ELECTRICAL ENGINEERING AND ENERGY

OPEN ACCESS

Estimating Solar Power Generation with RF, GB, and SVR Algorithms Based on Meteorological Data and Orientation Angles: Adıyaman Case Study

Abdurrahman Özbeyaz[1]

Department. of Electrical and Electronics Engineering, Adıyaman University, Adıyaman, Türkiye

Submitted: 3 August 2025
Accepted: 15 August 2025
Online First:18 August 2025

Corresponding author
Abdurrahman Özbeyaz
aozbeyaz@adiyaman.edu.tr

DOI:10.64470/elene.2025.1006

Abstract: This study attempts to develop precise machine learning algorithms for estimating solar power generation using meteorological data, including air pollution (PM2.5, PM10) for different orientation angles in the dual-axis solar panel tracking system. Our study focused on maximizing energy generation for multivariable input factors in Adıyaman, rather than optimizing environmental parameters for fixed panels as previously studied. Three machine learning algorithms—Random Forest (RF), Gradient Boosting (GB), and Support Vector Regression (SVR)—were studied. The experimental setup was designed to measure power, temperature, humidity, wind speed/direction, pressure, air pollution data, and to adjust orientation angles (tilt and azimuth angles). Algorithm performances were assessed using R^2, MSE, and RMSE metrics. RF yielded the best results (R^2: 0.83, MSE: 5.32, RMSE: 2.26). Including cumulative air pollution data improved prediction accuracy, implying particulate matter indirectly affects solar radiation and energy output.

Keywords Solar Energy, Machine learning, Energy Generation, Air pollution, Orientation Angle

1. Introduction

Energy consumption is increasing rapidly because of industrialization, technological advancements, and population growth. The continuous development and expansion of global infrastructure further amplifies energy demands, while the availability of fossil fuels to meet these needs is steadily declining. This growing imbalance underscores the urgent need to transition toward cleaner, more sustainable energy sources. Renewable energy has emerged as a critical solution due to its environmental sustainability and reduced ecological impact. Among these renewable resources, solar energy stands out for its abundant availability and immense potential to contribute significantly to the global energy mix (Bayrakçı & Gezer, 2019).

Photovoltaic panels, which enable the direct conversion of solar energy into electrical energy, are one of the basic technologies that utilize solar radiation most efficiently in renewable energy production. Various effective strategies are being developed today to minimize the negative impact of factors on the efficiency of photovoltaic (PV) panel electricity production. These factors include cloud cover, air pollution, extreme temperatures, and incorrect panel orientation and tilt angle (Arslan & Çunkaş, 2024). One strategy is to position the panel at the optimal tilt and orientation angles, considering the region's solar radiation data (Bakırcı, 2009). The correct positioning of PV panels is critical to achieving maximum annual energy production.

Another parameter that adversely impacts energy production in solar energy is represented by various forms of particulate matter (PM2.5 and PM10) (Zhou et al., 2016). These pollutants reduce solar radiation, which negatively affects the energy production capacity of photovoltaic (PV) panels (Shim et al., 2025). Many photovoltaic (PV) systems produce less energy than expected due to radiation losses caused by particulate matter (Wu & Zhou, 2019). The tilt angle of photovoltaic (PV) panels is a critical parameter for energy efficiency. For fixed systems, it must be adjusted to the optimal angle specific to the region, taking exposure to particulate matter into account. Recently, various studies have used solar tracking systems to determine the optimal tilt angle for fixed systems in specific regions (Arslan & Çunkaş, 2024; Yadav & Chandel, 2013). Solar tracking systems, particularly in regions with high solar energy potential such as Adıyaman, aim to optimize energy generation by adjusting orientation angles (Al-Mohamad, 2004). These systems have become a key focus in efforts to maximize the utilization of solar energy resources (Boyacı & Kocaman, 2018). In many studies, the adjustment of solar panel angles is typically based on variations in solar radiation intensity (Fahad et al., 2019; Sharma & Bhattacharya, 2020). However, research also indicates that atmospheric factors, especially particulate matter, and indirectly solar radiation, significantly influence solar energy efficiency (Demirtaş et al., 2019; Kılıç & Kumaş, 2016; Ozbeyaz & Demirci, 2019).

Recently, due to their potential to maximize solar energy generation, researchers have widely used artificial intelligence-based algorithms, including the data on pollutant particulate matter that negatively affects solar radiation and the panel tilt angle, to calculate the optimum energy generation, as we do in this study. In the literature, these AI-based studies generally have preferred the commonly known AI methods, such as neural networks (Oviedo et al., 2013, 2014) and multi-layer perceptron (AL-Rousan et al., 2021). In this context, AI-based methodologies are well-positioned to dynamically adapt PV panels to changing meteorological conditions, thereby

In this study, the machine learning algorithms were developed to estimate the energy generation at the various angles by using different meteorologic data inputs (Ozbeyaz & Demirci, 2019; Uz et al., 2022). A solar tracker system was specifically designed for this purpose, featuring the ability to adjust to different panel angles and perform biaxial movements. During the experimental phase, the tilt and azimuth angles of the solar panels were systematically changed at predetermined intervals, and the resulting energy data, along with meteorological parameters, were meticulously recorded. These collected data were employed as inputs in machine learning models (AL-Rousan et al., 2021; Kaul & Weed, 2021; Pierce et al., 2022), enabling the prediction of energy production through multiple algorithms. This study uniquely incorporates orientation angles and particulate matter that adversely impact solar brightness within the machine learning framework, specifically conducted in Adıyaman, Türkiye.

2. Data Acquisition System

Türkiye has significant potential for solar energy (around 380 GWh), with an average annual sunshine duration of 7.5 hours and a daily solar energy density of 4.2 kWh/m² (Kaçan & Ülgen, 2012). Adıyaman, located in the Southeastern Anatolia Region, lies between 39° east longitude and 37°25' to 38°11' north latitude. With an average daily sunshine duration of 8.11 hours—exceeding the national average—Adıyaman is an important region in solar energy production (Aslan et al., 2021).

In the hardware system, the solar panel operates along two axes, and power, voltage, and current measurements are systematically recorded in relation to these angles. Additionally, meteorological data were synchronously collected alongside photovoltaic (PV) panel measurements. These data include concentrations of particulate matter (PM2.5 and PM10), wind direction and speed, humidity, temperature, and atmospheric pressure. The solar panel integrated into the system is a high-efficiency monocrystalline module with a power output of 20 watts. Figure 1 presents visual representations of the developed system.


