# Monthly Solar Prediction Using Machine Learning: Diyala Governorate, Iraq as a Case Study

**Noor Razzaq Abbas [1], Hussein Alkattan *[2], Hamidreza Rabiei-Dastjerdi [3], Mohamed Saber[4], Marwa M. Eid [5]**

[1] Al-Furat Al-Awsat Technical University, Technical Institute of Najaf, Iraq
[2] Department of System Programming, South Ural State University, Chelyabinsk 454080, Russia
[3] School of Architecture, Planning, and Environmental Policy & CeADAR, University College Dublin, Dublin, Ireland
[4]Electronics and Communications Engineering Department, Faculty of Engineering, Delta University for Science and Technology, Gamasa City 11152, Egypt
[5] Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura, Egypt
Emails: noor.hachame@atu.edu.iq; alkattan.hussein92@gmail.com; hamid.rabiei@UCD.ie; Mohamed.saber@deltauniv.edu.eg; mmm@ieee.org

**Abstract**

Solar radiation constitutes the Earth's primary energy source and is critical in regulating surface radiation equilibrium, vegetation photosynthesis, hydrological cycles, and extreme atmospheric. On the other hand, the depletion of global fossil fuel reserves mandates the power sector to adopt renewable energy-based sources, including photovoltaic and wind energy conversion systems. Therefore, the precise solar radiation prediction is imperative for climate research and the solar industry. This paper illustrates the use of two machine-learning approaches: random forest (RF) and support vector machine (SVM), to predict surface solar radiation in the Diyala governorate of Iraq for one step ahead, utilizing only lagged monthly time series data of the factor as input predictors. The findings were evaluated using three performance measures: coefficient of determination (R2), root mean square error (RMSE), and mean absolute error (MAE). The results showed that using 10 monthly lags time series as input predictors leads to the best prediction performance. Furthermore, in terms of the RMSE, the prediction performance of the RF algorithm was better than that of the SVM algorithm (RF's RMSE, MAE, and R2 were 181.398, 129.522, and 0.979, while for SVM were 240.149, 184.802, and 0.978, respectively).

**Keywords:** random forest; support vector machine; machine learning; solar radiation prediction

## 1. Introduction

In contemporary times, the escalating impact of global climatic alteration, the decline of the human ecological environment, the scarcity of renewable energy resources, and the spread of environmental contamination, have led to the recognition of solar radiation energy as an indispensable and enduring source of clean energy on a global scale. Therefore, solar photovoltaic power has demonstrated a remarkable growth rate in its development and utilization [1]. Photovoltaic power generation exhibits volatility and intermittency, which can be predominantly attributed to fluctuations in solar radiation. [2]. To enhance the proportion of photovoltaic-generated electric power circulating within the power grid system, a crucial mechanism lies in the implementation of proficient and expeditious control over power dispatching. Subsequently, it is important to predict solar radiation precisely, the comes about of which can provide critical decision support for control power dispatching systems and can successfully and effectively decrease the operational costs of the power control framework [3].



Doi: https://doi.org/10.54216/JAIM.050204
Received: February 06, 2023 Revised: May 22, 2023 Accepted: September 09, 2023

Numerous techniques are employed for predicting solar radiation. Conventional prediction approaches typically employ statistical techniques to establish the linkage between historical data and solar radiation [4]. The conventional statistical techniques, however, are inadequate to accurately elucidate the intricate non-linear nature of solar radiation, thereby impeding the enhancement of prediction precision. The field of machine learning is widely acknowledged as an exceptional and interdisciplinary domain [5], [6]. The proliferation of the aforementioned technology has prompted numerous scholars to employ its techniques toward the task of predicting solar radiation, resulting in notable accomplishments [7]. For example, support vector machine (SVM) algorithm [8], deep learning techniques [9], [10], and artificial neural network (ANN) methods [11], [12] have demonstrated superior performance over conventional linear regression prediction models in the context of solar radiation predicting. Furthermore, combining machine learning techniques with numerical weather prediction methods has yielded substantial advancements in the prediction [13-15]. The goal of this paper is to present a reliable predictive modelling framework for solar radiation in the Diyala governorate of Iraq at a monthly timescale based on two machine-learning techniques (RF and SVM). A skillful prediction of monthly solar radiation for this region can help facilitate the management and operation of the electricity system and can ultimately aid in the effort to integrate more wind and solar energy sources into the power system.

