Communications for Statistical Applications and Methods
2022, Vol. 29, No. 6, 709–719

https://doi.org/10.29220/CSAM.2022.29.6.709
Print ISSN 2287-7843 / Online ISSN 2383-4757

Solar radiation forecasting using boosting decision tree and recurrent neural networks

Hyojeoung Kim[a], Sujin Park[a], Sahm Kim[1,a]

[a]Department of Applied Statistics, University of Chung-Ang, Korea

Abstract

Recently, as the importance of environmental protection has emerged, interest in new and renewable energy is also increasing worldwide. In particular, the solar energy sector accounts for the highest production rate among new and renewable energy in Korea due to its infinite resources, easy installation and maintenance, and eco-friendly characteristics such as low noise emission levels and less pollutants during power generation. However, although climate prediction is essential since solar power is affected by weather and climate change, solar radiation, which is closely related to solar power, is not currently forecasted by the Korea Meteorological Administration. Solar radiation prediction can be the basis for establishing a reasonable new and renewable energy operation plan, and it is very important because it can be used not only in solar power but also in other fields such as power consumption prediction. Therefore, this study was conducted for the purpose of improving the accuracy of solar radiation. Solar radiation was predicted by a total of three weather variables, temperature, humidity, and cloudiness, and solar radiation outside the atmosphere, and the results were compared using various models. The CatBoost model was best obtained by fitting and comparing the Boosting series (XGB, CatBoost) and RNN series (Simple RNN, LSTM, GRU) models. In addition, the results were further improved through Time series cross-validation.

Keywords: solar radiation, XGBoost, CatBoost, simple RNN, LSTM, GRU, time series cross-validation

1. Introduction

Korea is highly dependent on fossil fuels, accounting for 85% of the country's total energy consumption. The fossil fuel-based energy supply system has low sustainability due to price volatility, limitation of fuel reserves, and environmental problems, which are spurring the development of the new renewable energy industry to generate steep growth. In particular, solar energy has the highest production rate of 38% for new and renewable energy in Korea due to its infinite resources, ease of installation, and eco-friendly features such as low noise emission levels and less pollutants. However, solar power generation requires advanced power generation prediction technology due to unstable energy supply from the influence of the weather.

Recently, as a solution to this, the new renewable energy industry is showing a rapid growth trend in quantity. In particular, the solar energy field has infinite resources, easy installation and maintenance, and eco-friendly characteristics such as low noise emission levels and less pollutants during

This research was funded by Korea Institute of Energy Technology Evaluation and Planning (20199710100060), and the National Research Foundation of Korea (2016R1D1A1B01014954).
This research was supported by the Chung-Ang University Research Grants in 2021.
[1] Corresponding author: Department of Statistics, Chungang University, 84 Heuksukro, Dongjak-Gu, Seoul 06974, Korea.
E-mail: sahm@cau.ac.kr

Published 30 November 2022 / journal homepage: http://csam.or.kr

power generation, accounting for the highest production rate of 38% of new renewable energy in Korea. Moreover, although the cost of solar power generation is the highest among the currently supplied renewable energy sources, the cost of solar power is decreasing every year through technological development, and the cost of fossil fuel is expected to be the same within a few years. (Lee *et al.*, 2018) This is expected to increase the proportion of solar energy generation in the future.

However, since solar power generation is affected by weather and climate change, stable energy supply and demand are impossible compared to fossil fuels. Climate prediction is essential to compensate for the shortcomings of this solar power supply system. Solar radiation, which is closely related to solar power generation, is not currently predicted by the Korea Meteorological Administration. Therefore, solar radiation prediction can be the basis for establishing a reasonable new renewable energy operation plan, and this is very important because this can also be used in other fields such as power consumption prediction.

Various methods are being attempted at home and abroad to predict solar radiation. Solar radiation is the amount of energy when solar energy arrives to the ground, and it is difficult to find a trend within solar radiation. Solar energy tends to decrease from the actual value due to various weather variables such as water vapor and cloudiness as it passes through the Earth's atmosphere, so it is necessary to consider exogenous variables to predict solar radiation.

