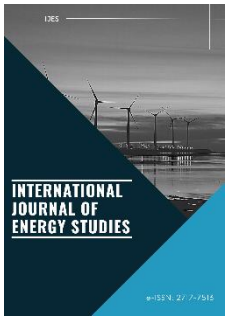
INTERNATIONAL JOURNAL OF ENERGY STUDIES

e-ISSN: 2717-7513 (ONLINE); homepage: https://dergipark.org.tr/en/pub/ijes



Research Article

Int J Energy Studies 2025; 10(3): 711-742

DOI: 10.58559/ijes.1633454

Received : 05 Feb 2025

Revised : 11 Apr 2025

Accepted : 03 July 2025

Machine learning-based prediction of solar radiation in the Southeastern Anatolia Region of Türkiye

Abdallah Adil Awad Bashir[a*], Abdulkadir Koçer[b], Ahmet Çoşgun[c], Afşin Güngör[d]

[a] *Department of Mechanical Engineering, Faculty of Engineering, Akdeniz University, Antalya, 07070, Türkiye, ORCID: 0000-0003-2372-8777*

[b] *Department of Energy, Vocational School of Technical Science, Akdeniz University, Antalya, 07070, Türkiye, ORCID: 0000-0002-5139-421X*

[c] *Department of Mechanical Engineering, Faculty of Engineering, Akdeniz University, Antalya, 07070, Türkiye, ORCID: 0000-0002-0243-5476*

[d] *Department of Mechanical Engineering, Faculty of Engineering, Akdeniz University, Antalya, 07070, Türkiye, ORCID: 0000-0002-4245-774*

*(*Corresponding Author: abduadil159@gmail.com)*

Highlights

- Data spanning 18 years from nine districts in Southern Anatolia were collected from the NSRDB database.
- Tree-based models outperformed other approaches in predicting both instantaneous and daily solar radiation.
- The Extra Trees model achieved the highest accuracy, with R^2 scores above 0.999 for daily GHI and 0.975 for daily DNI.
- The models performed better in forecasting GHI than DNI, indicating greater challenges in DNI prediction.

You can cite this article as: Bashir AAA, Kocer A, Çoşgun A, Güngör A. Machine learning-based prediction of solar radiation in the Southeastern Anatolia Region of Türkiye. Int J Energy Studies 2025; 10(3): 711-742.

ABSTRACT

Solar energy systems play a vital role in alleviating the potential environmental risks that arise from using conventional energy sources. Since the performance of these systems relies heavily on solar radiation, it is crucial to develop reliable tools for accurate solar radiation forecasting. This study investigates the utilization of supervised machine learning models for predicting solar radiation in the Southern Anatolian Region in Türkiye. Nine different models were used to predict both instantaneous and daily solar radiation in the study area, based on 18 years (2005–2022) of weather data obtained from the NSRDB database. The results showed that the tree-based models had better performance than other models evaluated. Moreover, the extra trees model was found to have the best performance, with R^2 scores above 0.999 for daily global horizontal irradiation, 0.975 for daily direct normal irradiation, 0.955 for instantaneous global horizontal irradiation, and 0.945 for instantaneous direct normal irradiation. Moreover, the extra trees model achieved its highest accuracy when predicting the daily global horizontal irradiation, with a station-wise average R^2 score of 0.9999, root mean squared error of 0.0244, mean absolute error of 0.0142, and mean absolute scaled error of 0.0047.

Keywords: Solar radiation, Machine learning, Weather forecasting

1. INTRODUCTION

Electricity consumption rates are rising rapidly worldwide, leading to critical environmental issues such as air pollution and global warming, primarily due to dependence on conventional energy sources. Consequently, transitioning to renewable energy sources is both crucial and inevitable [1]. Among these, solar energy represents an abundant and versatile option that can be used for electricity generation through photovoltaic systems or thermoelectric conversion plants. Additionally, solar energy has the potential to meet residential hot water demands and provide heat energy for industrial processes. Solar radiation constitutes the primary resource for solar systems, and therefore, accurate measurement or estimation is crucial for determining their output and efficiency [2].

Machine learning models consist of a variety of algorithms designed to identify patterns in provided data and utilize these relationships to predict desired outputs. Depending on the context of the problem, these models are categorized as supervised or unsupervised and can be applied to tasks such as classification or regression [3]. To enhance the performance of these models, techniques such as feature selection, encoding, and scaling are applied to preprocess and manipulate raw data. The models are then trained, validated, and evaluated using various statistical metrics [4-6]. Long-term weather data can be obtained from various sources, including satellite imagery, meteorological stations, and hybrid satellite-physical models. This data includes parameters representing weather conditions for each timestep, such as air temperature, relative humidity, and solar radiation [4].

A body of literature has explored the use of machine-learning models for forecasting solar radiation. Ercan U. and Kocer A. investigated the application of machine learning models to predict daily global solar irradiation in eight cities in the Mediterranean Region of Türkiye. Using satellite weather data, their study concluded that machine-learning methods outperform statistical methods in solar radiation forecasting [7]. Ağbulut Ü. et al. examined the performance of various models in forecasting daily global solar irradiation (GHI) using meteorological station data from four cities in Türkiye: Tokat, Kırklareli, Nevşehir, and Karaman. In their study, four models were evaluated: support vector machine, artificial neural network, k-nearest neighbors, and deep learning. The findings revealed that all the models performed well, with the artificial neural network achieving the best data fit [8]. A. B. Guher et al. conducted a study evaluating three machine learning models to predict hourly solar radiation in Kahramanmaraş and Isparta, Türkiye. Feature selection techniques were applied to identify the most relevant weather parameters, and the study concluded that the Multilayer Feed-Forward Neural Network outperformed the other

models [2]. In Zonguldak, a city in northern Türkiye, Hacıoğlu R. investigated the use of multiple and Gaussian linear regression models to predict hourly solar radiation based on meteorological station data. The findings confirmed the good performance of the models, as their prediction errors were relatively small [9]. Demirgül T. et al. employed a heuristic regression technique known as the M5-tree to estimate the monthly average solar radiation in Türkiye, using HELIOSTAT data from 2004 to 2018. The study demonstrated that the applied method could serve as a viable alternative to traditional approaches [10]. Demir V. and Citakoğlu H. used data from 163 meteorological stations across Türkiye to investigate the performance of machine learning models for predicting daily GHI. The results of various tests revealed that the Long Short-Term Memory and Gaussian Process Regression models outperformed the others, demonstrating their applicability in providing accurate estimates of solar radiation in Türkiye [11]. Demirgül T. et al. employed multivariate adaptive regression splines (MARS) and least-squares support vector regression (LSSVR) models to estimate annual solar radiation in Türkiye, utilizing data from 3,600 grid points provided by HELIOSTAT. Following the estimations, they produced a solar radiation map of Türkiye using interpolation techniques. The study concluded that the LSSVR model yielded highly satisfactory results, underscoring the effectiveness of machine learning methods in estimating solar radiation [12]. For Morocco, Mendyl A. et al. applied Long Short-Term Memory (LSTM), Support Vector Machine (SVM), and Multilayer Artificial Neural Network (MLANN) models to predict hourly solar radiation. The results showed that the LSTM model outperformed the other models, highlighting the potential of machine learning algorithms for solar radiation prediction [13].

Boosting and bagging tree-based ensemble models are known to provide higher accuracy than traditional decision tree models. In a study conducted by Toylan A., a bagging decision tree-based machine learning algorithm was used to predict hourly solar radiation at Kırklareli University in Türkiye. The proposed method outperformed traditional decision tree models, and the results indicated that it can be used for solar radiation prediction with acceptable accuracy [14]. Tercha W. et al. examined the applicability of machine learning algorithms for predicting solar radiation in Laghouat, Algeria. The evaluation showed that the ensemble models used, namely Random Forest and XGBoost, achieved higher accuracy than the Support Vector Machine (SVM) model in predicting both temperature and solar radiation [15]. Ahmad M. W. et al. conducted a comparative analysis of Random Forest (RF) and Extra Trees (ET) ensemble models against Support Vector Regression (SVR) for forecasting solar PV production, evaluating their accuracy, stability, and computational efficiency. The findings indicated that the ET model outperformed both RF and

SVR, making it a highly effective option for solar PV production prediction [16]. Hassan M. A. et al. explored the potential of gradient boosting, bagging, and Random Forest (RF) models for solar radiation prediction, comparing them with multi-layer perceptron (MLP) and support vector regression (SVR). The study concluded that bagging and RF models outperformed gradient boosting, with all three models surpassing MLP in performance [17]. Ahmad M. W. et al. conducted a similar study to compare four regression models: Support Vector Regression (SVR), Random Forest (RF), Extra Trees (ET), and Decision Trees (DT) for predicting the output of solar thermal systems. The results showed that RF and ET achieved comparable accuracy, outperforming the other models [18]. Based on this summary, it is concluded that using boosting and bagging techniques and utilizing the models that employ them such as XGBoost, Extra Trees, and Random Forest could result in high prediction accuracy for solar radiation.

Existing literature demonstrates the strong potential of machine learning models for forecasting both daily and instantaneous solar radiation. This study aims to build on this research by evaluating the capability of several models to predict both instantaneous and daily solar radiation in nine districts in southeastern Türkiye, providing valuable insights into the solar energy field. Applying machine learning models to both instantaneous and daily solar radiation can clearly highlight their strengths and weaknesses, offering useful guidance for developers and researchers. Moreover, this study utilizes the widely used NSRDB dataset for solar radiation prediction, making the proposed methodology generalizable to any location covered by the NSRDB service.

2. MATERIALS AND METHODS

To assess the capability of different machine learning models for the estimation of solar radiation, this study applied a four-stage methodology comprised of data acquisition, data processing and feature engineering, model training, and model evaluation. This methodology is widely used in literature, with some differences based on the context of the problem [3, 8]. Fig. 1 explains and summarizes the methodology followed.

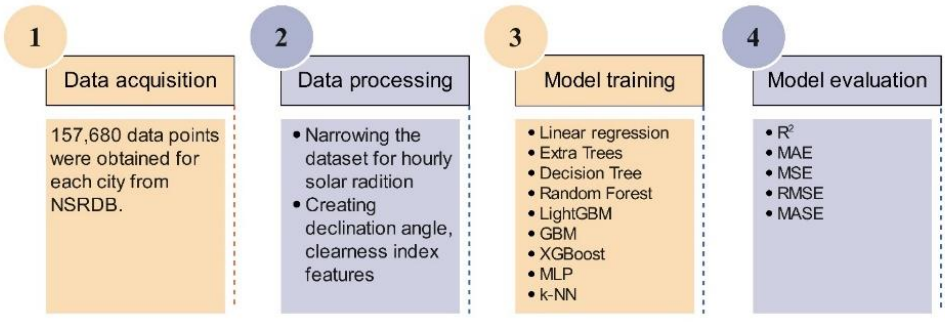


Figure 1. Methodology description

2.1. Data Acquisition

The study area (Southeastern Anatolian Region) consisted of nine districts, namely Adıyaman, Batman, Diyarbakır, Gaziantep, Kilis, Mardin, Siirt, Şanlıurfa, and Şırnak. As shown in Fig. 2, this region is distinguished by elevated direct normal irradiation (DNI) and global horizontal irradiation (GHI) values, which signifies its suitability for the development of solar energy infrastructure [19-22].