## 2. Solar Radiation Data Used

Monthly solar radiation data measured at the ground from 1981 to 2021 for Diyala governorate in Iraq were derived from the ERA5 atmospheric reanalysis dataset. This dataset is the fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) [16]. The ERA5 meteorological data are available from 1979 to three months in real-time and cover the Earth on a 0.25°×0.25° (~30 km) grid and resolve the atmosphere while using 137 levels from the surface up to a height of 80 km.

## 3. Methods

### A. Random Forest (RF)

An RF is a tree-based ensemble supervised machine-learning technique. The algorithm is widely used to model high-dimensional data in regression and classification problems. Simply, the RF algorithm builds (randomly created) and merges the output of multiple decision trees together and forming an ensemble to get a more stable and accurate prediction [17]. The bias in all the trees is the same; however, the variances can be decreased by reducing the relationship's coefficients [18]. RF for regression problems is based on a random vector, and it is generated by growing trees, using the trees predictor $d(x,\varphi)$ to obtain numerical values. It is assumed in the RF algorithm that the training data sample is independent statistically. The generalized mean square error of the numerical predictor $d(x)$ could be represented as follows:

$$E_{X,Y}\big(Y-d(x)\big)^2 = arg\ minimum_\delta \sum_{i=1}^{k} \Psi\big(y_i, F_{m-1}(x_i)+\delta d_m(x_i)\big) \tag{1}$$

The RF prediction is then simply created via averaging $k$ of the trees $(d(x,\varphi))$. More detailed information regarding RF model theories could be found in [17].

### B. Support Vector Regression (SVR)

SVR is an extension algorithm to the supervised machine-learning technique known as SVM for regression problems (predicting continuous output values based on the input) [19]. In regression problems, the SVR models try to reduce the error rate between the target and predicted sample by fitting around the hyperplane (regression line) within a certain threshold value ($\varepsilon c$) such that all data points within $\varepsilon$ are not penalized for their error. The problem can be formulated as follows:



Doi: https://doi.org/10.54216/JAIM.050204
Received: February 06, 2023 Revised: May 22, 2023 Accepted: September 09, 2023

$$Min. \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}|\xi_i| \tag{2}$$
$$s.a. \quad |y_i - w_i x_i| \leq \varepsilon + |\xi_i| \quad i = 1,2,\dots,n$$

where $n$ is the number of training samples is, the slack variable $\xi$ is the deviation for any value that falls outside $\varepsilon$, and $C$ is the penalty factor that determines the tradeoff between minimizing the training error and minimizing model complexity. As $C$ increases, the tolerance for points outside $\varepsilon$ also increases. The performance of SVR then depends on the choice of parameters $\varepsilon$ and $C$.

**4. Methodology**

- The proposed machine learning-based prediction system generally consists of the following six stages:
- The first stage is reading the monthly solar radiation time series data;
- The second stage is data preprocessing which is an integral step, as choosing useful and correct information directly affects on model's capacity for learning. In this paper, we examined up to 12 consecutive lagged time series of the solar radiation as input predictions to make the prediction effective and efficient;
- The third stage devides data into two subsets: train and test. This procedure measures machine learning algorithms' performance when they are used to make predictions on a data sample that was not used to train the model. The data from 1981 to 2015 (85% of data) were allocated for the training (calibration) of the models, while the remaining 6 years (15% of data) were used for the verification of the models;
- The fourth stage is building machine learning models (based on used RF and SVM algorithms) to predict solar radiation one month ahead;
- The fifth stage is performance evaluation, where three statistical measures based on the coefficient of determination, denoted R2 (Eq. 3), Root Mean Square Error (RMSE) (Eq. 4), and the Mean Absolute Error (MAE) (Eq. 5) to calculate the efficiency of the prediction models were used [33].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}\left(y_i - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i\right)\right)} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

Finally, in the last stage, the statistical performance evaluation measurements are compared between the two prediction models to choose the model with the best performance.
The flowchart of the proposed system used in this study is shown in Figure 1.