Figure 1 Sample images of the data collection system

Geomatic experts determined the system's orientation based on its geographical location. After establishing the system's true north and south orientations, the tilt angle of the solar panel was adjusted on specific days according to both the azimuth angle (γs, the angle between the south direction and the incident solar rays on the panel) and the inclination angle (β, the angle between the panel surface and the horizontal plane). During this process, power, voltage, and current measurements obtained from the photovoltaic panel were systematically recorded on an SD card, along with relevant meteorological data. Specialized electronic hardware was designed to facilitate automatic data recording. This hardware was integrated into a dedicated compartment located at the base of the system. Figure 2 presents the hardware components of the designed system.

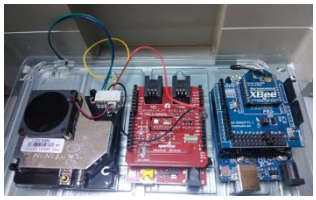
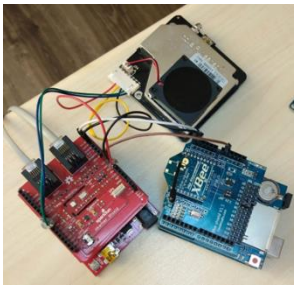




Figure 2. The images of the electronic hardware controlling the system.

The system consists of two microcontroller boards along with various sensor components. The primary microcontroller is responsible for managing the SparkFun Weather Shield and the Nova air quality sensor. The Sparkfun Weather Shield integrates multiple environmental sensors, including the Si7021 humidity and temperature sensor, the MPL3115A2 barometric pressure sensor, and the ALS-PT19 light sensors, while rain and wind sensors are connected via an RJ11 interface. The Nova air quality sensor measures the concentration of PM2.5 and PM10 particulate matter. The secondary microcontroller controls the data logger shield, the LoRa wireless communication module, and the INA226 power monitoring module. The data logger shield includes an SD card interface and a real-time clock module to ensure precise time-stamped data recording. The system systematically logged all collected data, including temporal information, onto the SD card. Data acquisition occurred every two minutes, with wireless transmission via the LoRa module, while data storage on the SD card was performed at five-minute intervals. Figure 3 provides a detailed schematic of the system's electronic hardware.

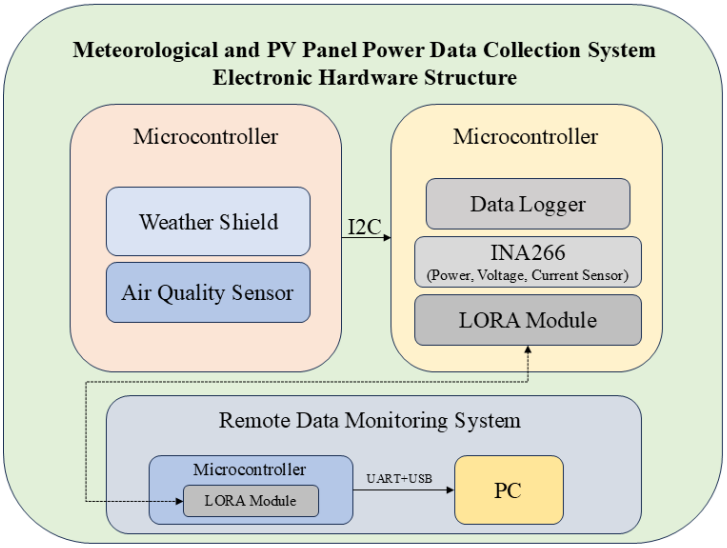


Figure 3. A schematic representation of the electronic hardware that controls the system.

Throughout the day, the solar panel was adjusted at specific angles in accordance with the sun's movement, and the generated power was recorded alongside meteorological data. The panel's stand angle was oriented southward, incrementally adjusted by 15-degree intervals each hour. At solar noon (12:00), the azimuth angle of the panel was aligned directly south (0 degrees). The same pattern was followed as the adjustment proceeded westward in the afternoon until sunset. Additionally, the tilt angle of the panel was set to 15, 30, and 45 degrees at different times of the day across various days. This approach allowed for the measurement of the varied energy generations attainable under different solar positions and meteorological conditions.

3. Methods

This study aims to develop machine learning models for estimating the energy output of photovoltaic (PV) panels by incorporating meteorological data and panel orientation parameters. The dataset was collected at 10-minute intervals using a custom-designed dual-axis solar tracking system capable of independently adjusting both azimuth and tilt angles. A total of 657 data samples were recorded between January and June 2024, representing a limited range of seasonal variation; therefore, future studies will focus on year-round data collection to improve model generalizability. Before model construction, the data underwent a preprocessing phase to eliminate irrelevant variables. In the modeling phase, both meteorological parameters and the physical panel orientation (tilt and azimuth angles) were used as independent variables. The dataset was split into separate training and test sets for model construction to ensure that training data was never used or seen during testing, thereby preventing bias in performance assessment. Regression-based forecasting models were then developed using three machine learning algorithms: Gradient Boosting (GB), Random Forest (RF), and Support Vector Regression (SVR).

3.1 Gradient Boosting Algorithm

Boosting is an ensemble learning method designed to enhance the predictive performance of classification and regression techniques. The boosting algorithm, originally introduced by Schapire in 1990 (Schapire, 1990), enables improved performance in classification and regression trees. One of the most prominent implementations of this approach is the AdaBoost algorithm. In addition, a variant known as gradient boosting, proposed by Friedman in 1999 (Ankarali et al., 2012), has also been widely adopted. Both methods incrementally incorporate new models to enhance predictive accuracy. By leveraging different loss functions, these methods offer effective solutions for both classification and regression tasks. Gradient boosting iteratively improves model performance by concentrating on reducing residual errors at each stage. The

mathematical representation of the gradient boosting model is shown in the following equation (Temel et al., 2014).

$$F(x) = F_0 + \beta_1 h_1(x) + \beta_2 h_2(x) + \cdots + \beta_M h_M(x) \quad (1)$$

In Equation (1), F_0 denotes the initial model, while $h_M(x)$ and β_M represent, respectively, the learner added at each stage and its corresponding weight. This framework proves highly effective not only for generating robust predictions but also for enhancing the model's generalization capacity.

