as solar power plant energy predictions. GBR operates through a sequential process where each model corrects the errors of its predecessor, a technique known as boosting (31). It was selected for its capability to iteratively improve prediction accuracy by minimizing previous model errors. This algorithm is highly efficient in recognizing intricate patterns within data and often outperforms traditional methods, especially when handling imbalanced datasets.

Light gradient boosting machine (LGBM)

The light gradient boosting machine (LGBM) is a cutting-edge machine learning algorithm known for its exceptional efficiency and performance in both classification and regression tasks. Developed by Microsoft, LGBM is specifically designed for scalability, making it well-suited for large datasets with numerous features while ensuring fast training and

prediction. Unlike traditional gradient boosting algorithms that use a level-wise tree growth strategy, LGBM employs a leaf-wise (or best-first) approach. This method involves splitting the leaf with the highest loss, resulting in deeper trees with fewer splits, which enhances accuracy and efficiency. Due to its speed, scalability and precision, LGBM is widely favoured by ML practitioners for handling large-scale datasets and complex predictive tasks.

Extreme gradient boosting machine (XGBM)

XGBM is a powerful and flexible distributed gradient-boosting library designed for high efficiency and portability. It applies the GBM framework to execute ML algorithms. XGBM efficiently solves various data science problems using parallel tree boosting. As an enhanced version of gradient-boosted decision trees (GBDT), it is specifically optimized for faster performance and improved effectiveness. Studies demonstrated the successful application of the XGBM algorithm in predicting solar power with minimal error, outperforming other ML models (32, 33).

Model evaluation

To evaluate the predictive accuracy of the ML models across various time intervals, three commonly adopted performance metrics were used: root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). RMSE quantifies the square root of the average of squared differences between actual and predicted values, serving as a measure of the model's residual variance. MAE calculates the average magnitude of the errors in a set of predictions, without considering their direction. MAPE expresses the mean absolute error as a percentage of the actual values, allowing for relative comparison across models. The mathematical expressions for these metrics are as follows (Eqn. 13-15):

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2} \qquad \text{(Eqn. 13)}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{N}|Y_t - \hat{Y}_t| \qquad \text{(Eqn. 14)}$$

$$MAPE = \left[\frac{1}{n}\left(\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|\right)\right] * 100 \qquad \text{(Eqn. 15)}$$

Here, Y_t represents the actual value, $\hat{Y}_t$ denotes the predicted value and n is the total number of observations. These metrics provided a comprehensive assessment of how closely the model outputs aligned with real solar power values, guiding model comparison and selection.

Results and Discussions

Detection of outliers

To improve the reliability of model predictions, outlier detection was performed during the data pre-processing phase. The refined time series data after outlier treatment is visualized in Fig. 5. The chart clearly shows a smoother and more coherent pattern in daily solar power output, which is essential for effective model training and evaluation.

Stationarity test

The stationarity of the solar power time series data was assessed using two widely accepted statistical tests: the Augmented Dickey-Fuller (ADF) and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests. The ADF test evaluates the null hypothesis that the series possesses a unit root (i.e., it is non-stationary). The resulting p-value of 0.0004 is significantly below the 0.05 threshold, leading to the rejection of the null hypothesis and indicating that the solar power series is stationary.

Conversely, the KPSS test assumes stationarity as the null hypothesis. In this analysis, the KPSS test yielded a p-value of 0.1000, which is above the 0.05 level, thus failing to reject the null hypothesis and further confirming that the series is stationary. The agreement between both tests reinforces the stationarity of the solar power data, validating its suitability for time series modelling without the need for differencing or transformation. The result of stationarity tests is given in Table 2.

Random forest based feature importance

To further understand the relative influence of different input variables on solar power generation, feature importance analysis was carried out using the random forest model. *Irradiance* is the most significant predictor, contributing nearly 38 % to the model's performance (Fig. 6). This is followed by

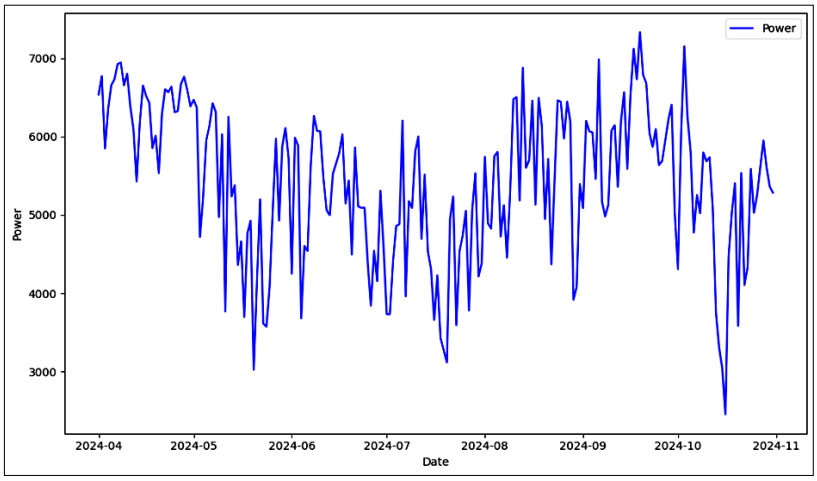


Fig. 5. Temporal variation in daily solar power output after outlier removal.