The meteorological variables mainly considered in predicting solar radiation include temperature, relative humidity, and cloudiness. Suh *et al.* (2018) compared the time series models ARIMA, ARIMAX, SEASONAL ARIMA, SEASONAL ARIMA, SEASONAL ARIMAX, ARIMA-GARCH, SEASONAL ARIMA-GARCH, SEASONAL ARIMA-GARCH, SEASONAL ARIMAX-GARCH, and SEASONAL ARIMAX-GARCH. It was confirmed that ARIMAX, seasonal ARIMAX, and seasonal ARIMAX-GARCH considering temperature and humidity as exogenous variables showed better performance than time series models without considering exogenous variables. Suh *et al.* (2018); Elizabeth *et al.* (1994) predicted global solar radiation by using various weather variables including temperature and humidity. Elizondo *et al.* (1994) and Rehman *et al.* (2008) predicted solar radiation using an ARN (Artificial Neural Network) model using the average temperature, maximum temperature, and relative humidity of a day to predict solar radiation in AbhCity, Saudi Arabia. (Rehman and Mohandes, 2008) Alam *et al.* (2009) considered the latitude (lat), longitude (long), altitude (alt), time, months of the year (moy), air temperature (at), relative humidity (rh), rainfall (rf), wind speed (ws), and net long wavelength (lw) as input parameters for ANN model. Alam *et al.* (2009); Sahm (2017) compared the prediction accuracy of the time series model by generating variables of out-of-the-atmosphere solar radiation using meteorological elements, latitudes, and longitudes for solar radiation prediction. Sahm (2017)

Juliang (2018) compared the accuracy of the Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost) using the Daily global solar radiation, maximum and minimum air temperatures (Tmax and Tmin) and transformed precipitation (Pt, 1 for rainfall $\neq$ 0 and 0 for rainfall = 0) in estimating daily solar radiation. As a result, the XGBoost model had comparable prediction accuracy with the SVM model. (Fan *et al.*, 2018) Also, Fan *et al.* (2020) assessed the performance of SVM and four tree-based soft computing models (M5Tree, RF, XGBoost and CatBoost) for predicting daily horizontal radiation (Rd) on in various climatic rons (Scenario 1: local, Scenarios 2 and 3: extrinsic) of China. Comprehensively considering prediction accuracy, generalization capability and computational efficiency, CatBoost is the best model to develop general models. Fan *et al.* (2020); Pang *et al.* (2020) figured out that the solar radiation prediction using the RNN model has a higher accuracy compared with the ANN model. Besides, in the case of relatively high cloud cover during the training period, results were severely biased from the measured values in the test period, suggesting that cloud

cover could have a significant impact on prediction accuracy (Pang *et al.*, 2020).

As with the above-mentioned models, various methods have been applied to improve the accuracy of predicting solar radiation in recent studies. In this paper, we would like to compare a total of seven models, which were frequently used in the above paper: Boosting series models (CatBoost, XGBoost), and RNN series models (Simple RNN, LSTM, GRU). The data used were ASOS (Automated Synoptic Observation System) from March 1, 2017 to February 28, 2022 at 42 points in Korea. A model was selected by applying various hyper parameters, and the final model performance was evaluated by MAE and RMSE.

Section 2 introduces the ARIMAX, XGBoost, CatBoost, Simple RNN, LSTM, and GRU models used for solar radiation prediction. Section 3 explains the weather data and preprocessing method used in this study, and compares and analyzes the prediction results by applying the above- mentioned model. Section 4 will propose conclusions and future research directions.

2. Methodology

This section introduces the various methods used in the study. In this paper, we use Autoregressive Integrated Moving Average Exogenous Variable Models (ARIMAX), the classic time series models, to compare how well the finally considered model predicts over the classical model. Compare the Boosting family models XGBoost, CatBoost, and the existing RNN family-based models Simple RNN, LSTM, GRU.

This section introduces the various methods used in the study. In this paper, we use Autoregressive Integrated Moving Average Exogenous Variable Models (ARIMAX), the classic time series models to compare how well the models considered predict than the classic models. Finally, we compare the Boosting family models XGBoost, CatBoost, and the existing RNN family-based models Simple RNN, LSTM, GRU.

2.1. ARIMAX (Auto-regressive integrated moving average with exogeneous variable)

The ARIMAX model adds exogenous variables to the ARIMA model. Like the ARIMA model, it has been used as a prediction model in various fields. When the degree of ARIMA is p, d, q and the number of exogenous variables is k, the exogenous variables are denoted as x_{it}, and the ARIMAX (p, d, q) model is as follows

$$\phi_p(B)(1-B)^d(Y_t-\mu) = \theta_q(B)\epsilon_t + \sum_{i=1}^{k}\beta_i x_{it} \quad \text{where } \epsilon_t \sim \mathrm{WN}\left(0, \sigma^2\right), \tag{2.1}$$

where $\phi_p(B)$ corresponds to the equation for the autoregressive model, p denotes the order of the current model, $\theta_q(B)$ is the equation for the moving average model, q represents the order of the current model, d denotes the equation containing the first difference, ϵ_t corresponds to an error term or white noise, and β_i is a coefficient of the exogenous variable, x_{it}.

