IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 January 2023, accepted 30 January 2023, date of publication 8 February 2023, date of current version 13 February 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3243252*

RESEARCH ARTICLE

# Solar Radiation Forecasting Based on the Hybrid CNN-CatBoost Model

**HYOJEOUNG KIM[1], SUJIN PARK[1], HEE-JUN PARK[2], HEUNG-GU SON[2], AND SAHM KIM[1], (Member, IEEE)**

[1]Department of Applied Statistics, Chung-Ang University, Seoul 06974, South Korea
[2]Department of Short-Term Demand Forecasting, Korea Power Exchange, Naju 58322, South Korea

Corresponding author: Sahm Kim (sahm@cau.ac.kr)

This work was supported by the Korea Power eXchange Research Grants in 2023.

**ABSTRACT** The renewable energy industry is rapidly expanding due to environmental pollution from fossil fuels and continued price hikes. In particular, the solar energy sector accounts for about 48.7% of renewable energy, at the highest production ratio. Therefore, climate prediction is essential because solar power is affected by weather and climate change. However, solar radiation, which is most closely related to solar power, is not currently predicted by the Korea Meteorological Administration; therefore, solar radiation prediction technology is needed. In this study, we predict solar radiation using extra-atmospheric solar radiation and three weather variables: temperature, relative humidity, and total cloud volume. We compared the performance of single models of machine and deep learning in previous work. For the single-model comparison, we used boosting techniques, such as extreme gradient boosting and categorical boosting (CatBoost) in machine learning, and the recurrent neural network (RNN) family (long short-term memory and gated recurrent units). In this paper, we compare CatBoost (previously the best model) with CNN and present a CNN-CatBoost hybrid model prediction method that combines CatBoost in machine learning and CNN in deep learning for the best predictive performance for a single-model comparison. In addition, we checked the accuracy change when adding wind speed and precipitation to the hybrid model. The model that considers wind speed and precipitation improved at all but three (Gangneung, Suwon, and Cheongju) of the 18 locations.

**INDEX TERMS** Solar radiation forecasting, weather variable, categorical boosting (CatBoost), convolutional neural network (CNN), hybrid model (CNN-CatBoost).

## I. INTRODUCTION

The fossil fuel-based energy supply system has low sustainability due to price volatility, limited fuel reserves, and environmental problems, spurring the development of the renewable energy industry to generate sharp growth. In addition, several countries, including Germany and Australia, have already achieved grid parity, as fossil fuel prices are soaring and technology development is lowering renewable energy production costs. While the added fossil fuel decreased from 64 GW in 2019 to 60 GW in 2020 [1], the produced renewable energy was 261 GW worldwide, and solar power generation increased to 127 GW—the most among renewable energy facilities [2]. Solar power generation occupies a high proportion of new and renewable energy due to its infinite resources, ease of installation, and eco-friendly characteristics that do not emit noise or pollutants. These advantages are expected to increase the proportion of solar energy generation further.

Solar photovoltaic (PV) systems are a primary renewable energy source and are simply panels that convert sunlight into electricity. However, solar power generation requires advanced prediction technology due to the unstable energy supply under the influence of the weather. The output of PV is highly dependent on solar irradiance, solar radiation, temperature, and various weather variables. Predicting solar radiation means that the output of PV is predicted one or more steps ahead of time [3]. Therefore,

The associate editor coordinating the review of this manuscript and approving it for publication was Rosalia Maglietta.

various weather variables are used to predict solar radiation accurately.

In addition, previous studies have used various weather variables to predict solar radiation. Alluhaidah et al. examined studies using various weather variables to identify the root mean square error (RMSE) and mean absolute percentage error (MAPE) and revealed that cloud cover, humidity, and temperature contribute the most to prediction [4]. Kwon et al. attempted to predict the global horizontal irradiance (GHI) using the temperature, relative humidity, dewpoint, and sky-coverage values [5]. Kisi et al. attempted to predict solar radiation in the Antakya and Adana areas using the lowest temperature, highest temperature, wind speed, relative humidity, and sunshine hours using an artificial neural network (ANN) and the extreme learning machine [6]. McCandless et al. used cloud cover, dewpoint temperature, categorical precipitation in the last hour (1 = precipitation did not occur), and a more accurate k-nearest neighbors cluster model with six meteorological parameters [7]. Wojtkiewicz et al. used weather variables and cloud cover as exogenous variables to predict solar radiation and fit the long short-term memory (LSTM) network and gated recurrent unit (GRU) models, confirming that LSTM provides a predicted performance of 23.79% based on the MAPE [8]. Qing and Niu used such variables as temperature, dewpoint, humidity, visibility, wind speed, and weather type (13 types of weather) to argue that the data fit the LSTM model, which performs much better than the other benchmark models [9].