Table 2. Stationarity test results at 5 % significance level.

Test	*p*-value	Stationarity status
ADF test	0.0004	Stationary
KPSS test	0.1000	Stationary

power_lag1 (17 %), *power_lag2* (7 %) and *power_lag3* (6 %), indicating that recent historical power values play a notable role in prediction. Interaction terms such as irr*humidity and irr*temp_diff also showed moderate importance, while variables like max_temp and temp_diff contributed the least. These interaction features were manually engineered during the pre-processing stage by combining existing meteorological variables to enhance the model's predictive power. These findings highlight the strong dependence of solar power output on real-time irradiance and past power trends.

Predicting solar power

The nine models developed during the training phase were evaluated using the test dataset. The actual and predicted values are illustrated in Fig. 7, effectively demonstrating how well the model captures the variations and trends in daily solar power generation based on point predictions. Model performance was assessed using three error metrics-RMSE, MAE and MAPE-as presented in Table 3.

The performance comparison of machine learning models using RMSE, MAE and MAPE metrics clearly highlights that the ANN model outperforms all other models in predicting daily solar power generation. ANN achieves the lowest RMSE (274.84 kWh), MAE (245.93 kWh) and MAPE (5.26 %), indicating its superior ability to capture the complex, non-linear patterns present in the power generation data. This enhanced performance can be attributed to ANN's capacity to model intricate relationships between solar power output and multiple weather parameters, such as irradiance, temperature and humidity, especially in high-resolution (daily) datasets.

The random forest (RF) model also shows strong performance, particularly in terms of MAE (384.58 kWh) and MAPE (7.86 %), due to its robustness against overfitting and its ensemble nature, which averages over multiple decision trees to reduce variance. RF's capability to handle both linear and non-linear relationships and rank features effectively makes it a

Table 3. Performance of ML models for solar power prediction.

MODEL	RMSE (kWh)	MAE (kWh)	MAPE (%)
Linear regression	554.17	424.48	8.23
ARIMA	494.37	413.81	9.00
ANN	274.84	245.93	5.26
SVM	493.92	410.57	8.00
Decision tree	531.99	418.90	8.94
RF	504.28	384.58	7.86
GBM	542.14	433.69	8.90
LGBM	554.03	462.65	9.92
XGBM	622.47	486.53	9.75

reliable choice for time series prediction with limited but diverse input features. The SVM model demonstrates relatively good predictive ability, with an RMSE of 493.92 kWh, MAE of 410.57 kWh, and MAPE of 8.00 %. The use of a linear kernel (optimized with hyperparameters C = 29.98, gamma = 0.57 and epsilon = 0.74) proved effective, likely due to the near-linear relationship between the engineered features and the target variable, as well as the model's ability to generalize well on the limited dataset. Its strength lies in its ability to model non-linear patterns using kernel functions, which proves useful in solar power prediction where complex interactions between variables exist. However, its performance may be slightly limited by sensitivity to noise and its dependence on proper kernel and parameter selection, especially in a dataset with temporal variability and seasonality. In contrast, models such as ARIMA. Gradient Boosting Models (LGBM and XGBM) and linear regression exhibit relatively lower performance. The ARIMA model, being a univariate approach, is limited by its dependence on historical power values alone, without the inclusion of exogenous weather factors. Boosting models, though highly effective in capturing complex patterns, can exhibit variability in performance due to their iterative learning process. In certain scenarios, this may lead to challenges in generalization, resulting in either overfitting or underfitting, especially when dealing with intricate relationships within the data.

