IJAMEC
e-ISSN: 2147-8228

INTERNATIONAL JOURNAL OF APPLIED MATHEMATICS ELECTRONICS AND COMPUTERS

www.dergipark.org.tr/ijamec

International Open Access

Volume 09
Issue 04

December, 2021

*Research Article*

# A comparison of machine learning algorithms for forecasting solar irradiance in Eskişehir, Turkey

***Ozan AYKO*** [a], ***Sinem BOZKURT KESER*** [a,*]

[a]*Eskisehir Osmangazi University, Department of Computer Engineering, Eskişehir, Turkey*

ARTICLE INFO

*Article history:*
Received 14 September 2021
Accepted 14 December 2021

*Keywords:*
Solar irradiance prediction,
Machine learning,
Satellite data,
Ensemble learning.

ABSTRACT

This work compares the efficiency of 45 different machine learning (ML) algorithms to provide a comprehensive and most accurate model for global horizontal solar irradiance (GHSI) prediction in Eskişehir, Turkey. The dataset is provided by NASA Prediction of Worldwide Energy Resource (POWER) as satellite data that involves some characteristic weather condition variables such as temperature, precipitation, humidity etc. over 35 years. Some ML algorithms such as Extra Trees, LightGBM, HistGB, Random Forest (RF), Bagging and Decision Tree exhibit better performance among the others with commonly used statistical evaluation metrics in literature such as coefficient of determination ($R^2$), root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). In addition, Extra Tress regression slightly outperformed the rest of ensemble learning methods with $R^2$ of 0.99, RMSE of 8.05, MAE of 5.67, MAPE of 4%. Finally, the outcome demonstrates that the ML algorithms belonging to ensemble learning family achieved great results in GHSI prediction at specific location.

## 1. Introduction

Solar irradiance is the amount of light energy comes from Sun, hitting per square meter of Earth's surface. As this energy is renewable and clean, it becomes urgent more than ever for living a sustainable life in this consumerist age. The ever-growing air pollution, limited fossil energy that is very damaging to nature, climate changing and global warming are the factors threating our future. Because of this gloomy and depressing situation, scientists are seeking alternative energy sources such as solar energy, wind energy, tidal energy etc. Sun is the most promising and inexhaustible energy resource for us among these green and renewable energies and its applications should be initiated immediately [1].

Based on these facts, one of the ways to use solar energy most efficiently is to calculate solar irradiance accurately for places where it cannot be measured due to lack of systems. There are some challenges in observing and calculating this energy such as cost of supplies, their maintenance and calibration, data collection and storage issues [2, 3]. Statistical and data-driven approaches come into play at that point using machine learning (ML) and deep learning (DL) methods. There are several characteristic weather condition variables for global horizontal solar irradiance (GHSI) prediction such as temperature, humidity, precipitation, cloudiness and so on [4]. There are different types of solar irradiance: GHSI, spectral solar irradiance and total solar irradiance. So as to forecast solar irradiance, there are also numerous measurement techniques such as empirical models, image-based prediction and statistical methods [5]. Many studies have been and continue to be done on empirical models dating back to the 1920s to calculate solar irradiance [6]. In a study conducted in Kocaeli, Turkey in 2016, 30 empirical models that exist in literature were examined for solar irradiance prediction. However, even the best predictive models have performed significant deviations in summer [7]. Another study used empirical models is done in Şanlıurfa which is located in South-East in Turkey. 5 different models has applied and one of them has selected based on statistical evaluation metrics [8]. Empirical correlations were developed to build models for calculating monthly average daily solar radiation on the horizontal surface in Isparta, Turkey. Considering all the results, one model has been selected among them [3]. Mengeş, Sonmete and Ertekin has compared 50 empirical models to estimate global solar radiation on a horizontal surface in Konya, Turkey. Best model predicted the monthly average daily

* Corresponding author. E-mail address: sbozkurt@ogu.edu.tr
DOI: 10.18100/ijamec.995506

global radiation via multiple linear regression with 0.99 coefficient of determination [9].