2.2. XGBoost (extreme gradient boosting)

The XGBoost model is a decision tree-based algorithm that improves gradient boosting and is used in various studies. Obiora *et al.* (2021); Huang *et al.* (2021); Rocha and Santos (2022) Gradient boosting is a machine learning technique that increases predictive power by sequentially generating a model by supplementing the predictive error of the previous tree with the slope-lowering method using gradient descent. Through the repetition process of creating a new prediction model by focusing

on poorly predicted individuals, a strong model is generated through a combination of several weak models. The XGBoost model supports parallel operations, so it is fast, can be used for very large models, and has the advantage that overfitting does not occur well. However, there is a disadvantage that it is complicated due to the large number of parameters (Kim and Kim, 2022). The XGBoost model consists of M decision trees, as in the following expression, where f denotes one decision tree, and F denotes a function of all decision trees:

$$Y_i = \sum_{m=1}^{M} f_m\left(x_{(i)}\right), \quad f_m \in F. \tag{2.2}$$

In the regression process, the model is expressed by the following equation:

$$\mathrm{Obj}\left(\theta\right) = \sum_{i=1}^{n} l\left(y_i, Y_i\right) + \sum_{m=1}^{M} \Omega\left(f_m\right), \quad \theta = \left(f_1, f_2, \ldots, f_m\right) e, \tag{2.3}$$

where l represents the loss function, and Ω indicates the regulation to prevent overfitting. The regulation equation follows

$$\Omega\left(f\right) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2, \tag{2.4}$$

where T denotes the number of nodes in the decision tree, Ω represents a weight vector, and γ and λ are penalty elements. In this paper, the hyperparameters are optimized to have excellent predictive power by adjusting the maximum depth, learning rate, and parameters.

2.3. CatBoost (unbiased boosting with categorical features)

The CatBoost model is a boosting-based model along with XGBoost that creates a tree with Level-wise. In addition, the overall aspect of the existing boosting process is similar, but there are some differences. If the existing boosting model performed residual calculations on all training data collectively, the model takes some of the residual calculations and then makes a model, and the residual of the subsequent data goes through the process of using the predicted value with this model. In addition, it has the advantage of preventing Overfitting by randomly selecting data through Random Permutation and making trees diverse. (Fan *et al.*, 2020) This is as follows:

$$\hat{x_k}^i = \frac{\sum_{j=1}^{n} \left[x_j^i = x_k^j\right] \circ y_j + \alpha P}{\sum_{j=1}^{n} \left[x_j^i = x_k^j\right] + \alpha}, \tag{2.5}$$

where α is corresponding weight, P denotes a prior value, $x_k = (x_k^1, \ldots, x_k^m)$ denotes random vector of m features an $y_k IN\mathbb{R}$ denotes corresponding label.

Unlike other ensemble algorithms that use GridSearchCV or RandomizedSearchCV to find optimal hyperparameters, this model has optimized initial hyperparameter values, so no separate parameter tuning procedure is required.

2.4. RNN (recurrent neural network)

A recurrent neural network refers to a neural network having one or more cyclic layers within a hidden layer. It is mainly used for ordered time series data, and the corresponding data is input and output is

obtained in order through one network. Weights and biases are repeatedly used for each time zone's data. In addition, previous data affects the results because the output of the layer is received and used as input again. These RNNs are characterized by unlimited input and output lengths, and can form various types of networks by changing their structures. Simple RNN, the model used in this paper, is the simplest form of RNN layer, and the structure is shown in the figure below (Sorkun *et al.*, 2017).

However, RNNs have disadvantages such as Gradient Vanishing or Exploding because the models are backpropagation algorithms over time, and the problem is Song-Term Dependency, where learning ability decreases as input data grows.