3.2 Random Forest Algorithm

The Random Forest algorithm, introduced by Leo Breiman in 2001, constitutes an enhanced version of the tree bagging method (Breiman, 2001). By constructing an ensemble of decision trees, it seeks to improve predictive accuracy. Each decision tree in the Random Forest is generated via the bootstrap method, which involves sampling random subsets from the training dataset. Furthermore, the Random Subspace Method is applied at each node, whereby only a small and randomly selected subset of attributes is considered for potential splits. This strategy fosters diversity among the trees and, consequently, enhances the model's generalization ability. Empirical research indicates that the Random Forest algorithm generally generations higher accuracy than tree bagging and other random tree ensemble methods. Notably, when the size of the attribute subset (K) is relatively small compared to the total number of attributes (n), the algorithm becomes more robust. In addition, by leveraging bootstrap sampling, the Random Forest outperforms models that rely solely on the Random Subspace Method. Fundamentally, the Random Forest comprises multiple decision trees, the predictions of which are aggregated to produce the outcome. The formula for the regression prediction is presented in the following equation.

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} h_b(x) \quad (2)$$

In Equation (2), $\hat{y}$ represents the final regression prediction, while B denotes the total number of trees, and $h_b(x)$ corresponds to the prediction made by the b_{th} decision tree for the input x. Through this approach, the Random Forest algorithm demonstrates strong performance in terms of both predictive accuracy and generalization capability.

3.3 Support Vector Regression

Support Vector Regression (SVR) is a supervised learning-based machine learning approach derived from the Support Vector Machine (SVM) algorithm (Jia et al., 2019). The fundamental components of SVR include a hyperplane supported by support vectors, as well as minimum and maximum margin lines, as illustrated in the following equation.

$$Y = f(x) = \omega\varphi(x) + b \quad (3)$$

In Equation (3), x represents the independent variable, while ω and b denote the weight vectors, and $\varphi(x)$ refers to the mapping function. When dealing with a multidimensional dataset, Y can have an infinite number of possible predictions. Therefore, to address the optimization problem presented in the following equation, a tolerance margin is introduced.

$$\min\left(\omega\frac{1}{2}\omega^2\right) + C\sum_{i=1}^{l}(\xi_i + \xi_i^*),$$
$$subjectto\begin{cases} y_i - \omega^T x_i - b \le \varepsilon + \xi_i \\ \omega^T x_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i\xi_i^* \ge 0 \end{cases} \quad (4)$$

Here, C represents a positive regularization parameter that balances the trade-off between prediction error and function smoothness, while ω denotes the penalty parameter, and ξ, ξ^* are slack variables used to minimize errors within the margin of the hyperplane. To solve the dual nonlinear problem, optimization methods can be reformulated as shown in the following equation.

$$\min : \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i.x_j) + \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{n} y_i(\alpha_i + \alpha_i^*)$$

$$subjectto\begin{cases}\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ \omega^T x_i + b - y_i \le \varepsilon + \xi_i^* \qquad for\ i = 1,2,\ldots.,n \\ 0 \le \alpha_i, \alpha_i^* \le C\end{cases} \tag{5}$$

Here, α_i and α_i^* denote the Lagrange multipliers, while $k(x_i.x_j)$ represents the kernel function used to address the nonlinear problem (Jia et al., 2019).

3.4 Cross Validation

Cross-validation is a computationally intensive technique that utilizes all available samples as both training and test data. In this method, the dataset is partitioned K times, ensuring the restructuring of training and test sets in each iteration. The algorithm undergoes repeated training K times, simulating the process by leaving out 1/K of the training samples in each step for testing purposes. Although such performance evaluation methods are computationally costly due to their iterative training processes, they aim to reduce the variance of predictions (Stone, 1974).

In practical applications, the dataset D is partitioned into K mutually exclusive subsets (or blocks) of approximately equal size ($m \approx n/K$). Here, Tk denotes the test set corresponding to the kth block, while Dk represents the training set obtained by excluding the elements in Tk. The cross-validation estimator is defined as the average error computed on the test set Tk for a model trained on the dataset formed by excluding each respective block Tk. The following equation formally represents this process.

$$CV(D) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m}\sum_{z_i \in T_k} L(A(D_k), z_i) \tag{6}$$

In this context, CV(D) denotes the mean cross-validation error for the dataset D. The parameter K represents the total number of cross-validation folds (e.g., K=10 for 10-fold cross-validation). Tk refers to the kth test set, meaning that each Tk in K iterations is a distinct subset used for testing. Correspondingly, Dk represents the training set obtained by excluding Tk, effectively comprising the remaining data after removing Tk from D. The variable mmm indicates the number of samples in each test set (i.e., in each block Tk). Finally, $L(A(D_k), z_i)$ denotes the error or loss function of the trained model $A(D_k)$ evaluated on the sample z_i within the test set (Bengio & Grandvalet, 2004).

3.5 Error Metrics

In this study, various metrics were employed to assess the performance of machine learning algorithms. Among these, the correlation coefficient (*R*) quantifies the strength of the linear relationship between predicted and actual values, ranging from [−1,1]. An *R* value close to 1 indicates a strong positive correlation, whereas a value near -1 signifies a strong negative correlation, and an *R* value of 0 suggests no correlation. The coefficient of determination (R^2) measures the proportion of the total variance in the target variable

explained by the model, taking values within the range [0,1]; a value approaching 1 indicates higher model performance (Magee, 1990). The mean squared error (MSE) represents the average of the squared differences between predicted and actual values, where a lower MSE suggests better predictive accuracy. The root mean square error (RMSE), obtained by taking the square root of the MSE, is more sensitive to larger deviations, making it a useful metric for evaluating error magnitude. The mean absolute error (MAE) is the mean of the absolute differences between predicted and actual values, providing a straightforward measure of prediction accuracy. The relative absolute error (RAE) expresses the MAE as a percentage by normalizing it with respect to the sum of deviations from the mean of the actual values. Similarly, the relative root square error (RRSE) normalizes the RMSE by dividing it by the standard deviation of the actual values, serving as a standardized error metric for model evaluation. These performance metrics play a crucial role in assessing both the accuracy and efficiency of machine learning models (James et al., 2021).

4. Results

As part of this study, a prototype scale solar tracking system was developed in Adıyaman, Türkiye. The system features dual-axis mobility, allowing adjustments of both azimuth and tilt angles, and it is equipped with additional hardware capable of recording meteorological data. Using this system, various environmental and operational parameters were recorded at different time intervals, including PM2.5, PM10, wind direction (hourly), wind speed (hourly), humidity, temperature, pressure, azimuth and tilt angles, and the generated voltage, current, and power. This comprehensive data collection system enabled an in-depth analysis of the relationship between solar power generation and meteorological variables.

