

SHORT-TERM SOLAR RADIATION FORECASTING WITH DISCRETE WAVELET TRANSFORM BASED BOOSTED MACHINE LEARNING METHODS

Burak Arseven[✉], Said Mahmut Çınar

Received on January 27, 2025

Presented by P. Petkov, Member of BAS, on March 25, 2025

Abstract

In this article, details are given about the study of forecasting the radiation value of the next hour, which was carried out with the data obtained from the meteorological station in Afyonkarahisar. The hourly radiation data were first decomposed with the one-dimensional discrete wavelet transform (DWT) method at different levels. Ridge regression (RR) and lasso regression (LR) models were developed for each sub-signal. In order to reduce errors, the determined models were boosted using gradient boosting machines (GBM) on the residuals. The main radiation forecast signal was created with the DWT reconstruction of sub-signals' forecasts. Statistical and graphical results showed that the best radiation forecast was obtained with the proposed DWT-BLR model.

Key words: solar radiation forecast, discrete wavelet transform, ridge regression, lasso regression, gradient boosting machines

Introduction. The use of solar energy systems in electrical energy production is increasing. The effective use of the solar systems depends fundamentally on accurately forecasting solar radiation. In this context, many radiation forecast studies have been conducted using various variables and approaches. JI and CHAN [1] used the auto regressive moving average (ARMA) and time delay neural network (TDNN) methods for solar radiation forecast. NADEEM et al. [2] compared the radiation forecast performances of eight different machine learning

(ML) methods. They obtained the most accurate forecast results with support vector regression (SVR) among the eight different ML methods. ALAY et al. [3] compared the performances of convolutional neural networks (CNN), long short-term memory (LSTM), artificial neural network (ANN), and SVR models for solar radiation forecast. The findings showed that the most accurate forecasts were obtained with the CNN model. Some researchers have used regression-based ML methods for solar radiation forecast. ARSEVEN and ÇINAR [4] forecasted next hour's radiation by using auto regressive integrated moving average (ARIMA), ridge regression (RR) and lasso regression (LR) methods and the most accurate results were obtained with the ARIMA(2,1,1) model. ZHANG et al. [5] used XGBoost algorithm for solar radiation forecast. It was observed that more successful forecasts were obtained by using the air quality index variable in the structure of the XGBoost model. IVANOVA et al. [6] used two conventional regression methods for forecasting total solar radiation. The developed models with regression methods included extraterrestrial solar radiation, total cloudiness, relative humidity, minimum-maximum air temperatures, and atmospheric pressure variables.

In solar radiation forecast studies, signal decomposition methods have been used to reduce noise and increase the accuracy of the models. EŞLİK et al. [7] decomposed the radiation data into sub-levels using the empirical mode decomposition (EMD) method and performed forecasts on the sub-signals using the SVR method. MAJUMDER et al. [8] used variational mode decomposition (VMD) and robust kernel extreme learning machine (RKELM) methods in the structure of the hybrid method they proposed for solar radiation forecast. MOHAMMADI et al. [9] proposed an innovative hybrid model called SVM-WT based on support vector machine (SVM) and discrete wavelet transform (DWT) to forecast solar radiation. The study's findings revealed that the SVM-WT model was more successful than other models such as ANN, genetic programming (GP), and ARMA. BAGHEL et al. [10] proposed a new hybrid model for solar radiation forecast by combining DWT with regression learning methods. The model provided significant improvements in radiation forecast by integrating with NN. ARSEVEN and ÇINAR [11] utilized one-dimensional DWT to decompose hourly solar radiation, relative humidity, and temperature data, then developed multivariate RR (MRR) and LR (MLR) models based on different sub-signal combinations. By hybridizing these models within clearness index (CI)-based clusters, their proposed approach achieved high accuracy in adapting to varying cloudiness conditions for solar radiation forecasting.

In this study, a new hybrid radiation forecast algorithm was developed by integrating different methods in the literature. Regression-based ML methods RR and LR, signal decomposition method DWT, and model boosting tool gradient boosting machines (GBM) were used for the proposed algorithm. One of the strengths of the study is the comprehensive analysis of how different decomposition levels of DWT influence forecast performance. The best-performing decomposi-

tion level was determined by applying DWT at various levels. After that, the modelling continued to optimize DWT's effectiveness in solar radiation forecasting at that level. More effective forecasting models were developed by boosting the models created for the sub-signals with the GBM method. The GBM, used in the structures of regression-based methods in previous studies, was not applied to sub-signals obtained through such decomposition. This is another contribution of the study.