Pedro et al. used machine learning algorithms named Gradient Boosting (GB) and *k*-Nearest-Neighbors (kNN) to calculate direct normal irradiance (DNI) and intra-hour global horizontal irradiance (GHI) ranging in a period 5-30 minutes. Image-based prediction was evaluated in this study and probabilistic metrics were used to measure error ratio. As a result, GB performs better when including sky images [10]. For estimating hourly global solar radiation in Eskişehir, Turkey, an empirical model Collares-Pereira & Rabl modified by Gueymard (CPRG) and machine learning (ML) methods such as ANN, Regression Tree and support vector regression (SVR) were applied. It is stated that the ML algorithms can be used instead empirical models and they gave, especially SVR, better results with the average of 0.97 coefficient of determination [11]. An artificial neural network was used to estimate daily total global sun radiation values for Mersin province in Turkey between April 2017 and March 2018. ANN was compared with other models available in the literature that predict global solar radiation. The best performing model has surpassed the ANN and has been the empirical model with the coefficient of determination value of 0.83 [2]. American Meteorology Society (AMS) organized a competition in 2013 to predict solar energy at 98 Oklahoma Mesonet sites. Linear and non-linear models such as least-square regression (LSR), regularized LSR, artificial neural network (ANN) has been compared. It is highlighted that the ensemble model of ANN and LSR models perform best among the other models [4]. In order to calculate monthly average daily global solar radiation values, Adaptive-Network Based Fuzzy Inference Systems (ANFIS), which is a hybrid AI method via ANN, and fuzzy logic have used and it is proved that ANFIS performs well with the mean absolute percentage error (MAPE) of 6% and the coefficient of determination ($R^2$) of 0.99 [12]. Kumari and Toshniwal proposed the ensemble XGBF-DNN model (extreme gradient boosting forest and deep neural network) to forecast hourly global horizontal irradiance in Jaipur, New Delhi and Gangtok. These models were integrated using Ridge regression to avoid overfitting and collinearity. Smart Persistennce (SP), SVR and Random Forest were used as benchmark models [1]. Recently, Aliyu et al. studied an artificial neural network (ANN) for estimating daily solar radiation in Nortwest Nigeria. It is showed that the proposed model ANN performed excellent to predict daily solar irradiance [13].

This study used long-short term memory (LSTM) for estimating global solar radiation by hour, day and month. The proposed LSTM model was compared with models such as recurrent neural network (RNN), convolutional neural network (CNN), backpropagation neural network (BPNN) according to whether the air is open or closed. LSTM model exhibited better performance than other models and gave an $R^2$ of 0.99 in sunny weather and 0.95 in cloudy water for all 3 cities Atlanta, New York and Hawaii [14]. Brahma and Wadhvani used LSTM model along with gated recurrent unit (GRU), CNN LSTM, attention LSTM and bidirectional LSTM in 2 India locations over 36 years. The forecasting tasks have performed better in shorter horizons [5]. Another LSTM work is done by Kara in Çorum, Turkey. The approach was compared some benchmark machine learning algorithms such as Random Forest, Gradient Boosting, kNN and Decision Tree. The outcome is that proposed LSTM model performed slightly better than other models in every statistical evaluation metrics [15]. Classic empirical models, deep neural network (DNN), ANN and time-series model were compared to find the best model for predicting daily global solar radiation in Eskişehir, Turkey. The obtained results showed that the best result is found by the DNN model [16]. Studies available in the literature are given below as shown Table 1.