4.1 Statistical evaluation of data

The primary motivation of this study was to model the power output of a photovoltaic (PV) panel under varying meteorological conditions and different azimuth and tilt angles for the Adıyaman province, Türkiye. In this context, the aim was to verify or recalculate the optimal tilt angle for improving solar energy performance by incorporating meteorological parameters into the analysis. Thus, 657 data points were collected. A statistical summary of the data is presented in Table 1.

Table 1 Statistics of the collected data.

Variables	Mean	Std. Dev.	Min	25th Percentile	Median	75th Percentile	Max
PM2.5 (μg/m³)	2.01	2.22	0.30	0.80	1.40	2.45	19.30
PM10 (μg/m³)	7.12	4.49	0.80	3.80	6.10	9.20	31.30
Wind Direction (°)	178.50	97.67	0.00	135.00	180.00	270.00	338.00
Wind Speed (m/s)	3.50	1.62	0.11	2.76	3.07	3.64	8.92
Humidity (%)	13.92	7.73	6.74	10.25	11.21	15.24	65.89
Temperature (°C)	40.83	8.34	10.13	40.75	43.13	45.63	47.50
Pressure (hPa)	930.15	2.24	926.32	928.34	930.16	931.56	938.98
Azimuth Angle (°)	181.32	12.63	145.00	180.00	180.00	180.00	255.00
Tilt Angle (°)	27.81	13.82	15.00	15.00	15.00	45.00	45.00
Voltage (V)	13.12	5.72	0.93	7.67	16.61	17.59	19.11
Current (A)	0.59	0.26	0.04	0.34	0.74	0.78	0.85
Power (W)	9.14	5.59	0.04	2.61	12.30	13.81	16.20

In Table 1, low levels of PM2.5 and PM10 positively influenced sunlight penetration to the panel, thereby supporting energy generation. However, to fully understand this effect, it was recognized that data should be collected over longer periods and across different seasons. Conversely, the high average temperature (40.83°C) had a negative impact on solar panel efficiency and energy output. The recorded wind speed (3.50 m/s) has contributed to panel cooling, potentially enhancing energy production; however, additional data

are required to confirm this relationship. Furthermore, low humidity levels (13.92%) provided favorable conditions for energy generation.

Looking at the table again, the azimuth angle has a positive impact on energy generation. This finding illustrates the importance of proper panel orientation in maximizing energy output. Similarly, the tilt angle was observed as a crucial parameter influencing energy production. A more thorough analysis of the relationship between tilt angle and energy output, however, necessitates a larger dataset. Figure 4 presents a graphical representation of energy production across varying panel angles.

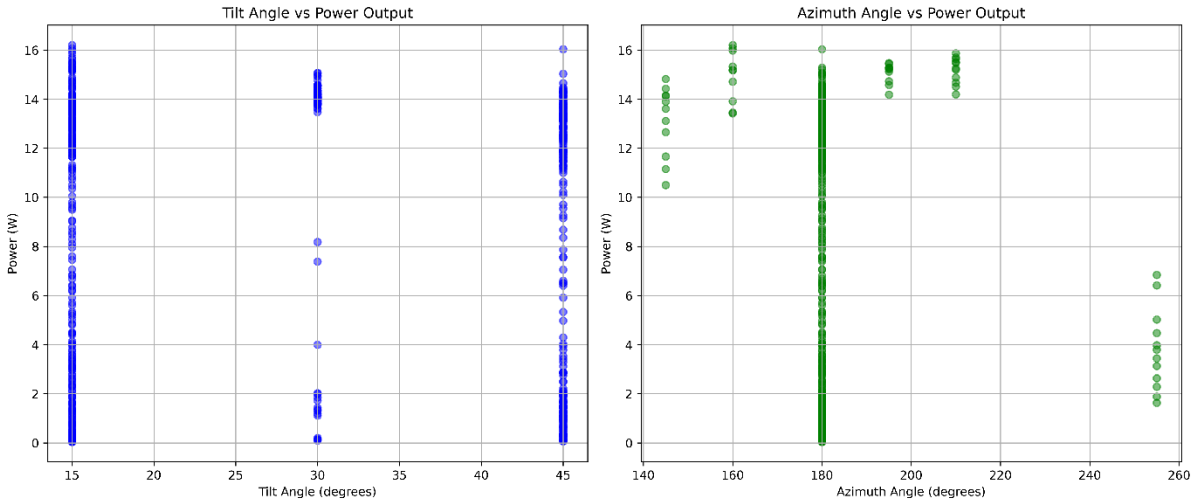


Figure 4 Energy densities produced at different angles.

In Figure 4, power values are predominantly concentrated at 15° and 45°, suggesting that the panel tilt angle was maintained within an optimal range, thereby enhancing energy production at specific angles. Similarly, in the azimuth angle-power graph, most generated power values are clustered around 180°, indicating that the panels were oriented southward, maximizing solar energy capture at this angle. Although energy production decreases as the azimuth angle deviates from 180°, the dataset generally suggests that the panels were operated at optimal angles, ensuring high efficiency. Overall, a strong correlation between power generation and panel angles is observed in both graphs, emphasizing the critical role of panel orientation in energy generation.

4.2 Modelling Results

In the study, regression analysis was basically conducted using three different machine learning algorithms (Gradient Boosting (GB), Random Forest (RF), and Support Vector Regression (SVR)) to estimate the energy output of the solar panel. The machine learning process involved two distinct modeling scenarios. In Scenario 1, meteorological parameters—including PM2.5, PM10, wind direction, wind speed, humidity, temperature, and pressure—along with panel tilt and azimuth angles were used as independent variables. This approach aimed to analyze the relationship between key meteorological factors influencing energy production and panel orientation. In Scenario 2, it was hypothesized that elevated particulate matter levels could negatively impact energy generation. To analyze this effect more thoroughly, aggregated PM2.5 and PM10 inputs were included as distinct, independent variables in the modeling methodologies. This approach allowed for a more comprehensive evaluation of the impact of particulate matter accumulation on energy production. Both scenarios were performed using the three selected machine learning algorithms, with independent variables structured differently for each approach. The results of Scenario 1 and Scenario 2 are presented in Table 2 and Table 3, respectively.

