International Journal for Holistic Research, Vol. 2, No. 2. Jan 2025

OPEN ACCESS

Print ISSN: 3009-6898
Online ISSN: 3009-6790

Machine Learning-Based Prediction of Illuminance and Ultraviolet Irradiance in Photovoltaic Systems

Montaser Abdelsattar [1,*], Ahmed AbdelMoety [1], Ahmed Emad-Eldeen [2]

Montaser.A.Elsattar@eng.svu.edu.eg , Ahmed.AbdelMoety@eng.svu.edu.eg , ahmed.emad@psas.bsu.edu.eg

[1] Electrical Engineering Department, Faculty of Engineering, South Valley University, Qena 83523, Egypt

[2] Renewable Energy Science and Engineering Department, Faculty of Postgraduate Studies for Advanced Sciences (PSAS), Beni-Suef University, Beni-Suef 62511, Egypt

Abstract

Photovoltaic (PV) systems are indispensable in the renewable energy industry as they convert sunlight into electricity. Accurate determination of important factors such as illuminance and Ultraviolet (UV) irradiation is essential for optimizing the effectiveness and maintenance of these systems. The objective of this work is to evaluate the predictive performance of several Machine Learning (ML) models in estimating the amounts of light and UV radiation in PV systems, by comparing and contrasting their effectiveness. The models that were assessed include Support Vector Classification (SVC), Linear Regression (LR), eXtreme Gradient Boosting (XGBoost), Gradient Boosting (GB), Random Forest (RF), and CatBoost. The study employed a comprehensive dataset that encompassed measurements for temperature, humidity, UV, voltage, current, and illuminance. The data was preprocessed to remove invalid values and align indices. Afterwards, it was divided into separate training and testing sets. The main metrics used to train and evaluate each model were Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2). The findings suggest that the Categorical Boosting (CatBoost) and RF models demonstrate greater performance in comparison to other models. This is evidenced by their ability to obtain the lowest RMSE and highest R^2 values for both illuminance and UV forecasts. More precisely, CatBoost algorithm obtained a RMSE of 16.088 and a R^2 of 0.999 for illuminance. Additionally, it achieved a RMSE of 0.228 and a R^2 of 0.990 for UV. However, LR and SVC had notably inferior results. The results offered valuable perspectives for enhancing decision-making procedures.

Keywords: Photovoltaic systems, machine learning, illuminance prediction, UV irradiance prediction, renewable energy.

1. Introduction

Photovoltaic (PV) systems utilize semiconductor materials to directly turn sunlight into electricity. This process relies on the PV effect, which involves the conversion of light energy into electrical energy at the atomic scale. PV systems play a crucial role in generating renewable energy by effectively utilizing the ample solar resources that are accessible worldwide. Deploying these systems is crucial for diminishing reliance on finite fossil fuels, which make a substantial contribution to environmental degradation and climate change [1]. The increasing global deployment of PV systems highlights their significance in the energy environment. They provide a multitude of environmental and economic advantages, including the reduction of greenhouse gas emissions, the decrease in air pollution, and the provision of energy security. Technological breakthroughs and economies of scale have made solar energy increasingly cost-effective, positioning it as one of the most promising sources of renewable energy [2, 3].

The swift proliferation of PV systems is propelled by their capacity to alleviate the consequences of climate change and offer sustainable energy alternatives. Nations worldwide are making significant investments in solar power infrastructure, bolstered by advantageous governmental policies and incentives designed to encourage the adoption of renewable energy [4, 5]. Monitoring and forecasting the performance of PV systems present numerous difficulties. Precise performance prediction is essential for enhancing the efficiency and dependability of PV systems. Nevertheless, various environmental conditions, including temperature, humidity, Ultraviolet (UV) irradiation, dust deposition, and shadowing, can have a substantial influence on the efficiency and output of PV panels [6, 7]. High temperatures, for example, might lower the voltage output and increase internal resistance, therefore compromising the efficiency of PV cells. Dust collection and humidity can cause PV panels to soil, therefore compromising their capacity to efficiently absorb sunlight. Shading, from surrounding buildings, trees, or passing clouds, can drastically lower power output. Studies have indicated that dust can lower efficiency by about 11.86% [8, 9]. Shade can reduce power production by up to 92.6%.

Furthermore, the dynamic character of solar irradiation and the different climatic circumstances call for advanced monitoring and forecasting models to sustain the best PV performance. Artificial neural networks and other cutting-edge Machine Learning (ML) methods have shown promise in enhancing the accuracy of performance predictions and in PV system anomaly detection [10, 11].

Cite this article: Montaser Abdelsattar, Ahmed AbdelMoety, Ahmed Emad-Eldeen" Machine Learning-Based Prediction of Illuminance and Ultraviolet Irradiance in Photovoltaic Systems ", International Journal for Holistic Research, Vol. 2, No. 2. Jan 2025. DOI: 10.21608/ijhr.2024.308523.1025

Maximum performance and maintenance of PV systems depend on accurate predictive models. By allowing timely maintenance and so reducing downtime, these models can greatly improve the efficiency and sustainability of PV systems so guaranteeing that the systems run at their best

[12, 13]. Better energy management, significant cost savings, and increased system dependability can follow from accurate forecasts of PV system performance. Predictive maintenance models, for example, can foresee failures and schedule repairs before problems get more severe, therefore lowering running costs and extending the lifetime of PV installations [14, 15]. To improve the forecasting accuracy for certain parameters in PV systems, ML techniques have been progressively used. Particularly helpful for forecasting solar irradiation, system performance, and maintenance needs [16, 17]. these algorithms can process enormous volumes of data and find intricate patterns that older approaches might overlook. ML has demonstrated significant potential in renewable energy research for optimizing systems, detecting defects, and predicting energy output. Support vector regression (SVR) and Adaptive Neuro-Fuzzy Inference System (ANFIS) have been utilized to accurately forecast PV power generation. This application has resulted in enhanced grid stability and improved integration of solar energy into the electrical grid [18, 19].

The primary aim of this work is to assess and contrast the efficacy of several ML models in forecasting illuminance and UV irradiance in PV systems. The research tries to determine the most successful models for crucial parameters by analyzing the performance of several algorithms, such as SVC, Linear Regression (LR), eXtreme Gradient Boosting (XGBoost), Gradient Boosting (GB), Random Forest (RF), and CatBoost. In addition, the study aims to comprehend the impact of several environmental conditions, such as temperature, humidity, and voltage, on the precision of these forecasts. The study will yield significant information that may be used to make recommendations for enhancing the monitoring and performance optimization of PV systems in the future. The importance of this study is in its capacity to enhance the field of solar performance monitoring. The results can provide significant advantages to professionals in the industry, researchers, and policymakers by improving the forecasting skills of various ML models. This will result in more precise and dependable monitoring of PV systems. Moreover, the study's data-driven insights might assist in making well-informed selections about the implementation, administration, and upkeep of PV systems.

This work proposes the utilization of ML models to forecast important factors, such as illuminance and UV irradiation, in order to enhance the monitoring and optimization of PV system performance. Existing approaches to forecast the performance of PV systems are frequently constrained by their dependence on linear models or oversimplified assumptions, neglecting the many environmental and operational factors that impact PV efficiency. This project intends to enhance the accuracy of forecasts by utilising modern machine learning techniques. The goal is to provide more trustworthy insights for improving energy production and system maintenance. This work is significant because it tackles the increasing worldwide need for renewable energy by providing creative ideas to enhance the efficiency of solar energy systems.