Material and methodology. This study focuses on the city of Afyonkarahisar in Turkey. Afyonkarahisar has a high solar energy potential. The annual solar radiation is approximately 1600–1700 kWh/m² [12]. The average daily sunshine duration of the city is around 7–8 h. The solar radiation data set used in the study was obtained from a station located at 38°44'16.8" N 30°33'37.4" E and covers four years, including a total of 35 064 h of measurement between January 1, 2018, and December 31, 2021. The first three and a half years of the data were used as training data, and the remaining six months were used as test data.

Solar radiation signals usually have a time-varying and non-stationary structure. The analysis of the components at varying frequencies over time is of critical importance in the context of non-stationary signals. At this point, signal components at different frequencies can be decomposed using time-frequency analysis methods such as Wavelet Transform (WT). The WT prevents the loss of time data of the signal while performing decomposition in the frequency domain [13]. In its most fundamental form, WT decomposes a signal into low and high-frequency components. While doing this, it uses wavelet structures in specific formats like Morlet, Daubechies (db), and Symlets. The WT is categorized into two primary temporal types: CWT allows the observation of temporal changes in all frequency components by applying a time shift for each frequency. However, since CWT requires high computational power, DWT is generally preferred in practical applications, because in most applications, analysis is not required for all frequency components [14].

The RR is a regression-based ML method. The line equation that does not entirely fit the training data is determined using the RR method. This low fit in the training data allows for a higher fit in the test data. Thus, more accurate and reliable results can be obtained in the testing and forecast stages with the determined RR model [15]. The LR method, quite similar to RR, is another ML method used in modelling processes to increase variable selection and forecast accuracy [16]. The most fundamental difference between the LR method, which is similar to the RR method, and RR is the penalty term. While the penalty term in LR uses the absolute values of the coefficients, the penalty term in RR uses the squares of the coefficients. Since RR and LR models primarily capture the underlying patterns in the data, they are expected to maintain their forecasting performance even when trained on data from different time periods recorded at the same location. As long as the general pattern remains consistent, the coefficients

of these models will exhibit minimal variation, ensuring their applicability across various time frames.

Boosting is a technique that allows weak learners of ML-based methods to be trained sequentially to evolve into a stronger and more accurate model. Gradient boosting machines (GBM) are one of these techniques that are frequently used in ML problems [17]. Decision trees (DT) used in GBM structures are generally known as non-deep or weak learners. These trees do not grow too much and do not form too complex structures. The DT divides the data into subsets by making the best split at each node, and at the end of this process, a forecast of the target variable is made at a leaf node. In this way, the model's overfitting is prevented, increasing overall performance. Each DT learns the errors of the previous trees and tries to correct these errors, thus improving the model's accuracy. A widely used variation of GBMs is the least squares boosting (LSBoost) algorithm. The LSBoost algorithm is designed for regression problems and uses an approach that minimizes the sum of squares of the error terms (residuals). In the LSBoost, the forecast errors made by the model are calculated at each step, and the next model is trained to minimize these errors. LSBoost minimizes the error using DTs; thus, the ML-based model is also boosted. The proposed method's comprehensive flowchart is presented in Fig. 1.

The method included the following steps: First, one-dimensional DWT was applied to the radiation data from the first to the fourth level. This process enabled in-depth analysis by revealing the frequency components of the data and dividing the data into smaller pieces (Step 1). Then, forecast models were created for all sub-signals using the training data of each sub-signal with RR and LR methods (DWT-RR and DWT-LR models). These created models reflected the most appropriate forecast models determined according to sub-signal patterns (Step 2). After the forecasts were made with these models, all models were boosted with LSBoost by using the residuals, which are the difference between the forecasted and actual values, as DT (Step 3). The forecasts made with the boosted models were combined using DWT reconstruction, and the final forecast results were obtained (Step 4). Finally, to examine the performances of the proposed forecast models (DWT-BRR and DWT-BLR), statistical metrics whose formulas are given in Table 1 are used (Step 5).

In this study, root mean square error (RMSE), normalized root mean square error (NRMSE), normalized mean absolute error (NMAE), and $\%R^2$ are used to determine the performances of the forecast models. Moreover, the Akaike Information Criterion (AIC) values are utilized to identify the optimal decomposition level for DWT. The formula of AIC is as follows

$$(1) \quad AIC = 2k - 2 \ln(\hat{L}).$$

The formulas of other statistical metrics are given in (2). This study used histogram and Kernel density estimation (KDE) methods to examine the distribution



Fig. 1. Flowchart of the proposed forecast methods

of forecast errors visually as well. The KDE eliminates sharp boundaries by producing a continuous and smooth density curve, thus visualizing the density distribution of data more precisely and providing smoother results than histograms [18].