**Table 1.** Some related studies in literature

| Reference | Best Model | Metrics | RMSE | $R^2$ (%) |
|---|---|---|---|---|
| [1] | XGBF-DNN | RMSE, MBE | 51.35 W/m² | - |
| [14] | LSTM | 90+ | 30 W/m² | |
| [15] | LSTM | $R^2$, RMSE, MAE, MAPE | 20.34 W/m² | 75.88 |
| [5] | Bidirectional LSTM & attention-based LSTM | $R^2$, MSE, RMSE | 10.25 & 11.16 W/m² | 63.21 & 56.4 |
| [2] | ANN | $R^2$, RMSE, MBE, MABE, MAE, MAPE | 1.10 kWh/m² | 75 |
| [13] | ANN | $R^2$, RMSE | 0.48 kWh/m2/day | 78 |
| [11] | SVR | $R^2$, RMSE, MBE | 63.86 W/m² | 97.95 |
| [16] | DNN | $R^2$, RMSE, rRMSE, MAE, MBE, tStatistic | 0.08 | 85.66 |
| [10] | GB & *k*NN | RMSE, Skill Score, CRPS, CRPSS, PICP, PINAW | 32.7 W/m² | - |

In this study, various machine learning algorithms are employed to forecast GHSI and best performing models are selected among them. Contribution of the study is that evaluating 45 different ML algorithms and determining the best ones for GHSI prediction and what family these algorithms belong to if there is a trend in that way. The paper

is organized in this flow: Section 2 states the methodology involving used dataset along with the variables, data pre-processing step, used machine learning algorithms and evaluation metrics. In Section 3, test environment that the study conducted on, obtained result along with the running time and error analysis are discussed. The conclusion is given in Section 4.

## 2. Methodology

In this section, the process of collecting and pre-processing of data is described. Moreover, the linear relationship and importance between independent variables and target variable is explained by Pearson correlation. After that, used scaling technique for machine learning algorithms is explained elaborately. Finally, statistical evaluation metrics that exist in the literature are given with their formulas. The schematic outline of the study was shown in Figure 1.

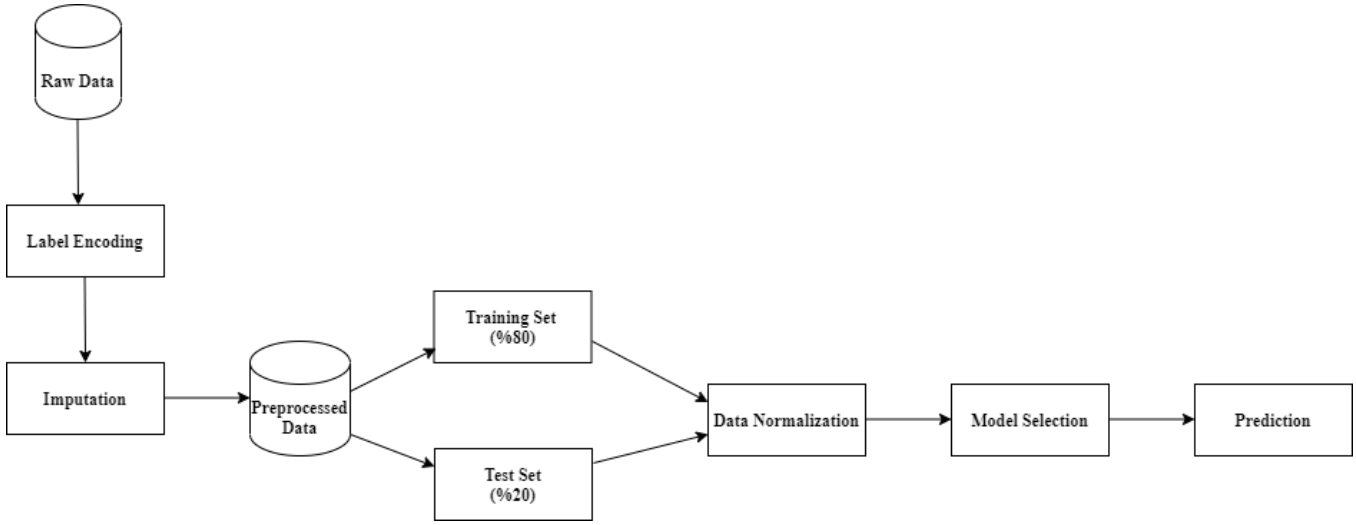


**Figure 1.** Framework of the study

### *2.1. Data Description*

The satellite-based data is provided by NASA Prediction of Worldwide Energy Resource (POWER) (https://power.larc.nasa.gov) [5]. It contains characteristic meteorological variables for GHSI prediction between January 1984 and December 2019 in Eskişehir, Turkey as shown in Table 2 [14].