Table 1 presents a thorough comparison of many research that concentrate on predicting solar irradiance using different ML techniques. Demir et al. [20] utilised Transformer and Rolling LSTMs to analyze a decade's worth of data from the Texas Mesonet Data Archive. Their approach yielded precise predictions for the short-term, but encountered difficulties in accurately forecasting long-term outcomes, mostly owing to the need for fine-tuning hyperparameters. Li and He [21] utilised an Enhanced Incremental Extreme Learning Machine to analyse historical irradiance data. This approach resulted in minimal prediction errors, but its effectiveness was constrained by the exclusive reliance on irradiance data as input.

Alzahrani [22] employed an Adaptive Extreme Learning Machine utilising hourly meteorological data from Najran University. The model achieved a high level of prediction accuracy, but its applicability was restricted due to its concentration on a single geographical region. Viscondi and Alves-Souza [23] applied a hybrid approach using Support Vector Machine (SVM), Artificial Neural Network (ANN), and Extreme Learning Machine (ELM) models to analyse a dataset from São Paulo, Brazil. Their investigation demonstrated that Support Vector Machine (SVM) yielded the lowest Root Mean Square Error (RMSE), notwithstanding the influence of regional weather variability on the models' performance. Aliberti et al. [24] performed a comparative investigation of neural networks, namely Non-Linear Autoregressive and LSTM models, utilising GHI data collected at 15-minute intervals. The study revealed that the implementation of the Echo State Network and clear-sky index filtering had a substantial positive impact on the accuracy of forecasting. It is important to note that the study specifically concentrated on neural networks.

Finally, Huang et al. [25] proposed a hybrid deep neural model for hourly solar irradiance forecasting based on weather and irradiance data, achieving strong results but requiring significant computational resources. Maitanova et al. [26] explored a ML approach using publicly available weather reports, emphasizing the cost-effective nature of their model but acknowledging the limitations posed by the availability and quality of public data sources.

Table 1 underscores the diverse methodologies and datasets used, highlighting both the advancements and limitations in solar irradiance prediction.

Table 1. Recent studies on ML-Based solar irradiance prediction and PV system performance optimization.

Study	Year	Algorithms Used	Dataset Characteristics	Key Findings	Limitations
[20]	2022	Transformer, Rolling LSTMs	10 years of data from Texas Mesonet Data Archive	Transformer model shows accurate short-term irradiance prediction but less accurate in the long term.	Long-term prediction requires hyperparameter tuning.
[21]	2022	Enhanced Incremental Extreme Learning Machine	Historical irradiance data	Enhanced extreme learning machine offers smaller prediction errors than standard models.	Only irradiance data is used as input, limiting prediction scope.
[22]	2022	Adaptive Extreme Learning Machine	Hourly weather data from Najran University	High accuracy in predicting solar irradiance with low MSE and MAE values.	Focused on a specific geographical area, limiting generalization.
[23]	2021	SVM, ANN, Extreme Learning Machine	São Paulo, Brazil dataset (1933-2014) with 10 meteorological parameters	SVM produced the lowest RMSE, while ELM showed faster training rates.	Meteorological variability across regions affects model accuracy.
[24]	2021	Non-Linear Autoregressive, Feed-Forward, LSTM, Echo State Network	GHI values sampled every 15 minutes	Echo State Network and clear-sky index filtering showed best accuracy for GHI predictions.	Focused on specific neural networks, limiting comparisons with other machine learning techniques.
[25]	2021	WPD-CNN-LSTM-MLP (Wavelet Packet Decomposition + CNN + LSTM + MLP)	Hourly solar irradiance and three climate variables (temperature, humidity, wind speed)	Hybrid model outperforms standalone models in irradiance forecasting, achieving more accurate results.	Computational complexity increases significantly.
[26]	2020	Long Short-Term Memory	Publicly available weather data without solar irradiance values	The model can predict PV power with reasonable accuracy using only publicly available weather data.	Requires large training sets for adequate predictions with publicly available data.

2. Methodology

2.1 Data Presentation

In this section, we present the key visualizations derived from the dataset to provide an in-depth understanding of the features and their relationships. Fig. 1 showcases histograms of all features to illustrate the frequency distribution of each variable. The features include Fig. 1(a) Temperature, Fig. 1(b) Humidity, Fig. 1(c) UV, Fig. 1(d) Voltage, Fig. 1(e) Current, and Fig. 1(f) Illuminance. These histograms help to visualize the distribution and range of values for each feature. Fig. 2 presents box plots for all features, which highlight the spread and potential outliers in the dataset. The features include Fig. 2(a) Temperature, Fig. 2(b) Humidity, Fig. 2(c) UV, Fig. 2(d) Voltage, Fig. 2(e) Current, and Fig. 2(f) Illuminance. Box plots are essential for understanding the central tendency and variability of the data.

Fig. 3 displays the correlation matrix as a heatmap. Fig. 3 indicates the strength and direction of linear relationships between pairs of features, including Temperature, Humidity, UV, Voltage, Current, and Illuminance. High correlation values (close to 1 or -1) suggest a strong relationship, while values close to 0 suggest a weak or no relationship. Fig. 4 shows the scatter plot of Voltage vs. Current. This plot helps in visualizing the relationship between these two electrical parameters, revealing any underlying patterns or trends. Fig. 5 presents the scatter plot of Temperature vs. Voltage. This visualization helps in understanding how voltage varies with temperature changes, providing insights into thermal effects on electrical performance. Fig. 6 depicts the scatter plot of UV vs. Illuminance. This plot illustrates the relationship between UV irradiance and illuminance, which are both critical factors influencing the performance of PV panels.

The dataset comprises 5801 records collected from a PV panel monitoring system over several days. Each record includes measurements of various environmental and electrical parameters: Temperature (in degrees Celsius), Humidity (in percentage), UV irradiance (in mW/cm²), Voltage (in volts), Current (in amperes), and Illuminance (in lux).