In addition to using assorted variables, techniques for solar radiation prediction have also been studied. Research on a single machine-learning model has actively been conducted [10], [11], [12], [13]. For example, Yadav and Behera applied the recurrent neural network (RNN) and wavelet transform by adding variables, such as temperature, humidity, wind speed, wind direction, dewpoint temperature, and pressure to predict solar radiation values. The wavelet deformation technique was excellent regarding the mean absolute error (MAE) at 9.62% and RMSE at 14.96% [10]. Kim proposed multiple regression models with an accuracy of 0.1553 based on the MAE, suitable for the autoregressive integrated moving average (ARIMA), ARIMA exogenous (ARIMAX), and multiple regression models [11]. Fan et al. applied the support vector machine, M% model tree, random forest (RF), extreme gradient boosting (XGBoost), and categorical boosting (CatBoost) models using the data from three stations in humid subtropical China. Comprehensively considering prediction accuracy, generalizability, and computational efficiency, CatBoost is the best model to develop general models. [12]. Pang et al. determined that solar radiation prediction using the RNN model has higher accuracy than the ANN model [13].

Recently, a hybrid model combining the two models was also developed [13], [14], [15], [16]. Some studies on the hybrid model have combined single machine-learning models. Ghimire et al. introduced convolutional neural network (CNN)-LSTM techniques combining the CNN, LSTM, deep neural network, decision tree, and multilayer perceptron. When the CNN-LSTM was used for prediction for a month, the results confirmed that the MAE (%) value was superior at 13.131 [14]. Agga et al. evaluated two hybrid models (CNN-LSTM and convolutional LSTM) that incorporate an LSTM layer using two types of datasets (univariate and multivariate, with weather features, such as wind speed, temperature, humidity, and cloud cover) [15]. The LSTM method is used as a baseline to evaluate the performance and efficiency of the models. Both hybrid models predicted the one-day-ahead power output well, using only the single-variable dataset with MAE values of 5.04 and 5.18 for the convolutional LSTM and CNN-LSTM models, respectively [15].

Lai et al. used a hybrid model applied to various previous studies, such as LSTM, GRU, CNN-LSTM, CNN-GRU, and the CNN with bidirectional LSTM (BiLSTM). The CNN-BiLSTM model outperformed other models in univariate and multivariate predictions in terms of the MAE [16]. In addition, Gala et al. used a single machine model, such as support vector regression (SVR), gradient boosted regression (GBR), and RF regression, and a hybrid SVR-GBR-RFR model, as a weighted linear combination of the SVR, GBR, and RFR outputs [17].

Based on previous studies, machine-learning models, including RNN and boosting, are used in a single model, but the hybrid model has a research spectrum with limited combinations of the CNN and RNN. Therefore, we demonstrate the performance of the CatBoost model by comparing the time-series model (ARIMA), RNN series (LSTM, GRU, simple RNN) and boosting series (XGBost and CatBoost) models [18]. This study proposes an accurate solar radiation prediction technique by expanding on the previous study. By applying basic weather variables, such as temperature, relative humidity, solar radiation, and total cloud volume, we applied a hybrid model combining the CNN with CatBoost and confirmed the results by adding two variables: precipitation and wind speed. This work proposes a rarely used CNN-CatBoost technique and display the best performance. This work contributes to predicting solar power generation in the future.

Section II describes the single machine-learning and hybrid models used as prediction techniques. Next, Section III discusses the solar insolation and meteorological variables for model training and fit, and Section IV details the performance after fitting the model. Finally, Section V proposes that the hybrid model is superior to a single model and that these models should be actively studied for accurate solar radiation and solar power generation prediction.