**Table 2.** Meteorological variables. HS: Horizontal surface

| Feature | Definition | Unit |
|---|---|---|
| All Sky Insolation HS | Under all sky conditions, the average solar irradiance for j months at the horizontal surface of the earth's surface. | W/m² |
| Clear Sky Insolation HS | When the cloud content is less than 10%, the monthly average solar irradiance that hits the earth's surface on the horizontal plane. | W/m² |
| Relative Humidity | The relationship between the actual partial pressure of water vapour and the saturation partial pressure, expressed as a percentage. | % |
| Precipitation | Monthly average of daily precipitation rate. | mm/day |
| Surface Pressure | The average of surface pressure at the surface of the earth. | kPa |
| Dwn. Radiative Flux | Under all sky conditions, the j-month long-wave radiant flux average over several years at the horizontal surface of the earth's surface. | W/m² |
| Earth Skin Temp. | The average temperature at the earth's surface. | °C |
| Temp. Range | The minimum and maximum hourly temperature range (dry bulb) at 10 meters above the earth's surface during the relevant period. | °C |
| Wind Speed | The average of wind speed at 10 meters above the surface of the earth. | m/s |
| Clear Sky Clearness | The score representing the clarity of the atmosphere; the average of the total solar irradiance of the upper atmosphere divided by the clear sky sunshine that passes through the atmosphere and reaches the surface of the earth. | unitless |

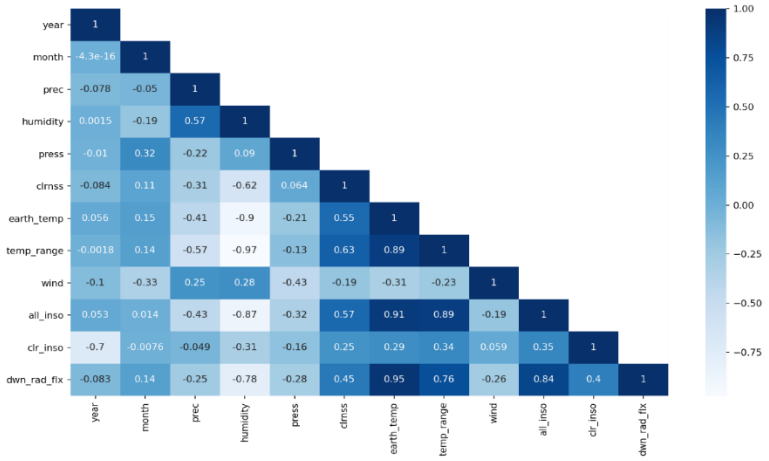


**Figure 2.** Pearson correlations in dataset

### *2.2. Data Pre-processing*

Data pre-processing is the process that is used to convert raw data into more interpretable and reliable format. Data cleaning and data transformation techniques are utilized in this study as a pre-processing step. Other approaches are imputation technique for handling missing values and encoding technique for converting categorical data into numerical data. There is only one categorical variable called month and label encoding is applied to it randomly regardless of ordering. As shown in Figure 2, correlation between variables in dataset is determined by Pearson correlation using heatmap in Python's seaborn library. Independent variables, especially several temperature and wind variables, are extracted from dataset due to multicollinearity issue to enhance the performance of model and its scores.

Figure 3 representing scatter plot tells us that there is a linear relationship between insolation incident and temperature by nature.

Data normalization is not applied to dataset directly. Instead, it is done after splitting the dataset into training and test set considering reproducibility.