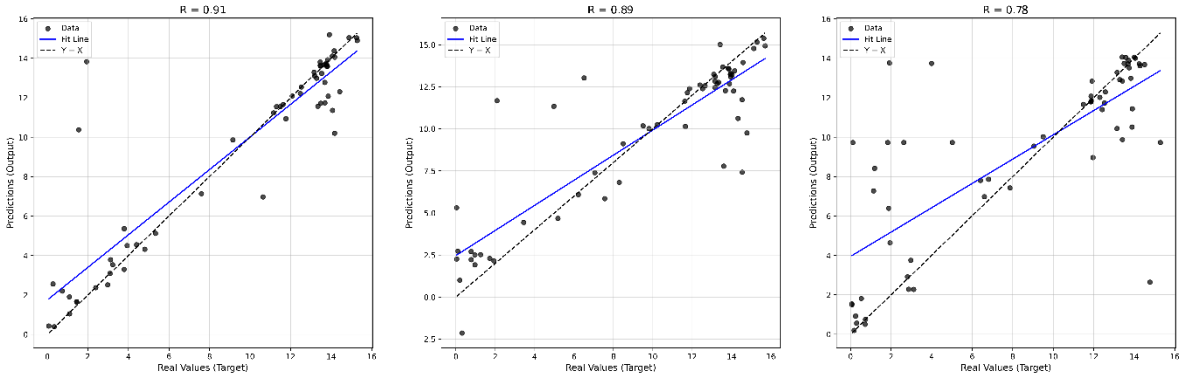
Table 2 Regression Analysis Results Using Meteorological Data & Panel Position Angles as Independent Variables: Scenario 1

	Random Forest (RF)		Gradient Boosting (GB)		Support Vector Regression (SVR)	
	Mean	Std	Mean	Std	Mean	Std
MSE	5.95	2.69	6.77	2.30	13.15	3.40
R	0.90	0.05	0.89	0.04	0.76	0.06
R2	0.81	0.09	0.78	0.07	0.57	0.11
MAE	1.19	0.29	1.56	0.23	2.10	0.34
RMSE	2.38	0.55	2.56	0.44	3.59	0.48
RAE	0.24	0.06	0.31	0.04	0.41	0.07
RRSE	0.48	0.06	0.55	0.04	0.65	0.08

Table 3 Regression Analysis Results with Cumulative Air Pollution Values Included as Independent Variables: Scenario 2

	Random Forest (RF)		Gradient Boosting (GB)		Support Vector Regression (SVR)	
	Mean	Std	Mean	Std	Mean	Std
MSE	5.32	2.13	6.24	2.63	12.00	3.47
R	0.91	0.04	0.89	0.05	0.78	0.07
R2	0.83	0.06	0.79	0.09	0.60	0.12
MAE	1.10	0.24	1.49	0.27	1.97	0.34
RMSE	2.26	0.48	2.44	0.52	3.43	0.51
RAE	0.21	0.04	0.30	0.05	0.39	0.07
RRSE	0.46	0.05	0.54	0.05	0.62	0.10

The results of Scenario 2 demonstrated better performance compared to Scenario 1, indicating that the inclusion of cumulative particulate matter (PM2.5 and PM10) data enhanced the model's predictive accuracy. As shown in Table 3, the R^2 values were calculated as 0.83, 0.79, and 0.60 for the RF, GB, and SVR algorithms, respectively. Among these, the RF algorithm exhibited the highest predictive performance, achieving an R^2 value of 0.83 and a correlation coefficient (R) of 0.91. A visual representation of R values is shown in Figure 5.



(a) (b) (c)

Figure 5 R values of (a) RF, (b) GB, and SVR algorithms for Scenario 2.

The RF algorithm's success may be abstracted as follows: the algorithm used the M5 tree model to automatically remove redundant linear characteristics and processed the dataset hierarchically from highest to lowest levels. Furthermore, linear functions were computed independently for each node in the RF tree structure, resulting in optimal prediction performance.

Although time-series data frequently need chronological splitting, the study used ten-fold cross-validation since the data were not chronologically continuous or autocorrelated but instead reflected a wide range of separate environmental circumstances. The successful results showed that including cumulative air pollution data marginally enhanced model performance. The metric findings show that the RF algorithm performs better during the modeling process. Figure 6 illustrates the performance of the three machine learning methods employed in Scenario 2.

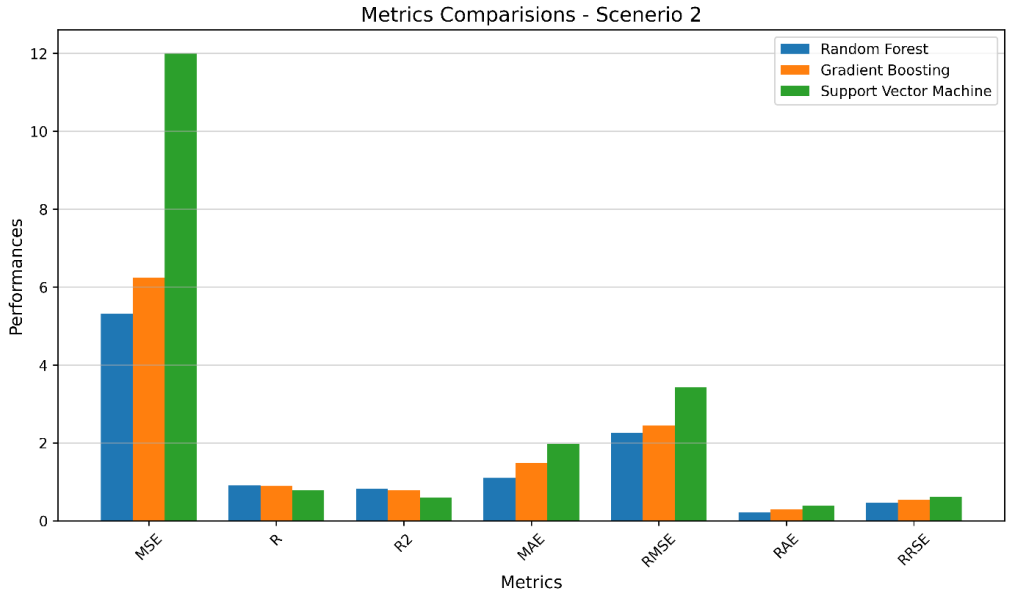


Figure 6 Metric comparisons.

Figure 7 depicts a comparison of the actual test data to the predicted values, including the related error values. This comparison demonstrates the RF algorithm's prediction accuracy in energy modeling, highlighting its usefulness as a machine learning tool for energy forecasting.