Doi: https://doi.org/10.54216/JAIM.050204
Received: February 06, 2023 Revised: May 22, 2023 Accepted: September 09, 2023

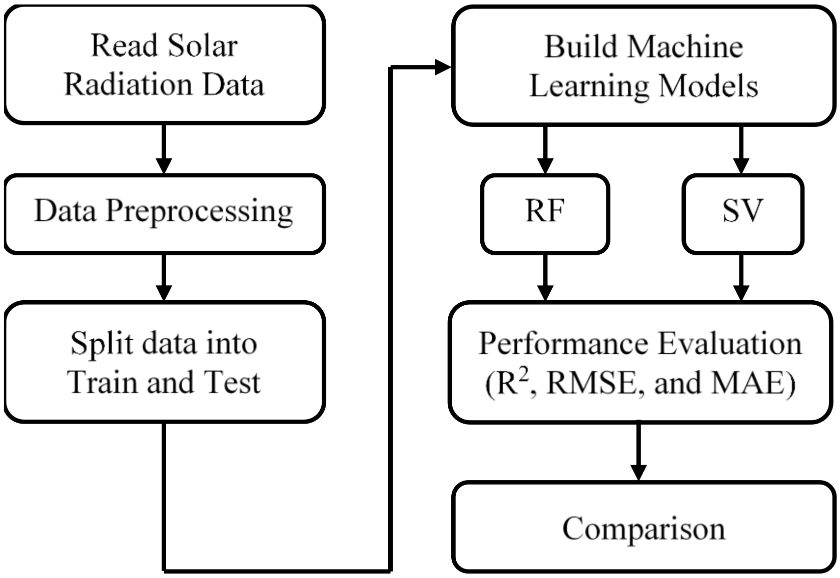


Figure 1: General block diagram of the proposed prediction system

## 5. Results and Discussion

The models' prediction performance during the verification period is presented in Table 1 for the two machine learning algorithms. It can be seen that as the number of lag time series predictors of solar radiation increases, the prediction performance of the two algorithms improves. Using 10 of lags time series, the algorithms showed the best prediction performance. It can also be seen that the RF algorithm is the dominant algorithm with an RMSE of 181.398, MAE of 129.522, and R2 of 0.979 compared with the SVM algorithm RMSE is 240,149, MAE is 184,802, and R2 is 0.979. Figures 2 and 3 graphically compare the observed time series solar radiation using the two prediction algorithms. Most predicted values are pretty close to the actual value. Sometimes, the results have an exact match with the actual value. Thus, we are pleased with the results.

Table 1: Models' performance for the different number of lags input during the verification period

| **No. Lags** | **RF** | | | **SVM** | | |
|---|---|---|---|---|---|---|
| | ***RMSE*** | ***MAE*** | ***$R^2$*** | ***RMSE*** | ***MAE*** | ***$R^2$*** |
| 1 | 449.048 | 309.400 | 0.873 | 660.623 | 452.740 | 0.842 |
| 2 | 323.135 | 191.382 | 0.935 | 312.860 | 207.185 | 0.959 |
| 3 | 287.772 | 156.862 | 0.948 | 305.720 | 199.545 | 0.960 |
| 4 | 194.184 | 131.889 | 0.976 | 259.700 | 201.635 | 0.976 |
| 5 | 181.541 | 123.830 | 0.979 | 236.533 | 184.852 | 0.980 |
| 6 | 203.082 | 137.419 | 0.974 | 253.186 | 189.641 | 0.974 |
| 7 | 188.553 | 134.141 | 0.978 | 249.784 | 189.514 | 0.974 |
| 8 | 188.466 | 134.457 | 0.978 | 249.462 | 190.801 | 0.973 |
| 9 | 191.575 | 132.594 | 0.977 | 230.906 | 176.982 | 0.978 |
| **10** | **181.398** | **129.522** | **0.979** | **240.149** | **184.802** | **0.978** |
| 11 | 195.332 | 136.663 | 0.976 | 249.266 | 190.871 | 0.977 |
| 12 | 182.616 | 122.586 | 0.972 | 282.314 | 182.536 | 0.972 |



Doi: https://doi.org/10.54216/JAIM.050204
Received: February 06, 2023 Revised: May 22, 2023 Accepted: September 09, 2023