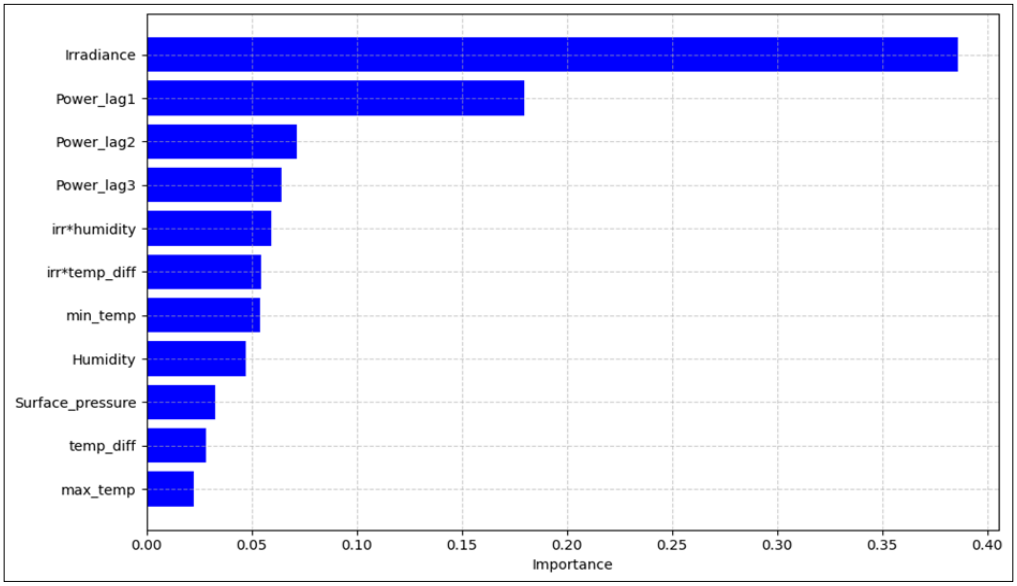
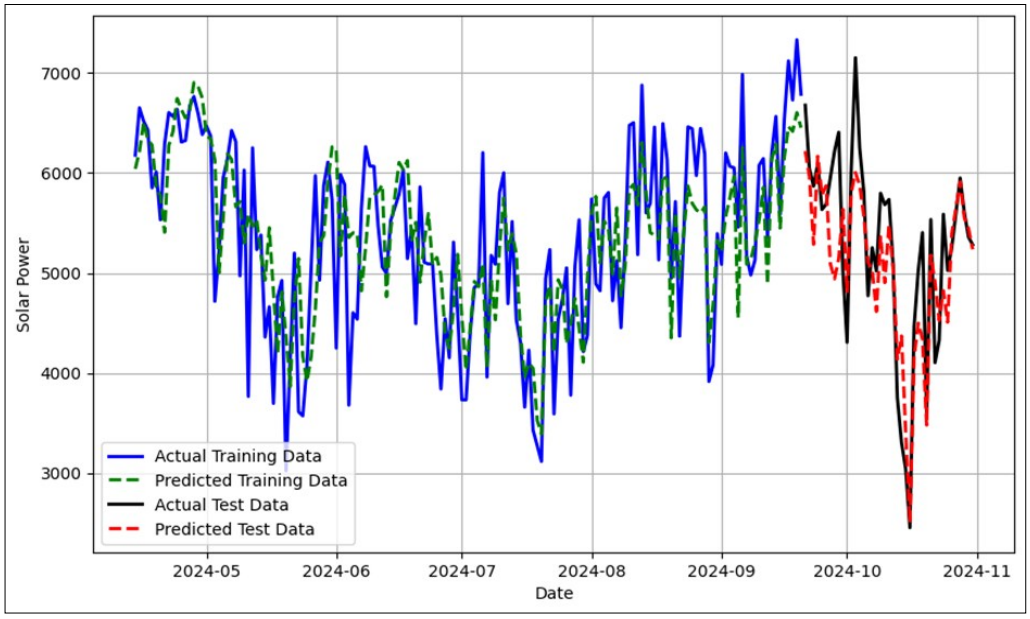
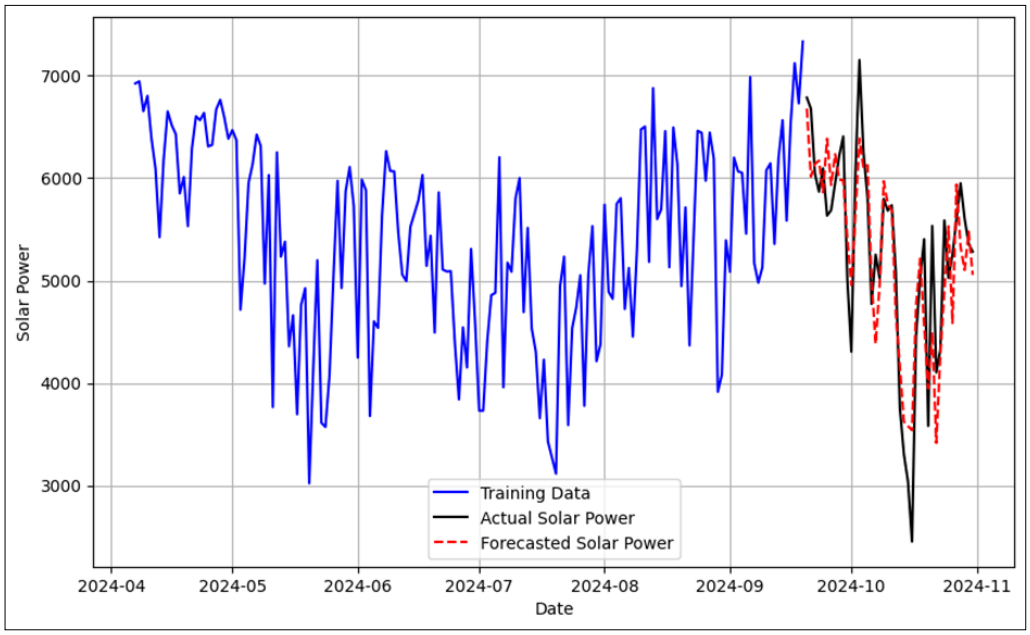