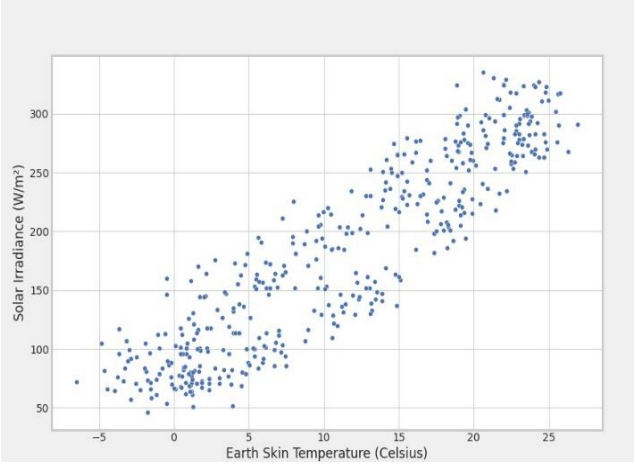


**Figure 3.** Linear relationship between earth skin temperature (°C) and solar irradiance (W/m²)

Another correlation chart the way more specific is that the correlation bar between target variable named as all sky insolation horizontal surface and other predictor variables. Temperature, earth skin temperature, downward radiative flux and clear sky insolation clearness variables are highly important for forecasting GHSI with the importance of 0.91, 0.88, 0.83 and 0.57, respectively, by Pearson correlation (see Figure 4). Relative humidity and precipitation are negatively correlated with the target variable all sky insolation horizontal surface with the importance of -0.86 and -0.43, respectively. However, wind speed, month and year variables look like unimportant.

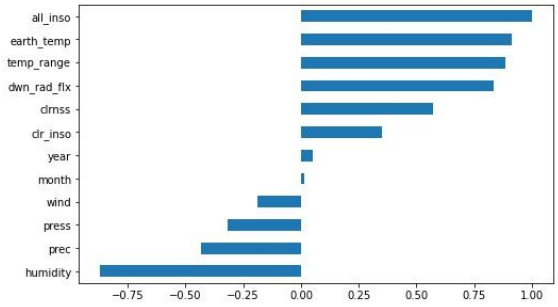


**Figure 4**. Correlation between the target variable and independent variables

### *2.3 Machine Learning Algorithms*

In this research, we used 45 different machine learning algorithms to compare each other based on the study [17]. The main goal is determining which machine learning algorithm will exhibit more robust performance for forecasting GHSI. The dataset is divided into two parts: training set and test set, and their ratio is determined as %80 and %20, respectively. For each machine learning algorithms, normalization technique is applied to numerical values. Min-max normalization approach that reduces numbers into range between 0 and 1 is selected as a feature scaling step, denoted as [14]:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

### *2.4 Performance Evaluation Metrics*

Commonly used statistical metrics are coefficient of determination ($R^2$), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) in literature. In this study, these evaluation metrics are utilized to verify the performance of proposed models to forecast GHSI. RMSE gives us an insight about the deviation error difference between the actual value and the predicted value in short term. The closer RMSE value is zero, the better the model performs. It is mathematically expressed as:

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

Another metric is coefficient of determination which determines the linear relationship between the actual value and the predicted value. It gives an information about how well the model fit the data. R2 value ranges between 0 and 1 and 1 indicates a strong linear relationship. This value can be calculated Equation (3):

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (3)$$

Mean absolute error (MAE) is the average of absolute error. It is calculated by Equation (4):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (4)$$

Mean absolute percentage error (MAPE) expresses the accuracy as a ratio with Equation (5):

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} x\ 100\% \quad (5)$$

where N is the number of samples, y is the actual value, yˆs the predicted value, y¯is the mean of predicted values.

## 3. Experiment Results and Discussions

In this section, obtained results of all applied machine learning algorithms and top 7 selected models among them for the best are discussed. Along with that, information of the environment in which the experiment was conducted is presented. Some commonly used models in machine learning are evaluated for giving better insight such as $R^2$, RMSE, MAE and MAPE values with their running time analysis.

### *3.1. Test Environment*

All experiments in this paper are conducted on Windows 10. Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz with 8 GB RAM. All the regression tasks are performed on Kaggle notebook using Python 3.7 programming language. Scikit-learn, one of the most popular machine learning algorithms library, is utilized on present Kaggle's packages for evaluating the experimental results.