## II. METHODOLOGY

This section proposes the RNN series models (LSTM and GRU), boosting series models (XGBoost and CatBoost), and a CNN model. The hybrid model combines the CNN and CatBoost models, and the results are compared by differentiating the number of convolutional layers.

### A. UNBIASED BOOSTING WITH CATEGORICAL FEATURES

The CatBoost model is an improved version of the gradient-boosting decision tree algorithm that can handle categorical features well. This algorithm has two main advantages: (1) dealing with categorical features during training time instead of preprocessing time and (2) using a new but more efficient schema to calculate leaf values during tree structure selection, reducing overfitting. We performed a random permutation of the dataset. We computed the average label value for the example with the same category value before computing the given label in the permutation [19]. We let $\mathbf{\Theta} = [\sigma_1, \cdots, \sigma_n]_n^T$ be a permutation, which is substituted as follows [20]:

$$x_{\sigma_{p,k}} = \frac{\sum_{j=1}^{p-1}\left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right]\cdot Y_{\sigma,j} + \beta\cdot P}{\sum_{j=1}^{p-1}\left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right]\cdot Y_{\sigma,j} + \beta}, \tag{1}$$

where $P$ is a prior value, and a parameter is the weight of the prior. For regression tasks, the standard technique for calculating the prior is to average the label value in the dataset. Another advantage of the CatBoost model is that oblivious trees are used as base predictors. In such trees, the same splitting criterion is used for an entire tree level [21].

The primary property of CatBoost is that the feature sample permutations maintain the diversity of the coupled inputs and prevent overfitting. The average values are classified into the same category and converted into numerical values. This stage copes with noisy low-frequency categories. The feature combination passes through greedy subtree splitting for terms that the initial trees do not consider in the first generation [22].

### B. CONVOLUTIONAL NEURAL NETWORK

The CNN is the most successful deep-learning algorithm for extracting image features at a good resolution by assigning weights [23]. These features become complex at a coarser resolution as the network becomes deeper. The CNN architecture can be divided into three layers. The initial layer is convolutional and extracts features by convolving images with weights known as kernels, which are randomly initialized. The kernel slides over the image with a certain stride value, extracting low-level features, such as shapes and edges, in the initial layer. After applying these kernels, the final output at each layer is known as a feature map.

To extract distinct features with various weights, we can vary the number of kernels according to the model requirements. More convolutional layers help extract high-level features [24], and the CNN is an effective technology for automatic feature extraction and has achieved remarkable success in image vision. Moreover, the CNN has a strong potential for dealing with time series, such as automatic speech recognition and wind speed forecasting [25]. The design of a CNN is determined by the types and number of layers it comprises, such as convolutional layers, pooling layers, and fully connected layers, and is inspired by the genetic structure of the visual cortex, which has configurations of simple and complicated cells [26]. Initially, input data are entered into the input layer to process the model for feature transformation. Then, features are extracted in the convolutional and pooling layers. Afterward, the extracted information from the convolutional and pooling layers is assimilated using the fully connected layers. Finally, the result is communicated through the output layer.

Each convolutional layer is targeted to extract spatial patterns from the target variable (i.e., GHI) and its related input variables (i.e., meteorological data and historical GHI values), demonstrated as follows:

$$y_{ik}^{k} = f\left(\left(W^{k} \times h\right)_{i,j} + b_k\right), \tag{2}$$

where $f$ represents an activation function, $W^k$ denotes the kernel weight, and × indicates a convolutional process operator. In the present work, the efficient recurrent linear unit activation function is used [27]

$$f(x) = \max(0, x) \tag{3}$$

In this study, the CNN layers were applied to the first and second layers, respectively, and the accuracy was checked. This model was combined with CatBoost to generate the proposed model.

### C. HYBRID MODEL(CNN-CatBoost)

In this study, the hybrid (CNN-CatBoost) model was constructed for the most accurate solar radiation prediction. This model is divided into feature extraction and prediction parts. We compare the results by attempting one-layer one-dimensional convolution (Conv1D) and two-layer Conv1D in feature extraction. This section addresses temperature, relative humidity, and cloudiness, and the predicted part of the CatBoost model yields GHI predictions [28]. The basic structure of the one-layer Conv1D hybrid model is shown in Figure 1. For the two-layer Conv1D hybrid model, one more Conv1D is added for the feature extraction.