2.5. LSTM (long short term memory)

The Long Short-Term Memory (LSTM) model was developed in 1997 by Hawkrite and Schmitthuber Hochreiter and Schmidhuber (1997); Kumari and Toshniwal (2019) This model is proposed to solve the Long-Term Dependency problem of the Basic RNN model and used in several time series prediction studies (Sorkun *et al.*, 2020). LSTM also has a chain-like structure like Basic RNN, but each iteration module has a different structure. That is, instead of a simple layer, four layers are configured to exchange information with each other in a special manner. The core of LSTM is the cell state, which corresponds to the 'memory' part. The cell state is controlled by a gate, and the model consists of an input gate, a forget gate, and an output gate. In other words, if this is summarized as one equation, it is as follows:

$$\begin{aligned} f &= \sigma\left(x_t w_x^{(f)} + h_{t-1} w_h^{(f)} + b^{(f)}\right), \\ g &= \tanh\left(x_t w_x^{(g)} + h_{t-1} w_h^{(g)} + b^{(g)}\right), \\ i &= \sigma\left(x_t w_x^{(i)} + h_{t-1} w_i^{(f)} + b^{(i)}\right), \\ o &= \sigma\left(x_t w_x^{(o)} + h_{t-1} w_h^{(o)} + b^{(o)}\right), \\ c_t &= f \bigodot c_{t-1} + g \bigodot i, \\ h_t &= o \bigodot \tanh\left(c_t\right). \end{aligned} \tag{2.6}$$

o is output gate, h_t is output gate multiplied by tanh (c_t) to generate h_t, which is hidden state. f is a forget gate, which is multiplied by c_{t-1} to generate a memory cell c_t. It then adds g, a new information to remember, and selectively accepts memories rather than acceptings all memories at the input gate, i. The LSTM receives information from the previous hidden state and the previous input from the forget gate, which is a sigmoid layer that determines which information to discard and transmits it to the cell state. The cell state is then updated based on the information received from the input gate and the forget gate. But LSTM has the limitation of too-heavy structures and many learning parameters.

2.6. GRU (Gated Recurrent Unit)

The Gated Recurrent Unit (GRU) was proposed by Cho *et al* in 2014 for similar reasons as the LSTM model, which further simplified the Cell of the Time-Step constituting the LSTM to reduce the computation volume (Cho *et al.*, 2014). This model has a slightly faster learning rate than LSTM, and when the amount of data is small, GRUs with a small amount of parameters are known to have better predictive performance than LSTM (Boubaker *et al.*, 2021). GRU integrates the three gates of LSTM into two, and the cell state and the hidden state into one hidden state. As a result, the number of learning parameters is reduced, making it relatively light. The computation process of GRU is performed

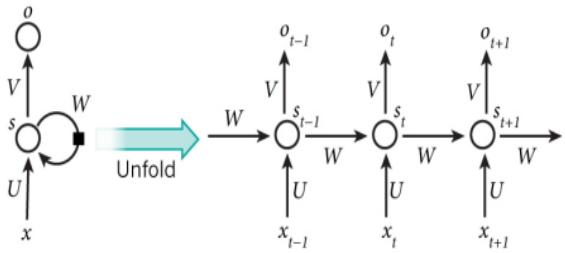

Figure 1: *Gulli and Pal (2017).*

through the following computation equation.

$$
\begin{aligned}
z_t &= \sigma\left(W_z x_t + U_z h_{(t-1)} + b_z\right), \\
r_t &= \sigma\left(W_z x_t + U_r h_{(t-1)} + b_r\right), \\
\hat{h}_t &= \phi(W_z x_t + U_z(r_t \odot h_{(t-1)}) + b_r).
\end{aligned} \tag{2.7}
$$

The $h_{(t-1)}$ transferred from the cell at the previous time point and the x_t value, which is a new input value, are combined and divided into two directions and transmitted. One serves as the information to be delivered to the gate, and the other serves as the LSTM's candidate state. In the above formula, z_t represents the update gate, r_t represents the reset gate, and h_t represents the candidate state, and r_t resets the information of h_t, and z_t determines which information to pass on and on the hidden state.

3. Application of models

3.1. Data collection and preprocessing

The data applied to this study are weather data provided from the weather portal data every hour from March 1, 2017 to February 28, 2022. From March 1, 2017, to February 28, 2021, the 5-year data were used as training data to fit the model. In addition, the remaining data from March 1, 2021 to February 28, 2022 were used to evaluate the performance of the model as test data. A total of four weather variables such as temperature, humidity, cloud, and wind speed and the out-of-atmosphere solar radiation proposed by Sahm (2017) were used as input variables of the model.

