

Predictive Modeling and Forecasting of Solar Power Generation Using Machine Learning Techniques

Prashis Raghuwanshi*

Senior Software Engineer and Researcher, (Associate Vice President), Dallas, Texas, USA

Email: prashish14@gmail.com

Abstract - The purpose of this research study is to investigate the predictive modelling and forecasting of solar power generation by utilising a variety of machine learning approaches. The paper addresses the significant obstacle that is presented by the variable production of solar energy, which is a barrier to the effective incorporation of solar power into the network of electrical power distribution systems. This research investigates the application of machine learning models such as Bayesian Ridge Regression, Gradient Boosting, and Linear Regression. The research makes use of historical meteorological data and solar power output to investigate the application of these models. The effectiveness of these models in estimating the amount of solar energy that will be produced under a variety of different weather situations is being investigated. Based on the findings, it is evident that the application of machine learning techniques has the potential to considerably improve the accuracy of solar power projections. This, in turn, will facilitate better grid integration and promote wider adoption of solar energy products. Through the demonstration of the potential of advanced machine learning approaches in enhancing the reliability and efficiency of solar power generation, this study makes a contribution to the expanding body of literature on the forecasting of renewable energy sources.

Keywords: Predictive Modeling, Solar Power Generation, Machine Learning Techniques, Renewable Energy, Forecasting.

-----X-----

INTRODUCTION

Within the context of the global transition towards sustainable energy sources, solar energy has emerged as a crucial component, delivering enormous advantages to both the environment and the economy. Due to the fact that it is renewable and has the ability to reduce emissions of greenhouse gases, solar power stands out as an alternative to fossil fuels, which are becoming increasingly sought after around the world. The broad use of solar power, on the other hand, is hampered by considerable hurdles, despite the fact that it has a great deal of promise. One of the most significant of these is the substantial initial expenditure that is necessary for the installation of photovoltaic (PV) systems. The considerable initial investment continues to discourage widespread adoption of solar panels, despite the fact that the cost of solar panels has been reducing as a result of government incentives and technological developments (Gupta et al., 2017; Alaraj et al., 2021). In addition, the return on investment continues to be a primary concern, particularly because of the inherent fluctuation in solar

energy generation that is induced by the varying weather conditions (Shrestha et al., 2019; Zhang et al., 2019).

There are a number of elements that can have an effect on photovoltaic cells, which are responsible for converting solar energy into electrical power. These factors include the geographical location, the time of day, and the weather conditions that are prevalent (Mohan et al., 2021). Solar irradiance, which is defined as the amount of power received from the Sun per unit area, plays a significant role in determining the efficiency of these cells and the energy output that arises from them (Bhowmik et al., 2020). Not only does the quantity of solar irradiance change throughout the day, but it also shifts throughout the seasons, which results in considerable variations in the amount of energy that may be created (Shrestha et al., 2019). When it comes to integrating solar power into the electrical grid, which is not yet completely equipped to deal with the unpredictable and uncontrollable nature of renewable energy sources, this fluctuation presents

a significant obstacle that must be overcome (Zhang et al., 2019).

In light of these obstacles, it is absolutely necessary to have an accurate prediction of solar power generation in order to maximize the utilization of solar energy and improve its integration into the grid (Alaraj et al., 2021). When grid operators have access to accurate predictions, they are able to better manage the supply and demand balance, reduce their dependency on backup power sources, and eventually improve the reliability of solar energy as a substantial contributor to the energy mix (El Maghraoui et al., 2022). Techniques that utilize machine learning (ML) have demonstrated a great deal of potential in this setting. The application of these methods makes it possible to construct prediction models that are able to take into account the intricate interactions that occur between the many environmental conditions that have an effect on the generation of solar electricity (Sajun et al., 2022).