### *3.2. Test Results*

#### *3.2.1. Running Time Analysis*

Computational time of the machine learning models is analysed in this section. The order of running time of the best 7 selected models are in this order: Decision Tree < Bagging = LightGBM < GB < ExtraTrees < HistGB < RF as shown in Figure 4. However, best performing models having the lowest running time – Bagging regression, LightGBM, Decision Tree – do not perform as well as ExtraTrees and LightGBM in terms of $R^2$ value (see Figure 5). Error analysis is mentioned in the next section.

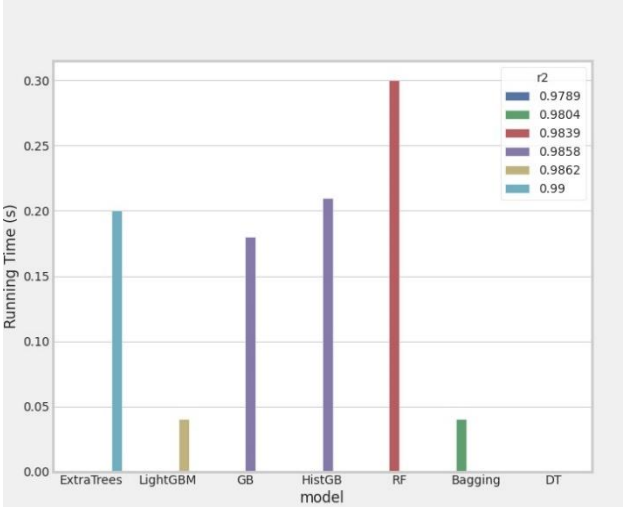


**Figure 5.** The running times of top 7 models

#### *3.2.2. Error Analysis*

Error analysis is the core evaluation approach in machine learning to compare efficiency of the algorithms. It helps us to select the best model in future predictions. In this study, prevalent error analysis techniques such as RMSE, MAE, MAPE are utilized to determine the accuracy of the models. 7 best performing models are given in Table 3 with their scores. In addition, RMSE values of all applied ML algorithms and selected top 7 models are given in Figure 6 and Figure 7, respectively. Moreover, coefficient determination ($R^2$) of top 7 algorithms is shown in Figure 8.

**Table 3.** Top 7 regression models. GB: Gradient Boosting

| Model | $R^2$ | RMSE | MAE | MAPE |
|---|---|---|---|---|
| ExtraTrees | **0.99** | **8.05** | **5.67** | **0.04** |
| LightGBM | 0.9862 | 9.465 | 6.961 | 0.049 |
| HistGB | 0.9858 | 9.576 | 6.96 | 0.05 |
| GB | 0.9858 | 9.598 | 7.018 | 0.052 |
| RF | 0.9839 | 10.217 | 6.90 | 0.05 |
| Bagging | 0.9804 | 11.261 | 7.669 | 0.056 |
| DT | 0.9789 | 11.698 | 6.886 | 0.05 |

Different ML algorithms such as Support Vector Regression (SVR), Gamma Regression, ElasticNet and Multiple Layer Perceptron (MLP) give high error as shown in Figure 6. Hyper-parameter optimization was not applied to any of the algorithms because results are excellent. Therefore, there was no improvement in the results when it was done. In addition, this process is not preferred because it takes time and keep processor busy unnecessarily.