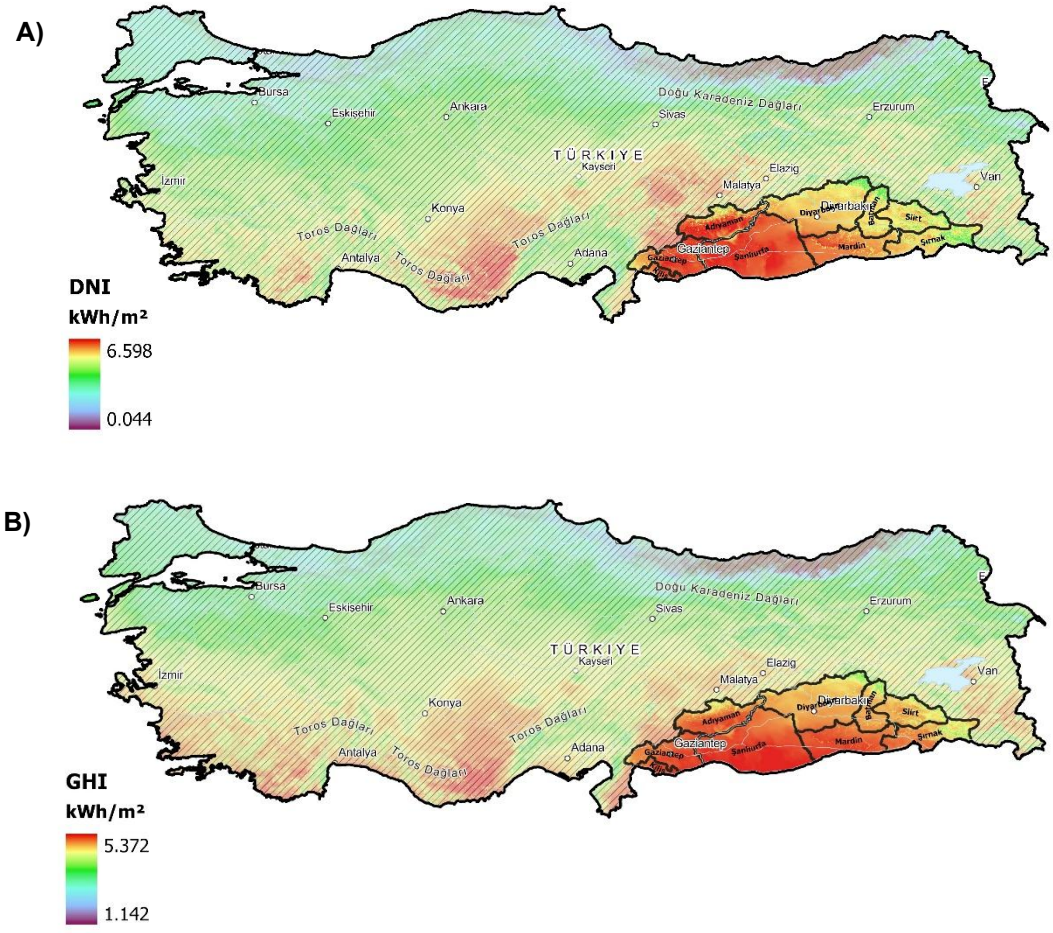


Figure 2. Daily average solar radiation in the study area [19]. A) DNI, B) GHI

The hourly meteorological data for each district was obtained from the National Solar Radiation Database (NSRDB) over a period of time covering 18 years, from 2005 to 2022 [23]. The year 2022 was used as it is the most recent year currently available in the NSRDB database. The data were generated by the NSRDB model, which combines satellite images and sophisticated algorithms to calculate solar radiation components [24]. The METEOSAT IODC satellite, which was used to generate this data, has a temporal resolution of 60 minutes and a spatial resolution of 4 km [23]. The NSRDB model was validated against actual surface observations, and the findings

indicated its capacity to forecast GHI with a mean percentage bias of 5% and DNI with a mean percentage bias of 10%. The NSRDB model's reputation for robustness and reliability has led to the adoption of its data by numerous prominent energy modelling and solar system design software programs. Examples include PVSyst, EnergyPlus, SAM, and PVWatts [25]. Furthermore, the weather data generated by this model is widely used for forecasting solar radiation with different machine-learning algorithms. Allal Z. et al. utilized NSRDB data to predict solar radiation in Izmir, Türkiye, employing a set of supervised machine learning models [3]. Moreover, Mukherjee A. et al. utilized the NSRDB data to predict the solar radiation in Kharagpur, India using deep neural networks [26]. Another example of NSRDB data usage is the study by Narvaez G. et al., where it was combined with in-situ radiation measurements to forecast solar radiation in Nariño, Colombia using deep learning techniques. The proposed method outperformed traditional approaches in terms of prediction accuracy [27].

A total of 157,680 observations (24 hours * 365 days * 18 years) were obtained from the NSRDB website for each city. The datasets were retrieved using the specific latitude and longitude coordinates of each city, ensuring that the data represents point-based measurements rather than regional aggregates. The observations included information on clear sky radiation, all-sky radiation, dew point and dry bulb temperatures, wind speed and direction, precipitable water, relative humidity, atmospheric pressure, hourly solar zenith angle, surface albedo, and cloud type. Based on their properties, clouds are classified into 12 types, as follows: 0 for clear, 1 for probably clear, 2 for fog, 3 for water, 4 for super-cooled water, 5 for mixed, 6 for opaque ice, 7 for cirrus, 8 for overlapping, 9 for overshooting, 10 for unknown, 11 for dust, and 12 for smoke [28]. For more information about the data sources and variables, refer to Refs. [23, 29].

2.2. Data processing and feature engineering

To improve the accuracy of machine learning models, raw data needs to be processed before being fed into the models. This process involves removing outliers, performing feature engineering to create more informative features, and encoding categorical variables. In this study, data preprocessing included narrowing the dataset used for predicting instantaneous solar radiation components to mitigate the impact of zero values on prediction accuracy. It also involved generating features such as declination angle and clearness index and encoding the categorical variable of cloud type.

The features used for predicting instantaneous solar radiation include temperature (T), wind speed (W), declination angle (δ), solar zenith angle (z), atmospheric pressure (P), relative humidity (H),

and cloud type. For daily solar radiation prediction, the daily averages of these variables are utilized, with cloud type replaced by the clearness index (K_T). The selected features are based on domain expertise and the characteristics of the datasets used. They represent both meteorological conditions (T, W, P, H, K_T, and cloud type) and temporal information (z and δ). As a result, these features effectively reflect the variations in instantaneous and daily solar radiation, given the nature of the dataset.

Rather than delineating the timestamp of data by month and day, the declination angle, which has a unique value for each day of the year, can be utilized. This angle is defined as the angular distance of the sun's rays north or south to the equator. It ranges from 23.45º (north) to -23.45º and depends only on the day number (n), as Eq.1 shows [30].

$$\delta = 23.45 \sin\left(\frac{360}{365}(284 + n)\right) \quad (1)$$

For instantaneous solar radiation, the categorical cloud type variable was encoded with a widely used one-hot encoding method. This technique can lead to an improvement in the performance of the machine-learning models by increasing the interpretability of the training data [31]. For the daily solar irradiation, on the other hand, the clearness of the sky can be determined using the daily clearness index (K_T), which is the ratio of the daily GHI (H) to the daily extraterrestrial horizontal irradiation (H_o), as explained by Eq. 2 [30].

$$K_T = \frac{H}{H_o} \quad (2)$$

The daily extraterrestrial horizontal irradiation (H_o) is defined as the daily cumulative amount of solar radiation that is incident on a horizontal plane outside the atmosphere [32]. It can be calculated using Eq. 3, which is applicable when specific values of day number, latitude, sunset hour angle, and declination angle are given [32].

$$H_o = \frac{24 \times 3600 G_{sc}}{\pi}\left(1 + 0.033 \cos\left(\frac{360n}{365}\right)\right) \times \left(\cos\phi \cos\delta \sin\omega_s + \frac{\pi\omega_s}{180}\sin\phi \sin\delta\right) \quad (3)$$

where G_{sc} is the solar constant (1367 W/m^2), ϕ is the latitude angle, ω_s is the sunset hour angle, which can be calculated using Eq. 4 [32].

$$\omega_s = \cos^{-1}(\tan\phi \tan\delta) \quad (4)$$

In order to enhance the performance of machine learning models for predicting instantaneous solar radiation, the analysis was limited to the period between hours 6-18. This restriction was necessitated by the fact that the weather data for cities included a high number of zero values for GHI and DNI, with the majority falling outside the specified period. By applying this limitation, a total of 85,410 observations for each city were used for training the models to predict instantaneous solar radiation components (GHI and DNI). For the daily solar irradiation components (GHI_{daily} and DNI_{daily}), the weather dataset consisted of 6,570 observations (365 days * 18 years). For the purposes of model evaluation, a total of 30% of the dataset was utilized for testing, with the remaining 70% allocated for training. A detailed description of the data used is provided in Tables 1 and 2.

Table 1. Description of the dataset of the instantaneous solar radiation prediction

City		T °C	W m/s	δ	z	P mbar	H %	GHI (W/m^2)	DNI (W/m^2)
Kilis (36.71° N, 37.11° E)	*Mean*	19.45	2.91	0	61.55	933.52	49.17	388.79	424.18
	Min	-8.70	0.20	-23.45	13.31	913	4.11	0	0
	Max	44	10.10	23.45	116.39	952	100	1084	1039
	Std	10.84	1.40	16.58	23.33	5.41	24.45	325.21	371.15
Gaziantep (37.06° N, 37.37° E)	*Mean*	18.49	3.01	0	61.72	913.94	47.71	382.83	427.56
	Min	-13.20	0	-23.45	13.61	893	3.80	0	0
	Max	42.90	11.70	23.45	116.65	932	100	1079	1047
	Std	11.36	1.49	16.58	23.31	5.31	26.09	324.26	379.99
Adıyaman (37.76° N, 38.27 ° E)	*Mean*	19.70	2.72	0	62.15	934	40.88	381.72	432.92
	Min	-12.90	0.20	-23.45	14.33	912	4.11	0	0
	Max	44.70	13	23.45	117.45	950	96.12	1081	1035
	Std	11.70	1.54	16.58	23.32	4.91	22.43	325.69	378.40
Şanlurfa (37.16° N, 38.79° E)	*Mean*	21.45	2.91	0	61.95	949.77	40.66	384.92	422.25
	Min	-8.10	0.10	-23.45	13.75	931	3.62	0	0
	Max	46.40	10.10	23.45	117.78	969	100	1076	1037
	Std	11.72	1.48	16.58	23.54	6.12	23.91	323.38	372.41
Diyarbakır (37.92° N, 40.21° E)	*Mean*	20.60	1.78	0	62.47	936.19	40.48	374.52	420.86
	Min	-14.10	0.10	-23.45	14.61	918	2.43	0	0
	Max	46.30	7.40	23.45	119	955	100	1069	1033
	Std	12.36	0.96	16.58	23.65	5.69	24.54	322.88	378.53
Mardin (37.31° N, 40.73° E)	*Mean*	19.61	1.45	0	62.29	911.07	41.71	386.07	429.98
	Min	-14.40	0.10	-23.45	14.09	895	2.63	0	0
	Max	44.50	5.80	23.45	119.35	929	100	1078	1043
	Std	12.13	0.76	16.58	23.88	5.71	26.32	327.69	380.26
Batman (37.88° N, 41.12° E)	*Mean*	20.91	1.49	0	62.59	937.35	40.26	375.43	417.81
	Min	-14.10	0	-23.45	14.70	920	3.40	0	0
	Max	46.30	6.60	23.45	119.72	956	100	1064	1022
	Std	12.29	0.83	16.58	23.84	5.67	23.95	324.49	375.93
Siirt (37.92° N, 41.94° E)	*Mean*	18.81	1.78	0	62.72	899.33	42.52	375.75	412.29
	Min	-16.80	0	-23.45	14.88	880	4.76	0	0
	Max	44.10	8.80	23.45	120.35	916	100	1074	1039

	Std	12.01	0.98	16.58	24	4.96	24.46	326.21	382.78
Şırnak (37.51° N, 42.45° E)	*Mean*	15.24	2.26	0	62.63	831.95	46.91	388.16	436.07
	Min	-19.70	0.10	-23.45	14.60	814	3.52	0	0
	Max	40.20	10.20	23.45	120.72	847	100	1098	1085
	Std	11.88	1.25	16.58	24.20	4.53	28.89	331.92	398.05

* Cloud Type is a categorical feature and takes values from 0 to 12, as mentioned before