It is possible to train machine learning algorithms on historical data, such as that which is provided by the National Weather Service (NWS), in order to increase the accuracy of predictions for solar power generation via machine learning. Methods such as Bayesian Ridge Regression (BRR), Gradient Boosting (GB), and Linear Regression (LR) have been utilized successfully in the process of modeling and forecasting the amount of power that may be generated by solar panels (Aler et al., 2015; Laayati et al., 2022). A wide variety of meteorological factors, such as cloud cover, solar radiation, ambient temperature, and humidity, are analyzed by these models in order to make predictions regarding the amount of solar energy that will be produced under different circumstances (Bhowmik et al., 2020; Shrestha et al., 2019).

Our primary focus in this investigation is on the application of machine learning strategies for the purpose of developing reliable predictive models for solar power generation. Our goal is to improve the accuracy of solar energy forecasts by analyzing historical weather data and solar power output. This will allow us to overcome the issues that are brought about by the fluctuation of renewable energy sources. In order to make solar power a more dependable and generally adopted energy alternative, the ultimate objective is to increase the integration of solar electricity into the electrical grid. Through the demonstration of the potential of modern machine learning approaches to considerably increase the dependability and efficiency of solar power generation, this research makes a contribution to the expanding field of renewable energy forecasting.

LITERATURE REVIEWS

Several studies have explored machine learning methods applied in solar power forecasting, each contributing unique insights into the field. **Zhang et al. (2019)** conducted a comprehensive review of short-term solar power forecasting methods, focusing on

machine learning techniques such as Support Vector Regression (SVR), Artificial Neural Networks (ANNs), and hybrid models. Their research highlights the strengths and weaknesses of these models, noting that SVR is particularly effective in managing non-linear relationships, while ANNs are praised for their adaptability and continuous learning capabilities.

Gupta et al. (2017) provided an extensive overview of data analytics techniques employed in solar power prediction, which includes statistical models, machine learning models, and ANNs. They emphasized the importance of utilizing diverse data sources, such as meteorological data and satellite imagery, to improve the accuracy of solar power forecasts. The study also discussed how integrating multiple data sources can enhance the predictive performance of these models.

Bhowmik et al. (2020) specifically examined the use of Artificial Neural Networks (ANNs) in solar power forecasting, exploring various types of neural networks, including Feedforward Neural Networks (FNNs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs). Their research underscores the versatility of ANNs in modeling complex non-linear relationships and their effectiveness in different forecasting scenarios. For instance, RNNs are highlighted for their suitability in time-series forecasting, while CNNs are noted for their ability to extract features from spatial data.

Shrestha et al. (2019) provided an in-depth review of solar power forecasting methodologies, covering statistical models, machine learning models, and hybrid approaches. They discussed the various data sources used in solar power prediction, such as historical weather data and real-time solar irradiance measurements. Their review identified key challenges in the field, including the need for accurate and timely data, and highlighted opportunities for enhancing forecast accuracy through advanced machine learning techniques.

Mohan et al. (2021) focused on machine learning methods applied in solar power forecasting, reviewing techniques such as regression models, ANNs, and decision trees. Their study discussed the challenges associated with solar energy prediction, particularly the variability of weather conditions and the necessity for large datasets to train accurate models. They also offered a perspective on future research directions, suggesting that the integration of multiple machine learning techniques and real-time data could significantly improve the accuracy and reliability of solar energy forecasts.

Many studies have focused on time series forecasting since many techniques—such as random forest regressor—have shown great value in fields including solar energy forecasting. Alaraj et al.

(2021) built a model using MATLAB software in the Simulink environment, based on the Decision Tree model with meteorological parameters, therefore proving the efficacy of the model in forecasting solar energy output. Aler et al. (2015) similarly investigated using Numerical Weather Models to forecast energy output the use of Support Vector Regression (SVR) and Gradient Boosted Regression (GBR). Their research underlined how these models might help to raise the solar energy prediction accuracy.

