

Enhancing Photovoltaic System Cleaning Using Machine Learning Algorithms

Haneen Abuzaid¹, Mahmoud Awad², Abdulrahim Shamayleh³

^{1, 2, 3} Department of industrial engineering, American University of Sharjah, Sharjah, UAE

Abstract:- The study focuses on the importance of maintaining photovoltaic (PV) systems for optimal performance in sustainable energy generation. It highlights the impact of dust accumulation on reducing system efficiency and proposes a method to predict system performance, aiding in scheduling cleaning activities effectively. Two prediction models are developed: one using time-series prediction techniques (LSTM, ARIMA, SARIMAX) to forecast Performance Ratio (PR), and another employing ensemble voting classifiers (RF, Log, GBM) to predict the need for cleaning. The SARIMAX model performs best, achieving high accuracy in PR prediction ($R^2 = 92.12\%$), while the classification model accurately predicts cleaning needs (91%). The research provides valuable insights for improving maintenance strategies and enhancing the efficiency and sustainability of PV systems.

Keywords: Time-series prediction; PV cleaning; Performance ratio; PV systems; Machine learning.

1. Introduction

Photovoltaic (PV) systems have witnessed substantial global growth due to their economic, environmental, and social sustainability benefits [1]. Factors such as cost reduction, abundant solar energy, efficiency upgrades, regulatory support, and ease of installation have contributed to their widespread adoption [2]. The performance of PV systems is influenced by various meteorological, design, and operational factors, including solar irradiance, temperature, wind speed, shading, and dust accumulation [2], [3],[4]. Dust accumulation can significantly reduce system performance, leading to losses of up to 50% [5], [6], emphasizing the importance of regular cleaning to maintain efficiency and enhance performance. Cleaning PV systems is crucial for optimal energy generation, economic viability, and mitigating the negative effects of dust accumulation [7].

Hence, maintaining cleanliness in PV systems during operation is crucial due to its significant impact on performance, especially considering their long lifespan of 25-30 years [8]. Various cleaning techniques are available, tailored to factors such as system size, type, installation, design, geographical location, reliability, and resource availability [9]. These techniques range from natural cleaning using rain, wind, and gravity [10], to manual cleaning requiring labor [11], automated cleaning controlled remotely or by schedule [12], semi-automated cleaning with labor oversight [13], to self-cleaning techniques integrated during manufacturing to prevent dust accumulation [14].

Accurately predicting the performance of PV systems is crucial for evaluating efficiency and scheduling cleaning processes effectively, aligning with economic viability thresholds. The Performance Ratio (PR) serves as a standardized metric widely used in performance-guaranteed contracts and regulatory frameworks [15]. Despite extensive research, no study has comprehensively considered all influential factors affecting PV system performance while using PR as an output. This research aims to fill this gap by employing advanced machine learning algorithms like LSTM, ARIMA, and SARIMAX, utilizing a large time-series dataset from various PV systems. Additionally, a threshold-based classification model will be developed using an ensemble voting classifier to assist PV system operators, particularly non-experts, in optimizing performance and maintenance strategies. Thus, this research aims to provide valuable insights into enhancing PV system efficiency.

Accordingly, this study aims to fill research gaps by incorporating all influential factors on PV system performance to predict the Performance Ratio (PR) and provide cleaning recommendations based on predefined thresholds. Using a large dataset encompassing meteorological and operational factors, time-series machine learning algorithms such as LSTM, ARIMA, and SARIMAX will be employed to construct PR prediction models. Evaluation will include metrics like MSE, RMSE, MAPE, SMAPE, and R^2 . Additionally, a threshold-based predictive model will be developed using ensemble voting classifiers like Log, RF, and GBM to predict the cleaning process based on contractual thresholds. Model performance will be assessed using accuracy, recall, precision, and F1-score, aiming to identify the optimal model for PV system operators, especially non-experts.

2. Methodology

The methodology utilized in this study commences with a comprehensive literature review. This review covers case studies focusing on predicting PV system performance and scheduling PV cleaning processes. Following this, multiple discussions were conducted with PV system experts, aiming to validate extracted influential factors and obtain feedback on the importance of study outcomes.

The subsequent step requires developing a prediction model that is capable of handling time-series datasets with high correlations, intending to forecast PV system performance. Among various performance metrics for PV systems, the Performance Ratio (PR) is prioritized, given its widespread use in contractual agreements and regulatory contracts [16].

Prediction models selected for PR include ARIMA, SARIMAX, and LSTM, with their proven capability of handling complex time-series datasets. The main steps in utilizing these models include collecting data, preprocessing, training the model, testing, and evaluating the prediction performance of each model. Additionally, assessing the features' importance is significant for improving the prediction, thereby, enhancing the performance.