Table 2. Description of the dataset of the daily solar irradiation prediction

City		**T_{avg} °C**	**W_{avg} m/s**	**δ**	**P_{avg} mbar**	**H_{avg} %**	**K_T**	**GHI_{daily} (kWh/m²)**	**DNI_{daily} (kWh/m²)**
Kilis (36.71° N, 37.11° E)	*Mean*	15.99	2.48	0	933.62	62.99	0.66	5.10	5.65
	Min	-4.30	0.72	-23.45	916.92	12.11	0.01	0.07	0
	Max	34.77	6.75	23.45	950.25	99.58	1.33	9.32	12.11
	Std	9.26	0.93	16.58	5.17	18	0.36	2.60	3.60
Gaziantep (37.06° N, 37.37° E)	*Mean*	15.02	2.82	0	914.02	60.78	0.66	5.03	5.71
	Min	-7.18	0.72	-23.45	897.17	12.21	0.01	0.07	0
	Max	33.80	7.88	23.45	929.92	100	1.33	9.31	12.19
	Std	9.79	1.04	16.58	5.08	20.46	0.36	2.61	3.68
Adıyaman (37.76° N, 38.27 ° E)	*Mean*	16.44	2.48	0	934.03	49.91	0.66	5.02	5.79
	Min	-7.06	0.74	-23.45	916.92	9.65	0.01	0.06	0
	Max	35.97	10.42	23.45	948.83	89.92	1.34	9.25	11.96
	Std	10.30	1.12	16.58	4.70	19.17	0.37	2.62	3.63
Şanlurfa (37.16° N, 38.79° E)	*Mean*	17.94	2.67	0	949.83	51.45	0.66	5.07	5.66
	Min	-3.97	0.71	-23.45	935.42	9.66	0.01	0.07	0
	Max	37	7.86	23.45	967.71	98.71	1.32	9.21	11.92
	Std	10.37	1.03	16.58	5.90	21.04	0.36	2.57	3.58
Diyarbakır (37.92° N, 40.21° E)	*Mean*	16.93	1.58	0	936.25	50.24	0.65	4.94	5.66
	Min	-8.58	0.48	-23.45	922	7.70	0.01	0.07	0
	Max	36.93	5.68	23.45	953.04	95.90	1.34	9.20	12.01
	Std	10.90	0.56	16.58	5.47	22.24	0.36	2.58	3.59
Mardin (37.31° N, 40.73° E)	*Mean*	15.89	1.32	0	911.09	52.01	0.67	5.10	5.80
	Min	-11.40	0.40	-23.45	896.38	8.77	0.01	0.07	0
	Max	35.51	4.71	23.45	927.71	100	1.33	9.27	12.21
	Std	10.63	0.46	16.58	5.51	23.97	0.36	2.61	3.61
Batman (37.88° N, 41.12° E)	*Mean*	17.45	1.33	0	937.39	48.88	0.66	4.96	5.63
	Min	-10.42	0.43	-23.45	922.71	7.37	0.01	0.06	0
	Max	37.54	5.49	23.45	954.17	96.17	1.33	9.19	11.97
	Std	10.88	0.46	16.58	5.46	21.88	0.37	2.60	3.60
Siirt (37.92° N, 41.94° E)	*Mean*	15.68	1.54	0	899.31	50.47	0.66	4.97	5.58
	Min	-11.86	0.41	-23.45	883.58	9.18	0.01	0.07	0
	Max	35.73	6.53	23.45	914.25	98.46	1.35	9.31	12.26
	Std	10.77	0.50	16.58	4.78	22.21	0.37	2.61	3.68
Şırnak (37.51° N, 42.45° E)	*Mean*	12.1	1.79	0	831.89	54.55	0.67	5.14	5.92
	Min	-15.4	0.56	-23.45	816.13	6.61	0.01	0.07	0
	Max	31.93	6.76	23.45	845.46	100	1.37	9.52	12.62
	Std	10.61	0.52	16.58	4.36	26.48	0.36	2.61	3.80

The correlation analysis presented in Fig. 3 illustrates the linear relationships between the input features and the target variables. For instantaneous DNI (Fig. 3A), the cloud type variable exhibits the strongest correlation, followed by relative humidity (H) and the zenith angle (z). In contrast, for GHI (Fig. 3B), the zenith angle (z) shows the highest correlation, followed by relative humidity (H) and cloud type. Wind speed (W) and atmospheric pressure (P) are found to have the weakest correlations with both target variables (DNI and GHI). A positive correlation indicates a direct (proportional) relationship between two variables, while a negative correlation implies an inverse (opposite-proportional) relationship.

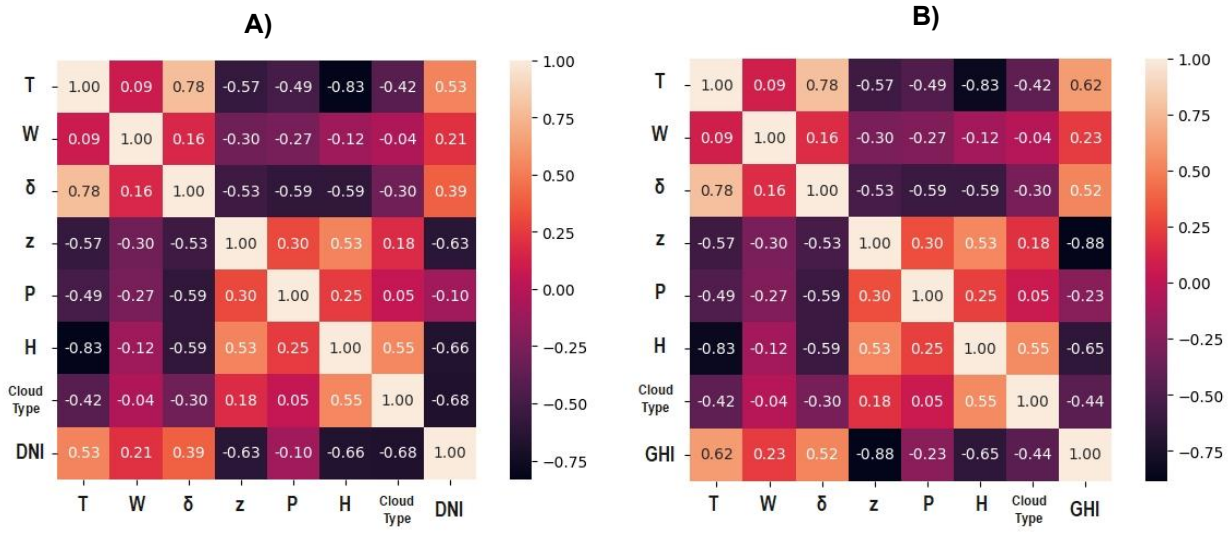


Figure 3. Correlation heat maps for instantaneous solar radiation components (Adıyaman city). A) DNI, B) GHI

Fig. 4 demonstrates the linear relationships between the input features and the daily radiation components. For daily DNI (DNI_{daily}), the clearness index (K_T) shows the strongest positive correlation (0.86), followed by the average temperature (T_{avg}) (0.61) the solar declination angle (δ) (0.58). Relative humidity (H_{avg}) and atmospheric pressure (P_{avg}) exhibit the most negative correlations with DNI_{daily}, at –0.78 and –0.19, respectively. In the case of daily GHI (GHI_{daily}), the highest correlation is also observed with K_T (0.99), followed by δ (0.87) and T_{avg} (0.79). Similar to DNI_{daily}, H_{avg} and P_{avg} show the most negative correlations with GHI_{daily}, with values of –0.77 and –0.42, respectively. For both the target variables (DNI_{daily} and GHI_{daily}), the average wind speed (W_{avg}) has a low correlation compared to other features.

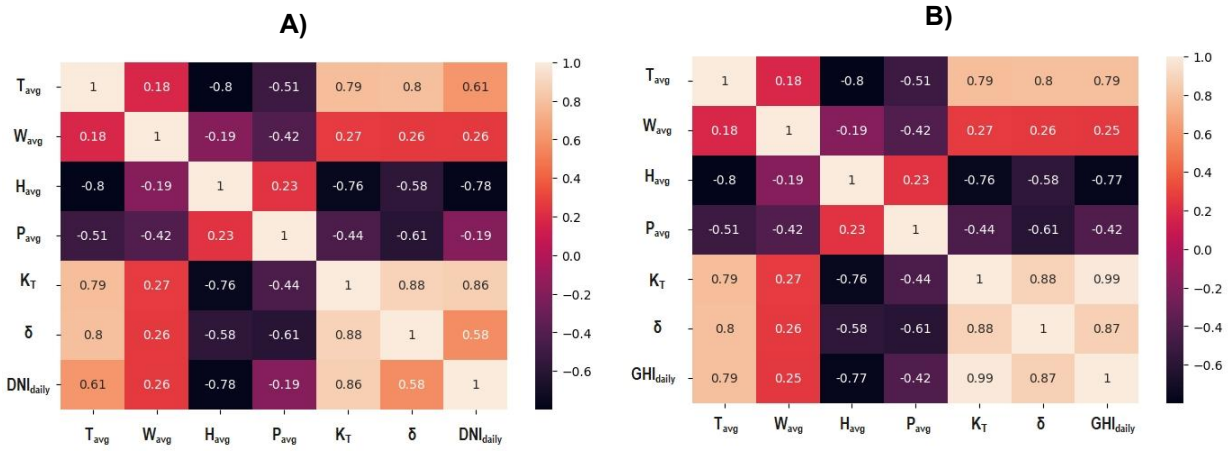


Figure 4. Correlation heat maps for daily solar radiation components (Adıyaman city). A) DNI, B) GHI

2.3. Model Selection and Training

In this study, a total of nine supervised machine-learning models categorized in four groups were evaluated to assess their ability to forecast instantaneous and daily solar radiation components (GHI, GHI_{daily}, DNI, and DNI_{daily}). The models included linear regression (LR), decision tree (DT), random forest (RF), extra trees (ET), gradient boosting machine (GBM), light gradient boosting machine (LGBM), eXtreme gradient boosting (XGBoost), k-Nearest Neighbours (k-NN), and multilayer perceptron (MLP). The scikit-learn Python package was used to train the models and evaluate their performance.

2.3.1. Linear regression (LR)

Linear regression (LR) is the most straightforward, widely applied, supervised machine learning model. This model obtains the correlation between the features and the target variable in such a way that this correlation represents the best fit. It minimizes the error by adjusting the coefficients of features and the bias using the gradient descent algorithm. The general formula of the linear regression model is explained by Eq. 5, where a_i refers to the coefficient of the ith feature, and b represents the bias [9].

$$y = b + \sum_{i=1}^{n} a_i x_i \quad (5)$$

2.3.2. Tree-based models

- **Decision tree (DT)**

Decision Tree (DT) is a regression model that splits the data into leaves based on simple binary decision rules. For each leaf, the model calculates the mean of the target variable's values and utilizes this mean as a prediction for feature data that satisfies the decision rule of the leaf. The determination of these decision rules is made in a manner that ensures a minimum error [5].

- **Ensemble tree-based models**

Due to the tendency of the decision tree (DT) model to overfit, the Random Forest (RF) method is employed to provide more realistic predictions by randomly modelling multiple decision trees and averaging their outputs. This technique is called ensemble learning. The Extra Trees (ET) regression model follows the same principle as RF model, but it randomly selects a subset of features to train each tree [18]. In this model, a dataset where each tree is assigned a unique sample is generated. Furthermore, a random subset of features is selected for each tree. The defining characteristic of Extra Trees (ET) lies in the random selection of the splitting value for a given feature. Instead of splitting the data and determining an optimal value based on criteria such as entropy or Gini index, this approach randomly selects a split value. This randomization results in trees that are diverse and uncorrelated, while also mitigating the risk of overfitting [16]. As a result, this technique is generally faster than Random Forest (RF) model [33]. Fig. 5 illustrates the extra trees (ET) model [33].

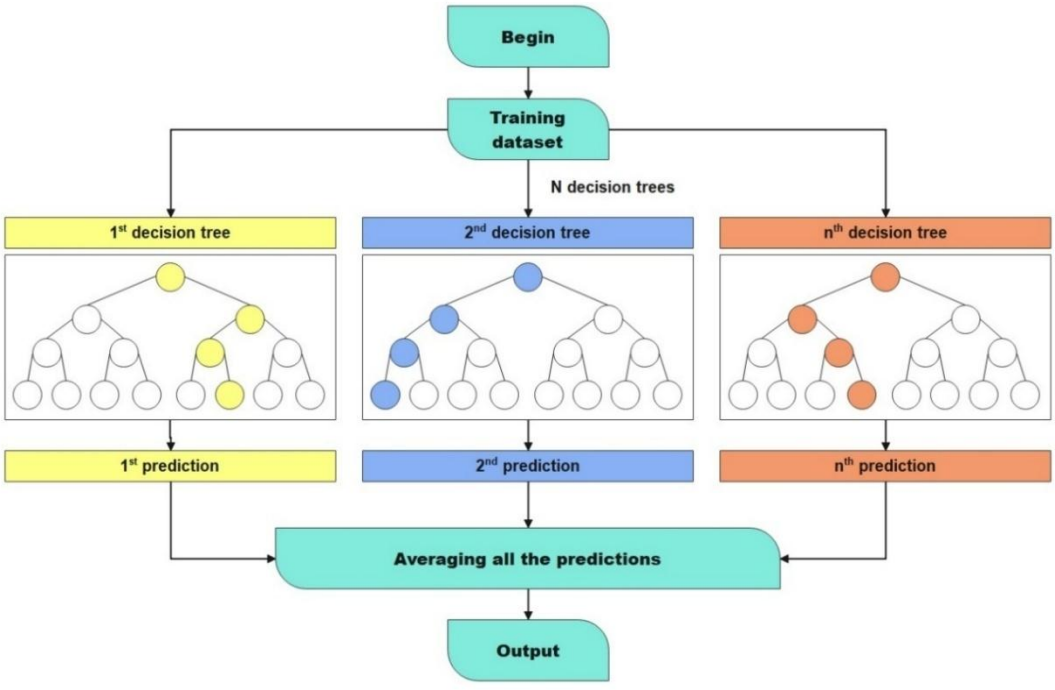


Figure 5. Extra Trees (ET) regression model [33]

Boosting is another technique that aims to construct a robust model by integrating multiple weak learners, where each learner focuses on correcting the errors of its predecessors [5]. Gradient Boosting Machine (GBM) is a highly flexible and customizable tree-based boosting algorithm that minimizes error by iteratively fitting decision trees using the gradient descent method [34]. GBM has several variants, including Light Gradient Boosting Machine (LGBM) and eXtreme Gradient Boosting (XGBoost) models. The LGBM model addresses the issue of high computational costs associated with large datasets by using a leaf-wise tree growth strategy, which accelerates the training process and improves efficiency [35]. On the other hand, the XGBoost model includes regularization terms in the objective function to prevent overfitting and enhance the model's generalizability [36].