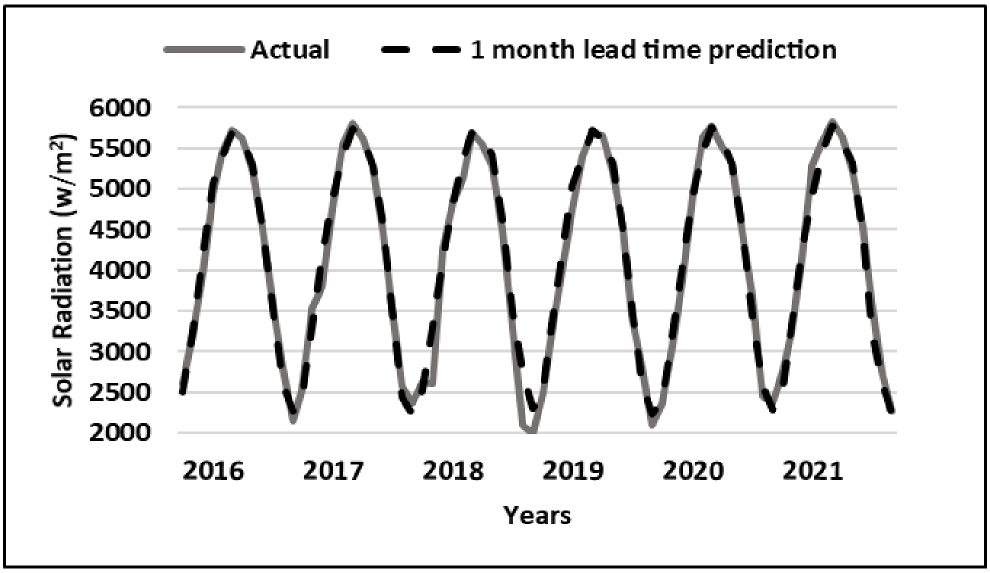


Figure 2: Observed and predicted time series of solar radiation during the verification period using the RF model

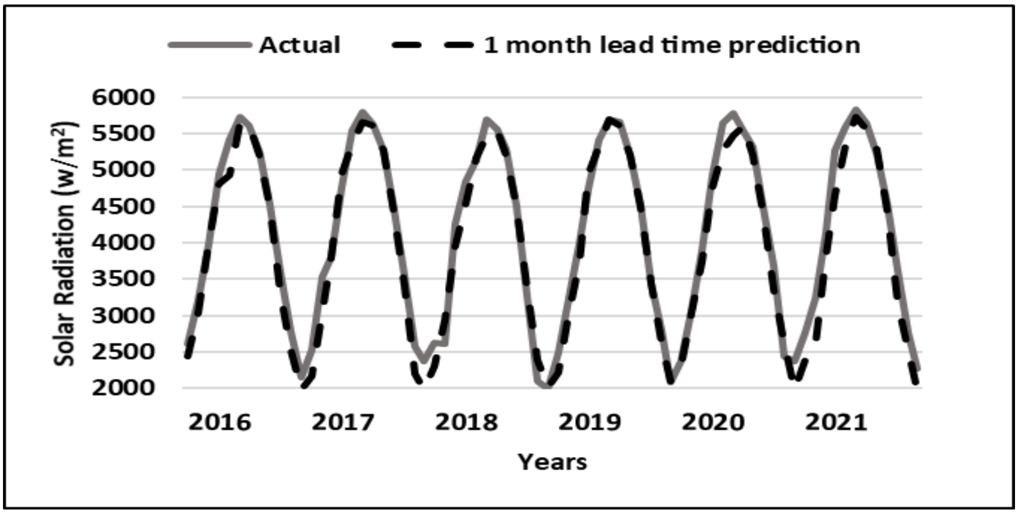


Figure 3: Observed and predicted time series of solar radiation during the verification period using the SVM model

## 6. Conclusions

The prediction of solar radiation has become a significant area of interest for many researchers, primarily due to the surging demand for renewable energy, and the increasing global concern over climate change. The present study proffered the employment of the RF and SVM algorithms as a means of anticipating solar radiation for the governorate of Diyala in Iraq a month prior. In order to determine the optimal prediction performance model, a series of up to twelve lagged time series values of the factor were utilized and tested. The best models were defined based on the RMSE values. Also, other statistical parameters such as R2 and MAE were calculated to evaluate the skillfulness of the models employed better. The research findings indicate that both predictive machine-learning algorithms exhibit promising potential in accurately predicting monthly solar radiation. However, in terms of the RMSE, the RF algorithm exhibited a superior prediction performance as compared to the SVM algorithm. This empirical analysis has facilitated favorable inspiration for our forthcoming research endeavors. The prediction of solar radiation data may be facilitated through the study of weather patterns or cloud classifications.