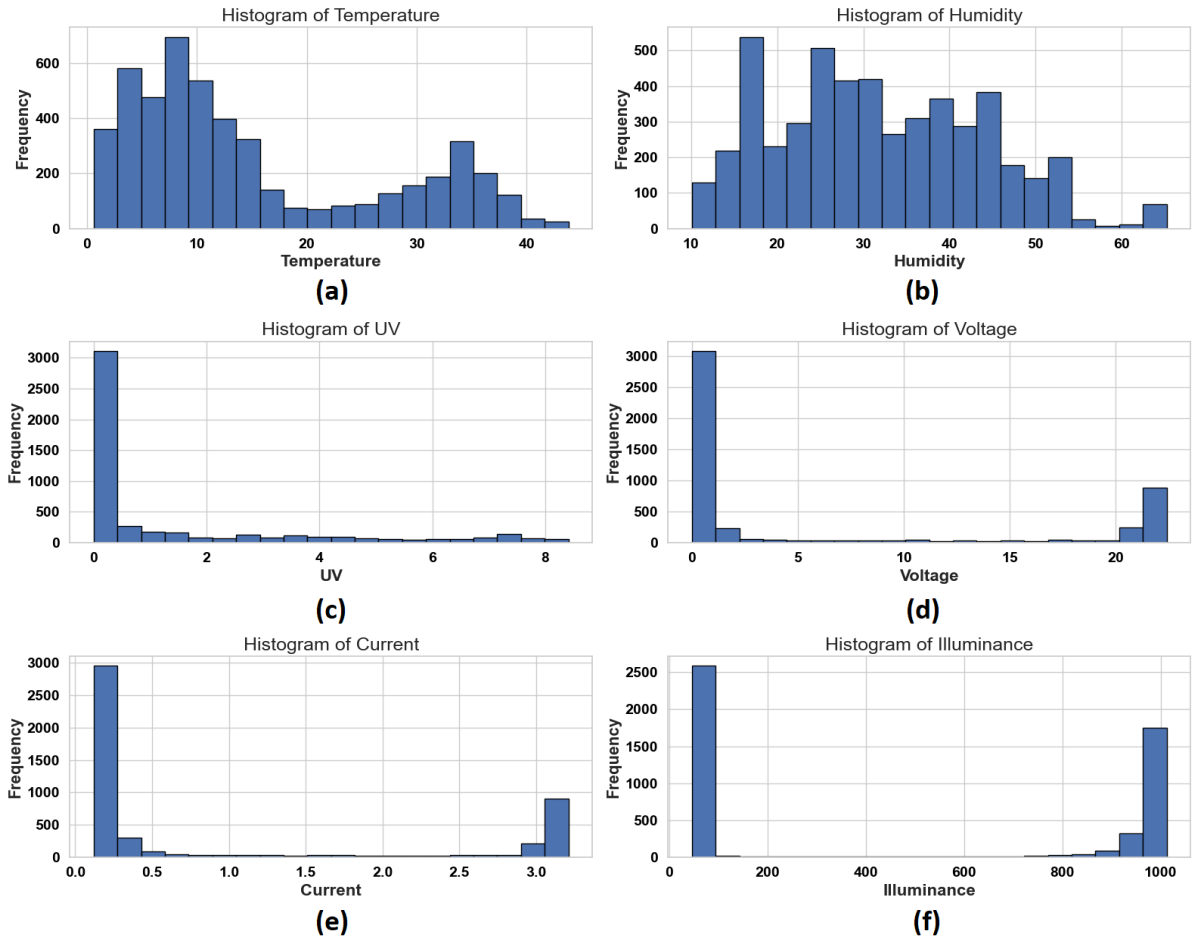

Fig. 1. Histograms of all features - (a) Temperature, (b) Humidity, (c) UV, (d) Voltage, (e) Current, (f) Illuminance.

Fig. 1 shows the histograms of every feature to underline the frequency distribution of every dataset variable. Fig. 1(a) shows the temperature histogram, and it is clear from the right-skewed distribution that most values occur between 0°C and 20°C and that frequency decreases clearly with increasing temperatures. Showed in Fig. 1(b), the humidity histogram displays a rather symmetric distribution with a peak between 30% and 40% humidity. Fig. 1(c) shows the UV irradiation histogram with a rather right-skewed distribution with most values close to 0 mW/cm². The behavior of the solar panel in several operational phases is suggested by the clustering of values near 0 V and roughly 20 V in Fig. 1(d). Fig. 1(e) shows the current histogram now, which once more shows a right-skewed distribution with a secondary peak about 3 A and most current values close to 0 A. Finally illustrated in Fig. 1(f), the histogram of illumination displays a distribution with most values around 0 and a secondary peak on the custom scale near 1000. These histograms clarify the variability and central patterns of every feature by showing its range of values.

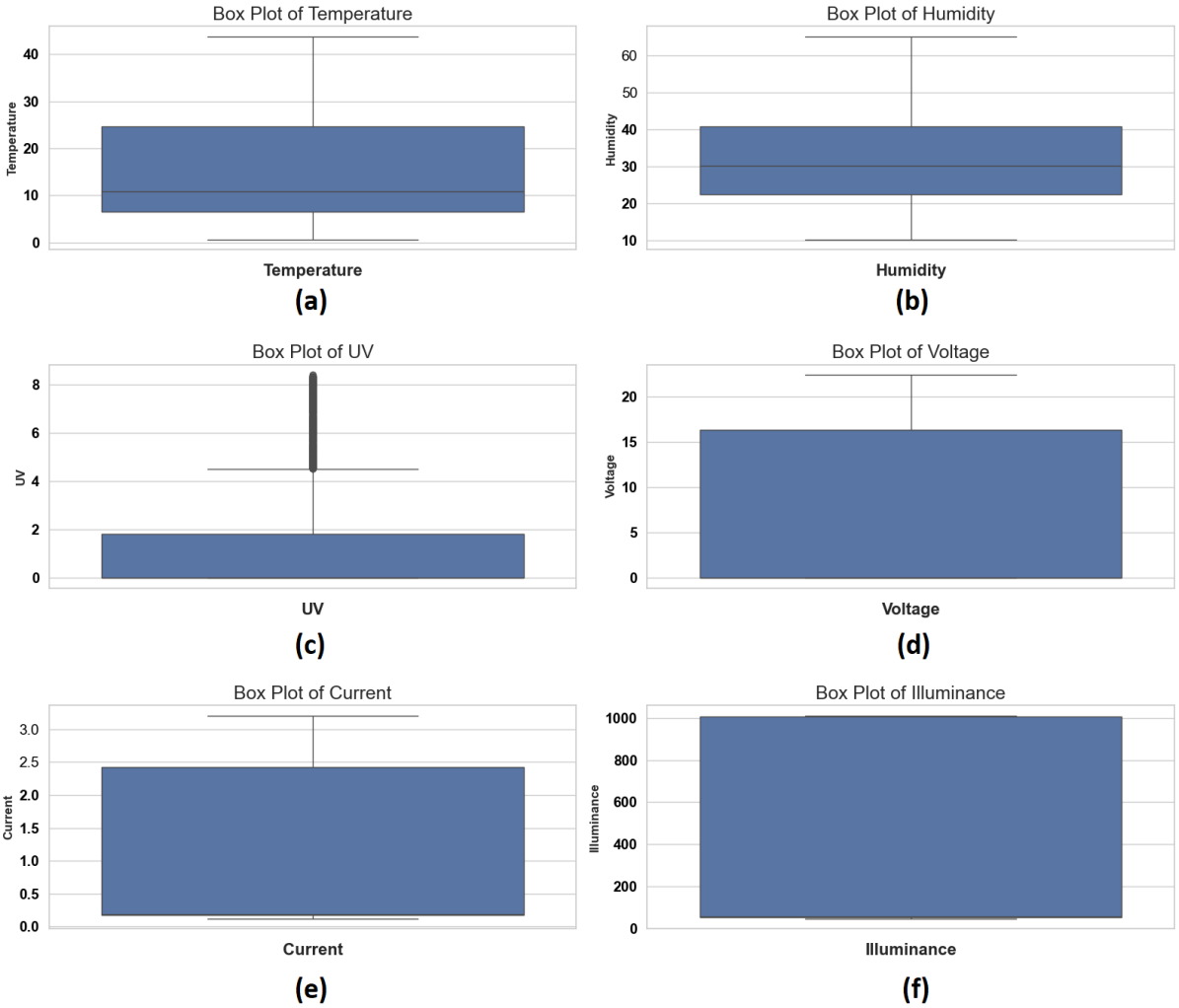

Fig. 2. Box plots of all features - (a) Temperature, (b) Humidity, (c) UV, (d) Voltage, (e) Current, (f) Illuminance.

Fig. 2 exhibits box plots for all features in the dataset to emphasize the dispersion and any anomalies. Fig. 2(a) displays the box plot of temperature, with the interquartile range (IQR) ranging from roughly 10°C to 20°C, and a median value of around 15°C. This suggests that the majority of temperature readings fall within this specific range, while a small number of exceptional results go as high as 40°C. Fig. 2(b) displays the box plot of Humidity, showing that the IQR spans from 20% to 40%, and the median is approximately 30%. This indicates that the bulk of humidity measurements fall within this range, with very few exceptions. Fig. 2(c) displays the box plot of UV irradiance, illustrating a significantly skewed distribution with the bulk of values falling between 0 and 2 mW/cm², and a few outliers reaching as high as 8 mW/cm². This suggests that although the majority of UV values are modest, there are sporadic instances of high readings. Fig. 2(d) presents a box plot of voltage, showing that the IQR spans from 5 V to 15 V. The median value is approximately 10 V, suggesting a broad distribution of voltage values, with a few values reaching as high as 20 V. Fig. 2(e) displays a box plot of the current variable, indicating an IQR ranging from around 0.5 A to 2 A, with a median value close to 1.5 A. This suggests that the majority of current values are inside this range, with only a few exceptional cases. Fig. 2(f) shows a box plot of Illuminance, with an IQR ranging from 0 to 1000 on a customized scale. This indicates that the illuminance values are distributed throughout the whole range of measurements. The box plots offer a comprehensive depiction of the central tendencies, variability, and probable outliers for each feature in the dataset, facilitating a lucid comprehension.