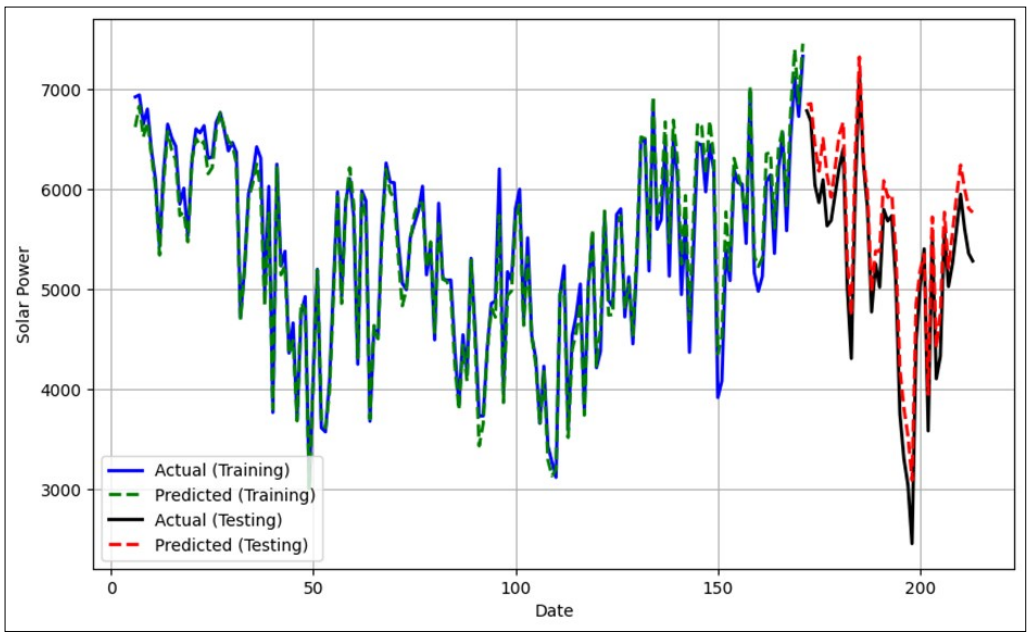
Fig. 6. Random forest-based feature importance of input variable on solar power.