In order to proceed with the analysis, national weather data were reconstructed through preprocessing. The points where the insolation values and transfer volumes were missing for a long time were removed, and analysis was conducted at a total of 18 points as shown in Figure 1. In addition, solar radiation starts after sunrise, forming the highest point, and after sunset, a pattern with a value of 0 is repeated every day. Due to this characteristic, there were many parts with 0 or NA-values after sunset and before sunrise, and the time when most points are observed for each season is set for sunrise and sunset, as shown in the Table 1, and the rest of the time is assumed not to be observed (sunshine value = 0).

In addition, days when solar radiation was not observed much for 24 hours for each point were excluded, and values were replaced using linear interpolation when temperature, humidity, cloudiness were missing. Using the final preprocessed data, the prediction results were compared by fitting the XGBoost, CatBoost, Simple RNN, LSTM, and GRU models.

Figure 2: *Solar radiation prediction point.*

Table 1: Sunrise and sunset time by season

Season	Month	Sunrise time	Sunset time
Spring	3, 4, 5	7 AM	7 PM
Summer	6, 7, 8	6 AM	8 PM
Autumn	9, 10, 11	7 AM	7 PM
Winter	12, 1, 2	8 AM	6 PM

3.2. Performance evaluations

Two measures of error were used: MAE and RMSE to compare suitable models in the previous section. In general, MAPE is widely used to evaluate the model, but it is difficult to apply a MAPE calculation because the solar radiation amount is often 0. Therefore, we evaluated accuracy on a scale of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which are defined as follows.

$$\begin{aligned} \text{MAE} &= \frac{\sum_{i=1}^{n} |Y_t - F_t|}{n}, \\ \text{RMSE} &= \sqrt{\frac{\sum_{t=1}^{n} (Y_t - F_t)^2}{n}}. \end{aligned} \tag{3.1}$$

Here, n is the number of data used for prediction, and is the observed value at time t, and is the predicted value through the model at time t. In both MAE and RMSE, the smaller the value, the higher the accuracy.

3.3. Performance results

ARIMAX was automatically designated using R's auto.arima function without cross-validation. Other models used Gridsearch CV for hyperparameter optimization. The Boosting family adjusted learning rates (0.05, 0.07, 0.1) and max depth (3, 8, 10), while the RNN family tuned num_unit (32, 64) and optimizer ('Nadam', 'Adam', 'RMSProp'). The hyperparameter was selected as the value with the smallest MAE average value, and the results are shown in Table 2. XGB had the smallest average

Table 2: Hyperparameter optimization results

Station	XGB (max depth / learning rate)		CatBoost (max depth / learning rate)		LSTM (n_unit / optimizer)		GRU (n_unit / optimizer)	
	Parameter	MAE Mean	Parameter	MAE Mean	Parameter	MAE Mean	Parameter	MAE Mean
Gwangju	3 / 0.07	0.110	10 / 0.1	0.101	64 / Nadam	0.142	64 / RMSProp	0.127
Daegu	3 / 0.07	0.102	10 / 0.1	0.093	64 / Adam	0.126	64 / Adam	0.117
Daejeon	3 / 0.07	0.118	10 / 0.1	0.108	64 / Nadam	0.148	64 / Nadam	0.131
Mokpo	3 / 0.07	0.116	10 / 0.1	0.106	64 / Adam	0.145	64 / Adam	0.130
Bukgangneung	3 / 0.07	0.102	10 / 0.1	0.092	64 / Adam	0.129	64 / Adam	0.120
Bukchuncheon	3 / 0.07	0.112	10 / 0.1	0.104	32 / Adam	0.155	64 / Adam	0.137
Seoul	3 / 0.07	0.114	8 / 0.1	0.105	64 / Adam	0.139	64 / Adam	0.128
Suwon	3 / 0.07	0.107	10 / 0.1	0.099	64 / Adam	0.139	64 / Adam	0.122
Andong	3 / 0.07	0.109	10 / 0.1	0.097	64 / Nadam	0.139	64 / Nadam	0.122
Yeosu	3 / 0.07	0.109	10 / 0.1	0.098	64 / Adam	0.131	64 / Adam	0.122
Ulleung Island	3 / 0.07	0.129	10 / 0.1	0.118	32 / RMSProp	0.143	64 / RMSProp	0.129
Incheon	3 / 0.07	0.119	10 / 0.1	0.110	64 / Nadam	0.138	64 / Adam	0.129
Jeonju	3 / 0.07	0.111	10 / 0.1	0.100	64 / Nadam	0.139	64 / Adam	0.127
Jeju Island	3 / 0.07	0.115	10 / 0.1	0.103	64 / Nadam	0.138	64 / RMSProp	0.123
Changwon	3 / 0.07	0.105	10 / 0.1	0.092	64 / Adam	0.132	64 / Nadam	0.118
Cheongju	3 / 0.07	0.108	10 / 0.1	0.099	64 / Adam	0.135	64 / Adam	0.122
Pohang	3 / 0.07	0.134	10 / 0.1	0.121	64 / RMSProp	0.150	64 / Nadam	0.139
Heuksan Islnad	3 / 0.07	0.147	10 / 0.1	0.133	64 / Nadam	0.169	64 / Nadam	0.169