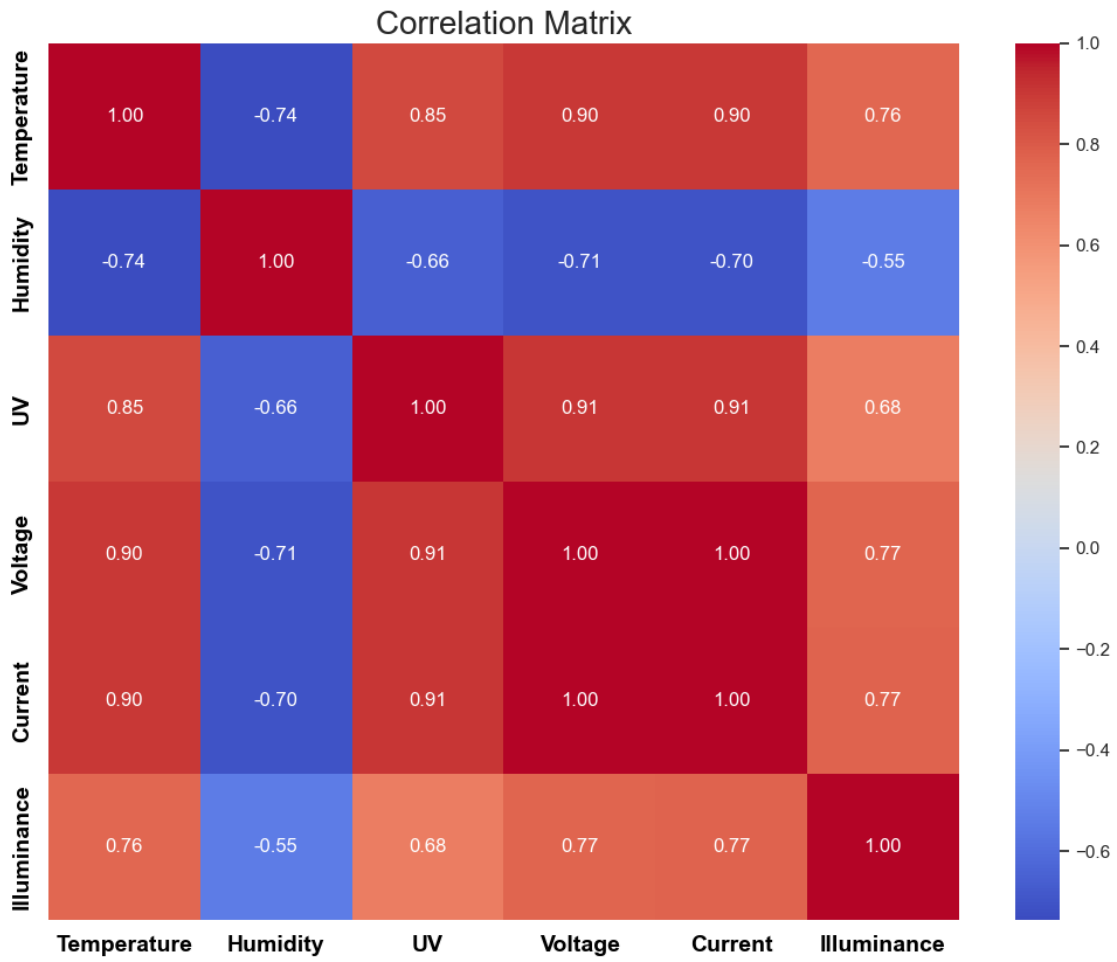

Fig. 3. Correlation matrix - Correlation heatmap of Temperature, Humidity, UV, Voltage, Current, and Illuminance.

Fig. 3 depicts a heatmap that illustrates the correlation matrix. This matrix depicts the magnitude and orientation of linear correlations among several attributes, such as temperature, humidity, UV, voltage, current, and illuminance. The correlation coefficients range from -1 to 1, with values near 1 indicating a strong positive relationship, values around -1 indicating a strong negative correlation, and values around 0 indicating no linear association. After examining this heatmap, it is clear that temperature is strongly positively correlated with UV (0.85), Voltage (0.90), and current (0.90). Therefore, a rise in temperatures will result in a proportional increase in these factors. On the other hand, there is a significant negative correlation (-0.74) between temperature and humidity, indicating that higher temperatures are associated with lower humidity levels. The correlation study demonstrates that humidity has significant negative relationships with UV (-0.66), Voltage (-0.71), and Current (-0.70), indicating that higher humidity levels are connected with lower values of these electrical parameters. UV irradiance has positive significant relationships with Voltage (0.91) and Current (0.91), suggesting that higher UV levels result in higher voltage and current outputs from solar panels. The correlation coefficient between voltage and current is 1.00, indicating a perfect relationship. This is in line with expectations in electrical systems, where these two variables are strongly interconnected. Ultimately, Illuminance has a notable and affirmative correlation with all variables, with the exception of Humidity. The most prominent connections are observed with Voltage (0.77) and Current (0.77), suggesting that higher levels of illuminance generally result in greater levels of electrical output. This heatmap provides a detailed depiction of the relationships between the features, which is crucial for understanding the core patterns and linkages within the dataset.

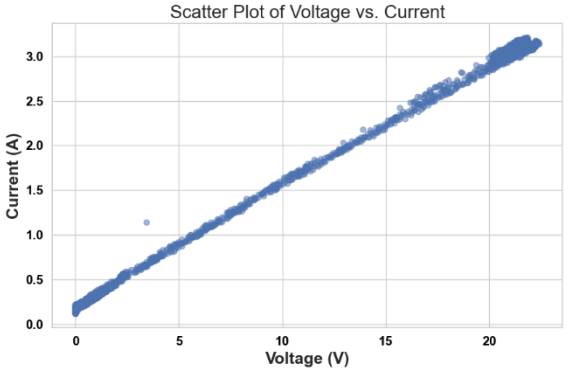

Fig. 4. Scatter plot of Voltage vs. Current.

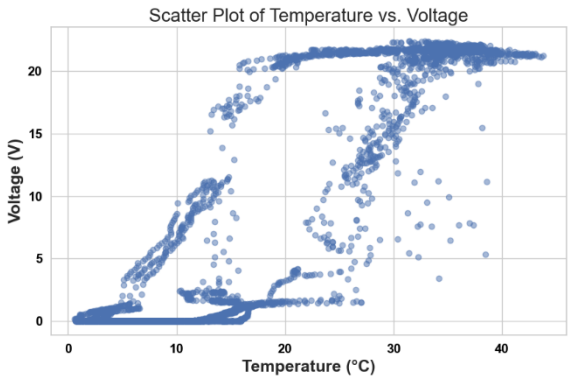

Fig. 5. Scatter plot of Temperature vs. Voltage.

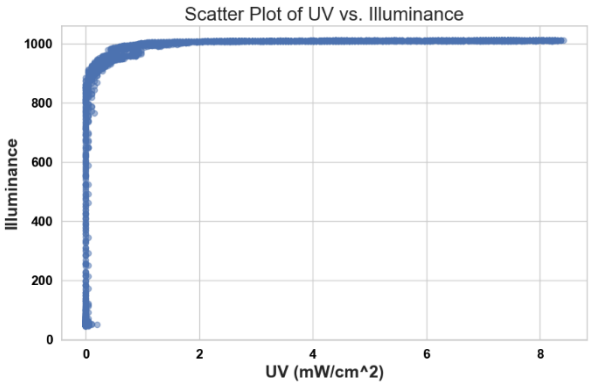

Fig. 6. Scatter plot of UV vs. Illuminance.