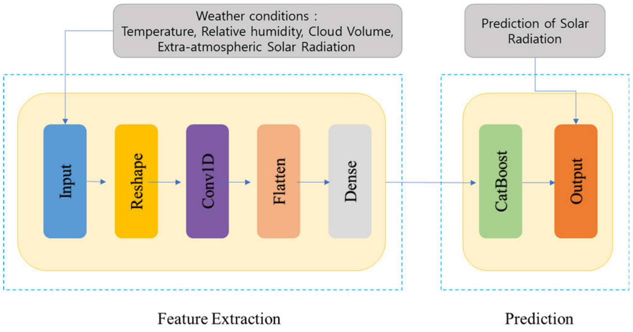

**FIGURE 1.** Model architecture of the convolutional neural network with categorical boosting (CNN-CatBoost).

## III. DATA

### A. DATA COLLECTION AND PREPROCESSING

The data applied to this study are 1-h weather observation data provided by the Weather Data Open Portal (https://data.kma.go.kr) from March 1, 2017, to February 28, 2022. The data cover 5 years, and the learning and testing data are divided into 8:2 so that all seasons can be tested. From March 1, 2017, to February 28, 2021, the model was used for training data, and the remaining data from March 1, 2021, to February 28, 2022, were used for testing data to evaluate model performance. The independent variables were applied to the model as basic input variables, including three weather variables with high correlation (temperature, humidity, and total cloud volume) and the out-of-atmosphere solar radiation (*ei*) proposed by He et al. [29].

The weather data were reconstructed through preprocessing. The points where the solar radiation value or total cloud amount was missing for a long time were removed, and the analysis was conducted at 18 points, as depicted in Figure 2. In addition, solar radiation starts to form a peak after sunrise, and a pattern with a value of zero is repeated daily after sunset. Due to these characteristics, many parts had zero or inapplicable values after sunset and before sunrise, and it was assumed that solar radiation was not observed for the remaining time, by setting the sunrise and sunset time, as listed in Table 1. Sunrise and sunset times are the same in spring and autumn, so there are only three time zones, although it displays four seasons.

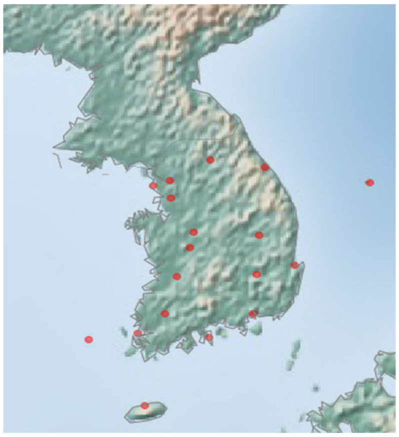

**FIGURE 2.** Solar radiation prediction point.

**TABLE 1.** Sunrise and sunset time by season.

| Season | Month | Sunrise Time | Sunset Time |
|---|---|---|---|
| **Spring** | 3, 4, 5 | 7:00 AM | 7:00 PM |
| **Summer** | 6, 7, 8 | 6:00 AM | 8:00 PM |
| **Autumn** | 9, 10, 11 | 7:00 AM | 7:00 PM |
| **Winter** | 12, 1, 2 | 8:00 AM | 6:00 PM |

In addition, days on which solar radiation was not observed for one day at each point were excluded, and values were replaced using linear interpolation if the temperature, humidity, cloudiness, and wind speed were missing. If precipitation was missing, the precipitation was judged to be the condition of not raining, and the value was replaced with zero. The prediction results were compared by fitting the LSTM, GRU, XGBoost, CatBoost, and CNN models using the final preprocessed data.

The predicted results were compared in the previous analysis by fitting the XGBoost, CatBoost, simple RNN, LSTM, GRU, and CNN models with the four input variables mentioned above [30]. Among them, the best CatBoost results were compared with the results of the CNN model recently used in several algorithms. Both models demonstrated high performance, incorporating a hybrid model that combines the CatBoost and CNN, and exhibited better performance, selecting the hybrid model as the final model. The accuracy change was confirmed when wind speed and precipitation variables were added with basic input variables to understand the influence of other weather variables on the selected model.