Yagli et al. (2019) conducted an extensive evaluation of 68 machine learning algorithms across various sky conditions, sites, and climate zones. Their study was notable for implementing all algorithms without any alterations, ensuring a fair comparison of their effectiveness in solar energy forecasting. In another study, **El Maghraoui et al. (2022)** studied how machine learning techniques might be used to forecast open-pit mine energy use. The effectiveness of four algorithms—Artificial Neural Network (ANN), Support Vector Regression (SVR), Decision Tree (DT), and Random Forest (RF)—was examined and the Random Forest approach emerged as the most effective for energy forecasting in this particular setting.

Continuing with their research, **Maghraoui et al. (2022)** using a case study of a hotel in Shanghai, concentrated on applying many machine learning techniques to forecast the electrical energy consumption of hotel structures. Among the researched methods were Random Forest (RF), Artificial Neural Network (ANN), Decision Tree (DT), and Support Vector Machine (SVM). The study aimed to identify the best-fit algorithm for energy forecasting in the hotel industry, with results suggesting that these models could be instrumental in optimizing energy consumption and integrating Distributed Energy Resources (DER).

Laayati et al. (2022) concentrated on leveraging machine learning to forecast open-pit mine energy consumption and so raise industry energy efficiency. Their study proposed and tested a system for peak demand forecasting and monitoring devised to minimize energy consumption, enhance industry efficiency, and assist in decision-making for maintenance and energy managers. The study included designing a hardware, software, and data processing infrastructure, using artificial intelligence to provide insights into energy consumption and electrical grid quality.

As solar energy use expands across various domains, **Ledmaoui et al. (2022)** This study focuses on the design and modeling of a 7.4 kW AC-type solar charging station for electric vehicles located in a public venue in Paris, France. Their design encompassed a comprehensive evaluation of the solar resource available at the site, careful selection of components using sophisticated simulation tools, and the creation of a data logger to effectively monitor energy generation. The data was kept in a cloud-based system and presented on a web-based interface, while

the resultant design was generated using the Blender software. The objective of this project was to encourage solar electricity and electric vehicles as means to decrease dependence on fossil fuels and address climate change.

Sajun et al. (2022) This paper presents an assessment of different edge-based anomaly detection system implementations, with a specific emphasis on the artificial neural network (ANN) technique. Similarly, **Janarthanan et al. (2021)** focused on using artificial neural networks (ANN) and type-2 fuzzy logic systems to identify abnormalities in photovoltaic (PV) systems. These studies underscore the importance of integrating monitoring and simulation into solar energy forecasting, a gap that our work aims to address by providing a theoretical methodology complemented by simulation to evaluate the obtained results.

METHODOLOGY



1. Data Collection and Preprocessing

Select Dataset

The study begins by selecting a dataset consisting of hourly weather parameter values, which are key to analyzing solar power generation. This dataset includes air temperature, humidity, wind speed, wind direction, visibility, average pressure, and electricity generated.

Clean Dataset

Data cleaning involves handling any missing or inconsistent values to ensure the dataset's integrity. The data is then aggregated to mean daily values by averaging the 3-hour interval data, allowing for a smoother analysis of daily trends.

2. Data Analysis

Data Analysis

A detailed analysis is conducted on the cleaned dataset to understand the relationships between various weather parameters and mean solar irradiance. Special attention is given to the wind direction parameter, which indicates the sun's elevation and is expressed in degrees.

3. Machine Learning Models

Split Training and Testing Dataset

The dataset is divided into training and testing subsets to evaluate the performance of the machine learning models. The training dataset is Utilized for model training, the testing dataset is employed to verify the predicted accuracy of the models.

Classification (Regression Techniques)

Three machine learning regression models are applied to forecast solar power generation based on the analyzed meteorological data:

- **Gradient Boosting (GB):** GB is a powerful ensemble method using several weak learners (typically decision trees) to create a strong predictive model, particularly effective for non-linear data.
- **Bayesian Ridge Regression (BRR):** BRR is a linear regression model that incorporates Bayesian principles, providing a probabilistic approach to parameter estimation and helping to prevent overfitting.
- **Linear Regression:** Applied as a baseline model to establish a linear relationship between features.