Fig. 4 is a scatter plot of voltage versus current, indicating a strong linear connection between these two electrical parameters. The diagram illustrates that when the voltage increases, so does the current, suggesting a straight proportionality compatible with Ohm's rule. This linear trend demonstrates the solar panel's predictable performance under different voltage circumstances. Fig. 5 depicts a more intricate relationship between temperature and voltage using a scatter plot. The graphic shows that voltage fluctuates dramatically with temperature variations, with many clusters of data points. This suggests that temperature has a nonlinear effect on the voltage output of the solar panel, most likely due to varied operating states and external conditions influencing the panel's performance. Fig. 6 depicts a scatter plot of UV vs. Illuminance, demonstrating that increased UV irradiance corresponds to higher illuminance values. The plot shows a quick rise in illuminance as UV irradiance increases, followed by a plateau at higher UV levels. This suggests that, whereas UV irradiance has a substantial impact on illuminance, other variables may play a role at higher UV levels.

2.2 ML Algorithms

The procedure starts with loading and prepping the information. This involves importing data from a CSV file into a pandas "DataFrame" and eliminating superfluous columns to concentrate on the pertinent characteristics and objectives. The characteristics, with the exception of 'Illuminance' and 'UV', are clearly specified, and the objectives are established as 'Illuminance' and 'UV'. In order to maintain the accuracy and consistency of the data, any rows that contain incorrect values such as infinity or NaN are eliminated. Additionally, the indices of the features and targets are adjusted to guarantee proper alignment once these rows are dropped. Subsequently, the dataset is partitioned into separate training and testing sets for both illuminance and UV goals, guaranteeing that the model may be assessed on data that it has not been exposed to previously. Several ML models, such as XGBoost, RF, LR, SVC, GB, and CatBoost, are initialized. Subsequently, each model undergoes training on the training set and assessment on the test set, considering both illuminance and UV goals. This evaluation process employs metrics like Root Mean Square Error (RMSE) and Coefficient of Determination (R^2). The outcomes from the several models are merged into a unified "DataFrame" and stored in a CSV file for subsequent study. The outcomes are represented graphically using several graphs, such as RMSE and R^2 for each model, along with a heatmap illustrating the correlation matrix. Ultimately, the outcomes are presented for examination. This systematic methodology utilizes several ML models to forecast important output variables by analyzing environmental and electrical input characteristics. This process yields significant information on the efficiency and effectiveness of solar panels. The flowchart depicted in Fig. 7 succinctly represents each stage of this process, guaranteeing lucidity and facilitating comprehension.

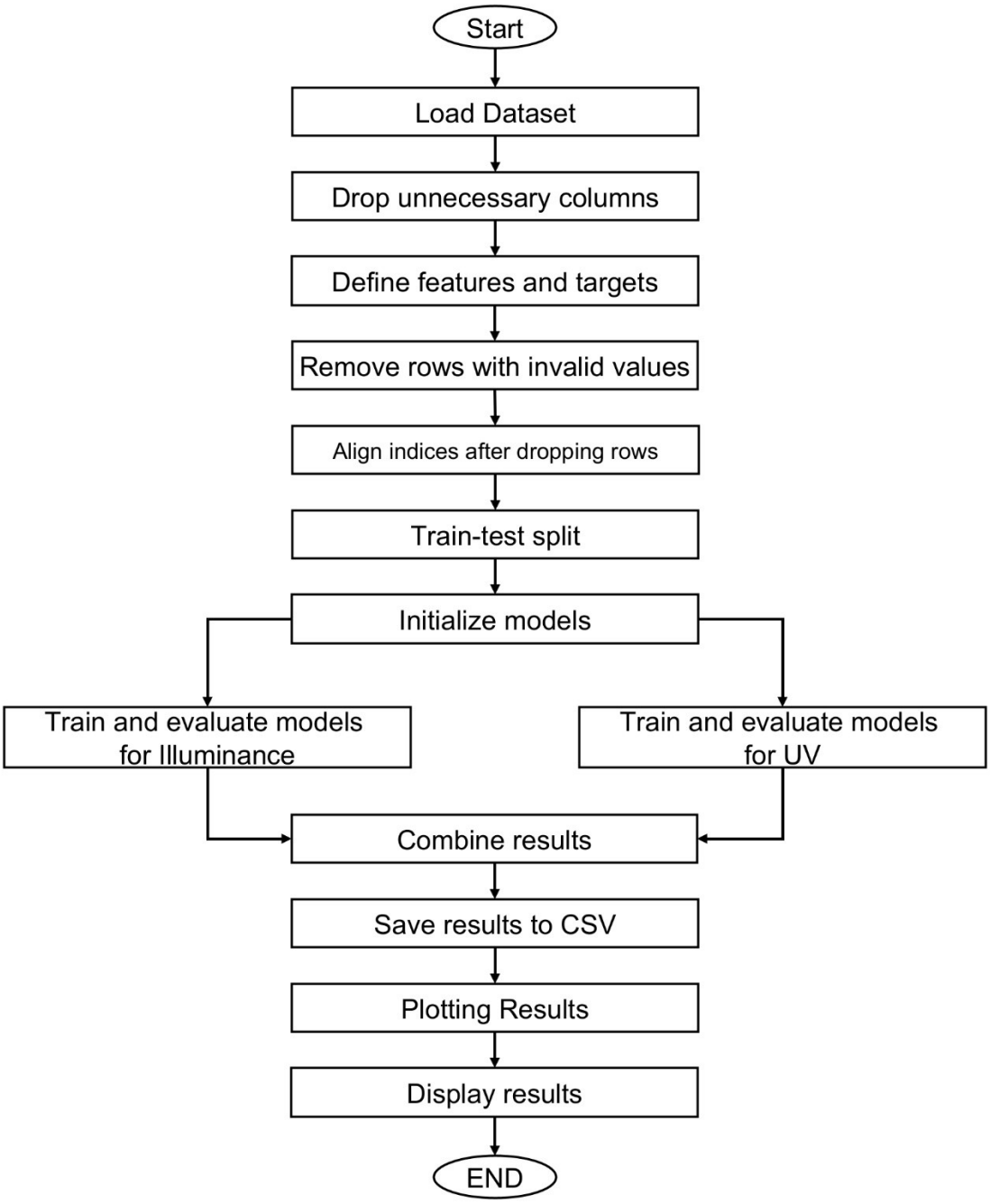

Fig. 7. Flowchart of the methodology for predicting illuminance and UV irradiance in PV systems.

Table 2 presents a comprehensive comparison of the ML techniques employed in the prediction of illuminance and UV irradiance in PV systems. The examined algorithms are SVC, LR, XGBoost, GB, RF, and CatBoost. Every algorithm possesses unique attributes, advantages, and limitations. For example, Support Vector Classification (SVC) is highly efficient in spaces with a large number of dimensions and may be used with various kernel functions. However, it necessitates meticulous parameter adjustment and is computationally demanding when dealing with extensive datasets. LR is a straightforward and efficient method, although it relies on the assumption of linearity and is highly responsive to outliers [27]. However, ensemble approaches such as XGBoost and RF exhibit superior accuracy and resilience to noise, but at the cost of increased complexity in terms of tuning and interpretation. Categorical Boosting (CatBoost) has exceptional performance in managing categorical data and training efficiency, while it may require significant memory resources. This comparison elucidates the appropriateness of each algorithm for distinct categories of data and prediction assignments, providing guidance for the selection process in particular applications of PV system performance monitoring.