2.3.3. Instance-Based Models

Only the k-nearest neighbours (k-NN) model is evaluated in this group. k-NN algorithm is a simple and interpretable machine learning model that identifies the most similar data points (neighbours) based on feature similarity. For regression tasks, it predicts the target variable by averaging the target values of these neighbours [8].

2.3.4. Neural Network Models

From this group, only the Multilayer Perceptron (MLP) model will be evaluated. The MLP model is a type of artificial neural network consisting of several layers of neurons, where each layer uses activation functions to transform input features into higher-level abstractions. This model adjusts the weights and biases of neurons in each layer to minimize prediction error using the backpropagation technique [35].

For model training, a set of hyperparameters was defined for each individual model. Table 3 summarizes the hyperparameters used in the training process.

Table 3. Hyperparameters used for training and fitting the models

Model(s)	Hyperparameters
LR	• fit_intercept=True: Specifies whether to calculate the intercept for the model.
DT	• criterion='gini': Function to measure the quality of a split. • max_depth=None: Maximum depth of the tree (None means no limit).
RF and ET	• n_estimators=100: Number of trees in the forest. • criterion='gini': Function to measure the quality of a split.

	• max_depth=None: Maximum depth of the trees (None means no limit).
GBM	• loss='squared_error': Loss function to be optimized. • learning_rate=0.1: Shrinks contribution of each tree. • n_estimators=100: Number of boosting stages. • max_depth=3: Maximum depth of the individual estimators.
LGBM	• num_leaves=31: Maximum number of leaves in one tree. • learning_rate=0.1: Boosting learning rate. • n_estimators=100: Number of boosting iterations. • max_depth=-1: Maximum tree depth for base learners; -1 means no limit.
XGBoost	• eta=0.3: Step size shrinkage used in update to prevent overfitting. • max_depth=6: Maximum depth of a tree. • n_estimators=100: Number of boosting rounds. • objective='reg:squarederror': learning task and corresponding objective function.
k-NN	• n_neighbors=5: Number of neighbors to use. • weights='uniform': Weight function used in prediction. • algorithm='auto': Algorithm used to compute the nearest neighbors.
MLP	• hidden_layer_sizes=(100,): Number of neurons in each hidden layer. • activation='relu': Activation function for the hidden layer. • solver='adam': The solver for weight optimization. • learning_rate_init=0.001: The initial learning rate used. • max_iter=200: Maximum number of iterations.

2.4. Model Evaluation

The performance of machine learning regression models is measured by the difference between the actual and predicted values. Among many evaluation metrics used for solar radiation estimation, this study utilized five of them, namely mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), mean absolute scaled error (MASE), and R^2 score. MAE is the averaged sum of absolute differences between actual and predicted values. It can be mathematically expressed by Eq. 6, where n is the number of samples, y refers to actual data, and $\hat{y}$ refers to the predicted data [37].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

MSE is also a popular evaluation metric, which is defined as the averaged sum of the squared differences between actual and predicted values. Although this metric is popular and widely used, it has the disadvantage of not being robust to outliers. It also results in large numbers in some cases, making it difficult to compare different models. For a specific model and dataset, MSE can be calculated using Eq. 7 [37].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$

RMSE is the square root of MSE, and it takes the same units as data, making it more interpretable. RMSE can be calculated using Eq. 8 [37].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

R^2 score or coefficient of determination is a measure of the prediction capability of models, and it can be defined as the ratio of the variance that can be explained by the specified model to the overall total variance in the dataset. R^2 score can be calculated using Eq. 9, and it takes values ranging from zero (the worst case) to one (the best performance) [37].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i - y_{mean}|} \quad (9)$$

Some studies preferred using mean absolute percentage error (MAPE), which can be defined as the averaged sum of the ratio of the difference between actual and predicted values to actual values [7, 8]. Due to the presence of a substantial number of zeros in the target variables of the datasets utilized in this study, the mean absolute percentage error (MAPE) would, in effect, approach infinity, as the resultant value would be divided by zero or an infinitesimal number. To overcome this issue, the mean absolute scaled error (MASE) can be used for forecasting when datasets may contain zeros [2, 38]. MASE is defined as the ratio of the MAE of the model to a naïve forecast in which each prediction is the value from a previous period. It can be obtained using Eq. 10, where m refers to the timestep (24 for instantaneous solar radiation and 365 for daily solar irradiation) [39].

$$MASE = \frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\frac{1}{n-m}\sum_{i=m+1}^{n}|y_i - y_{i-m}|} \quad (10)$$

K-Folds cross-validation is a popular validation technique for machine learning models. This method helps ensure that model evaluation is robust, providing realistic predictions and avoiding overly optimistic error estimates by leveraging multiple train-test splits. As shown in Fig. 6, this method evaluates model performance over n iterations by using 1/n of the data for testing and the remaining n−1/n for training during each iteration. The overall cross-validation performance is calculated as the average of the evaluation metrics across all iterations [7]. In this study, this technique was applied to validate the performance of different models using a total of 10 folds.

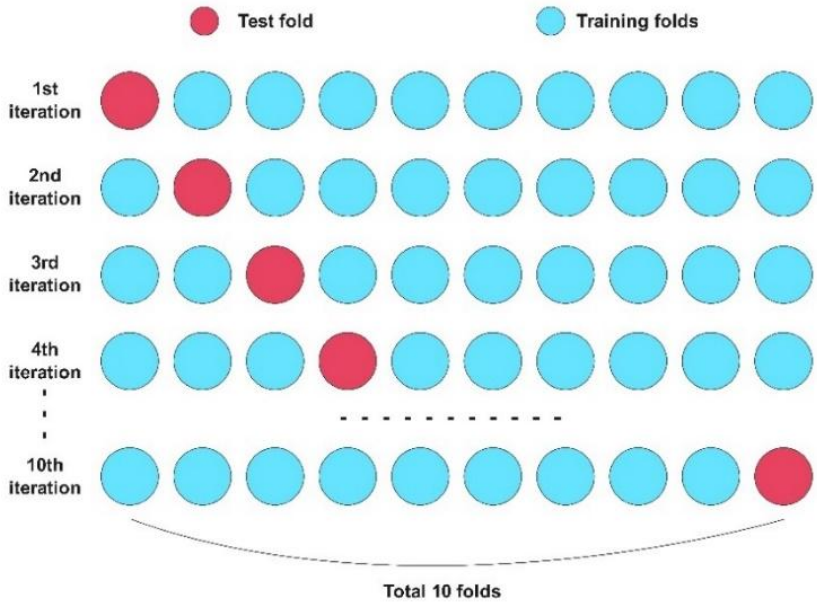


Figure 6. K-Folds cross validation method

Various methods can be used to assess feature importance and identify which features have the greatest impact on predictions. In this study, the widely adopted SHAP (SHapley Additive exPlanations) method is utilized. This method estimates the contribution of each feature to the model's output using Shapley values, which represent the marginal impact of that feature on the prediction relative to a baseline [36]. Furthermore, SHAP can be used to assess whether increasing the value of a feature contributes positively or negatively to the predicted outcome [35].

3. RESULTS AND DISCUSSION

In this study, raw data were processed, and the supervised machine learning models were trained and tested using Python 3.13 with the widely used scikit-learn package. The results for instantaneous solar radiation showed that all models performed well in predicting GHI, with R^2 scores greater than 0.85. When averaging the performance of all models across all cities, the ET

model demonstrated the best results for GHI, with an average R^2 score of 0.965, MSE of 3689.9, RMSE of 60.66, MAE of 29.35, and MASE of 0.079. It was closely followed by the LGBM and XGBoost models, with average R^2 of 0.964 and 0.9637, MSE of 3806.24 and 3845.9, RMSE of 61.61 and 61.91, MAE of 32 and 31.41, and MASE of 0.086 and 0.0846, respectively. The RF model showed comparable performance, with an average R^2 of 0.9632, MSE of 3897.48, RMSE of 62.35, MAE of 30.29, and MASE of 0.0815. In contrast, the k-NN and LR models had the lowest performance, with average R^2 scores of 0.896 and 0.8788, respectively. Therefore, it is evident that the ensemble tree-based models performed better than other models for predicting instantaneous GHI. The comparison between the models for predicting instantaneous GHI is illustrated in Fig. 7.

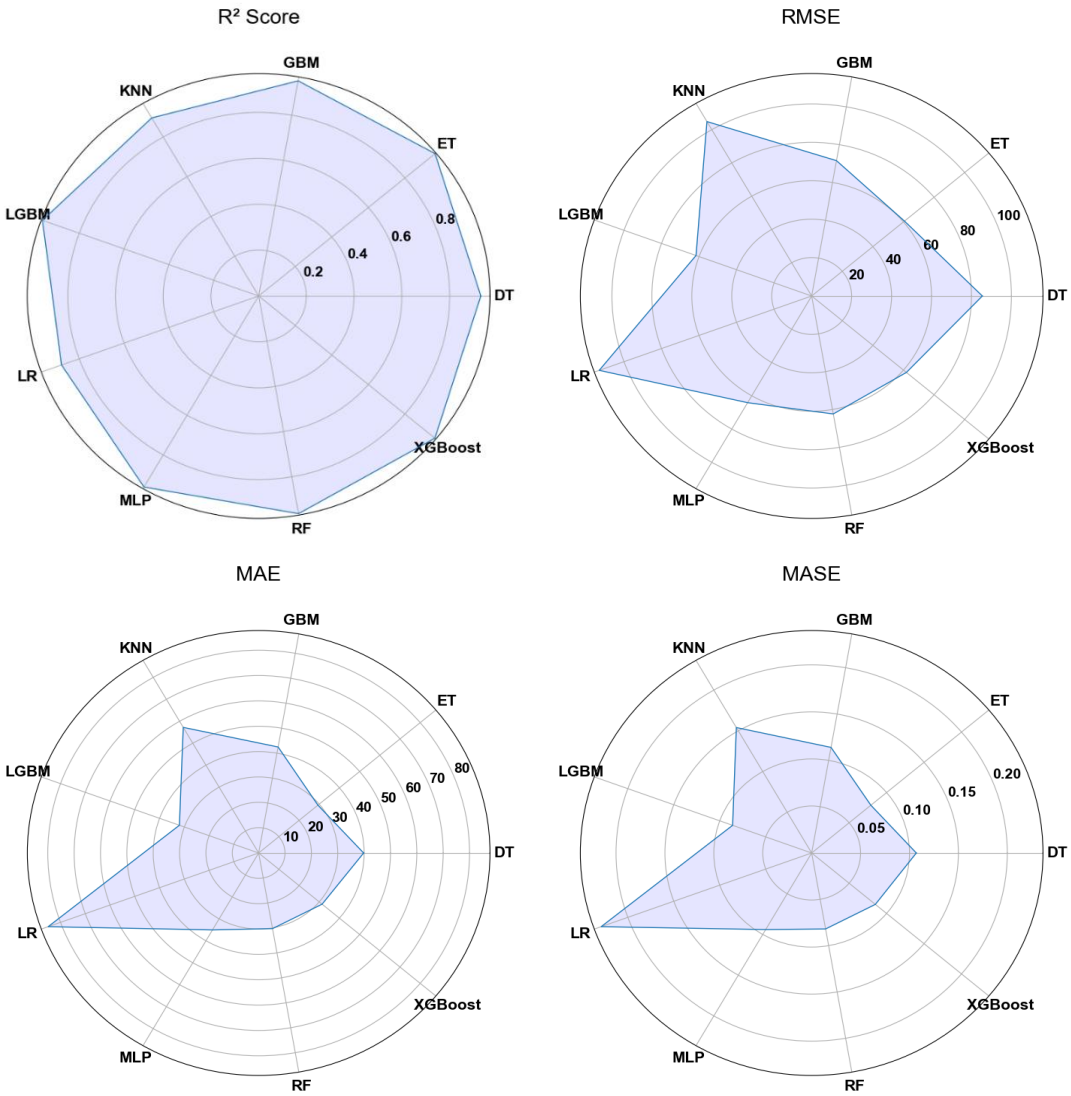


Figure 7. Comparison of model performance for predicting instantaneous GHI, based on evaluation metrics averaged across all cities.