Furthermore, a rule-based classification model utilizing ensemble techniques was developed using the same dataset to predict the cleaning process based on a PR threshold. This ensemble model combines three classification models: Log regression, RF, and GBM, with predictions combined through voting to attain optimal classification performance.

3. Results

A case study (CS) was employed from a PV project in the UAE. The PV project is connected to the grid. A summary of the project is provided in Table 1. The project undergoes regular cleaning cycles as per the contractor's pre-established schedule to ensure that the overall Performance Ratio (PR) remains above a predefined threshold, as is customary in performance-guaranteed contracts for PV projects.

However, this threshold is not fixed and varies depending on factors such as installation settings, system complexity, additional equipment like storage systems, system availability, and shading levels at the location. Contractors set the threshold upon contract award, which is subsequently validated by a third-party consultant before signing the performance-guaranteed agreement.

Table 1. Case study summary information.

		CS
Location		UAE
Size		648kWp
Type of installation		Rooftop – corrugated sheet
Tilt angle		5°
PV modules	Type	Mono crystalline
	Power	470Wp
	Quantity	1379
	Efficiency	20.93%

The features for predicting the PR include years, season, month, day, atmospheric pressure, irradiance, relative humidity, ambient temperature, modules temperature, wind speed, cloud coverage, cleaning process, and time since last cleaning. The dataset encompasses the period from November 2021 to December 2023.

Meteorological factors were gathered from weather sensors and the Supervisory Control and Data Acquisition (SCADA) system onsite. Performance Ratio (PR) data for each project was obtained from the data logger, while information regarding the timing of cleaning processes was extracted from archived documents maintained by the Operation and Maintenance (O&M) team for each project.

Missing values were addressed in two ways: those occurring during periods of no generation, such as from sunset to sunrise, were eliminated as the focus is solely on system operation. For missing values in the web-based monitoring system due to internet disruptions, data was retrieved from onsite data loggers. Subsequently, numerical features underwent normalization to ensure uniform scale, thereby enhancing convergence stability, interpretability, and model performance. Data preprocessing and modeling were conducted using RStudio software.

Moreover, feature encoding was carried out for years, seasons, months, days, cleaning process, and time since the last cleaning. Different encoding techniques were utilized based on the nature of each feature. For instance, binary encoding was employed for cleaning processes, while cyclical encoding captured recurring trends and patterns in time-series datasets for days and months.

In the final preprocessing step, the dataset was divided into training (75%) and testing (25%) sets, utilizing a 10-fold cross-validation technique. This crucial step allows for model evaluation, prevents overfitting, tunes hyperparameters, and ensures the machine learning model generalizes effectively to new datasets, thereby enhancing prediction performance reliability.

Subsequently, prediction models were trained individually for each PV system's Performance Ratio (PR) using ARIMA, SARIMAX, and LSTM algorithms. Each model possesses distinct parameters contributing to its unique predictive abilities. To optimize model performance, parameters were fine-tuned using the grid search method within the RStudio package. This involved evaluating all combinations of hyperparameters across a predefined grid, with 5-fold cross-validation ensuring robustness and reliability in performance assessment.

Table 2 displays the evaluation metrics, such as MSE, RMSE, MAPE, SMAPE, and R^2 , for each PV project's testing dataset. The results indicate that the SARIMAX model consistently surpassed other models in predicting the Performance Ratio (PR). Figure 1 depicts the feature importance for the PV project. Notably, the most significant features in the prediction model comprise module temperature, irradiance, ambient temperature, season, and relative humidity.

Table 2. Evaluation metrics of prediction models using testing dataset.

Measure	ARIMA	SARIMAX	LSTM
MSE	0.0723	0.0312	0.0715
RMSE	0.2689	0.1766	0.2674
MAPE	0.1812	0.1421	0.1765
SMAPE	0.1672	0.1345	0.1623
R^2	76.26	92.12	88.74