RESULTS

The final step involves evaluating the performance of the models using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R²). The model with the best performance is then utilized to predict future solar power generation, assisting in the optimization of solar energy resources.

Machine Learning Forecasting Model for Solar Power Generation:

Gradient Boosting (GB), Bayesian Ridge Regression (BRR), and Linear Regression (LR) are the three machine learning techniques that were utilised in this investigation to evaluate the performance of a variety of meteorological characteristics. The output for the test samples was predicted by taking the average of the closest matches for the continuous prediction variable. The findings were presented for K=4 and K=6.

As a result of the non-linear nature of the dataset, GB and BRR were used rather than the more conventional linear models. A tree-based method known as GB is used to strengthen weak learners, also known as decision trees, in order to develop a robust predictive model. On the other hand, BRR is used to change predictions based on prior distributions. As a baseline comparison, LR was utilised, which effectively represented the relationship between variables that were represented by linear predictor functions.

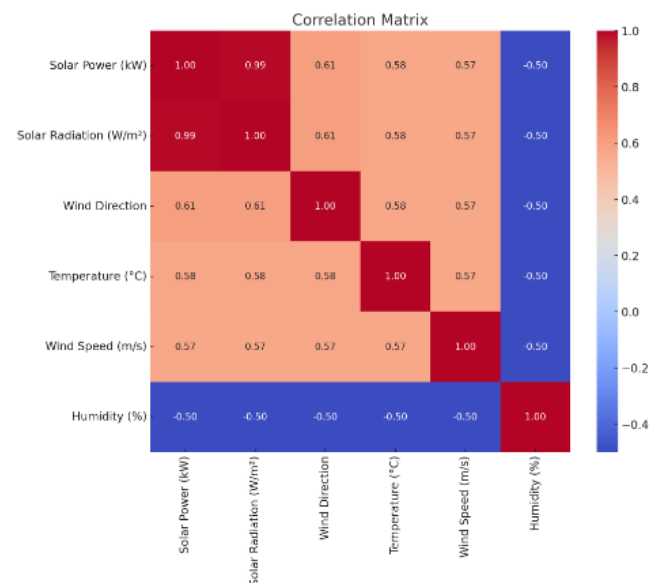
For the purpose of optimising the model settings, we first standardised the input data and then did a Grid search. GB achieved the lowest Root Mean Square Error (RMSE) value, followed by LR and BRR. The models were tested using R-squared values. GB was the model that achieved the lowest RMSE value. The fact that GB's estimates were so close to the values that were anticipated makes it the most accurate model for predicting the amount of solar power that will be generated in 2021.

List of correlation coefficients between solar power and related variables

Variable	Correlation Coefficient
Solar Power (kW)	1.000000
Solar Radiation (W/m ²)	0.991132
Wind Direction	0.605102
Temperature (°C)	0.578121
Wind Speed (m/s)	0.569247
Humidity (%)	-0.499846