**Funding:** "This research received no external funding"

**Conflicts of Interest:** "The authors declare no conflict of interest."



Doi: https://doi.org/10.54216/JAIM.050204
Received: February 06, 2023 Revised: May 22, 2023 Accepted: September 09, 2023

## References


[1] Y Yang, P E Campana, B Stridh, J Yan, Potential analysis of roof-mounted solar photovoltaics in Sweden. Applied Energy, 279, 115786, 2020.

[2] M O Moreira, P P Balestrassi, A P Paiva, P F Ribeiro, B D Bonatto, Design of experiments using artificial neural network ensemble for photovoltaic generation forecasting. Renewable and Sustainable Energy Reviews, 135, 110450, 2021.

[3] L Ge, Y Xian, J Yan, B Wang, Z Wang, A Hybrid Model for Short-term PV Output Forecasting Based on PCA-GWO-GRNN. Journal of Modern Power Systems and Clean Energy, 8(6), 1268–1275, 2020.

[4] H Sun, D Yan, N Zhao, J Zhou, Empirical investigation on modeling solar radiation series with ARMA–GARCH models. Energy Conversion and Management, 92, 385–395, 2015.

[5] B T Al-Nuaimi, H K Al-Mahdawi, Z Albadran, H Alkattan, M Abotaleb, E-S M El-kenawy, Solving of the Inverse Boundary Value Problem for the Heat Conduction Equation in Two Intervals of Time. Algorithms, 16(1), 33, 2023.

[6] H Alkattan, M Abotaleb, A A Subhi, O A Adelaja, A Kadi, H K I Al-Mahdawi, The prediction of students' academic performances with a classification model built using data mining techniques, 2022.

[7] A Alfadda, S Rahman, M Pipattanasomporn, Solar irradiance forecast using aerosols measurements: A data driven approach. Solar Energy, 170, 924–939, 2018.

[8] J Lin, H Li, A Short-Term PV Power Forecasting Method Using a Hybrid Kmeans-GRA-SVR Model under Ideal Weather Condition. Journal of Computer and Communications, 8(11), 102–119, 2020.

[9] T Zhu, Y Guo, Z Li, C Wang, Solar Radiation Prediction Based on Convolution Neural Network and Long Short-Term Memory. Energies, 14(24), 8498, 2021.

[10] S M Awan, Z A Khan, M Aslam, Solar Generation Forecasting by Recurrent Neural Networks Optimized by Levenberg-Marquardt Algorithm, in IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society, 276–281, 2018.

[11] T ZHU, H ZHOU, H WEI, X ZHAO, K ZHANG, J ZHANG, Inter-hour direct normal irradiance forecast with multiple data types and time-series. Journal of Modern Power Systems and Clean Energy, 7(5), 1319–1327, 2019.

[12] S Rahman, S Rahman, A K M Bahalul Haque, Prediction of Solar Radiation Using Artificial Neural Network. Journal of Physics: Conference Series, 1767(1), 012041, 2021.

[13] J Gari da Silva Fonseca, F Uno, H Ohtake, T Oozeki, K Ogimoto, Enhancements in Day-Ahead Forecasts of Solar Irradiation with Machine Learning: A Novel Analysis with the Japanese Mesoscale Model. Journal of Applied Meteorology and Climatology, 59(5), 1011–1028, 2020.

[14] S Ghimire, R C Deo, N J Downs, N Raj, Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. Journal of Cleaner Production, 216, 288–310, 2019.

[15] S Pereira, P Canhoto, R Salgado, M J Costa, Development of an ANN based corrective algorithm of the operational ECMWF global horizontal irradiation forecasts. Solar Energy, 185, 387–405, 2019.

[16] H Hersbach et al, The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999–2049, 2020.

[17] L Breiman, Random Forests. Machine Learning, 45(1), 5–32, 2001.

[18] T Hastie, R Tibshirani, J Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edi. Springer New York, NY, 2009.

[19] S Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. USA: Cambridge University Press, 2014.




Doi: https://doi.org/10.54216/JAIM.050204
Received: February 06, 2023 Revised: May 22, 2023 Accepted: September 09, 2023