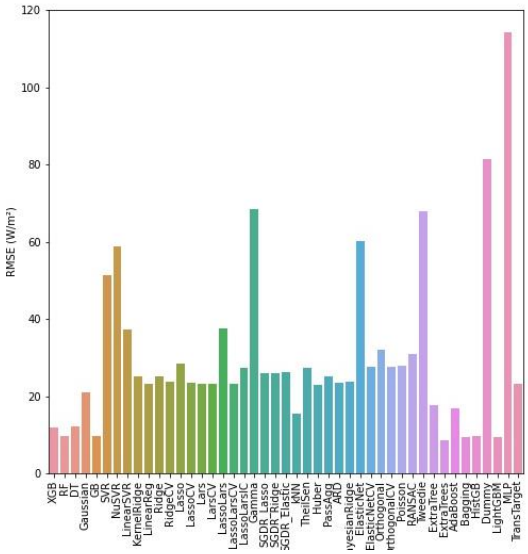


**Figure 6.** RMSE values of all ML algorithms

As one can see in Table 3, Figure 7 and Figure 8, Extra Trees Regression model performs slightly better than the other algorithms in terms of all evaluation metrics with the highest $R^2$ value (0.99), the lowest RMSE (8.05), MAE (5.67) and MAPE (4%). Since the difference of model performances are not too high, running time analysis should be considered as selecting a predictive solar irradiance model.

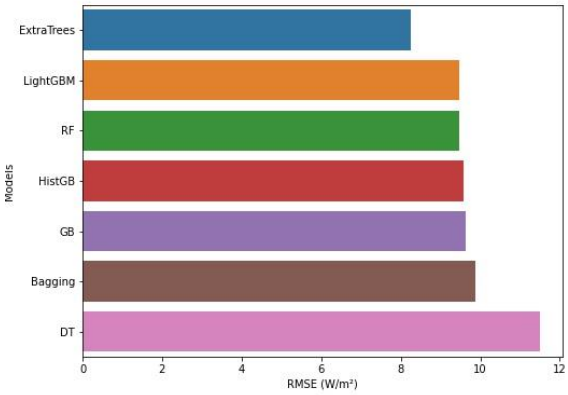


**Figure 7.** RMSE values of top 7 algorithms

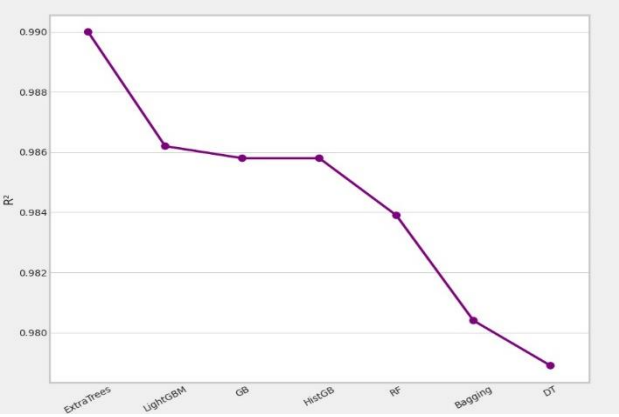


**Figure 8.** $R^2$ values of top 7 models.

Extra Trees Regression, which performs best, is selected for prediction model. It shows that how well predicted data fit to actual data as shown Figure 9.

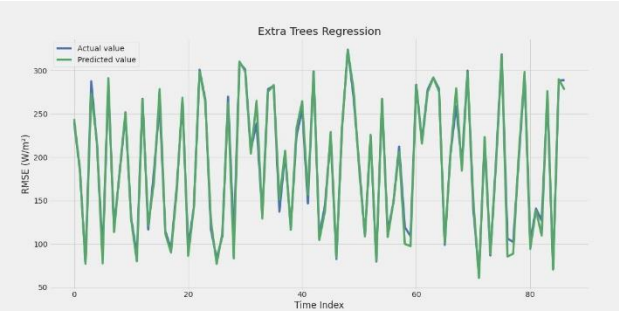


**Figure 9.** Solar irradiance prediction on randomly distributed data

## 4. Conclusion

Examining these results, one can infer from that ensemble learning methods give promising outputs in forecasting GHSI. 6 models out of 7 are ensemble learning algorithms except Decision Tree. The main outcome of this study is that the machine learning algorithms that belong to ensemble learning family achieved the highest accuracy and the lowest error values in implementing predictive solar irradiance model.

Moreover, it can be deduced that ensemble learning algorithms can be used for GHSI estimation instead empirical models that are used so far. Future research could focus on investigating time series solar irradiance and implementing the deep neural networks.