Table 2. Comparative overview of ML algorithms.

Algorithm	Type	Key Characteristics	Strengths	Weaknesses
SVC [28, 29].	Supervised (Classification & Regression)	Uses support vectors; kernel trick for non-linear data	Effective in high-dimensional spaces; versatile	Requires parameter tuning; computationally expensive for large datasets
LR [30-32].	Supervised (Regression)	Models linear relationships; minimizes squared errors	Simple and interpretable; fast	Assumes linearity; sensitive to outliers
XGBoost [33].	Ensemble (Boosting)	Optimized GB; uses decision trees	High accuracy; handles large datasets	Complex tuning; may overfit on small datasets
GB [34, 35].	Ensemble (Boosting)	Sequential error correction; uses decision trees	High predictive accuracy; handles complex data	Prone to overfitting; computationally intensive
RF [36, 37].	Ensemble (Bagging)	Multiple decision trees; merges predictions	Reduces overfitting; robust to noise	Less interpretable; slow for large datasets
CatBoost [38].	Ensemble (Boosting)	Categorical feature support; efficient Graphics Processing Unit (GPU) usage	Excellent on categorical data; fast training	Memory intensive; fewer resources available

2.3 Evaluation Metrics

For this work, we employed two main assessment metrics to gauge the effectiveness of the ML models: RMSE and the R^2. These metrics offer valuable information on the precision and ability of the models to forecast illuminance and UV irradiance in solar systems. RMSE is a frequently employed metric for quantifying the average magnitude of the discrepancies between expected and actual data [39]. The square root of the mean squared difference between expected and actual values. RMSE is defined Equation (1). Where n is the number of observations, y_i is the actual value, and $\hat{y}_i$ is the predicted value. RMSE is sensitive to outliers and gives a higher weight to larger errors, with a lower RMSE indicating a better fit of the model to the data [39]. On the other hand, the R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides an indication of how well the model's predictions match the actual data [40]. R^2 is defined by Equation (2). Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the actual values. R^2 values vary from 0 to 1, with values closer to 1 suggesting a higher proportion of the variance is accounted for by the model, indicating a more accurate fit. The selection of these measures was based on their capacity to offer a thorough assessment of model performance, striking a balance between minimizing prediction errors and maximizing the models' explanatory capability. The investigation involved calculating and analyzing the RMSE and R^2 values of the models for predicting both illuminance and UV. This was done to identify the algorithms that performed the best.

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

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

3. Results and Discussion

This study assessed the efficacy of several ML models in forecasting illuminance and UV irradiance in solar systems. The evaluated models consist of XGBoost, RF, LR, SVC, GB, and CatBoost. The models are evaluated using the performance metrics of RMSE and the R^2. The findings are succinctly presented in Table 3 and visually shown in Fig. 8, Fig. 9, Fig. 10, and Fig. 11.

Table 3 presents a comprehensive analysis of the performance metrics of the models. Table 3 clearly demonstrates that the CatBoost and RF models outperform others in predicting both illuminance and UV since they have the lowest RMSE and greatest R^2 values. CatBoost algorithm obtained an RMSE of 16.088 and an R^2 of 0.999 for illuminance. Additionally, it achieved an RMSE of 0.228 and an R^2 of 0.990 for UV. The RF model has a close correlation with an RMSE of 16.568 and an R^2 value of 0.999 for illuminance. Similarly, it exhibits an RMSE of 0.233 and an R^2 value of 0.990 for UV. LR and SVC, on the other hand, had noticeably worse results. The LR model achieved an RMSE of 259.031 and an R^2 of 0.683 for illuminance. For UV, the model produced an RMSE of 0.943 and an R^2 of 0.833. The Support Vector Classifier (SVC) had the greatest RMSE and the lowest R^2 values compared to the other models. Specifically, the RMSE for illuminance was 318.890 with an R^2 of 0.520, whereas the RMSE for UV was 0.756 with an R^2 of 0.893.

These findings suggest that non-linear models such as CatBoost and RF are more appropriate for predicting intricate relationships within the dataset, as opposed to linear models. Fig. 8 and Fig. 9 display the R^2 and RMSE values for several models used to forecast illuminance. The statistics clearly demonstrate that CatBoost and RF models outperform other models by a substantial margin. The models with higher R^2 values and lower RMSE values demonstrate a stronger fit and reduced prediction error, respectively. Fig. 10 and Fig. 11 display the R^2 and RMSE values for several models used to forecast UV irradiation. Like the illuminance predictions, CatBoost and RF exhibit exceptional performance, emphasizing their resilience and precision in capturing the fundamental patterns in the data. The impressive performance of the CatBoost and RF models may be ascribed to their adeptness in successfully managing non-linear connections and interactions among features. These models also have the advantage of being able to effectively handle missing values and outliers, which are frequently encountered in real-world datasets. The findings indicate that sophisticated ML models, such as CatBoost and RF, have improved forecast accuracy for both illuminance and UV irradiance in solar systems. The findings indicate that integrating these models into solar performance monitoring systems may greatly improve the precision and dependability of forecasts, therefore enhancing the overall efficiency of the system and decision-making procedures. Additional studies might investigate the incorporation of these models with real-time data gathering and adaptive learning methods to consistently enhance the accuracy of predictions over a period of time.

Table 3. Reordered model performance from less efficient to more efficient based on RMSE and R^2 metrics for illuminance and UV predictions.

Model	Illuminance RMSE	Illuminance R^2	UV RMSE	UV R^2
SVC	318.890	0.520	0.756	0.893
LR	259.031	0.683	0.943	0.833
XGBoost	19.290	0.998	0.258	0.988
GB	16.594	0.999	0.254	0.988
RF	16.568	0.999	0.233	0.990
CatBoost	16.088	0.999	0.228	0.990

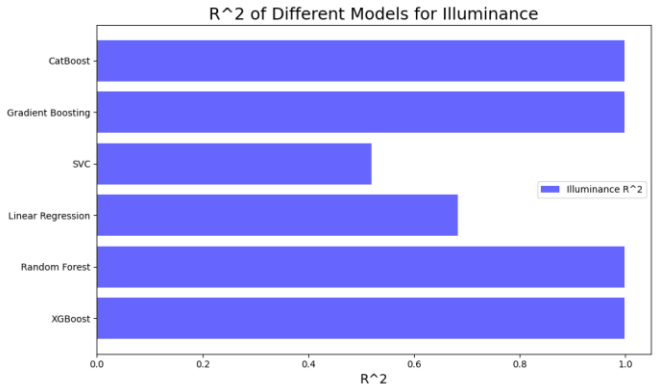

Fig. 8. R^2 of Different Models for Illuminance.

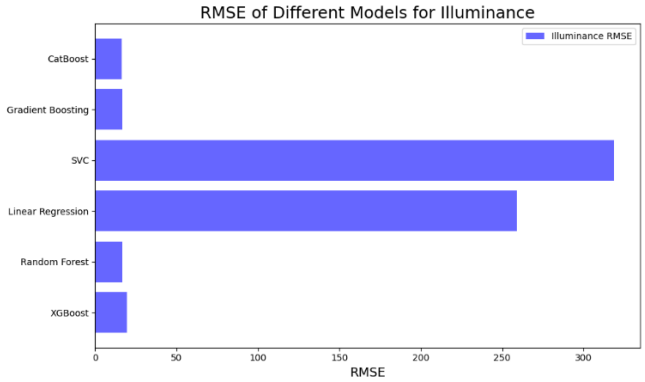

Fig. 9. RMSE of Different Models for Illuminance.