### B. EVALUATION METRICS

To compare suitable models, MAE and RMSE were used as error measures. In general, MAPE is widely used to evaluate models, but it was challenging to apply the MAPE calculation because the solar radiation value was often zero. Therefore, the accuracy was evaluated on the scale of the MAE and RMSE, defined as follows:

$$nMAE = \frac{1}{n}\sum_{t=1}^{n}|\frac{Y_t - F_t}{C_t}| \tag{4}$$

where $n$ indicates the number of data for prediction, $Y_t$ denotes the observation value at time $t$, and $F_t$ represents the prediction value through the model at time $t$. For MAE and RMSE, a smaller value indicates higher accuracy.

## IV. MODEL APPLICATION

### A. PERFORMANCE COMPARISON OF MACHINE-LEARNING AND HYBRID MODELS

The performance results of the machine learning(CNN and CatBoost) and hybrid models are presented in Table 2, and the values with the best MAE and RMSE values for each branch are underlined.

When comparing the boosting series (XGBoost and CatBoost) with the RNN series (LSTM and GRU) models in previous studies, the MAE of the boosting models was about 0.12, and that of the RNN models was 0.16, confirming that the boosting series was more accurate than the RNN series [18]. Among the boosting models, CatBoost results performed best at all points for the MAE standards [18]. The importance of the variables of the CatBoost model with good performance varies slightly from branch to branch, but the importance ranking was the same in the order of out-of-atmosphere solar radiation, total cloud volume, humidity, and temperature outside the atmosphere, as depicted in Figure 3 for the Seoul branch.

**TABLE 2.** Performance results using the CNN, categorical boosting [CatBoost] and hybrid models.

| Station | CatBoost | | CNN | | Hybrid (one layer) | | Hybrid (two layers) | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| Andong | 0.0995 | 0.2069 | 0.0876 | 0.1789 | 0.0838 | 0.1734 | 0.0835 | 0.1729 |
| Changwon | 0.0958 | 0.2049 | 0.0967 | 0.1963 | 0.0889 | 0.1868 | 0.089 | 0.1857 |
| Cheongju | 0.1076 | 0.2193 | 0.0858 | 0.1859 | 0.0926 | 0.1861 | 0.0891 | 0.1817 |
| Chuncheon | 0.1131 | 0.2338 | 0.0886 | 0.1858 | 0.094 | 0.1907 | 0.0928 | 0.1907 |
| Daegu | 0.101 | 0.2117 | 0.1024 | 0.2135 | 0.0921 | 0.1895 | 0.0889 | 0.186 |
| Daejeon | 0.0992 | 0.2075 | 0.1025 | 0.2021 | 0.0831 | 0.1696 | 0.0835 | 0.1698 |
| Gangneung | 0.1 | 0.2228 | 0.0932 | 0.2047 | 0.0923 | 0.2009 | 0.0906 | 0.2021 |
| Gwangju | 0.1071 | 0.2245 | 0.0944 | 0.1901 | 0.0927 | 0.1907 | 0.0904 | 0.1912 |
| Heuksan island | 0.1438 | 0.2899 | 0.1591 | 0.3269 | 0.1362 | 0.277 | 0.1348 | 0.274 |
| Incheon | 0.1113 | 0.226 | 0.0993 | 0.2167 | 0.0972 | 0.1933 | 0.0949 | 0.1927 |
| Jeju island | 0.1139 | 0.242 | 0.102 | 0.2208 | 0.1011 | 0.2123 | 0.1011 | 0.2138 |
| Jeonju | 0.1011 | 0.2157 | 0.0929 | 0.1958 | 0.0901 | 0.1912 | 0.0893 | 0.1887 |
| Mokpo | 0.1101 | 0.2181 | 0.1106 | 0.2276 | 0.1007 | 0.1992 | 0.0987 | 0.1987 |
| Pohang | 0.1354 | 0.2696 | 0.1124 | 0.2443 | 0.1266 | 0.2469 | 0.124 | 0.2489 |
| Seoul | 0.1203 | 0.2408 | 0.0882 | 0.1885 | 0.0962 | 0.192 | 0.0945 | 0.188 |
| Suwon | 0.1036 | 0.2123 | 0.1062 | 0.2302 | 0.0951 | 0.1913 | 0.0924 | 0.1901 |
| Ulleung Island | 0.1879 | 0.3682 | 0.2289 | 0.4485 | 0.1855 | 0.358 | 0.186 | 0.3634 |
| Yeosu | 0.1325 | 0.2612 | 0.1363 | 0.2689 | 0.1229 | 0.2413 | 0.1247 | 0.2479 |
| Average | 0.1157 | 0.2375 | 0.1104 | 0.2292 | 0.1040 | 0.2106 | 0.1027 | 0.2104 |