The models showed slightly worse results for predicting instantaneous DNI. Once again, the ET model turned out to be the best model predicting DNI, with an average R^2 score of 0.951, MSE of 7056.7, RMSE of 83.92, MAE of 50.15, and MASE of 0.11. Other tree-based models (GBM, RF, DT, LGBM and XGBoost) demonstrated good performance with an average R^2 score values greater than 0.9. Moreover, the LR model was found to be able to predict instantaneous DNI with an average R^2 score of 0.8069, MSE of 27804, RMSE of 166.64, MAE of 133.55, and MASE of 0.316. k-NN model performed the worst in predicting instantaneous DNI with an average R^2 score of 0.6501, which is below the acceptable limit. An R^2 score lower than 0.7 generally indicates a weak relationship between the predicted and actual values [40]. Therefore, the k-NN model should be considered for further optimization. The comparison between the models for predicting instantaneous DNI is illustrated in Fig. 8.

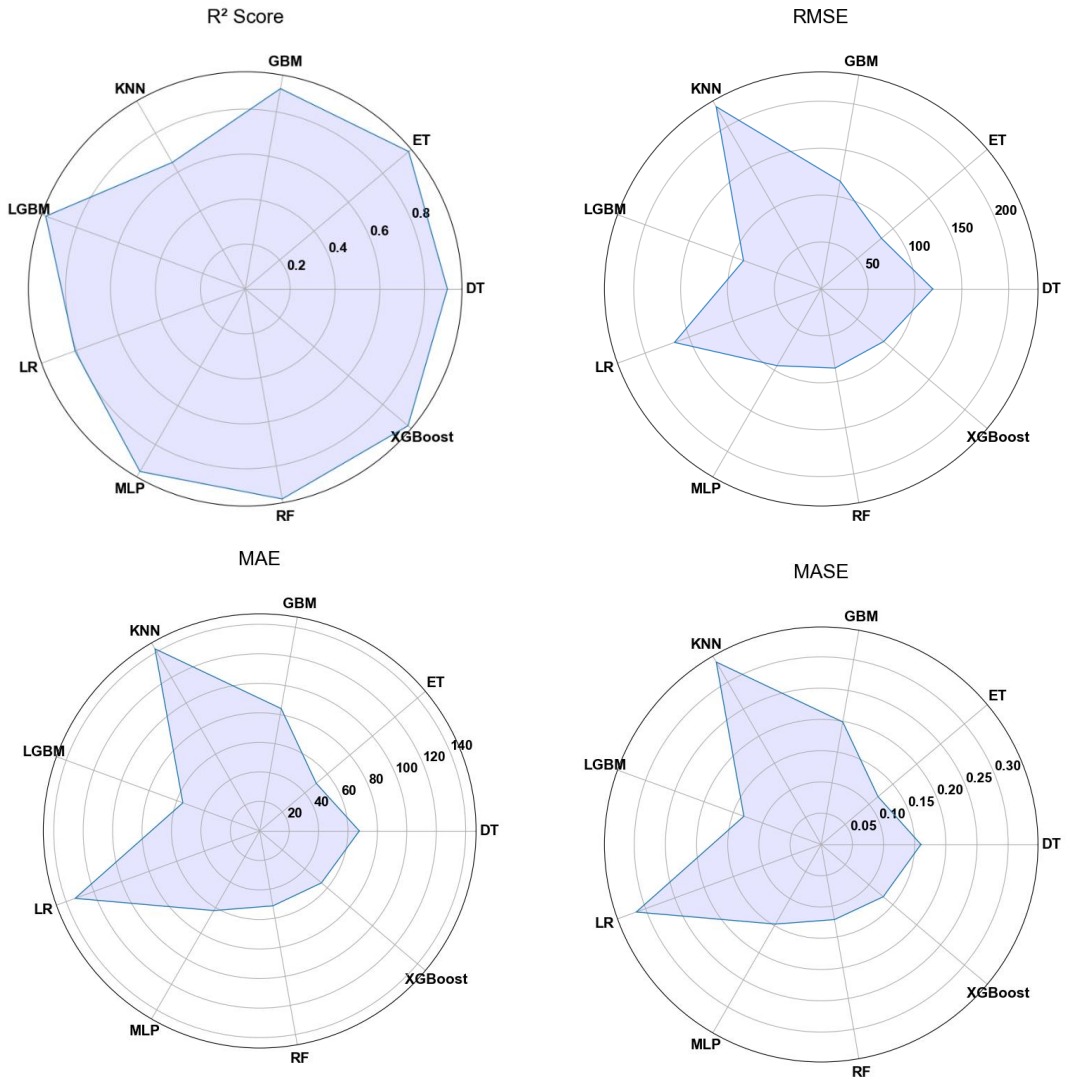


Figure 8. Comparison of model performance for predicting instantaneous DNI, based on evaluation metrics averaged across all cities.

For instantaneous solar radiation, the performance of all the models was slightly better in Adıyaman compared to other cities. Fig. 9 illustrates the comparison between different models for predicting both the instantaneous GHI and DNI in Adıyaman city.

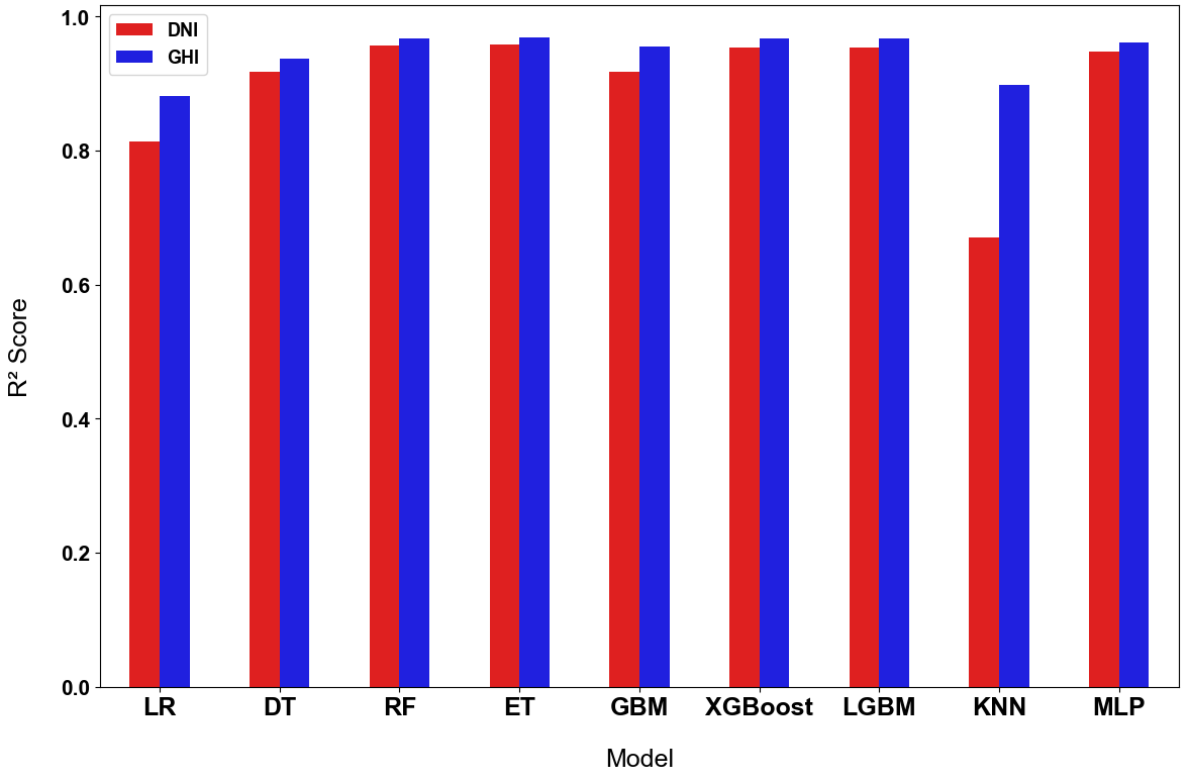


Figure 9. R^2 score of different models predicting instantaneous solar radiation in Adıyaman city

In terms of daily solar irradiation, the models consistently outperformed their results for instantaneous solar radiation. All tree-based models (GBM, RF, DT, LGBM, XGBoost, and ET) demonstrated excellent predictive performance, with R^2 values exceeding 0.999 for GHI_{daily} and 0.95 for DNI_{daily}, highlighting their strong ability to capture the variability in the data. Among these, the Extra Trees (ET) model achieved the highest accuracy for GHI_{daily} predictions across all cities, with an average R^2 of 0.9999, MSE of 0.0006, RMSE of 0.0244, MAE of 0.0142, and MASE of 0.0047. This was closely followed by the XGBoost model, which yielded an average R^2 of 0.9998, MSE of 0.0012, RMSE of 0.0343, MAE of 0.0245, and MASE of 0.0082. The Random Forest (RF) model also delivered a very similar performance, achieving an average R^2 of 0.9998. The MLP model showed reasonably strong predictive capability, with an average R^2 of 0.981, MSE of 0.123, RMSE of 0.342, MAE of 0.27, and MASE of 0.0906. In contrast, the k-NN model exhibited the weakest performance, with an average R^2 of 0.874, MSE of 0.853, RMSE of 0.923, MAE of 0.638, and MASE of 0.213. Fig. 10 provides a comparative summary of the models' performance in predicting GHI_{daily}.

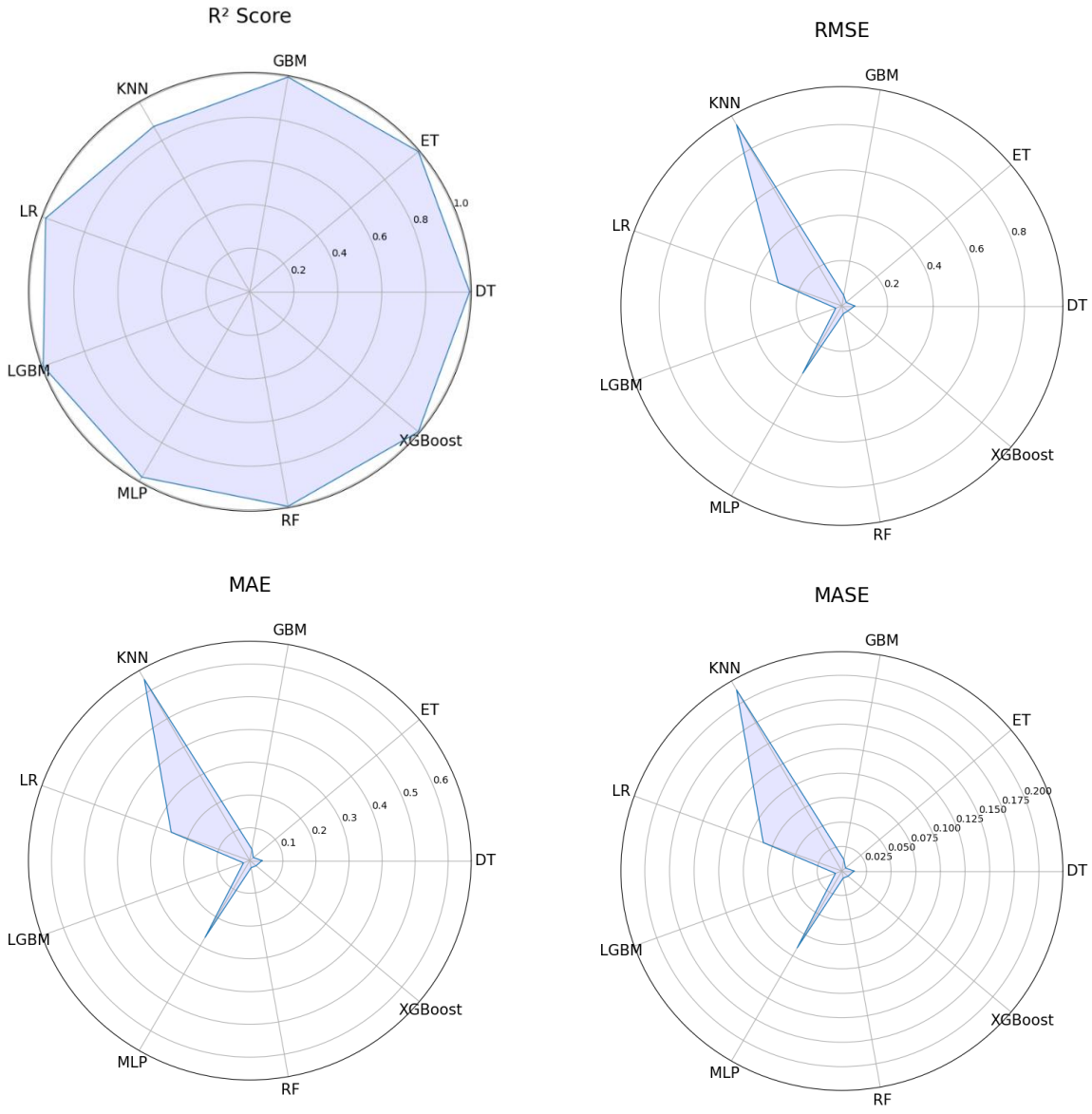


Figure 10. Comparison of model performance for predicting GHI_{daily}, based on evaluation metrics averaged across all cities.