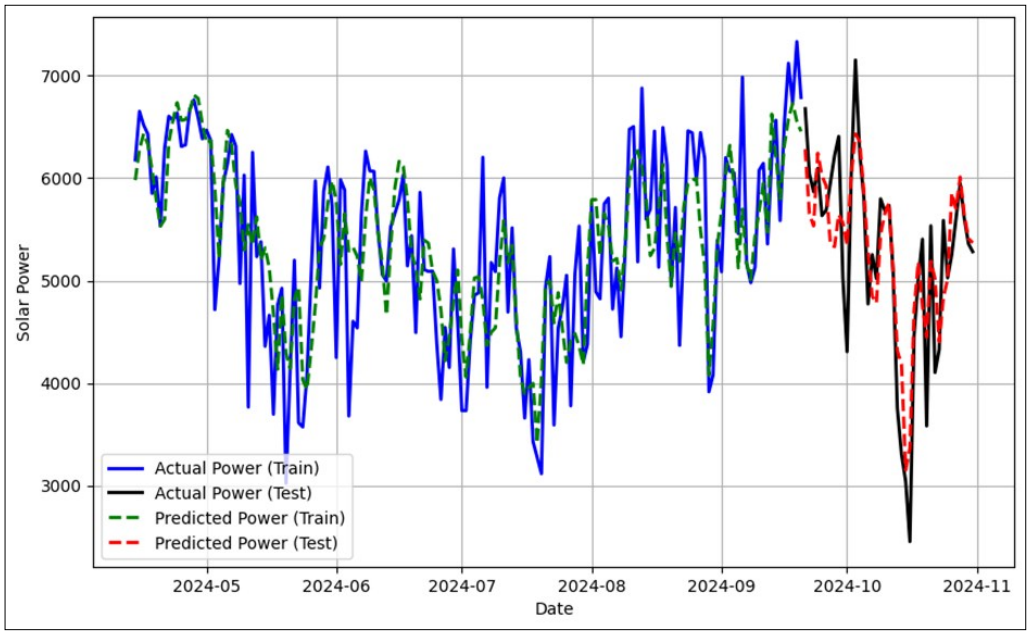
LINEAR REGRESSION



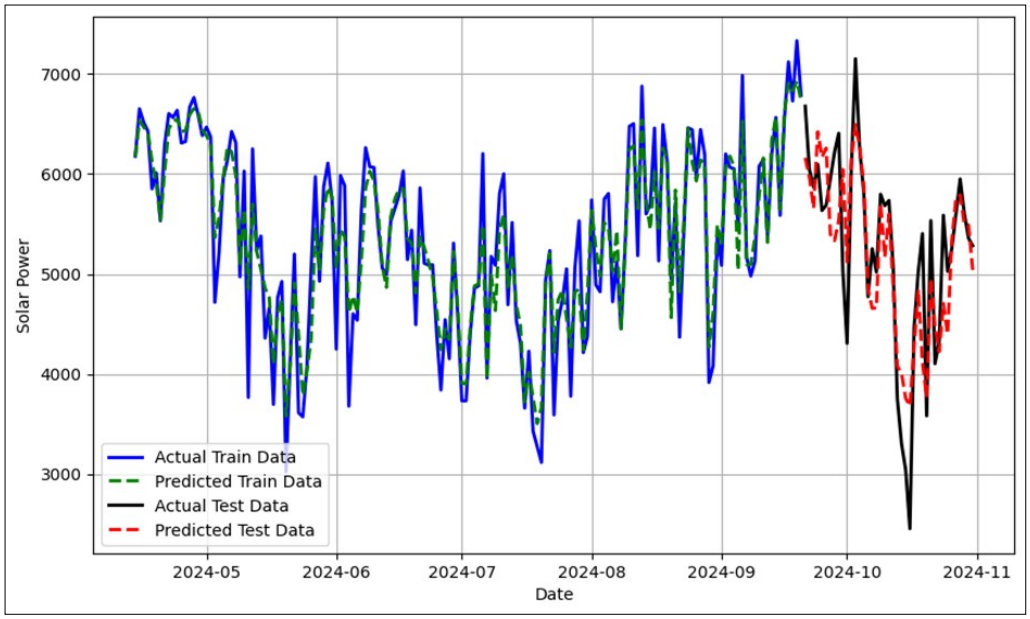
ARIMA



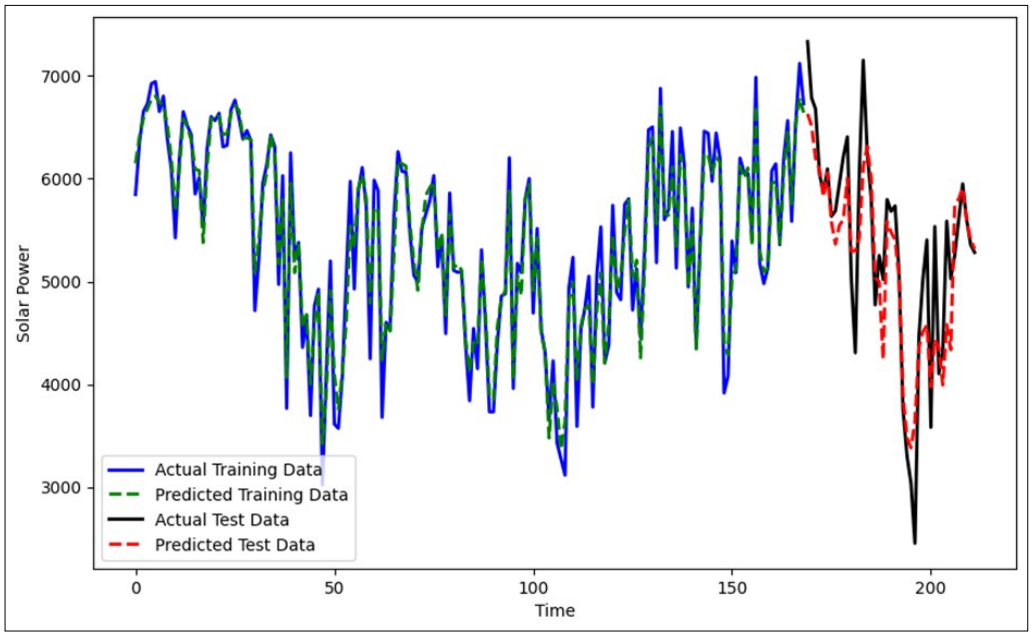
ANN



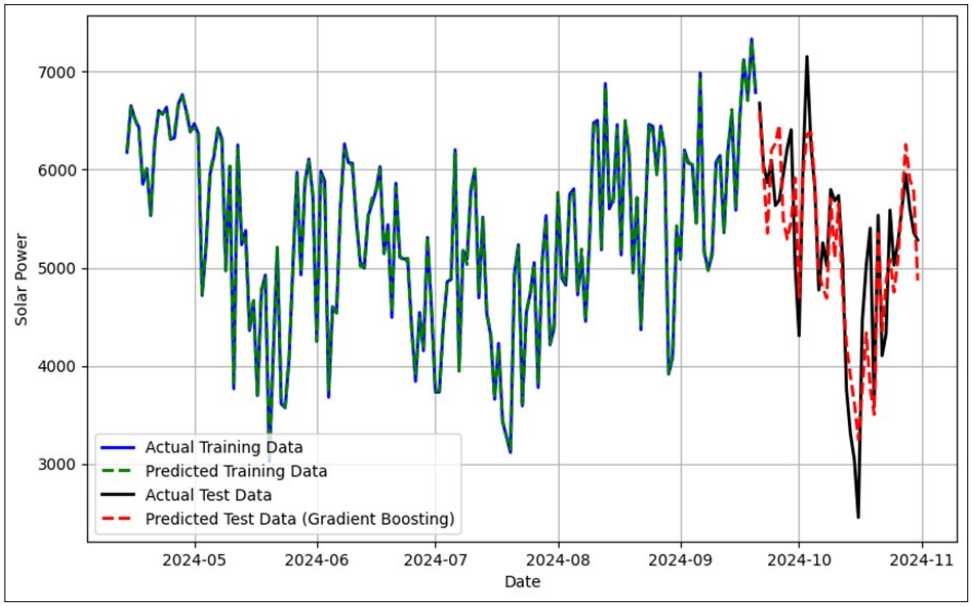
SVM



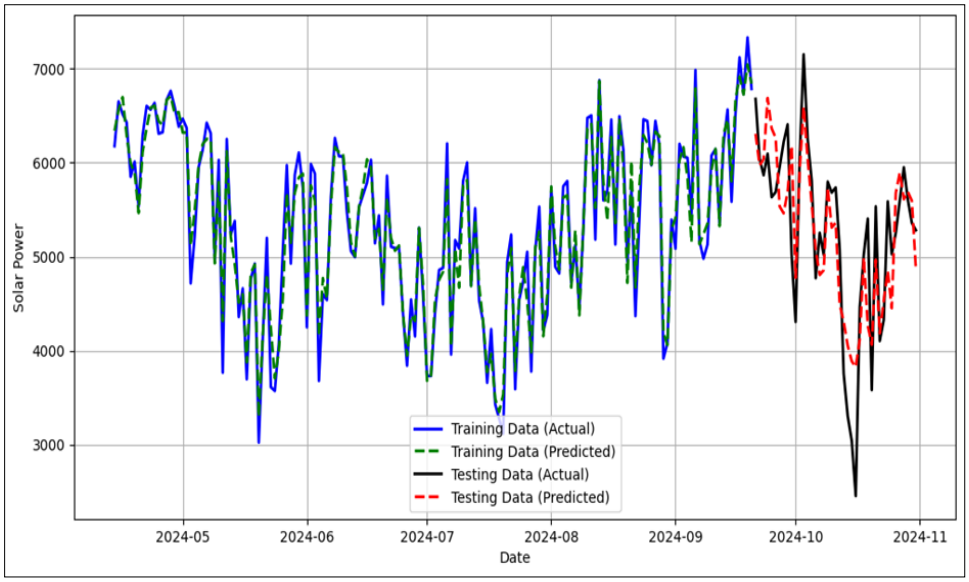
DECISION TREE



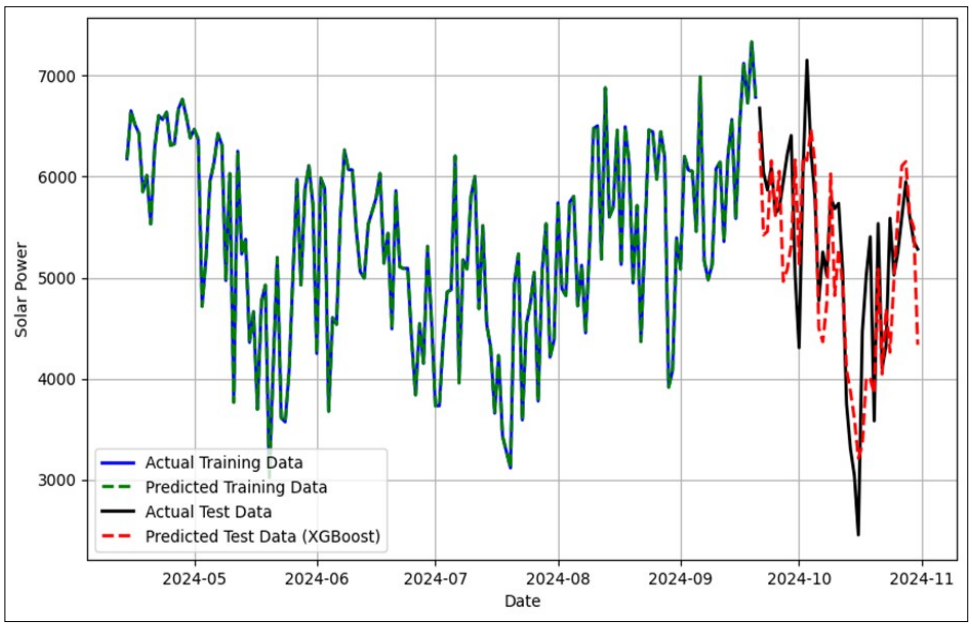
RANDOM FOREST



GBM



LGBM



XGBM

Fig. 7. Actual vs predicted solar power for different models.

These findings support the central hypothesis of this research that accurate solar power prediction is achievable even with a limited amount of data, provided that suitable machine learning models are employed. The success of ANN, RF and SVM demonstrates that complex models, when carefully chosen, can adapt well to smaller datasets and yield reliable results in real-world, short-horizon solar power prediction tasks.