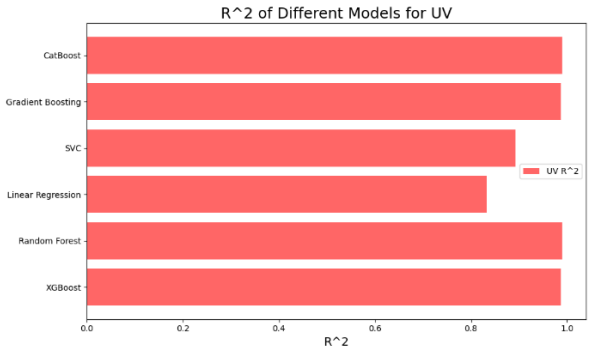

Fig. 10. R^2 of Different Models for UV.

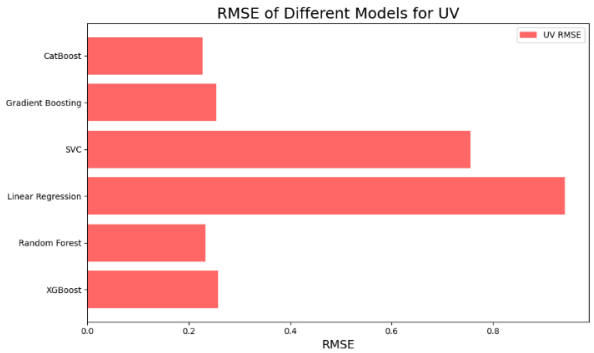

Fig. 11. RMSE of Different Models for UV.

4. Conclusion

This study has undertaken a comprehensive evaluation and comparison of various ML models for predicting the levels of illuminance and UV irradiation in solar systems. We have demonstrated the effectiveness of many models, including SVC, LR, XGBoost, GB, RF, and CatBoost, in accurately predicting important factors that affect the efficiency of PV panels. The results indicated that the CatBoost and RF models outperformed the other models. They obtained the lowest RMSE and the highest R^2 values for predicting both illuminance and UV. The results demonstrate the effectiveness of sophisticated ML techniques in enhancing the precision of forecasts and the productivity of PV systems. The preparation procedures, which involved data cleansing and alignment, together with a reliable division of data into training and testing sets, guaranteed that the models were trained and assessed using data of superior quality. The meticulous methodology employed enhanced the dependability of the findings and underscored the significance of comprehensive data preprocessing in ML endeavors. The study's findings are beneficial for enhancing system efficiency and facilitating decision-making processes for solar performance monitoring. By making precise forecasts of illuminance and UV irradiance, individuals involved may optimize the timing of maintenance tasks, improve techniques for managing energy, and eventually enhance the overall efficiency of PV installations. Moreover, this study adds to the expanding corpus of work on the utilization of ML in renewable energy, establishing a strong basis for future progress in this domain. With the increasing need for sustainable energy solutions, the incorporation of advanced predictive models will be essential in optimizing the effectiveness of renewable energy sources. To summarize, the results of this study confirm that CatBoost and RF models are effective in forecasting important parameters in solar systems. This provides a technique to enhance the efficiency and dependability of renewable energy solutions.

5. Future work

While this work revealed the efficacy of multiple ML models in forecasting illuminance and UV irradiance in PV systems, significant areas for future research remain. First, further research might look into integrating real-time data acquisition systems with predictive models to improve prediction timeliness and accuracy. Incorporating streaming data would enable continuous monitoring and fast modifications, increasing the operational efficiency of PV systems. Furthermore, future research could look into the use of more advanced ML techniques, such as Deep Learning (DL), which may perform better in capturing complex, non-linear relationships in data. It is worthwhile to investigate how Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) might improve prediction accuracy. Examining the effects of various geographic locations and environmental variables on the model's performance is another important subject for future research. Expanding the dataset to encompass additional climatic zones would offer a more thorough assessment of the models' resilience and suitability. To further improve the model's performance, feature engineering techniques should be investigated to produce new and more insightful features. By using existing characteristics or including more environmental information, such as wind speed or solar angle, one can gain a more profound understanding and enhance predictive abilities. The creation of hybrid models that combine the strengths of various algorithms could be pursued. Ensemble methods, for example, that combine DL and standard ML models may produce higher accuracy and resilience results. Finally, future research should focus on using real-time data, advanced ML techniques, and various datasets to improve the predictive capacities of PV systems models. These activities will help to improve the efficiency and reliability of renewable energy solutions, thereby facilitating the ongoing shift to sustainable energy sources.

Abbreviations

ANFIS	Adaptive Neuro-Fuzzy Inference System
CatBoost	Categorical Boosting
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
DL	Deep Learning
GB	Gradient Boosting
GPU	Graphics Processing Unit
IQR	Interquartile Range
LR	Linear Regression
ML	Machine Learning
PV	Photovoltaic
RF	Random Forest
R^2	Coefficient of Determination
RNN	Recurrent Neural Network
RMSE	Root Mean Squared Error
SVR	Support Vector Regression
SVC	Support Vector Classification
UV	Ultraviolet
XGBoost	eXtreme Gradient Boosting

References

[1] Abdelsattar, M., AbdelMoety, A., & Emad-Eldeen, A. (2024). Comparative Analysis of Machine Learning Techniques for Fault Detection in Solar Panel Systems. SVU-

International Journal of Engineering Sciences and Applications, 5(2), 140-152. doi: https://doi.org/10.21608/svusrc.2024.279389.1198

[2] Abdelsattar, M., AbdelMoety, A., & Emad-Eldeen, A. (2023). A review on detection of solar PV panels failures using image processing techniques. In 2023 24th International Middle East Power System Conference (MEPCON) (pp. 1-6). Mansoura, Egypt. https://doi.org/10.1109/MEPCON58725.2023.10462371

[3] Jiang, S., Wan, C., Chen, C., Cao, E., & Song, Y. (2018). Distributed photovoltaic generation in the electricity market: status, mode and strategy. CSEE Journal of Power and Energy Systems. https://doi.org/10.17775/CSEEJPES.2018.00600.

[4] Sygletou, M., Petridis, C., Kymakis, E., & Stratakis, E. (2017). Advanced Photonic Processes for Photovoltaic and Energy Storage Systems. Advanced Materials, 29. https://doi.org/10.1002/adma.201700335.

[5] Wu, Y., Lin, J., & Lin, H. (2016). Standards and Guidelines for Grid-Connected Photovoltaic Generation Systems: A Review and Comparison. IEEE Transactions on Industry Applications, 53, 3205-3216. https://doi.org/10.1109/IAS.2016.7731810.

[6] Souza, G., Santos, R., & Saraiva, E. (2022). A Log-Logistic Predictor for Power Generation in Photovoltaic Systems. Energies. https://doi.org/10.3390/en15165973.

[7] Ma, T., Guo, Z., Shen, L., Liu, X., Chen, Z., Zhou, Y., & Zhang, X. (2021). Performance modelling of photovoltaic modules under actual operating conditions considering loss mechanism and energy distribution. Applied Energy, 298, 117205. https://doi.org/10.1016/J.APENERGY.2021.117205.

[8] Mustafa, R., Gomaa, M., Al-Dhaifallah, M., & Rezk, H. (2020). Environmental Impacts on the Performance of Solar Photovoltaic Systems. Sustainability. https://doi.org/10.3390/su12020608.

[9] Ramli, M., Prasetyono, E., Wicaksana, R., Windarko, N., Sedraoui, K., & Al-Turki, Y. (2016). On the investigation of photovoltaic output power reduction due to dust accumulation and weather conditions. Renewable Energy, 99, 836-844. https://doi.org/10.1016/J.RENENE.2016.07.063.