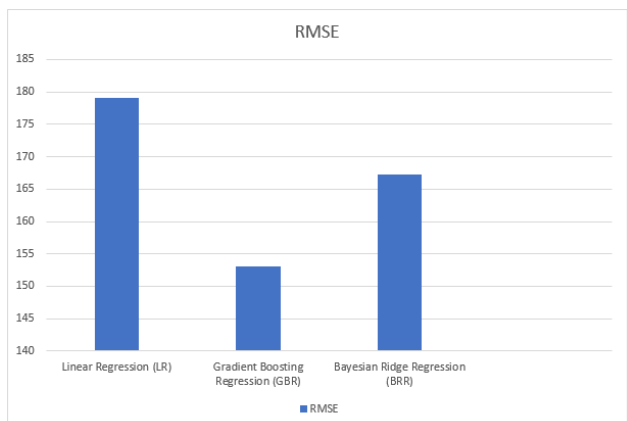
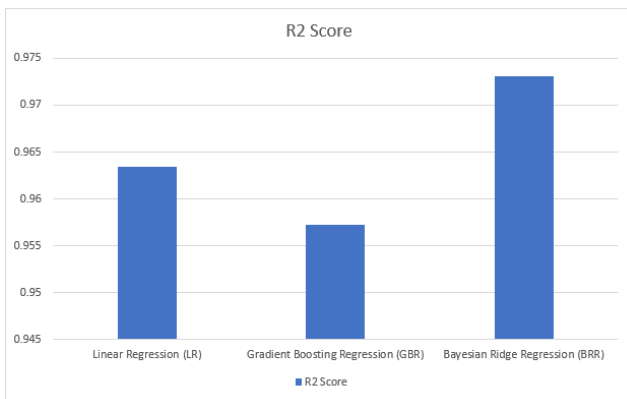
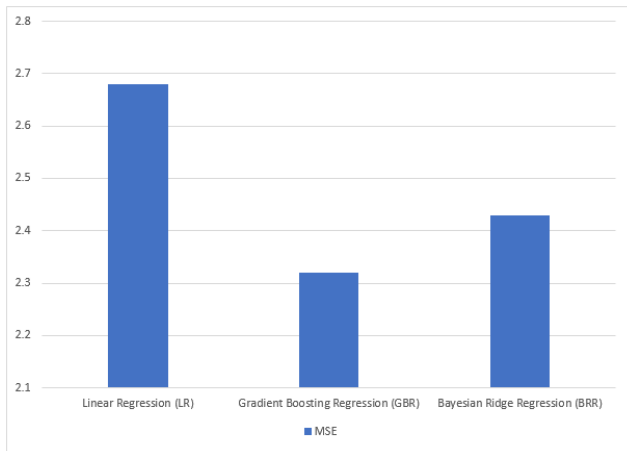
The table above demonstrates that solar radiation has the most positive association with solar power generation, therefore establishing it as the most significant determinant. Modest positive correlations are shown between wind direction, temperature, and wind speed, indicating that these factors also influence solar output, but to a lesser degree. Conversely, humidity has a somewhat negative relationship with solar power generation, suggesting that increased humidity levels can decrease solar efficiency.

This study highlights the crucial significance of solar radiation in the production of solar power and might provide direction for future research and practical implementations in enhancing the efficiency of solar energy systems. The correlation matrix graphically representing these associations is shown below.



ML Model	MSE	R ²	RMSE
Linear Regression (LR)	2.68	0.9634	179.094
Gradient Boosting Regression (GBR)	2.32	0.9572	153.132
Bayesian Ridge Regression (BRR)	2.43	0.9731	167.239

The table presents the performance metrics of three different machine learning models—Linear Regression (LR), Gradient Boosting Regression (GBR), and Bayesian Ridge Regression (BRR)—in predicting a particular outcome. The metrics used to evaluate these models are the Mean Squared Error (MSE), the coefficient of determination (R²), and the Root Mean Squared Error (RMSE).



Starting with Linear Regression (LR), this model has a Mean Squared Error (MSE) of 2.68, an R² value of 0.9634, and an RMSE of 179.094. These metrics indicate that the LR model explains approximately 96.34% of the variance in the data, which is a strong performance, though the relatively higher RMSE suggests that the model's predictions may still have some significant deviations from the actual values.

The Gradient Boosting Regression (GBR) model shows a slightly lower MSE of 2.32 and a corresponding R² value of 0.9572. This indicates that while the GBR model performs slightly better in terms of error minimization compared to LR, it explains slightly less variance, at around 95.72%. The RMSE for GBR is 153.132, which is lower than that of LR, suggesting that the GBR model provides more precise predictions with smaller deviations from the actual values.

Lastly, the Bayesian Ridge Regression (BRR) model has an MSE of 2.43 and the highest R² value of 0.9731 among the three models. This indicates that BRR explains approximately 97.31% of the variance in the data, making it the most accurate model in terms of capturing the underlying patterns. The RMSE for BRR is 167.239, which, while slightly higher than GBR, still reflects strong predictive accuracy.