Implementation and results of the proposed methods. Firstly, the main signal consisting of four-year hourly solar radiation data was decomposed into sub-signals using one-dimensional DWT from first to fourth levels. During decomposition, the db10 wavelet function was selected due to its similarity to the main signal and orthogonal structure. Using the training data of the approximation and detail signals, optimal DWT-RR and DWT-LR models were determined for each decomposition level. After determining the most suitable DWT-RR and DWT-LR models for the sub-signals at each level, forecasts for the next hours of all these sub-signals were performed for a six-month test period. The optimal

DWT level was determined based on the AIC values. The AIC values were examined for A_n and D_n signals from the first to the fourth level. Using DWT-RR models, AIC values for A_n were obtained as 8.71, 9.39, 9.73, and 10.44, respectively, for the decomposition levels. Again, using DWT-RR models, AIC values for D_n were obtained as 8.79, 8.98, 9.76, and 10.95, respectively, for the separation levels. Similarly, using DWT-LR models, AIC values very close to these were obtained. Upon examining the values, it was concluded that the optimal decomposition level is the first level. Thus, excessive decomposition and unnecessary model generation were avoided, eliminating additional computational burden.

The LSBoost technique was used to minimize the errors in the forecasts made with the models (DWT-RR and DWT-LR) determined for the sub-signals obtained due to the first-level DWT decomposition. The LSBoost aimed to improve the forecasts gradually by training a series of DTs sequentially in each iteration. The number of DTs and the number of leaves in each tree were determined using the trial-and-error method. In this context, the number of DTs was set to 3, and the number of leaves of each tree was set to 5. In this way, the number of DTs was kept at the optimum level while the complexity of the model was reduced. By setting the number of leaves at the optimum level, the trees were prevented from growing too much and causing the risk of overfitting. The sub-forecast signals obtained for approximation and detail signals using the proposed models (DWT-BRR and DWT-BLR) were combined through DWT reconstruction, thereby generating the main forecast signals. The comparison graphs, histograms, and KDE plots of the residuals obtained with DWT-RR and DWT-LR models, as well as those obtained with DWT-BRR and DWT-BLR models, are presented in Fig. 2. Figure 2 shows a notable increase in the concentration of residuals around the zero value, corresponding to the improved performance of forecast models boosted with GBM.

Table 1 presents the values

$$(2) \quad \begin{aligned} \text{RMSE} &= \sqrt{\frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{N}} \text{ W/m}^2, & \text{NRMSE} &= \frac{\sqrt{\frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{n}}}{\max(y_i)}, \\ \text{NMAE} &= \frac{1}{N} \sum_{i=1}^N \frac{|\hat{y}_i - y_i|}{\max y_i} & \text{and} & \quad R^2 = 100 \cdot \left(1 - \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (\bar{y}_i - y_i)^2} \right) \% \end{aligned}$$

of forecasts performed using RR, LR, DWT-RR, DWT-LR, DWT-BRR, and DWT-BLR methods. As a result, it was determined that the most accurate forecasts were achieved using the DWT-BLR method. In addition, the comparison graph of the forecasted results obtained with the DWT-BLR method and the actual values is presented in Fig. 3.

Conclusion. In this study, new hybrid models were proposed to forecast the radiation value with high accuracy for the next hour. The study examined