[10] Spataru, S., Sera, D., Kerekes, T., & Teodorescu, R. (2013). Photovoltaic array condition monitoring based on online regression of performance model. 2013 IEEE 39th Photovoltaic Specialists Conference (PVSC), 0815-0820. https://doi.org/10.1109/PVSC.2013.6744271.

[11] Spiliotis, E., Legaki, N., Assimakopoulos, V., Doukas, H., & Moursi, M. (2018). Tracking the performance of photovoltaic systems: a tool for minimising the risk of malfunctions and deterioration. Iet Renewable Power Generation, 12, 815-822. https://doi.org/10.1049/IET-RPG.2017.0596.

[12] Mofidul, R., Alam, S., Chakma, A., Chung, B., & Jang, Y. (2023). Predictive Maintenance in Photovoltaic Systems Using Ensemble ML Empirical Analysis. 2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN), 636-638. https://doi.org/10.1109/ICUFN57995.2023.10199326.

[13] Livera, A., Paphitis, G., Theristis, M., Lopez-Lorente, J., Makrides, G., & Georghiou, G. (2022). Photovoltaic System Health-State Architecture for Data-Driven Failure Detection. Solar. https://doi.org/10.3390/solar2010006.

[14] Hopwood, M., & Gunda, T. (2022). Generation of Data-Driven Expected Energy Models for Photovoltaic Systems. Applied Sciences. https://doi.org/10.3390/app12041872.

[15] Goudelis, G., Lazaridis, P., & Dhimish, M. (2022). A Review of Models for Photovoltaic Crack and Hotspot Prediction. Energies. https://doi.org/10.3390/en15124303.

[16] Achouri, F., Damou, M., Harrou, F., Sun, Y., & Bouyeddou, B. (2023). Gaussian Processes for Efficient Photovoltaic Power Prediction. 2023 International Conference on Decision Aid Sciences and Applications (DASA), 290-295. https://doi.org/10.1109/DASA59624.2023.10286780.

[17] Lara-Cerecedo, L., Hinojosa, J., Pitalua-Diaz, N., Matsumoto, Y., & González-Ángeles, Á. (2023). Prediction of the Electricity Generation of a 60-kW Photovoltaic System with Intelligent Models ANFIS and Optimized ANFIS-PSO. Energies. https://doi.org/10.3390/en16166050.

[18] Das, U., Tey, K., Idris, M., Mekhilef, S., Seyedmahmoudian, M., Stojcevski, A., & Horan, B. (2022). Optimized Support Vector Regression-Based Model for Solar Power Generation Forecasting on the Basis of Online Weather Reports. IEEE Access, PP, 1-1. https://doi.org/10.1109/ACCESS.2022.3148821.

[19] Abdellatif, A., Mubarak, H., Ahmad, S., Ahmed, T., Shafiullah, G., Hammoudeh, A., Abdellatef, H., Rahman, M., & Gheni, H. (2022). Forecasting Photovoltaic Power Generation with a Stacking Ensemble Model. Sustainability. https://doi.org/10.3390/su141711083.

[20] Demir, A., Gutiérrez, L., Namin, A., & Bayne, S. (2022). Solar Irradiance Prediction Using Transformer-based Machine Learning Models. 2022 IEEE International Conference on Big Data (Big Data), 2833-2840. https://doi.org/10.1109/BigData55660.2022.10020615.

[21] Li, Q., & He, Q. (2022). Hourly solar irradiance prediction based on enhanced incremental extreme learning machine. , 12255, 1225519 - 1225519-8. https://doi.org/10.1117/12.2639376.

[22] Alzahrani, A. (2022). Short-Term Solar Irradiance Prediction Based on Adaptive Extreme Learning Machine and Weather Data. Sensors (Basel, Switzerland), 22. https://doi.org/10.3390/s22218218.

[23] Viscondi, G., & Alves-Souza, S. (2021). Solar Irradiance Prediction with Machine Learning Algorithms: A Brazilian Case Study on Photovoltaic Electricity Generation. Energies. https://doi.org/10.3390/en14185657.

[24] Aliberti, A., Fucini, D., Bottaccioli, L., Macii, E., Acquaviva, A., & Patti, E. (2021). Comparative Analysis of Neural Networks Techniques to Forecast Global Horizontal Irradiance. IEEE Access, 9, 122829-122846. https://doi.org/10.1109/ACCESS.2021.3110167.

[25] Huang, X., Li, Q., Tai, Y., Zaiqing, C., Zhang, J., Shi, J., Gao, B., & Liu, W. (2021). Hybrid deep neural model for hourly solar irradiance forecasting. Renewable Energy, 171, 1041-1060. https://doi.org/10.1016/J.RENENE.2021.02.161.

[26] Maitanova, N., Telle, J., Hanke, B., Grottke, M., Schmidt, T., Maydell, K., & Agert, C. (2020). A Machine Learning Approach to Low-Cost Photovoltaic Power Prediction Based on Publicly Available Weather Reports. Energies. https://doi.org/10.3390/en13030735.

[27] Peña, D. (2023). Detecting Outliers and Influential and Sensitive Observations in Linear Regression. In: Pham, H. (eds) Springer Handbook of Engineering Statistics. Springer Handbooks. Springer, London. https://doi.org/10.1007/978-1-4471-7503-2_31

[28] Jun, Z. (2021). The Development and Application of Support Vector Machine. Journal of Physics: Conference Series, 1748. https://doi.org/10.1088/1742-6596/1748/5/052006.

[29] Shakibian, H., & Nasiri, J. (2022). Probabilistic Twin Support Vector Machine for Solving Unclassifiable Region Problem. International Journal of Engineering. https://doi.org/10.5829/IJE.2022.35.01A.01.

[30] Geche, F., Mulesa, O., Hrynenko, V., & Smolanka, V. (2019). Search for impact factor characteristics in construction of linear regression models. Technology audit and production reserves. https://doi.org/10.15587/2312-8372.2019.175020.

[31] Lim, H. (2019). A Linear Regression Approach to Modeling Software Characteristics for Classifying Similar Software. 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 1, 942-943. https://doi.org/10.1109/COMPSAC.2019.00152.

[32] Luis, F., Zulima, F., & Denys, B. (2018). The uncertainty analysis in linear and nonlinear regression revisited: application to concrete strength estimation. Inverse Problems in Science and Engineering, 27, 1740 - 1764. https://doi.org/10.1080/17415977.2018.1553969.

[33] Du, M., Yu, Z., Wang, T., Wang, X., & Jiang, X. (2020). XGBoost Based Strategic Consumers Classification Model on E-commerce Platform. Proceedings of the 2020 The 6th International Conference on E-Business and Applications. https://doi.org/10.1145/3387263.3387284.

[34] Denuit, M., Hainaut, D., & Trufin, J. (2019). Gradient Boosting with Neural Networks. Springer Actuarial. https://doi.org/10.1007/978-3-030-25827-6_7.

[35] Lu, H., & Mazumder, R. (2018). Randomized Gradient Boosting Machine. SIAM J. Optim., 30, 2780-2808. https://doi.org/10.1137/18m1223277.

[36] Schonlau, M., & Zou, R. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20, 29 - 3. https://doi.org/10.1177/1536867X20909688.

[37] Olaniran, O., & Abdullah, M. (2019). BayesRandomForest: An R Implementation of Bayesian Random Forest for Regression Analysis of High-Dimensional Data. Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017). https://doi.org/10.1007/978-981-13-7279-7_33.

[38] Hancock, J., & Khoshgoftaar, T. (2020). CatBoost for big data: an interdisciplinary review. Journal of Big Data. https://doi.org/10.1186/s40537-020-00369-8.

[39] Hodson, T. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development. https://doi.org/10.5194/gmd-15-5481-2022.

[40] Chicco, D., Warrens, M., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7. https://doi.org/10.7717/peerj-cs.623.