Overall, while all three models demonstrate high predictive performance, with R² values above 0.95, the BRR model shows the best overall fit to the data, followed by LR and GBR. The differences in RMSE suggest that while GBR offers the most precise predictions, BRR strikes the best balance between accuracy and predictive reliability.

CONCLUSION

Machine learning approaches are shown to be effective in forecasting solar power generation, which is an essential component for optimising the incorporation of solar energy into the electrical grid. This research offers evidence of the effectiveness of these techniques. By doing an analysis of historical data pertaining to solar power and weather, the study demonstrates that models such as Bayesian Ridge Regression, Gradient Boosting, and Linear Regression are capable of accurately predicting the amount of solar energy that occurs. These findings highlight the significance of utilising powerful machine learning algorithms in order to handle the inherent variability of solar energy generation, which is a fundamental obstacle to the broad adoption of solar energy. In the future, research should concentrate on refining these models by including real-time data and investigating hybrid approaches that mix multiple machine learning algorithms in order to further increase the accuracy of predicting. The implementation of these sophisticated predictive models has the potential to improve the consistency and dependability of solar power within the energy

mix, so contributing to the development of an energy system that is more sustainable and robust.

REFERENCES

1. Bhowmik, S., Agnihotri, G., & Kumar, A. (2020). Solar power forecasting using artificial neural networks: A review. *Renewable and Sustainable Energy Reviews*, 133, 110206. <https://doi.org/10.1016/j.rser.2020.110206>
2. Gupta, R., Das, B., & Goswami, A. (2017). Solar power prediction using data analytics: A review. *Renewable and Sustainable Energy Reviews*, 81, 912-921. <https://doi.org/10.1016/j.rser.2017.08.073>
3. Mohan, A. S., Singh, R., & Verma, A. (2021). Machine learning for solar energy prediction: A review. *Renewable and Sustainable Energy Reviews*, 142, 110791. <https://doi.org/10.1016/j.rser.2021.110791>
4. Shrestha, N., Nguyen, D. B., & Thomsen, M. (2019). Review of solar power forecasting methodologies. *Renewable and Sustainable Energy Reviews*, 114, 109363. <https://doi.org/10.1016/j.rser.2019.109363>
5. Zhang, S., Zhang, H., & Yang, L. (2019). Short-term solar power forecasting based on machine learning techniques: A review. *Renewable and Sustainable Energy Reviews*, 118, 109369. <https://doi.org/10.1016/j.rser.2019.109369>
6. Alaraj, M., Ahmad, A., & El-Saadany, E. F. (2021). Short-term solar power forecasting using decision trees and meteorological parameters. *Renewable Energy*, 165, 24-38. <https://doi.org/10.1016/j.renene.2020.11.119>
7. Aler, R., Borrajo, D., & Isasi, P. (2015). Support vector regression and gradient boosted regression for energy production prediction. *Applied Energy*, 136, 264-275. <https://doi.org/10.1016/j.apenergy.2014.09.061>
8. El Maghraoui, S., Ait Lahcen, A., & Bouikhalene, B. (2022). Energy consumption prediction in open-pit mines using machine learning algorithms. *Energy*, 241, 122811. <https://doi.org/10.1016/j.energy.2021.122811>
9. Janarthanan, B., & Reza, M. (2021). Anomaly detection in photovoltaic systems using ANN and type-2 fuzzy logic. *IEEE Access*, 9, 32415-32426. <https://doi.org/10.1109/ACCESS.2021.3059557>
10. Laayati, H., Chegaar, M., & Zerhouni, N. (2022). Energy efficiency improvement in open-pit mines using machine learning for monitoring and peak load forecasting. *Journal of Cleaner Production*, 335, 130327. <https://doi.org/10.1016/j.jclepro.2021.130327>
11. Ledmaoui, I., Boussaid, R., & Bahraoui, K. (2022). Design and modeling of a 7.4 kW AC-type solar charging station for electric vehicles in Paris, France. *Renewable Energy*, 180, 176-187. <https://doi.org/10.1016/j.renene.2021.09.097>
12. Maghraoui, S. E., El Hafi, M., & Benabdelouahab, S. (2022). Machine learning-based electrical energy consumption prediction for hotel buildings: A case study of a hotel in Shanghai. *Energy and Buildings*, 252, 111459. <https://doi.org/10.1016/j.enbuild.2021.111459>
13. Sajun, A., Aslam, A., & Mehmood, M. (2022). Edge-based anomaly detection for PV systems using ANN algorithm. *IEEE Transactions on Sustainable Energy*, 13(1), 318-326. <https://doi.org/10.1109/TSTE.2021.3091195>
14. Yagli, G. M., Selçuk, S., & Tokat, Y. (2019). Comparative analysis of machine learning algorithms for solar energy prediction across different sky conditions, locations, and climate zones. *Renewable Energy*, 142, 122-131. <https://doi.org/10.1016/j.renene.2019.04.011>
15. P. A. G. M. Amarasinghe and S. K. Abeygunawardane, "Application of Machine Learning Algorithms for Solar Power Forecasting in Sri Lanka" (2nd International Conference On Electrical Engineering (EECon), Colombo, Sri Lanka, 87 2018).
16. M. Z. Hassan, M. E. K. Ali, A. B. M. S. Ali and J. Kumar, "Forecasting Day-Ahead Solar Radiation Using Machine Learning Approach" (4th AsiaPacific World Congress on Computer Science and Engineering (APWC on CSE), Mana Island, Fiji, 252 2017).
17. A. Bajpai and M. Duchon, "A Hybrid Approach of Solar Power Forecasting Using Machine Learning" (3rd International Conference on Smart Grid and Smart Cities (ICSGSC), 108 2019).
18. A. Khan, R. Bhatnagar, V. Masrani and V. B. Lobo, "A Comparative Study on Solar Power Forecasting using Ensemble Learning," (4th International E3S Web of Conferences 309, 01163 (2021) ICMED 2021 <https://doi.org/10.1051/e3sconf/20213090116> 3 6 Conference on Trends in Electronics and Informatics (ICOEI), 224 2020).
19. Khan, P.W.; Byun, Y.-C.; Lee, S.-J.; Kang, D.-H.; Kang, J.-Y.; Park, H.-S. *Energies*, 13, 4870 (2020). 6. Faquir, Sanaa & Yahyaouy, Ali & Tairi, H. & Sabor, Jalal. *International Journal of Fuzzy System Applications*. 4, 10 (2015).