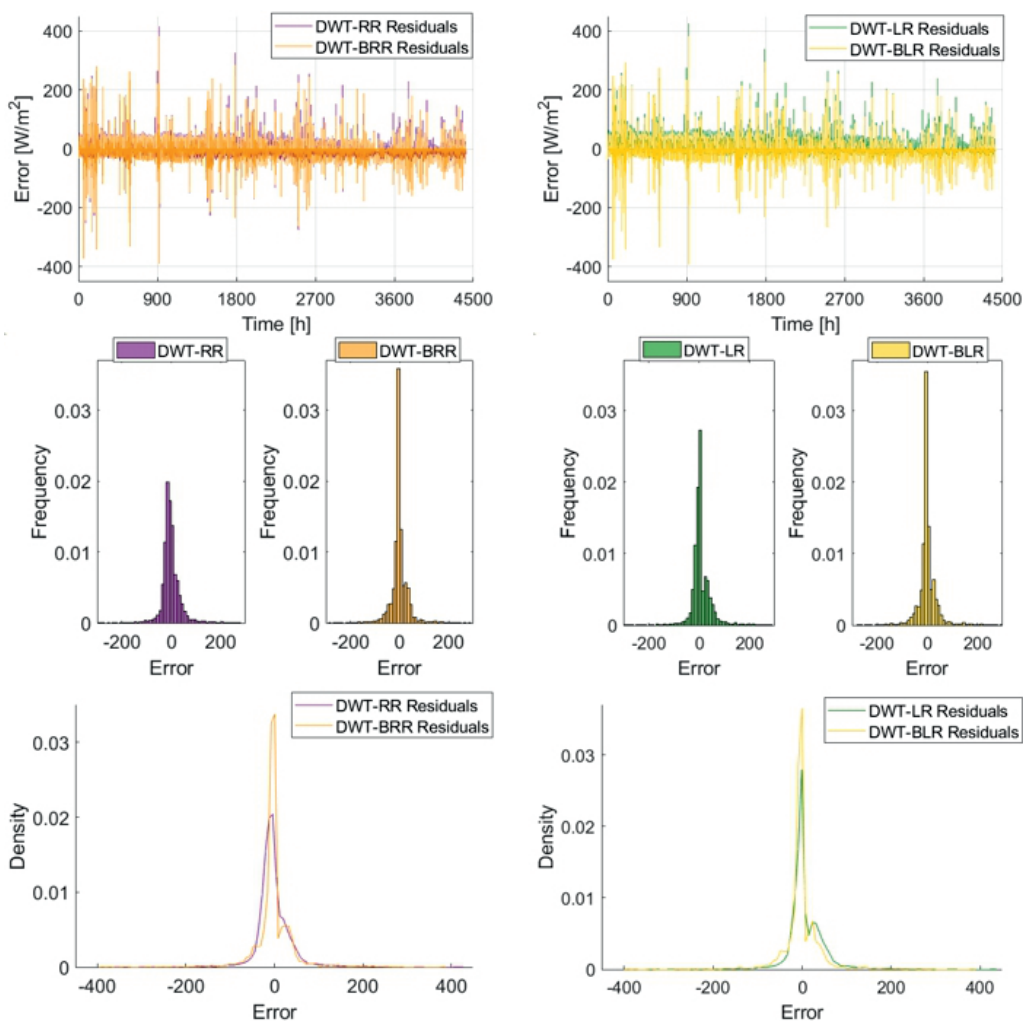


Fig. 2. Comparisons of residuals

T a b l e 1

Performance evaluations of methods

Method	RMSE (W/m2)	NRMSE	NMAE	R^2 %
RR	94.66	0.0890	0.0651	81.26
LR	95.52	0.0898	0.0598	80.89
DWT-RR	44.28	0.0416	0.0253	97.79
DWT-LR	44.02	0.0413	0.0234	97.82
DWT-BRR	41.92	0.0398	0.0225	98.03
DWT-BLR	40.24	0.0376	0.0204	98.19

at which level the best forecast result would be obtained using different DWT decomposition levels. When the forecast studies in the literature are examined,

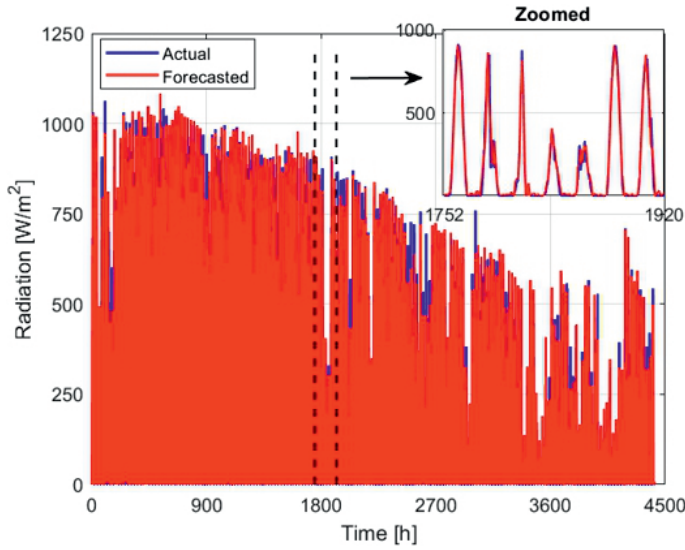


Fig. 3. Actual and forecasted radiations

it is seen that in the studies using DWT for decomposition, the main signal is usually decomposed at a single level, and the forecast is made (in some studies at the third level, in some at the fifth level). However, information must be provided to explain why the decomposition is done at these levels. In addition, after determining the most appropriate RR and LR models for the sub-signals, these models were boosted using GBM. This approach has not been used in the literature before. Consequently, it was inferred that the first level is the most suitable decomposition level. In addition, errors were significantly reduced by boosting the models. It is thought that more accurate radiation forecast results can be obtained by integrating different methods in the future.