Table 3: Test data fit performance results

Station	ARIMAX		XGB		CatBoost		LSTM		GRU		Simple RNN	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Gwangju	0.278	0.486	0.114	0.226	0.107	0.224	0.165	0.305	0.159	0.299	0.157	0.267
Daegu	0.659	1.116	0.108	0.214	0.101	0.212	0.134	0.255	0.131	0.260	0.126	0.250
Daejeon	0.705	1.170	0.106	0.210	0.099	0.208	0.123	0.242	0.138	0.269	0.139	0.237
Mokpo	0.357	0.626	0.119	0.222	0.110	0.218	0.157	0.294	0.147	0.281	0.150	0.247
Bukgangneung	0.266	0.457	0.108	0.226	0.100	0.223	0.143	0.276	0.156	0.299	0.138	0.286
Bukchuncheon	0.260	0.468	0.123	0.241	0.113	0.234	0.198	0.379	0.157	0.303	0.132	0.249
Seoul	0.533	0.915	0.132	0.248	0.120	0.241	0.143	0.274	0.146	0.277	0.154	0.291
Suwon	0.237	0.427	0.111	0.217	0.104	0.212	0.138	0.265	0.132	0.254	0.144	0.249
Andong	0.220	0.388	0.111	0.216	0.100	0.207	0.154	0.280	0.142	0.264	0.126	0.225
Yeosu	0.298	0.504	0.142	0.267	0.132	0.261	0.192	0.351	0.168	0.319	0.257	0.441
Ulleung Island	0.396	0.733	0.199	0.374	0.188	0.368	0.236	0.425	0.216	0.413	0.229	0.381
Incheon	0.303	0.536	0.122	0.230	0.111	0.226	0.153	0.283	0.140	0.286	0.118	0.232
Jeonju	0.625	1.071	0.110	0.221	0.101	0.216	0.180	0.300	0.129	0.256	0.120	0.228
Jeju Island	0.271	0.451	0.122	0.241	0.114	0.242	0.155	0.299	0.139	0.299	0.189	0.303
Changwon	0.246	0.426	0.108	0.213	0.096	0.205	0.118	0.226	0.141	0.262	0.136	0.232
Cheongju	0.294	0.557	0.115	0.225	0.108	0.219	0.146	0.262	0.141	0.266	0.155	0.281
Pohang	0.248	0.455	0.148	0.279	0.135	0.270	0.189	0.369	0.160	0.307	0.148	0.272
Heuksan Islnad	0.278	0.497	0.153	0.291	0.144	0.290	0.154	0.300	0.184	0.345	0.154	0.296

MAE when max depth = 3 and learning rate = 0.07, and CatBoost was found to be the most suitable hyperparameters for max depth = 10 and learning rate = 0.1 except for Seoul. In the case of LSTM and GRU, n_unit=64 was generally more accurate, and the average MAE value was small when 'Adam' and 'Nadam' were used as optimizers. The performance results are shown in Table 3 below with hyperparameters optimized for each point, and the values with the best MAE and RMSE for each point were underlined.

As a result of the model performance, it was confirmed that the MAE of the ARIMAX was over 0.25, the MAE of the RNN series models was about 0.16, and the Boosting series models were more accurate than the RNN series with the Boosting level of about 0.12. In particular, among the Boosting models, the CatBoost results showed the best performance at all points when viewed based on the MAE. In addition, based on RMSE, CatBoost results were good at all points except Jeju Island. The importance of the variables of XGB and CatBoost, which had good performance, varies slightly from station to station, but the importance ranking was all the same in the order of solar radiation outside