Once again, the ET model demonstrated superior performance in predicting DNI_{daily} compared to the other evaluated models. In terms of R^2, the ET model led with a score of 0.98, followed closely by the RF and XGBoost models, which achieved scores of 0.979 and 0.975, respectively. Regarding RMSE, the performance remained consistent, with the ET model achieving the lowest RMSE of 0.517, while the RF and XGBoost models recorded RMSE values of 0.529 and 0.57, respectively.

The DT model also exhibited strong accuracy in predicting DNI_{daily}, with an average R^2 of 0.957, MSE of 0.566, RMSE of 0.752, MAE of 0.518, and MASE of 0.124. In contrast, the MLP model performed less effectively for DNI_{daily} predictions compared to GHI_{daily}, with an average R^2 of

0.836, MSE of 2.17, RMSE of 1.473, MAE of 1.146, and MASE of 0.275. The k-NN model exhibited the poorest predictive accuracy, with an average R^2 of 0.654, falling below the acceptable threshold for R^2 values. This highlights the importance of optimizing this model by selecting the best hyperparameters that lead to better performance [41]. A detailed comparison of the models' performance in predicting DNI_{daily} is presented in Fig. 11. It is worth mentioning that the low MASE values of the models (< 1) indicate that the forecasts made by these models are better than a naïve forecast using one-step-ahead data, highlighting their usefulness in forecasting solar radiation [39].

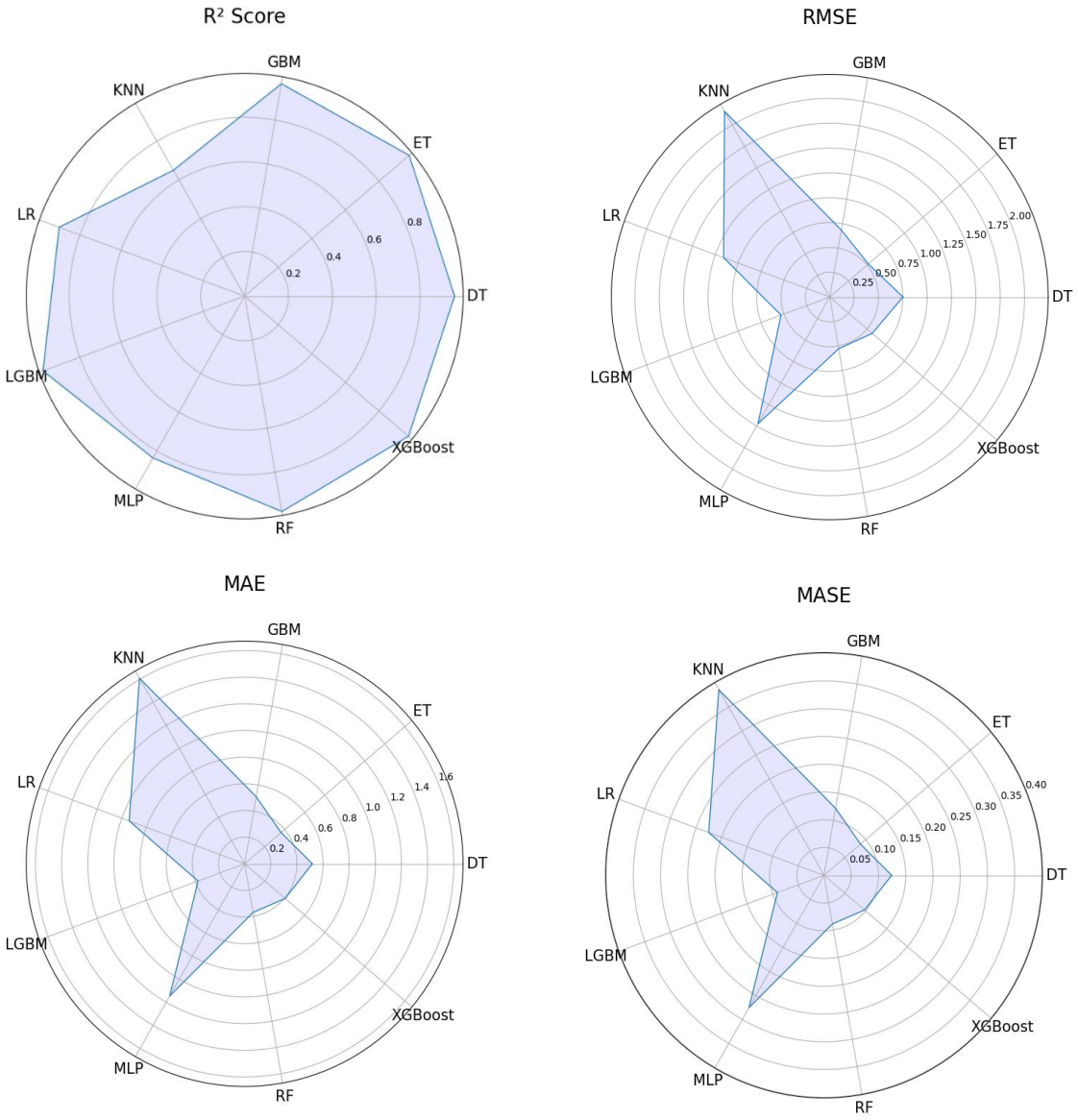


Figure 11. Comparison of model performance for predicting DNI_{daily}, based on evaluation metrics averaged across all cities.

Based on the results and the above analysis, it is evident that the ET model had a better overall performance than other models. Table 4 summarizes the ET model's results. The comparison of the ET model's performance for the instantaneous and daily solar radiation components in Adıyaman city (GHI, GHI_{daily}, DNI, DNI_{daily}) is illustrated in Fig. 12. As shown in Fig. 12, the ET model achieved the highest R^2 score for predicting GHI_{daily}, followed by DNI_{daily}, instantaneous GHI, and instantaneous DNI, in that order.

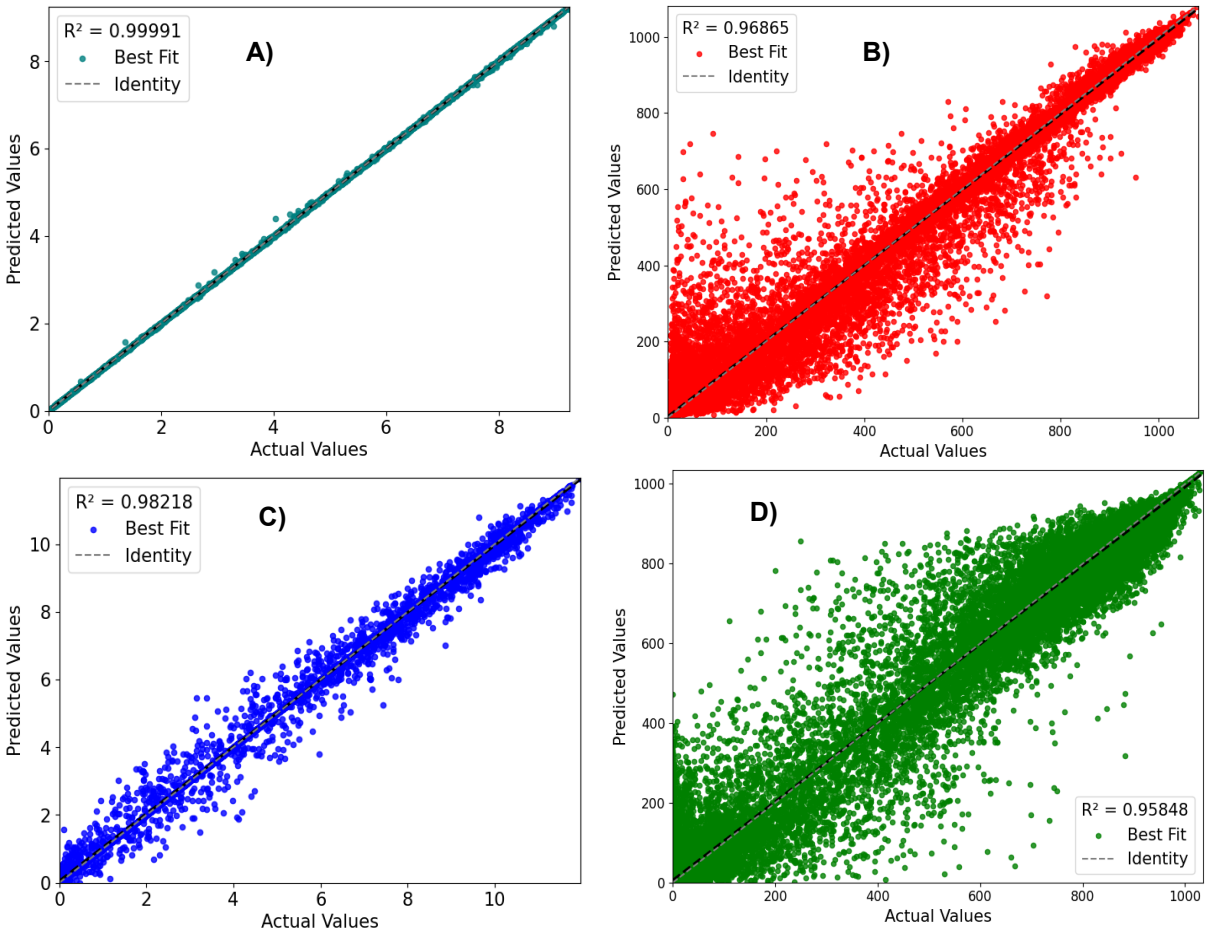


Figure 12. Prediction errors of ET Model for Adıyaman city. A) Daily GHI, B) Instantaneous GHI, C) Daily DNI, D) Instantaneous DNI

All the models demonstrated similar performance across different cities, with only slight variations attributed to the characteristics of the datasets, fluctuations in radiation values, and differences in weather patterns. For instance, when predicting GHI, the ET model achieved the highest R^2 score of 0.9685 for Adıyaman and the lowest R^2 score of 0.9589 for Şırnak. For GHI_{daily} predictions, the ET model consistently demonstrated very similar accuracy, with only slight variations (R^2 = 0.9999, MSE ≤ 0.0007, RMSE ≤ 0.0262, MAE ≤ 0.0155, MASE ≤ 0.0052).