REFERENCES

- [1] JI W., C. K. CHAN (2011) Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN, *J. Sol. Energy*, **85**(5), 808–817.
- [2] NADEEM T. B., S. U. ALI, M. ASIF, H. K. SUBERI (2024) Forecasting daily solar radiation: An evaluation and comparison of machine learning algorithms, *AIP Advances*, **14**(7), <https://doi.org/10.1063/5.0211723>.
- [3] ALAY F. D., N. İLHAN, M. T. GÜLLÜOĞLU (2024) Solar radiation prediction in PV power systems: a comparison of deep learning models using big data, *C. R. Acad. Bulg. Sci.*, **77**(9), 1347–1354.
- [4] ARSEVEN B., S. M. ÇINAR (2023) Use of ARIMA, ridge regression and lasso regression methods improved with extraterrestrial radiation in hourly radiation estimation, *NÖHÜ Eng. Sci. J.*, **12**(3), 733–742, (in Turkish).

- [5] ZHANG C., Y. ZHANG, J. PU et al. (2025) An hourly solar radiation prediction model using eXtreme gradient boosting algorithm with the effect of fog-haze, *Energy Built Environ.*, **6**(1), 18–26.
- [6] IVANOVA I. Y., V. SHAKIROV, N. KHALGAEVA (2023) Accuracy analysis of estimates of total solar radiation in databases and regression models for Eastern Russia, *Geogr. Nat. Resour.*, **44**(3), 278–283.
- [7] EŞLİK A. H., E. AKARSLAN, F. O. HOCAOĞLU (2021) Segregation-combination learning approach for solar radiation estimation, *JIST*, **11**(1), 132–144, (in Turkish).
- [8] MAJUMDER I., P. K. DASH, R. BISOI (2018) Variational mode decomposition based low rank robust kernel extreme learning machine for solar irradiation forecasting, *Energy Convers. Manage.*, **171**(6), 787–806.
- [9] MOHAMMADI K., S. SHAMSHIRBAND, W. T. CHONG et al. (2015) A new hybrid support vector machine-wavelet transform approach for estimation of horizontal global solar radiation, *Energy Convers. Manage.*, **92**(4), 162–171.
- [10] BAGHEL R. S., A. K. GOSH, Y. PATIDAR et al. (2023) SIF-DWTRL: solar irradiation forecasting using discrete wavelet transform and regression learning, *Energy Syst.*, **16**, 395–411.
- [11] ARSEVEN B., S. M. ÇINAR (2025) A novel hybrid solar radiation forecasting algorithm based on discrete wavelet transform and multivariate machine learning models integrated with clearness index clusters, *JASTP*, **267**, 106417.
- [12] DEMIRKOL Z., M. CUNKAŞ (2014) Renewable energy potential of Afyonkarahisar province, *Selcuk Univ. Eng. Sci. Technol. J.*, **2**(1), 45–56, (in Turkish).
- [13] DAUTOV C. P., M. S. ÖZERDEM (2018) Introduction to wavelets and their applications in signal denoising, *BEU J. Sci. Technol.*, **8**(1), 1–10.
- [14] MALLAT S. (1999) *A Wavelet Tour of Signal Processing*, San Diego, Elsevier.
- [15] TIBSHIRANI R. (1996) Regression shrinkage and selection via the lasso, *J. R. Stat. Soc., B: Stat. Methodol.*, **58**(1), 267–288.
- [16] ARLOT S., A. CELISSE (2010) A survey of cross-validation procedures for model selection, *Statist. Surv.*, **4**, 40–79.
- [17] FRIEDMAN J. H. (2001) Greedy function approximation: a gradient boosting machine, *Ann. Stat.*, **29**(5), 1189–1232.
- [18] SCOTT D. W. (2015) *Multivariate Density Estimation: Theory, Practice, and Visualization*, Hoboken, John Wiley & Sons.

Department of Electrical Engineering, Afyon Kocatepe University, Ahmet Necdet Sezer Campus, 03200, Afyonkarahisar, Türkiye
e-mails: arseven@aku.edu.tr, smcinar@aku.edu.tr