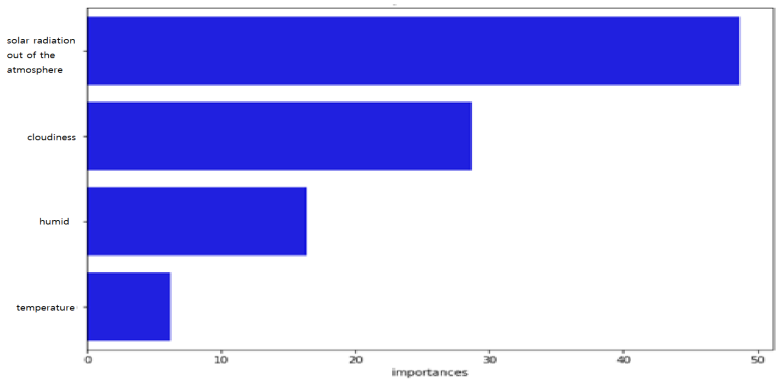

Figure 3: *The variables importance of CatBoost in Changwon.*

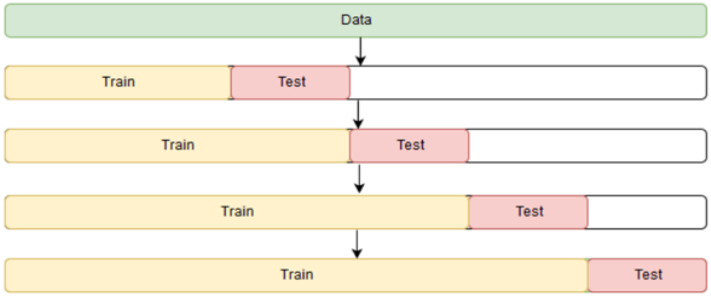

Figure 4: *Solar radiation prediction point.*

the atmosphere, cloudiness, humidity, and temperature, as shown in the example of Changwon (Figure 3).

To further improve the performance, the Boosting model was re-trained with the information from the very previous day to be predicted by applying the time series cross-validation method as shown in Figure 4 and then the model was re-fit. Table 4 below shows the results of comparing the observed and predicted solar radiation values by repeating this process.

the CatBoost model, the accuracy of the results was improved. However, considering that the existing CatBoost MAE point average was 0.116 and the MAE point average was 0.113 when timeseries cross-validation was applied, the calculation time to learn new algorithms and produce results per day may be large compared to performance improvement. It seems that it is necessary to choose whether or not timeseries cross validation is applied according to the purpose of use.

4. Conclusions

As interest in solar power prediction and power consumption prediction increases, the importance of solar radiation prediction is also increasing. Therefore, this study was conducted for the purpose of improving the accuracy of solar radiation prediction. Solar radiation was predicted by a total of three weather variables, temperature, humidity, and total cloud volume, and solar radiation outside the atmosphere, and the results were compared using various models. Various hyperparameters were applied to the Boosting series (XGB, CatBoost) and RNN series (Simple RNN, LSTM, GRU) models to find optimal hyperparameters, and the results were compared. Based on the average MAE by point, it was confirmed that XGB was 0.125, CatBoost was 0.116, simple RNN was 0.154, LSTM

Table 4: Test data fit performance results using timeseries cross-validation

Station	XGB		XGB (timeseries cross-validation)		CatBoost		CatBoost (timeseries cross-validation)	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Gwangju	0.114	0.226	0.115	0.226	0.107	0.224	0.107	0.225
Daegu	0.108	0.214	0.108	0.213	0.101	0.212	0.100	0.212
Daejeon	0.106	0.210	0.105	0.208	0.099	0.208	0.098	0.207
Mokpo	0.119	0.222	0.119	0.221	0.110	0.218	0.109	0.219
Bukgangneung	0.108	0.226	0.108	0.225	0.100	0.223	0.099	0.223
Bukchuncheon	0.123	0.241	0.120	0.236	0.113	0.234	0.109	0.226
Seoul	0.132	0.248	0.128	0.244	0.120	0.241	0.116	0.237
Suwon	0.111	0.217	0.112	0.215	0.104	0.212	0.102	0.211
Andong	0.111	0.216	0.109	0.214	0.100	0.207	0.099	0.208
Yeosu	0.142	0.267	0.138	0.259	0.132	0.261	0.128	0.253
Ulleung Island	0.199	0.374	0.189	0.356	0.188	0.368	0.177	0.350
Incheon	0.122	0.230	0.119	0.227	0.111	0.226	0.109	0.224
Jeonju	0.110	0.221	0.109	0.217	0.101	0.216	0.099	0.213
Jeju Island	0.122	0.241	0.122	0.241	0.114	0.242	0.113	0.242
Changwon	0.108	0.213	0.104	0.209	0.096	0.205	0.094	0.203
Cheongju	0.115	0.225	0.114	0.224	0.108	0.219	0.106	0.219
Pohang	0.148	0.279	0.142	0.270	0.135	0.270	0.130	0.260
Heuksan Islnad	0.153	0.291	0.147	0.283	0.144	0.290	0.139	0.284