Conclusion

This study has demonstrated the effectiveness of various ML models in predicting daily solar power generation using a limited dataset from a newly installed plant in Tamil Nadu. Among the nine models tested, the ANN emerged as a top performer, followed by random forest (RF) and SVM, highlighting the potential of these models in capturing non-linear relationships and making accurate predictions even with fewer observations. The research distinguishes itself by achieving high prediction accuracy using minimal data, offering a practical alternative to data-intensive approaches commonly found in existing literature. Looking ahead, future work can focus on developing real-time prediction systems using live weather data, integrating explainable AI for transparency and further improving accuracy with satellite-derived inputs. Advancements such as smart grid integration, energy storage optimization and decentralized frameworks like federated learning could further revolutionize solar power forecasting and management.

Acknowledgements

The authors are thankful to the Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu for supporting their research work.

Authors' contributions

PKF conceptualized the study, developed the models, and drafted the initial manuscript. MR supervised the research, refined the methodology and provided critical revisions. SVS contributed to the evaluation of model performance and assisted in interpreting the results. MND, GV and RM reviewed the draft manuscript and provided suggestions and corrections. All authors read and approved the final version of the manuscript.

Compliance with ethical standards

Conflict of interest: Authors do not have any conflict of interests to declare.

Ethical issues: None.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (OpenAI) solely to paraphrase certain sentences for improved clarity and readability. The authors reviewed and edited the content thoroughly and take full responsibility for the final content of the publication.

References

1. Hsu DD, O'Donoughue P, Fthenakis V, Heath GA, Kim HC, Sawyer P, et al. Life cycle greenhouse gas emissions of crystalline silicon photovoltaic electricity generation: Systematic review and harmonization. J Ind Ecol. 2012;16(s1):s122-35. https://doi.org/10.1111/j.1530-9290.2011.00439.x
2. Tajjour S, Chandel SS, Chandel R, Thakur N. Power generation enhancement analysis of a 400 kWp grid-connected rooftop photovoltaic power plant in a hilly terrain of India. Energy Sustain Dev. 2023;77:101333. https://doi.org/10.1016/j.esd.2023.101333
3. Chandel SS, Gupta A, Chandel R, Tajjour S. Review of deep learning techniques for power generation prediction of industrial solar photovoltaic plants. Sol Compass. 2023;8:100061. https://doi.org/10.1016/j.solcom.2023.100061
4. Tajjour S, Chandel SS, Alotaibi MA, Malik H, García Márquez FP, Afthanorhan A. Short-term solar irradiance forecasting using deep learning techniques: A comprehensive case study. IEEE Access. 2023;11:119851–61.
5. Rafati A, Joorabian M, Mashhr E, Shaker HR. High dimensional very short-term solar power forecasting based on a data-driven heuristic method. Energy. 2021;219:119647. https://doi.org/10.1016/j.energy.2020.119647
6. du Plessis AA, Strauss JM, Rix AJ. Short-term solar power forecasting: Investigating the ability of deep learning models to capture low-level utility-scale photovoltaic system behaviour. Appl Energy. 2021;285:116395. https://doi.org/10.1016/j.apenergy.2020.116395
7. Son N, Jung M. Analysis of meteorological factor multivariate models for medium- and long-term photovoltaic solar power forecasting using long short-term memory. Appl Sci. 2021;11:1:316. https://doi.org/10.3390/app11010316
8. Al-Dahidi S, Madhiarasan M, Al-Ghussain L, Abubaker AM, Ahmad AD, Alrbai M, et al. Forecasting solar photovoltaic power production: A comprehensive review and innovative data-driven modeling framework. Energies. 2024;17:16:4145. https://doi.org/10.3390/en17164145
9. Pieri E, Kyprianou A, Phinikarides A, Makrides G, Georghiou GE. Forecasting degradation rates of different photovoltaic systems using robust principal component analysis and ARIMA. IET Renew Power Gener. 2017;11(10):1245–52. https://doi.org/10.1049/iet-rpg.2017.0090
10. Tajjour S, Chandel SS, Malik H, Alotaibi MA, Ustun TS. A novel metaheuristic approach for solar photovoltaic parameter extraction using manufacturer data. Photonics. 2022;9:11:858. https://doi.org/10.3390/photonics9110858
11. Wang F, Lu X, Mei S, Su Y, Zhen Z, Zou Z, et al. A satellite image data based ultra-short-term solar PV power forecasting method considering cloud information from neighboring plant. Energy. 2022;238:121946. https://doi.org/10.1016/j.energy.2021.121946
12. Pasion C, Wagner T, Koschnick C, Schuldt S, Williams J, Hallinan K. Machine learning modeling of horizontal photovoltaics using weather and location data. Energies. 2020;13(10):2570.
13. Oladapo BI, Olawumi MA, Omigbodun FT. Machine learning for optimising renewable energy and grid efficiency. Atmosphere. 2024;15:10:1250. https://doi.org/10.3390/atmos15101250
14. Chang R, Bai L, Hsu CH. Solar power generation prediction based on deep Learning. Sustain Energy Technol Assess. 2021;47:101354. https://doi.org/10.1016/j.seta.2021.101354
15. Suanpang P, Jamjuntr P. Machine learning models for solar power generation forecasting in microgrid application implications for smart cities. Sustainability. 2024;16:14:6087. https://doi.org/10.3390/su16146087
16. Ibrar M, Hassan MA, Shaukat K, Alam TM, Khurshid KS, Hameed IA, et al. A machine learning-based model for stability prediction

of decentralized power grid linked with renewable energy resources. Wirel Commun Mob Comput. 2022;2022(1):2697303. https://doi.org/10.1155/2022/2697303