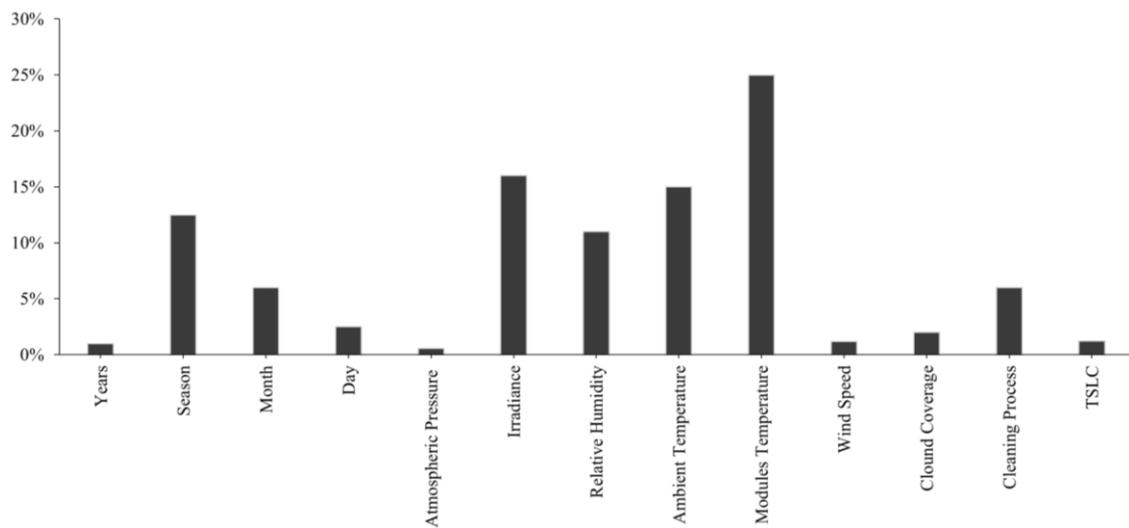


Figure 1. Features importance of SARIMAX model

An ensemble technique combining Logistic regression, RF, and GBM models predicted the cleaning process based on the contractually guaranteed thresholds (75%). The voting classifier showed superior performance, demonstrating high accuracy, recall, precision, and F1-score. Among individual classifiers, the gradient boost machines model performed the best, as shown in Table 3.

Table 3. Evaluation of ensemble technique classifier for PV cleaning process.

Performance measure	RF	GBM	Log	Voting classifier ensemble
Accuracy	0.86	0.90	0.84	0.91
Precision	0.83	0.86	0.82	0.88
Recall	0.81	0.83	0.83	0.85
F1-score	0.80	0.85	0.83	0.87

4. Discussion

This study constructs time-series predictive models to forecast PV system performance, focusing on the Performance Ratio (PR). ARIMA, SARIMAX, and LSTM models are employed, with SARIMAX

outperforming other models due to its ability to handle seasonal patterns and multicollinearity issues [17], [18], [19], [20]. Notably, cleaning-related features show low importance, suggesting the need for accurate scheduling to optimize PV system performance.

Predicting PR forms the basis for scheduling cleaning cycles, with thresholds guiding the process. An ensemble technique, including RF, GBM, and Log regression, predicts the cleaning process, with GBM demonstrating superior performance. While slightly lower than PR-based models, the ensemble classifier offers robust performance across both projects, highlighting its effectiveness for non-experts in optimizing PV system maintenance. Overall, this study provides a comprehensive approach for PV system optimization, offering valuable insights for stakeholders.

5. Conclusion

This study offers a methodological framework aimed at optimizing the efficiency of PV system cleaning processes and enhancing overall system performance and viability. The approach involves predicting the Performance Ratio (PR) of PV systems and subsequently forecasting cleaning cycles to meet contractual performance guarantees, while accounting for various influential factors including temporal, meteorological, and operational elements pertinent to existing PV systems. Time-related factors encompass year, month, day, and season, reflecting

degradation and seasonal performance trends. Meteorological factors include solar irradiance, ambient temperature, atmospheric pressure, relative humidity, wind speed, and cloud coverage, while operational factors include PV module temperature and the cleaning process.

To validate this methodology, a large dataset was employed. Results indicate that SARIMAX outperforms other prediction models, achieving R^2 values of 92.12%. Moreover, the classification performance for the cleaning process reached 91%, utilizing a threshold-based ensemble classification approach. Notably, predictive models for PR demonstrated significantly higher performance compared to classification models for predicting the cleaning process.

This study provides a tool for scheduling cleaning cycles based on predicted PR, aiding non-experts in effectively managing the PV cleaning process. For future research, additional parameters related to PV system design, such as tilt angle, tracking systems, and PV module types, along with location-specific weather conditions like precipitation, sandstorms, and dust characteristics, could be incorporated to further refine the prediction model. Additionally, conducting comparative analyses across different locations with similar PV system attributes would yield valuable insights into the impact of varying meteorological factors on similar PV systems.

Competing interest

The authors declare that they have no conflicts of interest related to this research.