was 0.16, and GRU was 0.152, and the CatBoost model was the best. Time series cross-validation also allowed XGB to improve performance from 0.125 to 0.123 and CatBoost from 0.116 to 0.113. However, although all branches have improved their overall results with Time series cross-validation, it is necessary to use it with this in mind because it may take a large amount of time to learn new algorithms and produce results every day compared to performance improvement.

References

Lee K, Son HG, and Kim S (2018). A study on solar energy forecasting based on time series models, *The Korean Journal of Applied Statistics*, **31**, 139–153.

Suh YM, Son H, and Kim S (2018). Solar radiation forecasting by time series models, *The Korean Journal of Applied Statistics*, **31**, 785–799.

Elizondo D, Hoogenboom G, and McClendon RW (1994). Development of a neural network model to predict daily solar radiation, *Agricultural and Forest Meteorology*, **71**, 115–132.

Rehman S and Mohandes M (2008). Artificial neural network estimation of global solar radiation using air temperature and relative humidity, *Energy Policy*, **36**, 571–576.

Alam S, Kaushik SC, and Garg SN (2009). Assessment of diffuse solar energy under general sky condition using artificial neural network, *Applied Energy*, **86**, 554–564.

Sahm Kim (2017). A study on solar irradiance forecasting with weather variables, *The Korean Journal of Applied Statistics*, **30**, 1005–1013.

Fan J, Wang X, Wu L, Zhou H, Zhang F, Yu X, Lu X, and Xiang Y (2018). Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China, *Energy Conversion and Management*, **164**, 102–111.

Fan J, Wang X, Zhang F, Ma X, and Wu L (2020). Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing

models with local and extrinsic climatic data, *Journal of Cleaner Production*, **248**, 119264.

Pang Z, Niu F, and O'Neill Z (2020). Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons, *Renewable Energy*, **156**, 279–289.

Obiora CN, Ali A, and Hasan AN (2021). Implementing extreme gradient boosting (XGBoost) algorithm in predicting solar irradiance, In *Proceeding of the 2021 IEEE PES/IAS PowerAfrica*, Nairobi, 1–5.

Huang L, Kang J, Wan M, Fang L, Zhang C, and Zeng Z (2021). Solar radiation prediction using different machine learning algorithms and implications for extreme climate events, *Frontiers in Earth Science*, **9**, 596860.

Rocha PA and Santos VO (2022). Global horizontal and direct normal solar irradiance modeling by the machine learning methods XGBoost and deep neural networks with CNN-LSTM layers: A case study using the GOES-16 satellite imagery, *International Journal of Energy and Environmental Engineering*, **13**, 1271–1286.

Kim H and Kim S (2022). A study on frost prediction model using machine learning, *The Korean Journal of Applied Statistics*, **35**, 543–552.

Sorkun MC, Paoli C, and Incel, ÖD (2017). Time series forecasting on solar irradiation using deep learning, In *Proceeding of the 2017 10th international conference on electrical and electronics engineering (ELECO), IEEE*, Bursa, 151–155.

Hochreiter S and Schmidhuber J (1997). Long short-term memory, *Neural Compute*, **9**, 1735–1780, Available from: 10.1162/neco.1997.9.8.1735. PMID: 9377276

Kumari P and Toshniwal D (2019). Hourly solar irradiance prediction from satellite data using lstm.

Sorkun MC, Incel ÖD, and Paoli C (2020). Time series forecasting on multivariate solar radiation data using deep learning (LSTM), *Turkish Journal of Electrical Engineering and Computer Sciences*, **28**, 211–223.

Cho K, Van Merriënboer B, Bahdanau D, and Bengio Y (2014). On the properties of neural machine translation: Encoder-decoder approaches, Available from: arXiv preprint arXiv:1409.1259

Boubaker S, Benghanem M, Mellit A, Lefza A, Kahouli O, and Kolsi L (2021). Deep neural networks for predicting solar radiation at Hail Region, Saudi Arabia, *IEEE Access*, **9**, 36719–36729.

Received July 28, 2022; Revised August 30, 2022; Accepted September 05, 2022