17. Yadav AK, Malik H, Chandel SS. ANN based prediction of daily global solar radiation for photovoltaics applications. In: 2015 Annual IEEE India Conference (INDICON). 2015. p.1–5.
18. Pereira S, Canhoto P, Salgado R, Costa MJ. Development of an ANN based corrective algorithm of the operational ECMWF global horizontal irradiation forecasts. Sol Energy. 2019;185:387-405. https://doi.org/10.1016/j.solener.2019.04.070
19. Gundu V, Simon SP. Short term solar power and temperature forecast using recurrent neural networks. Neural Process Lett. 2021;53(6):4407–18. https://doi.org/10.1007/s11063-021-10606-7
20. Long H, Zhang C, Geng R, Wu Z, Gu W. A combination interval prediction model based on biased convex cost function and auto-encoder in solar power prediction. IEEE Trans Sustain Energy. 2021;12(3):1561–70.
21. Yuan X, Liu S, Feng W, Dauphin G. Feature importance ranking of random forest-based end-to-end learning algorithm. Remote Sens. 2023;15(21):5203.
22. Hillmer SC, Tiao GC. An ARIMA-model-based approach to seasonal adjustment. J Am Stat Assoc. 1982;77(377):63-70. https://doi.org/10.1080/01621459.1982.10477767
23. Newbold P. ARIMA model building and the time series analysis approach to forecasting. J Forecast. 1983;2(1):23–35. https://doi.org/10.1002/for.3980020104
24. Sairamya NJ, Susmitha L, Thomas George S, Subathra MSP. Hybrid approach for classification of electroencephalographic signals using time-frequency images with wavelets and texture features. In: Hemanth DJ, Gupta D, Emilia Balas V, editors. Intelligent data analysis for biomedical applications. Academic Press; 2019. p.253-73. https://doi.org/10.1016/B978-0-12-815553-0.00013-6
25. Hamzaçebi C. Improving artificial neural networks' performance in seasonal time series forecasting. Spec Sect Genet Evol Comput. 2008;178(23):4550-59. https://doi.org/10.1016/j.ins.2008.07.024
26. Vandeginste BGM, Massart DL, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J. Artificial neural networks. In: Vandeginste BGM, Massart DL, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J, editors. Data handling in science and technology. Elsevier; 1998. p. 649-99. https://doi.org/10.1016/S0922-3487(98)80054-3
27. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. https://doi.org/10.1007/BF00994018
28. Mohammadi K, Shamshirband S, Anisi MH, Alam KA, Petković D. Support vector regression based prediction of global solar radiation on a horizontal surface. Energy Convers Manag. 2015;91:433-41. https://doi.org/10.1016/j.enconman.2014.12.015
29. Setiawati P, Karno ASB, Hastomo W, Sestri E, Kasoni D, Arif D, et al. Predicting solar power generation: A machine learning approach for grid stability and efficiency. J Pilar Nusa Mandiri. 2025;21(1):34–43.
30. Aning S, Przybyła-Kasperek M. Comparative study of twoing and entropy criterion for decision tree classification of dispersed data. Knowl-Based Intell Inf Eng Syst Proc 26th Int Conf KES2022. 2022;207:2434-43. https://doi.org/10.1016/j.procs.2022.09.301
31. Safari A, Kheirandish Gharehbagh H, Nazari Heris M. DeepVELOX: INVELOX wind turbine intelligent power forecasting using hybrid GWO-GBR algorithm. Energies. 2023;16:19:6889. https://doi.org/10.3390/en16196889
32. Sauer J, Mariani VC, dos Santos Coelho L, Ribeiro MHDM, Rampazzo M. Extreme gradient boosting model based on improved Jaya optimizer applied to forecasting energy consumption in residential buildings. Evol Syst. 2022;13(4):577–88. https://doi.org/10.1007/s12530-021-09404-2
33. Fan GF, Zhang LZ, Yu M, Hong WC, Dong SQ. Applications of random forest in multivariable response surface for short-term load forecasting. Int J Electr Power Energy Syst. 2022;139:108073. https://doi.org/10.1016/j.ijepes.2022.108073

Additional information

Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

Publisher information: Plant Science Today is published by HORIZON e-Publishing Group with support from Empirion Publishers Private Limited, Thiruvananthapuram, India.