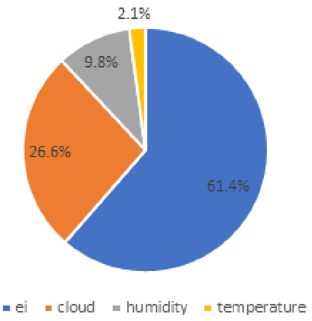

**FIGURE 3.** Variable importance of categorical boosting in Seoul.

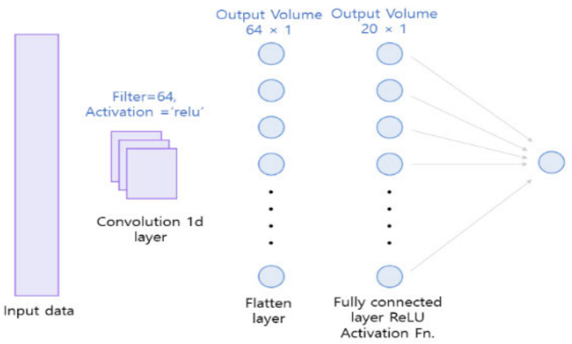

**FIGURE 4.** Convolutional neural network (CNN) structure.

Additional CNN models were assessed to improve performance further. For the CNN, weather observation values from 3 h prior to the prediction time point were used as input variables, and a rolling prediction of the hourly solar radiation was performed. The number of features and past values were not used much; thus, it is difficult to use multiple layers or a large kernel, and the CNN structure comprises Conv1D, flattened, and dense layers, as illustrated in Figure 4.

The calculated value of the fully connected layer in the CNN model was obtained, and a hybrid model was implemented with the value using the CatBoost model, with the best results in the last step. Although no significant difference exists in the average of the MAEs by point between the CNN and CatBoost, the CNN results were lower than the CatBoost results in Table 2. However, the accuracy was not comparatively improved; thus, a hybrid method was applied to extract features from the CNN and predict the final value using CatBoost. As a result, the MAE results improved at almost all points.

The graphs in Figures 5 and 6 present the learning and verification curves of the CNN and hybrid (CNN+CatBoost) models in Seoul according to the number of convolutional layers. The x and y-axes represent the number of epochs and

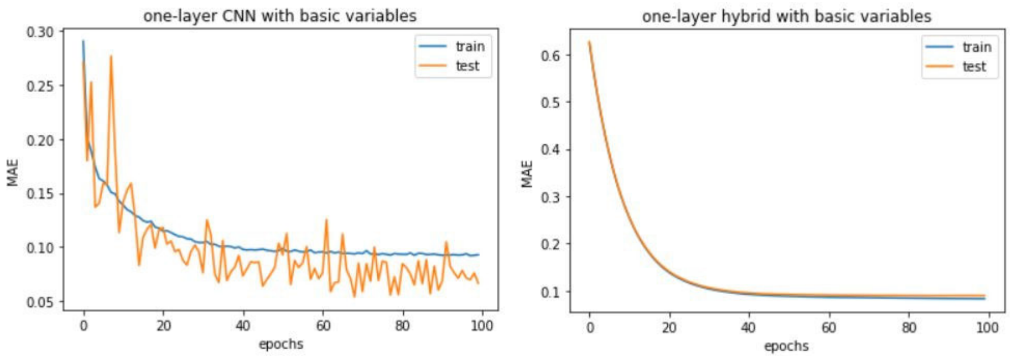

**FIGURE 5.** Training and validation curves for the one-layer convolutional neural network (CNN) and one-layer hybrid model for Seoul.