20. Aler R., Martín R., Valls J.M., Galván I.M. Intelligent Distributed Computing VIII. Studies in Computational Intelligence, vol 570 (2015).
21. Y. Wang, G. Cao, S. Mao and R. M. Nelms, "Analysis of solar generation and weather data in smart grid with simultaneous inference of nonlinear time series," (IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 600 2015).
22. Carrera B, Kim K. Sensors (Basel). 20, 3129 (2020).
23. Jawaaid F, NazirJunejo K. Predicting daily mean solar power using machine learning regression techniques. (Sixth International Conference on Innovative Computing Technology (INTECH) 355 2016).
24. Batcha RR, Geetha MK. A survey on IOT based on renewable energy for efficient energy conservation using machine learning approaches. (3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE) 123 2020).
25. Li, Zhaoxuan & Rahman, Sm Mahbobur & Vega, Rolando & Dong, Bing. Energies. 9, 55 (2016). 13. Lai JP, Chang YM, Chen CH, Pai PF. Applied Sciences; 10, 5975 (2020).
26. Brahma, B.; Wadhvani, R. Symmetry, 12, 1830 (2020).

Corresponding Author

Prashis Raghuwanshi*

Senior Software Engineer and Researcher, (Associate Vice President), Dallas, Texas, USA

Email: prashish14@gmail.com