Table 4. Results of ET model

City	Target	R^2	MSE	RMSE	MAE	MASE
Adıyaman	*GHI*	0.9685	3331.0143	57.7149	27.4514	0.0739
	DNI	0.9583	5962.5068	77.2173	46.8309	0.1114
	GHI_{daily}	0.9999	0.0006	0.0249	0.0141	0.0047
	DNI_{daily}	0.9822	0.2381	0.4880	0.3418	0.0830
Batman	*GHI*	0.9681	3352.9485	57.9047	28.4148	0.0770
	DNI	0.9525	6691.5842	81.8021	49.2006	0.1177
	GHI_{daily}	0.9999	0.0005	0.0217	0.0131	0.0044
	DNI_{daily}	0.9800	0.2603	0.5102	0.3587	0.0869
Diyarbakır	*GHI*	0.9676	3372.8091	58.0759	28.1709	0.0766
	DNI	0.9561	6289.9206	79.3090	47.7224	0.1134
	GHI_{daily}	0.9999	0.0005	0.0228	0.0137	0.0046
	DNI_{daily}	0.9794	0.2675	0.5172	0.3666	0.0895
Gaziantep	*GHI*	0.9629	3887.6609	62.3511	30.1307	0.0815
	DNI	0.9491	7313.2061	85.5173	50.4724	0.1191
	GHI_{daily}	0.9999	0.0007	0.0269	0.0155	0.0052
	DNI_{daily}	0.9818	0.2509	0.5009	0.3516	0.0836
Kilis	*GHI*	0.9664	3544.7217	59.5376	28.3098	0.0763
	DNI	0.9494	6950.1130	83.3673	49.8915	0.1206
	GHI_{daily}	0.9999	0.0006	0.0245	0.0144	0.0048
	DNI_{daily}	0.9805	0.2567	0.5067	0.3549	0.0863
Mardin	*GHI*	0.9656	3677.6097	60.6433	29.1996	0.0780
	DNI	0.9470	7632.4735	87.3640	52.4788	0.1233
	GHI_{daily}	0.9999	0.0007	0.0265	0.0142	0.0047
	DNI_{daily}	0.9792	0.2690	0.5187	0.3709	0.0896
Şanlıurfa	*GHI*	0.9669	3456.1493	58.7890	28.8131	0.0781
	DNI	0.9466	7392.4623	85.9794	52.0905	0.1252
	GHI_{daily}	0.9999	0.0006	0.0250	0.0145	0.0049
	DNI_{daily}	0.9812	0.2398	0.4897	0.3487	0.0847
Siirt	*GHI*	0.9616	4072.2818	63.8144	31.2479	0.0843
	DNI	0.9509	7157.9748	84.6048	50.2821	0.1184
	GHI_{daily}	0.9999	0.0005	0.0215	0.0139	0.0046
	DNI_{daily}	0.9794	0.2799	0.5290	0.3695	0.0877
Şırnak	*GHI*	0.9589	4513.8975	67.1855	32.4415	0.0858
	DNI	0.9486	8119.9127	90.1106	52.4271	0.1187
	GHI_{daily}	0.9999	0.0007	0.0262	0.0150	0.0050
	DNI_{daily}	0.9754	0.3576	0.5980	0.4144	0.0950

The SHAP summary plots in Fig. 13 provide a detailed view of how each feature influences the ET model's predictions for solar radiation. In this figure, red indicates high feature values and blue indicates low; points on the right increase the prediction (positive impact), while those on the left decrease it (negative impact). As shown in Fig. 13A, for daily GHI, the clearness index (K_T) was found to be the most impactful variable, with higher values leading to increased radiation

estimates. This was followed by the declination angle (δ), the average relative humidity (H_{avg}), and the average temperature (T_{avg}) which were determined to have moderate impact. The average wind speed (W_{avg}) was found to be the least important feature. Similar results could be observed for daily DNI (Fig. 9B). These results are consistent with the correlation analysis (Fig. 4), where K_T, H_{avg}, and δ show strong correlations with both GHI_{daily} and DNI_{daily}. In addition, based on domain knowledge, K_T, which represents atmospheric transparency, is known to have a significant effect on the amount of solar radiation that reaches the surface [35].

For the prediction of instantaneous GHI (Fig. 13C), the most influential feature was the solar zenith angle (z), followed by relative humidity (H) and cloud type. In the case of instantaneous DNI (Fig. 13D), cloud type had the strongest impact, with H and z also playing significant roles. The other features—temperature (T), declination angle (δ), wind speed (W), and pressure (P)—had a moderate influence on both DNI and GHI predictions. These findings are in line with the correlation analysis (Fig. 3), where the cloud type and z features showed the strongest correlation with the target variables DNI and GHI, respectively.

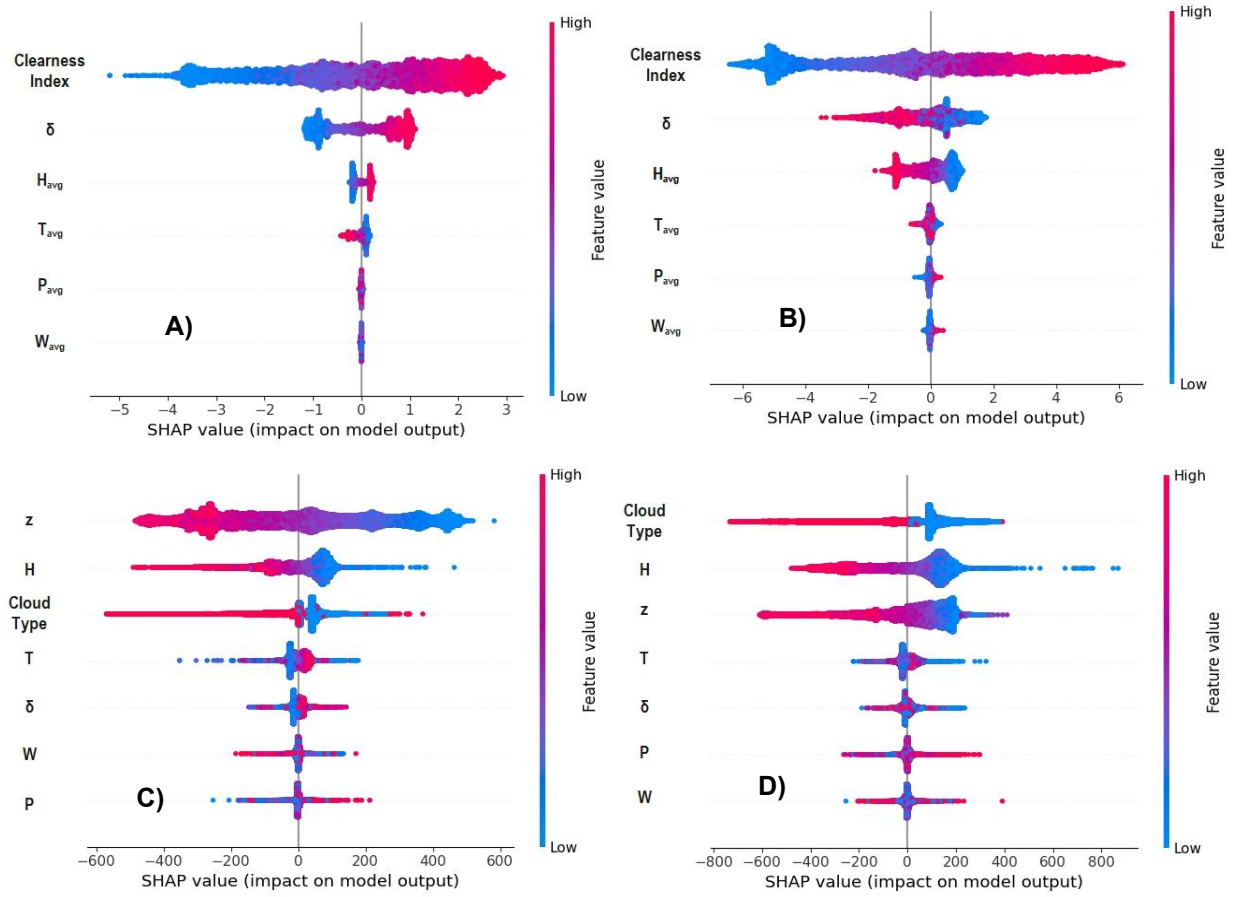


Figure 13. SHAP summary plot for the feature importance of the ET model. A) Daily GHI, B) Daily DNI, C) Instantaneous GHI, D) Instantaneous DNI.

Table 5 summarizes the evaluation results of LR model and correlation equations of daily GHI and DNI. As shown in Table 5, the LR model explained the variation in daily GHI data with R^2 scores greater than 0.986 for all cities. For daily DNI, the model also demonstrated good performance, with R^2 scores greater than 0.88 across all cities. Since the LR model performed well on daily solar irradiation data, the correlations obtained from it can be used to provide a rough estimate of daily solar irradiation in the study area.

Table 5. Results of LR model for predicting daily solar irradiation

City		R^2	MSE	RMSE	MAE	MASE	Correlation
Adıyaman	GHI_{daily}	0.9872	0.0908	0.3013	0.2589	0.0872	-1.8492 - 0.0029 T_{avg} - 0.0253 W_{avg} - 0.0073 H_{avg} + 0.0030 P_{avg} + 0.0005 δ + 6.8720 K_T
	DNI_{daily}	0.9102	1.2313	1.1097	0.8872	0.2172	-9.3237 - 0.0330 T_{avg} + 0.1414 W_{avg} - 0.0384 H_{avg} + 0.0087 P_{avg} - 0.1550 δ + 13.8147 K_T
Batman	GHI_{daily}	0.9874	0.0868	0.2947	0.2511	0.0840	-7.9459 - 0.0019 T_{avg} - 0.0286 W_{avg} - 0.0071 H_{avg} + 0.0095 P_{avg} + 0.0031 δ + 6.7330 K_T
	DNI_{daily}	0.9031	1.2890	1.1353	0.9185	0.2229	-0.9609 - 0.0551 T_{avg} + 0.0431 W_{avg} - 0.0406 H_{avg} + 0.0005 P_{avg} - 0.1444 δ + 13.7269 K_T
Diyarbakır	GHI_{daily}	0.9872	0.0872	0.2952	0.2515	0.0849	-6.1443 - 0.0028 T_{avg} - 0.0222 W_{avg} - 0.0072 H_{avg} + 0.0076 P_{avg} + 0.0019 δ + 6.7744 K_T
	DNI_{daily}	0.9066	1.2235	1.1061	0.8934	0.2160	-8.0924 - 0.0486 T_{avg} + 0.1162 W_{avg} - 0.0375 H_{avg} + 0.0077 P_{avg} - 0.1498 δ + 13.9336 K_T
Gaziantep	GHI_{daily}	0.9880	0.0838	0.2896	0.2478	0.0834	-1.9769 + 0.0055 T_{avg} - 0.0469 W_{avg} - 0.0037 H_{avg} + 0.0029 P_{avg} - 0.0020 δ + 7.0335 K_T
	DNI_{daily}	0.9105	1.2555	1.1205	0.9088	0.2160	4.1140 - 0.0223 Tavg + 0.0979 W_{avg} - 0.0296 H_{avg} - 0.0068 P_{avg} - 0.1678 δ + 14.6842 K_T
Kilis	GHI_{daily}	0.9879	0.0837	0.2893	0.2480	0.0840	-1.5214 + 0.0098 T_{avg} - 0.0854 W_{avg} - 0.0014 H_{avg} + 0.0022 P_{avg} - 0.0031 δ + 7.1352 K_T
	DNI_{daily}	0.9044	1.2868	1.1344	0.9213	0.2264	12.7217 - 0.0042 T_{avg} - 0.0766 W_{avg} - 0.0203 H_{avg} -

							0.0168 P_{avg} - 0.1747 δ + 15.2635 K_T
Mardin	GHI_{daily}	0.9869	0.0903	0.3004	0.2553	0.0853	-4.9887 - 0.0010 T_{avg} - 0.0750 W_{avg} - 0.0059 H_{avg} + 0.0065 P_{avg} - 0.0003 δ + 6.9224 K_T
	DNI_{daily}	0.8952	1.3897	1.1788	0.9524	0.2303	-0.9924 - 0.0470 T_{avg} + 0.2330 W_{avg} - 0.0301 H_{avg} - 0.0010 P_{avg} - 0.1609 δ + 14.5269 K_T
Şanlıurfa	GHI_{daily}	0.9871	0.0855	0.2924	0.2489	0.0847	-0.0972 + 0.0015 T_{avg} - 0.0599 W_{avg} - 0.0050 H_{avg} + 0.0010 P_{avg} - 0.0003 δ + 6.9463 K_T
	DNI_{daily}	0.9017	1.2790	1.1309	0.9236	0.2259	17.1293 - 0.0333 T_{avg} + 0.0323 W_{avg} -0.0272 H_{avg} - 0.0207 P_{avg} - 0.1776 δ + 15.2375 K_T
Siirt	GHI_{daily}	0.9871	0.0890	0.2983	0.2533	0.0843	-10.2056 - 0.0024 T_{avg} + 0.0056 W_{avg} - 0.0066 H_{avg} + 0.0123 P_{avg} + 0.0030 δ + 6.7736 K_T
	DNI_{daily}	0.8876	1.5166	1.2315	1.0005	0.2350	-13.8573 - 0.0526 T_{avg} + 0.1521 W_{avg} -0.0391 H_{avg} + 0.0144 P_{avg} - 0.1450 δ + 13.8188 K_T
Şırnak	GHI_{daily}	0.9861	0.0944	0.3072	0.2622	0.0866	-9.2220 + 0.0001 T_{avg} - 0.0295 W_{avg} - 0.0035 H_{avg} + 0.0119 P_{avg} + 0.0004 δ + 6.9625 K_T
	DNI_{daily}	0.8844	1.6539	1.2861	1.0506	0.2369	-12.1501 - 0.0582 T_{avg} + 0.0323 W_{avg} - 0.0368 H_{avg} + 0.0136 P_{avg} - 0.1509 δ + 14.0053 K_T

Although no ground-based validation was conducted in this study, the NSRDB dataset used is widely recognized and trusted in the literature. As previously noted, the NSRDB model has been validated against surface observations, showing a mean percentage bias of 5% for GHI and 10% for DNI, which supports its reliability for solar resource assessment [25]. However, we acknowledge that the use of ground-based measurements could lead to more realistic and locally accurate results.