References

- [1] H. Iftikhar, E. Sarquis, and P. C. Branco, "Why can simple operation and maintenance (O&M) practices in large-scale grid-connected PV power plants play a key role in improving its energy output?," *Energies*, vol. 14, no. 13, p. 3798, 2021.
- [2] A. Saseendran, C. Hartl, Y. Tian, and Y. Qin, "Development, Optimization, and Testing of a Hybrid Solar Panel Concept With Energy Harvesting Enhancement," *Journal of Physics Conference Series*, 2023, doi: 10.1088/1742-6596/2526/1/012033.
- [3] I. Al Siyabi, A. Al Mayasi, A. Al Shukaili, and S. Khanna, "Effect of Soiling on Solar Photovoltaic Performance under Desert Climatic Conditions," *Energies*, vol. 14, no. 3, p. 659, 2021.
- [4] H. Abuzaid, M. Awad, and A. Shamayleh, "Impact of dust accumulation on photovoltaic panels: a review paper," *International Journal of Sustainable Engineering*, vol. 15, no. 1, pp. 264-285, 2022.
- [5] A. A. Hachicha, I. Al-Sawafta, and D. Ben Hamadou, "Numerical and experimental investigations of dust effect on CSP performance under United Arab Emirates weather conditions," *Renewable Energy*, vol. 143, pp. 263-276, 2019, doi: 10.1016/j.renene.2019.04.144.
- [6] M. J. Adinoyi and S. A. M. Said, "Effect of dust accumulation on the power outputs of solar photovoltaic modules," *Renewable Energy*, vol. 60, pp. 633-636, 2013, doi: 10.1016/j.renene.2013.06.014.
- [7] L. Hernández-Callejo, S. Gallardo-Saavedra, and V. Alonso-Gómez, "A review of photovoltaic systems: Design, operation and maintenance," *Solar Energy*, vol. 188, pp. 426-440, 2019, doi: <https://doi.org/10.1016/j.solener.2019.06.017>.
- [8] A. Almufarrej and T. Erfani, "Modelling the Regional Effect of Transmittance Loss on Photovoltaic Systems Due to Dust," *International Journal of Energy and Environmental Engineering*, 2022, doi: 10.1007/s40095-022-00510-8.
- [9] H. M. Khalid *et al.*, "Dust accumulation and aggregation on PV panels: An integrated survey on impacts, mathematical models, cleaning mechanisms, and possible sustainable solution," *Solar Energy*, vol. 251, pp. 261-285, 2023, doi: <https://doi.org/10.1016/j.solener.2023.01.010>.
- [10] F. H. B. M. Noh *et al.*, "Development of Solar Panel Cleaning Robot Using Arduino," *Indonesian Journal of Electrical Engineering and Computer Science*, 2020, doi: 10.11591/ijeecs.v19.i3.pp1245-1250.
- [11] V. Gupta, M. Sharma, R. K. Pachauri, and K. D. Babu, "Comprehensive review on effect of dust on solar photovoltaic system and mitigation techniques," *Solar Energy*, vol. 191, pp. 596-622, 2019.
- [12] Y. N. Chanchangi, A. Ghosh, S. Sundaram, and T. K. Mallick, "Dust and PV Performance in Nigeria: A review," *Renewable and Sustainable Energy Reviews*, vol. 121, p. 109704, 2020.

- [13] T. Salamah *et al.*, "Effect of dust and methods of cleaning on the performance of solar PV module for different climate regions: Comprehensive review," *Science of The Total Environment*, vol. 827, p. 154050, 2022, doi: <https://doi.org/10.1016/j.scitotenv.2022.154050>.
- [14] Y. Wu *et al.*, "A review of self-cleaning technology to reduce dust and ice accumulation in photovoltaic power generation using superhydrophobic coating," *Renewable Energy*, vol. 185, pp. 1034-1061, 2022, doi: <https://doi.org/10.1016/j.renene.2021.12.123>.
- [15] A. Sundararajan and A. I. Sarwat, "Hybrid data-model method to improve generation estimation and performance assessment of grid-tied PV: a case study," *IET Renewable Power Generation*, vol. 13, no. 13, pp. 2480-2490, 2019.
- [16] S. Lindig, M. Theristis, and D. Moser, "Best practices for photovoltaic performance loss rate calculations," *Progress in Energy*, vol. 4, no. 2, p. 022003, 2022.
- [17] M. Lehna, F. Scheller, and H. Herwartz, "Forecasting day-ahead electricity prices: A comparison of time series and neural network models taking external regressors into account," *Energy Economics*, vol. 106, p. 105742, 2022.
- [18] C. S. Yarrington, *Review of forecasting univariate time-series data with application to water-energy nexus studies & proposal of parallel hybrid SARIMA-ANN model*. West Virginia University, 2021.
- [19] A. A. Karim, E. Pardede, and S. Mann, "A Model Selection Approach for Time Series Forecasting: Incorporating Google Trends Data in Australian Macro Indicators," *Entropy*, vol. 25, no. 8, p. 1144, 2023.
- [20] S. I. Doguwa and S. O. Alade, "On time series modeling of Nigeria's external reserves," *CBN Journal of Applied Statistics*, vol. 6, no. 1, pp. 1-28, 2015.