## References

[1] P. Kumari, D. Toshniwal, Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance, Journal of Cleaner Production 279 (2021) 123285.

[2] G. ARSLAN, B. BAYHAN, K. YAMAN, Mersin/Türkiye için Ölçülen Global Güneş Işınımının Yapay Sinir Ağları ile Tahmin Edilmesi ve Yaygın Işınım Modelleri ile Karşılaştırılması, Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji 7(1) (2019) 80-96.

[3] M. Öztürk, N. Özek, B. Berkama, Comparison of Some Existing Models for Estimating Monthly Average Daily Global Solar Radiation for Isparta, Pamukkale University Journal of Engineering Sciences 18(1) (2012) 13-27.

[4] S. Aggarwal, L. Saini, Solar energy prediction using linear and non-linear regularization models: A study on AMS (American Meteorological Society) 2013–14 Solar Energy Prediction Contest, Energy 78 (2014) 247-256.

[5] B. Brahma, R. Wadhvani, Solar irradiance forecasting based on deep learning methodologies and multi-site data, Symmetry 12(11) (2020) 1830.

[6] Y.S. GÜÇLÜ, Ş. Zekâi, Güneş Işınımı Tahmini için Yeni Bir Yaklaşım: OrtLin Modeli, İklim Değişikliği ve Çevre 5(1) (2020) 26-31.

[7] N. ARSLANOĞLU, KOCAELİ İÇİN MEVCUT GLOBAL GÜNEŞ IŞINIMI TAHMİN MODELLERİNİN UYGULANABİLİRLİĞİNİN DEĞERLENDİRİLMESİ, Uludağ University Journal of The Faculty of Engineering 21(1) (2016) 217-226.

[8] H. Karakaya, A.S. AVCI, U. Ercan, M.A. Kallioğlu, Şanlıurfa ilinde yatay yüzeye gelen anlık global güneş ışınımının modellenmesi, Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi 10(1) (2019) 147-155.

[9] H.O. Menges, C. Ertekin, M.H. Sonmete, Evaluation of global solar radiation models for Konya, Turkey, Energy Conversion and Management 47(18-19) (2006) 3149-3173.

[10] H.T. Pedro, C.F. Coimbra, M. David, P. Lauret, Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts, Renewable Energy 123 (2018) 191-203.

[11] M. ALSAFADI, Ü.B. FİLİK, HOURLY GLOBAL SOLAR RADIATION ESTIMATION BASED ON MACHINE LEARNING METHODS IN ESKISEHIR, Eskişehir Technical University Journal of Science and Technology A-Applied Sciences and Engineering 21(2) 294-313.

[12] G. Muhammet, E. Çelik, ANFIS kullanılarak Tunceli ili için global güneş radyasyonu tahmini, Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi 8(4) (2017) 891-899.

[13] S. Aliyu, A.S. Zakari, M. Ismail, M.A. Ahmed, An Artificial Neural Network Model for Estimating Daily Solar Radiation in Northwest Nigeria, FUOYE Journal of Engineering and Technology 5(2) (2020).

[14] Y. Yu, J. Cao, J. Zhu, An LSTM short-term solar irradiance forecasting under complicated weather conditions, IEEE Access 7 (2019) 145651-145666.

[15] K. Ahmet, Uzun-Kısa Süreli Bellek Ağı Kullanarak Global Güneş Işınımı Zaman Serileri Tahmini, Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji 7(4) (2019) 882-892.

[16] M. QASEM, Ü. BAŞARAN FİLİK, Solar radiation forecasting by using deep neural networks in Eskişehir, Sigma: Journal of Engineering & Natural Sciences/Mühendislik ve Fen Bilimleri Dergisi 39(2) (2021).

[17] S.S. Moustafa, M.S. Abdalzaher, M.H. Yassien, T. Wang, M. Elwekeil, H.E.A. Hafiez, Development of an optimized regression model to predict blast-driven ground vibrations, IEEE Access 9 (2021) 31826-31841.