4. CONCLUSION

This study aimed to explore the effectiveness of supervised machine learning models in predicting both instantaneous and daily solar radiation in the Southern Anatolian Region of Türkiye. For each of the nine districts in the region, 157,680 data points spanning 18 years (2005–2022) were

collected from the NSRDB database and processed for model training. Nine different machine learning algorithms were evaluated, including linear regression, decision tree, random forest, extra trees, gradient boosting machine, light gradient boosting machine, XGBoost, multilayer perceptron, and k-nearest neighbours. Their performance was assessed using five metrics: coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute scaled error (MASE).

The results demonstrated that the extra trees model outperformed all others in predicting both instantaneous and daily solar radiation. Overall, models showed higher accuracy in forecasting global horizontal irradiation (GHI) compared to direct normal irradiation (DNI), with tree-based models consistently delivering superior performance. Specifically, the extra trees model was capable of predicting instantaneous GHI and DNI with R^2 scores greater than 0.96 and 0.95, respectively. Regarding daily solar irradiation, the extra trees model exhibited remarkably high R^2 scores greater than 0.99 for GHI and 0.975 for DNI. Moreover, the ET model achieved its highest accuracy when predicting GHI_{daily}, with a station-wise average R^2 of 0.9999, MSE of 0.0006, RMSE of 0.0244, MAE of 0.0142, and MASE of 0.0047. Feature importance analysis revealed that the clearness index, declination angle, and relative humidity were key predictors for daily radiation, while cloud type and solar zenith angle played a major role in forecasting instantaneous DNI and GHI.

Despite the absence of optimization or hyperparameter tuning, the findings underscore the strong potential of machine learning techniques for solar radiation forecasting in Türkiye. However, there are limitations to note. First, the study relies on NSRDB model data rather than ground-based measurements, which typically offer more accurate and realistic representations of solar radiation. Second, only traditional machine learning models were used, which may not capture complex patterns as effectively as more sophisticated modeling approaches.

For future work, investigating more complex models could lead to further improvements in prediction accuracy. Incorporating ground-based measurement data could also enhance model reliability. Additionally, machine learning can be extended beyond radiation forecasting to predict the actual energy output of solar systems based on weather conditions, contributing to more efficient system planning and management.

NOMENCLATURE

DNI	Direct normal irradiation
GHI	Global horizontal irradiation
T	Ambient temperature
W	Wind speed
δ	Declination angle
z	Solar zenith angle
P	Atmospheric pressure
H	Relative humidity
K_T	Clearness index
MAE	Mean absolute error
MSE	Mean squared error
RMSE	Root mean squared error
MASE	Mean absolute scaled error
R^2	Coefficient of determination
NSRDB	National Solar Radiation Database
LR	Linear regression
DT	Decision tree
RF	Random forest
GBM	Gradient boosting machine
LGBM	Light gradient boosting machine
XGBoost	eXtreme gradient boosting
MLP	Multilayer perceptron
ET	Extra trees
k-NN	k-nearest neighbors

DECLARATION OF ETHICAL STANDARDS

The authors of this paper declare that nothing that is necessary for achieving the paper requires ethical committee and legal-special permissions.

CONTRIBUTION OF THE AUTHORS

Abdallah Adil Awad BASHIR: Analyzed the data, performed the experiments, and wrote the manuscript.

Abdulkadir KOCER: Conceptualized the idea, organized the work, and revised the manuscript.

Ahmet ÇOSGUN: Revised the manuscript.

Afşin GÜNGÖR: Supervised and assisted with the project.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] Obiora CN, Ali A, Hassan AN. Predicting Hourly Solar Irradiance Using Machine Learning Methods. 2020 11th International Renewable Energy Congress (IREC)2020. pp. 1-6.

[2] Guher AB, Tasdemir S, Yaniktepe B. Effective Estimation of Hourly Global Solar Radiation Using Machine Learning Algorithms. International Journal of Photoenergy. 2020;2020:8843620. https://doi.org/10.1155/2020/8843620.

[3] Allal Z, Noura HN, Chahine K. Machine Learning Algorithms for Solar Irradiance Prediction: A Recent Comparative Study. e-Prime - Advances in Electrical Engineering, Electronics and Energy. 2024;7:100453. https://doi.org/10.1016/j.prime.2024.100453.

[4] Zhou Y, Liu Y, Wang D, Liu X, Wang Y. A review on global solar radiation prediction with machine learning models in a comprehensive perspective. Energy Conversion and Management. 2021;235:113960. https://doi.org/10.1016/j.enconman.2021.113960.

[5] Voyant C, Notton G, Kalogirou S, Nivet M-L, Paoli C, Motte F, Fouilloy A. Machine learning methods for solar radiation forecasting: A review. Renewable Energy. 2017;105:569-82. https://doi.org/10.1016/j.renene.2016.12.095.

[6] Nematchoua MK, Orosa JA, Afaifia M. Prediction of daily global solar radiation and air temperature using six machine learning algorithms; a case of 27 European countries. Ecological Informatics. 2022;69:101643. https://doi.org/10.1016/j.ecoinf.2022.101643.

[7] Ercan U, Kocer A. Prediction of solar irradiance with machine learning methods using satellite data. International Journal of Green Energy. 2024;21:1174-83. 10.1080/15435075.2024.2305857.
[8] Ağbulut Ü, Gürel AE, Biçen Y. Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Renewable and Sustainable Energy Reviews. 2021;135:110114. https://doi.org/10.1016/j.rser.2020.110114.
[9] Hacioğlu R. Prediction of solar radiation based on machine learning methods. The journal of cognitive systems. 2017;2:16-20.
[10] Demir V, Demirgül T, Sevimli MF. Model-Ağacı (M5-tree) yaklaşımı ile HELIOSAT tabanlı güneş radyasyonu tahmini. Geomatik. 2023;8:124-35. 10.29128/geomatik.1137687.
[11] Demir V, Citakoglu H. Forecasting of solar radiation using different machine learning approaches. Neural Computing and Applications. 2023;35:887-906. 10.1007/s00521-022-07841-x.
[12] Demirgül T, Demir V, Sevimli MF. Farklı makine öğrenmesi yaklaşımları ile Türkiye'nin solar radyasyon tahmini. Geomatik. 2024;9:106-22. 10.29128/geomatik.1374383.
[13] Mendyl A, Demir V, Omar N, Orhan O, Weidinger T. Enhancing Solar Radiation Forecasting in Diverse Moroccan Climate Zones: A Comparative Study of Machine Learning Models with Sugeno Integral Aggregation. Atmosphere. 2024;15:103.
[14] Toylan H. SOLAR IRRADIANCE PREDICTION USING BAGGING DECISION TREE-BASED MACHINE LEARNING. Kirklareli University Journal of Engineering and Science. 2022;8:15-24. 10.34186/klujes.1106357.
[15] Tercha W, Tadjer SA, Chekired F, Canale L. Machine Learning-Based Forecasting of Temperature and Solar Irradiance for Photovoltaic Systems. Energies. 2024;17:1124.
[16] Ahmad MW, Mourshed M, Rezgui Y. Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. Energy. 2018;164:465-74. https://doi.org/10.1016/j.energy.2018.08.207.
[17] Hassan MA, Khalil A, Kaseb S, Kassem MA. Exploring the potential of tree-based ensemble methods in solar radiation modeling. Applied Energy. 2017;203:897-916. https://doi.org/10.1016/j.apenergy.2017.06.104.
[18] Ahmad MW, Reynolds J, Rezgui Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. Journal of Cleaner Production. 2018;203:810-21. https://doi.org/10.1016/j.jclepro.2018.08.207.
[19] SolarGIS. Solar resource maps & GIS data (Turkey), https://solargis.com/resources/free-maps-and-gis-data?locality=turkey; 2024 [accessed 15 November 2024].

[20] Kaygusuz K. Prospect of concentrating solar power in Turkey: The sustainable future. Renewable and Sustainable Energy Reviews. 2011;15:808-14. https://doi.org/10.1016/j.rser.2010.09.042.

[21] Bulut M. Integrated solar power project based on CSP and PV technologies for Southeast of Turkey. International Journal of Green Energy. 2022;19:603-13. 10.1080/15435075.2021.1954006.

[22] Güngör-Demirci G. Spatial analysis of renewable energy potential and use in Turkey. Journal of Renewable and Sustainable Energy. 2015;7:10.1063/1.4907921.

[23] NREL. NSRDB: National Solar Radiation Database, https://nsrdb.nrel.gov/; 2024 [accessed 29 December 2024].

[24] Buster G, Bannister M, Habte A, Hettinger D, Maclaurin G, Rossol M, et al. Physics-guided machine learning for improved accuracy of the National Solar Radiation Database. Solar Energy. 2022;232:483-92. https://doi.org/10.1016/j.solener.2022.01.004.

[25] Sengupta M, Xie Y, Lopez A, Habte A, Maclaurin G, Shelby J. The National Solar Radiation Data Base (NSRDB). Renewable and Sustainable Energy Reviews. 2018;89:51-60. https://doi.org/10.1016/j.rser.2018.03.003.

[26] Mukherjee A, Ain A, Dasgupta P. Solar Irradiance Prediction from Historical Trends Using Deep Neural Networks. 2018 IEEE International Conference on Smart Energy Grid Engineering (SEGE)2018. pp. 356-61.

[27] Narvaez G, Giraldo LF, Bressan M, Pantoja A. Machine learning for site-adaptation and solar radiation forecasting. Renewable Energy. 2021;167:333-42. https://doi.org/10.1016/j.renene.2020.11.089.

[28] Hinkelman L, Sengupta M. Relating Solar Resource Variability to Cloud Type. AGU Fall Meeting Abstracts2012. pp. A31F-0086.

[29] NREL. High-Performance Computing: National Solar Radiation Database, https://www.nrel.gov/hpc/nsrdb-dataset.html; 2024 [accessed 30 December 2024].

[30] Kalogirou SA. Chapter two - Environmental Characteristics. in: S.A. Kalogirou, (Ed.). Solar Energy Engineering. Academic Press, Boston, 2009. pp. 49-762.

[31] Poslavskaya E, Korolev A. Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding? arXiv preprint arXiv:2312.16930. 2023.

[32] Duffie JA, Beckman WA. Solar Engineering of Thermal Processes. ed. Wiley; 2013.

[33] Atiea MA, Shaheen AM, Alassaf A, Alsaleh I. Enhanced Solar Power Prediction Models With Integrating Meteorological Data Toward Sustainable Energy Forecasting. International Journal of Energy Research. 2024;2024:8022398. https://doi.org/10.1155/er/8022398.

[34] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics. 2013;7. https://doi.org/10.3389/fnbot.2013.00021

[35] Chaibi M, Benghoulam EM, Tarik L, Berrada M, Hmaidi AE. An Interpretable Machine Learning Model for Daily Global Solar Radiation Prediction. Energies. 2021;14:7367.

[36] Song Z, Cao S, Yang H. An interpretable framework for modeling global solar radiation using tree-based ensemble machine learning and Shapley additive explanations methods. Applied Energy. 2024;364:123238. https://doi.org/10.1016/j.apenergy.2024.123238.

[37] Plevris V, Solorzano G, Bakas NP, Ben Seghier MEA. Investigation of performance metrics in regression analysis and machine learning-based prediction models. 8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress 2022). European Community on Computational Methods in Applied Sciences. 2022.

[38] Hyndman RJ, Athanasopoulos G. Forecasting: Principles and Practice. ed. OTexts; 2014.

[39] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. International Journal of Forecasting. 2006;22:679-88. https://doi.org/10.1016/j.ijforecast.2006.03.001.

[40] Di Bucchianico A. Coefficient of Determination (R^2). Encyclopedia of Statistics in Quality and Reliability. 2007.

[41] Zhang S. Challenges in KNN Classification. IEEE Transactions on Knowledge and Data Engineering. 2022;34:4663-75. 10.1109/TKDE.2021.3049250.