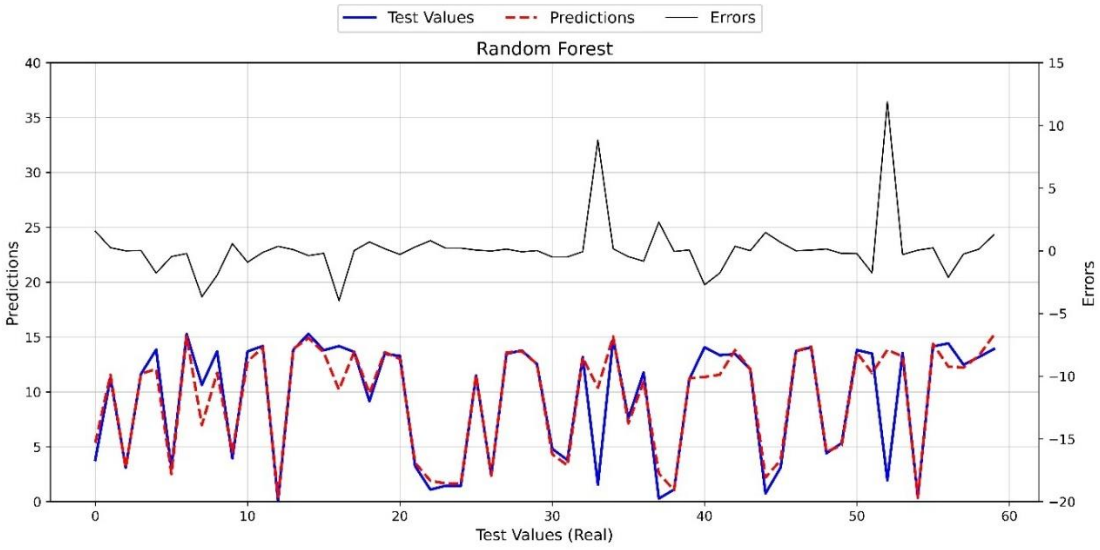


Figure 7 Comparison of actual-predicted and error values in the RF method.

Feature importance analysis and learning curve diagnostics were conducted to provide additional details about the learned relationships and to assess the physical interpretability of the succeeded model (Scenario 2). As shown in Figure 8, the feature importance plot reveals that pressure, temperature, and wind speed are the most influential variables affecting solar energy output, followed by humidity and azimuth angle. These results support the physical plausibility of the model's structure and confirm that meteorological and angular parameters are critical determinants in solar energy performance.

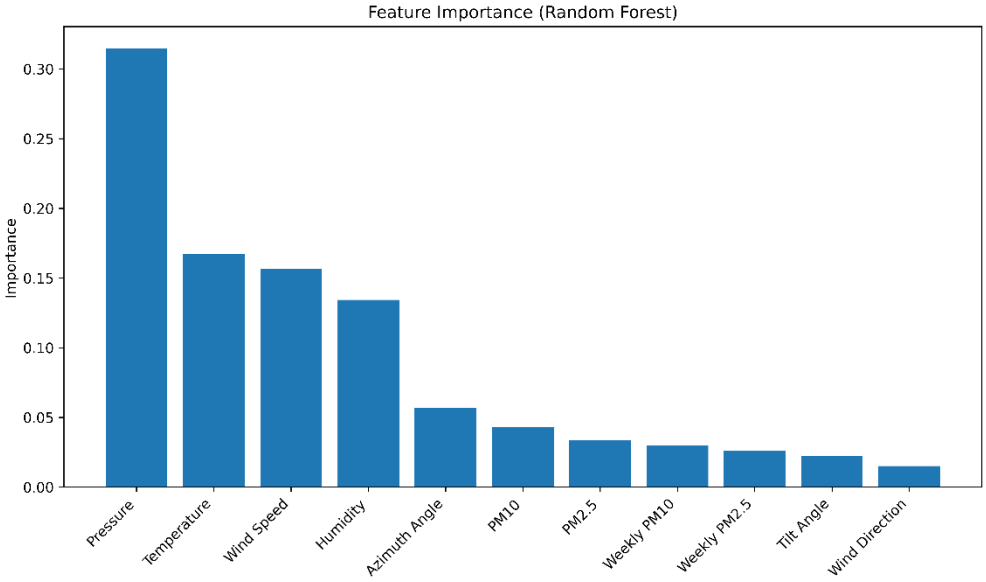


Figure 8 Feature importance (RF).

In addition, the learning curve presented in Figure 9 reveals the model's training and cross-validation R^2 scores across increasing training set sizes. The training performance remains consistently high, while the cross-validation scores indicate notable variance at lower sample sizes and modest improvement as more data are introduced. This suggests that although the model fits the training data well, generalization may be limited by the current dataset size, and additional data could enhance stability and performance.

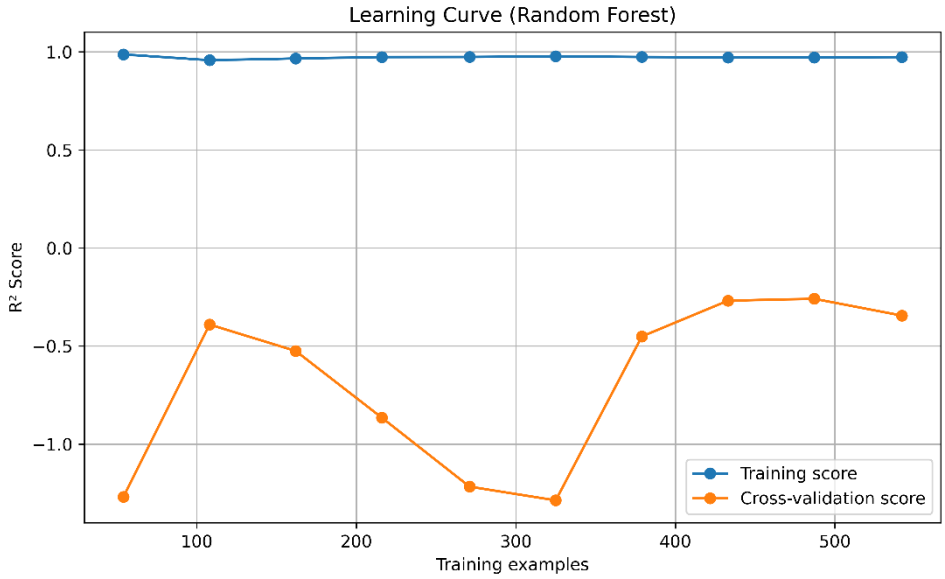


Figure 9 Learning curve (RF).

5. Discussion

In this study, the results indicated that Scenario 2 (cumulative particulate matter was included in the model as an independent variable) is better than Scenario 1, and the RF algorithm outperformed the others, demonstrating better predictive performance. The RF algorithm's ability to effectively capture complex and non-linear relationships within the dataset is primarily responsible for this success.

Although air pollution parameters do not have a direct significant impact on energy production, incorporating these variables into the model resulted in an increase in prediction accuracy. This suggests that particulate matter may indirectly influence solar radiation, thereby leading to variations in energy generation. The findings also reinforce the positive impact of physical parameters, such as panel tilt and azimuth angles, on energy generation. Furthermore, the inclusion of multiple meteorological variables significantly contributed to the model's success by enhancing prediction accuracy in solar energy forecasting.

This study's RF method achieved a high R^2 value of 0.83 with only 657 data points, in contrast to previous studies that relied on larger datasets and Artificial Neural Networks (ANN) (Kayri & Gencoglu, 2019; Keddouda et al., 2023; Postawa et al., 2024). Compared to ANN models, RF and GB provided advantages in terms of model interpretability and computational complexity for small-to-moderate datasets. Furthermore, these algorithms typically need less hyperparameter tuning effort. While ANN models are not necessarily more sophisticated than RF or GB models, their training frequently requires more computer resources and increases the danger of overfitting in the absence of sufficient data.

This study sets itself apart by integrating air pollution parameters (PM2.5, PM10) into the predictive framework, a feature often overlooked in comparable studies. Thus, despite its data set limitations, the methodology presented offers a novel and computationally efficient approach that can be extended and refined with more comprehensive data in future work.

Future studies will expand the dataset and conduct a more comprehensive evaluation by testing the models under varied environmental conditions. Additionally, the long-term relationships between air pollution and energy production will be analyzed, with a particular focus on assessing the impact of climate change in this context. To enhance the interpretability of model outputs, the inclusion of solar irradiance and panel temperature sensors is also planned, allowing for a clearer separation between particulate matter attenuation and temperature-induced efficiency losses.

The insights gained from these analyses will contribute to energy production, management, and policy development processes in Adıyaman. The findings of this investigation serve as an invaluable foundation for optimizing the efficient utilization of renewable energy resources and advancing the development of solar energy systems.

6. Conclusions

Recently, Adıyaman has hosted an increasing number of solar power plants that contribute to local energy production. Most of these plants are fixed systems. To improve solar energy production efficiency in the region while considering environmental conditions, there is a need to perform angle calculations for fixed panels using more innovative machine learning algorithms and a larger data set. The software and installations developed within the scope of this study will provide the infrastructure for these calculations, specifically for Adıyaman.

In this study, energy production forecasting models were developed by incorporating meteorological data and air pollution parameters specific to Adıyaman province, utilizing dual-axis solar panels. The data were analyzed under two distinct scenarios, and the model including cumulative air pollution factors as independent variables performed better. Three different machine learning algorithms—Random Forest (RF), Gradient Boosting (GB), and Support Vector Regression (SVR)—were evaluated, with RF demonstrating the highest predictive performance despite the limited data, achieving a high R^2 value and low error metrics.

This research stands out from many existing studies because it includes air pollution indicators in solar forecasting models and utilizes data from a specially designed dual-axis solar tracking system, which allows for a detailed assessment of how orientation affects energy output.

This study's ideas pave the way for more sustainable and efficient energy forecasting and renewable energy systems in Adıyaman. The methodological combination of underutilized ambient parameters, a localized dual axis tracking setup, and interpreted machine learning algorithms makes a new and useful contribution to the field. The findings provide useful insights that may be used to guide energy management and policy formation processes, as well as to contribute meaningfully to scientifically driven initiatives for encouraging the uptake and optimization of renewable energy resources.

Declaration of Ethical Standards

As the author of this study, I declare that I comply with all ethical standards.

Credit Authorship Contribution Statement

Software, Validation, Formal analysis, Writing -Original Draft, Visualization.

Declaration of Competing Interest

The author declared that he has no conflict of interest.

Funding / Acknowledgements

I am sincerely grateful to Research Assistant Mustafa Kaya and the students of the Department of Electrical and Electronics Engineering at Adıyaman University for their significant contributions to the hardware development and data collection stages of this research.

Data Availability

No datasets were generated or analyzed during the current study.

References

Al-Mohamad, A. (2004). Efficiency improvements of photo-voltaic panels using a Sun-tracking system. *Applied Energy*, *79*(3), 345–354. https://doi.org/10.1016/j.apenergy.2003.12.004

AL-Rousan, N., Mat Isa, N. A., Mat Desa, M. K., & AL-Najjar, H. (2021). Integration of logistic regression and multilayer perceptron for intelligent single and dual axis solar tracking systems. *International Journal of Intelligent Systems*, *36*(10), 5605–5669. https://doi.org/10.1002/int.22525

Ankarali, H., Temel, G. O., Tasdelen, B., & Ozge, A. (2012). Boosting Tree As a Stronger Approach in Classification: An Application of Carpal Tunnel Syndrome. *Journal of Turgut Ozal Medical Center*, *19*(4), 228–233. https://doi.org/10.7247/jtomc.19.4.5

Arslan, M., & Çunkaş, M. (2024). An experimental study on determination of optimal tilt and orientation angles in photovoltaic systems. *Journal of Engineering Research (Kuwait)*, *July*. https://doi.org/10.1016/j.jer.2024.07.015

Aslan, M., Ulum, T., & Türkmenler, H. (2021). Adıyaman İlinin Yenilenebilir Enerji Potansiyelinin Belirlenmesi Üzerine Bir Değerlendirme. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, *33*(1), 263–274. https://doi.org/10.35234/fumbd.791647

Bakırcı, K. (2009). A simple calculation method for estimation of instantaneous global solar radiation on horizontal surface. *Journal of Thermal Science and Technology*, *29*(2), 53–58.

Bayrakçı, H. C., & Gezer, T. (2019). Bir Güneş Enerjisi Santralinin Maliyet Analizi: Aydın İli Örneği. *Teknik Bilimler Dergisi*, *9*(2), 46–54. https://doi.org/10.35354/tbed.574190

Bengio, Y., & Grandvalet, Y. (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *Journal of Machine Learning Research*, *5*, 1089–1105.

Boyacı, Ö., & Kocaman, Ç. (2018). Matlab / Simulink Üzerinden Gerçek Zamanlı Gömülü Sistem Tabanlı Güneş Takip Sisteminin Tasarımı ve Uygulaması. *Anka E-Dergi*, *3*(1), 1–15. http://dergipark.gov.tr/anka/issue/36841/370502

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Demirtaş, M., Akkoyun, N., Akkoyun, E., & Çetinbaş, İ. (2019). Akıllı Şebekelerde Güneş Enerjisi Üretiminin Zamana Bağlı Olasılıksal Tahmini. *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji*, *7*(2), 411–424. https://doi.org/10.29109/gujsc.549704

Fahad, H. M., Islam, A., Islam, M., Hasan, M. F., Brishty, W. F., & Rahman, M. M. (2019). Comparative Analysis of Dual and Single Axis Solar Tracking System Considering Cloud Cover. *2019 International Conference on Energy and*

Power Engineering (ICEPE), 1–5. https://doi.org/10.1109/CEPE.2019.8726646

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in Python*. Springer Berlin Heidelberg.

Jia, W., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology*, *17*(1), 26–40. https://doi.org/https://doi.org/10.11989/JEST.1674-862X.80904120

Kaçan, E., & Ülgen, K. (2012). Güneş Enerjisi Toplayıcılarında Eğim ve Yönlendirmenin Yararlanabilirliğe Etkisi. *Journal of the Faculty of Engineering and Architecture of Gazi University*, *27*(4), 837–846.

Kaul, J. D., & Weed, G. D. (2021). Prototype Development for Adaptive Solar Tracking and Optimization of Data Communication Protocol. *American Society for Engineering Education*, 8.

Kayri, I., & Gencoglu, M. T. (2019). Predicting power production from a photovoltaic panel through artificial neural networks using atmospheric indicators. *Neural Computing and Applications*, *31*(8), 3573–3586. https://doi.org/10.1007/s00521-017-3271-6

Keddouda, A., Ihaddadene, R., Boukhari, A., Atia, A., Arıcı, M., Lebbihiat, N., & Ihaddadene, N. (2023). Solar photovoltaic power prediction using artificial neural network and multiple regression considering ambient and operating conditions. *Energy Conversion and Management*, *288*(April), 117186. https://doi.org/10.1016/j.enconman.2023.117186

Kılıç, B., & Kumaş, K. (2016). Burdur İli Güneşlenme Değerlerinin Yapay Sinir Ağları Metodu İle Tahmini. *SDÜ Teknik Bilimler Dergisi*, *6*(1), 38–44.

Magee, L. (1990). R2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests. *The American Statistician*, *44*(3), 250–253. https://doi.org/10.1080/00031305.1990.10475731

Oviedo, D., Romero-Ternero, M. C., Carrasco, A., Sivianes, F., Hernandez, M. D., & Escudero, J. I. (2013). Multiagent system powered by neural network for positioning control of solar panels. *IECON Proceedings (Industrial Electronics Conference)*, 3615–3620. https://doi.org/10.1109/IECON.2013.6699710

Oviedo, D., Romero-Ternero, M. C., Hernández, M. D., Sivianes, F., Carrasco, A., & Escudero, J. I. (2014). Multiple intelligences in a MultiAgent System applied to telecontrol. *Expert Systems with Applications*, *41*(15), 6688–6700. https://doi.org/10.1016/j.eswa.2014.04.048

Ozbeyaz, A., & Demirci, Y. (2019). Investigation of Correlation between the Solar Energy Efficiency And Air Pollution Using Decision Tree Algorithm. *4th International Engineering and Natural Sciences Conference*, 782–789.

Pierce, B. G., Braid, J. L., Stein, J. S., Augustyn, J., & Riley, D. (2022). Solar Transposition Modeling via Deep Neural Networks with Sky Images. *IEEE Journal of Photovoltaics*, *12*(1), 145–151. https://doi.org/10.1109/JPHOTOV.2021.3120508

Postawa, K., Czarnecki, M., Wrzesińska-Jędrusiak, E., Łyskawiński, W., & Kułażyński, M. (2024). Cascade-Forward, Multi-Parameter Artificial Neural Networks for Predicting the Energy Efficiency of Photovoltaic Modules in Temperate Climate. *Applied Sciences (Switzerland)*, *14*(7). https://doi.org/10.3390/app14072764

Schapire, R. E. (1990). The Strength of Weak Learnability. In *Machine Learning* (pp. 197–227).

Sharma, M. K., & Bhattacharya, J. (2020). A novel stationary concentrator to enhance solar intensity with absorber-only single axis tracking. *Renewable Energy*, *154*, 976–985. https://doi.org/10.1016/j.renene.2020.03.064

Shim, J., Park, S., & Song, D. (2025). Impact of particulate matter (PM10, PM2.5) on global horizontal irradiance and direct normal irradiance in urban areas. *Building and Environment*, *271*(September 2024), 112610. https://doi.org/10.1016/j.buildenv.2025.112610

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Temel, G. O., Ankaralı, H., Taşdelen, B., Erdoğan, S., & Özge, A. (2014). A Comparison of Boosting Tree and Gradient Treeboost Methods for Carpal Tunnel Syndrome. *Turkey Clinicians Journal of Biostatistics*, *6*(2), 67–73.

Uz, Ö., Özdemir, T., & Özmen, Ö. T. (2022). Estimating Energy Production of Solar Power Plant at the University of Bakırçay Using Artificial Neural Networks Based on Meteorological Conditions. *Artificial Intelligence Theory and Applications*, *2*(1), 27–40.

Wu, Y., & Zhou, J. (2019). Risk assessment of urban rooftop distributed PV in energy performance contracting (EPC) projects: An extended HFLTS-DEMATEL fuzzy synthetic evaluation analysis. *Sustainable Cities and Society*, *47*, 101524. https://doi.org/10.1016/j.scs.2019.101524

Yadav, A. K., & Chandel, S. S. (2013). Tilt angle optimization to maximize incident solar radiation: A review. *Renewable and Sustainable Energy Reviews*, *23*, 503–513. https://doi.org/10.1016/j.rser.2013.02.027

Zhou, X., Cao, Z., Ma, Y., Wang, L., Wu, R., & Wang, W. (2016). Concentrations, correlations and chemical species of PM2.5/PM10 based on published data in China: Potential implications for the revised particulate standard. *Chemosphere*, *144*, 518–526. https://doi.org/10.1016/j.chemosphere.2015.09.003