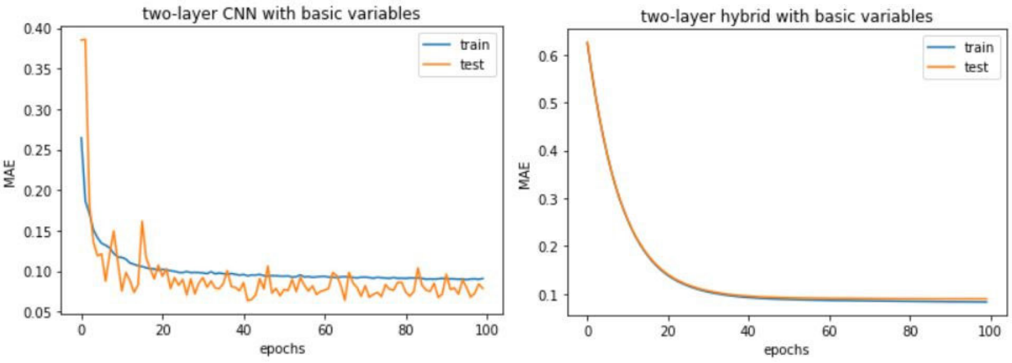

**FIGURE 6.** Training and validation curves for the two-layer convolutional neural network (CNN) and two-layer hybrid model for Seoul.

MAE values, respectively. The training and validation curves confirm that the MAE value for learning and validation gradually decreases as the epoch increases, and no overfitting or underfitting occurs. The MAE value is uneven in the case of a model using only the CNN, but the hybrid model using the CNN and CatBoost reveals that the MAE values are steadily decreasing to similar values in both training and validation; thus, it is a more stable model.

The Conv1D layer was split into two layers to fit the hybrid model. The MAE of the two-layer hybrid model remained similar or slightly lower at most points compared to the one-layer hybrid model. Although the average MAE value per point does not significantly differ, the two-layer hybrid model was more stable with better results at most points; therefore, it was selected as the final model.

Figure 7 present the graph of the solar radiation observations in Seoul and the predictions of the CatBoost and two-layer hybrid models from February 22 to 28, 2022. The blue line indicates the two-layer hybrid prediction. The green line marks the CatBoost single model, and the red line represents the actual prediction. In most cases, the peak points of the two-layer hybrid model were closer to the actual values than those of the model using only the CatBoost model. In particular, the gap in peak points can be observed on the fourth, fifth, and seventh days (February 25, 26, and 28). In addition, with CatBoost, the results often vary, such as on the afternoon of Days 4 and 7. However, the hybrid model seems to calibrate these parts to affect the accuracy.

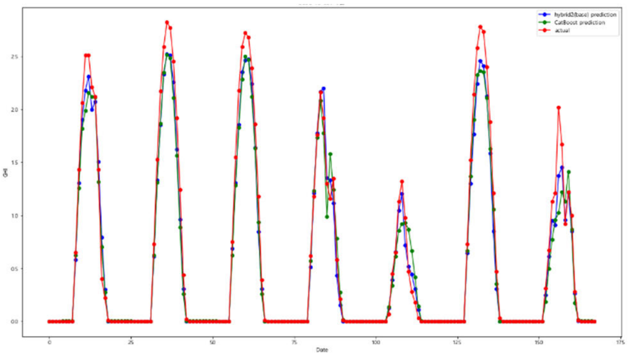

**FIGURE 7.** Forecast graph of the categorical boosting (CatBoost) and hybrid convolutional neural network for Seoul.

### B. RESULTS OF ADDING VARIABLES

In addition to temperature, humidity, total cloud volume, and out-of-atmosphere solar radiation, changes in accuracy were confirmed when wind speed and precipitation variables were added to understand the effects of other weather variables. The accuracy of adding variables in the selected hybrid (two Conv1D layers) model is presented in Table 3. Adding wind speed or precipitation rather than the basic variables made the average MAE for each point slightly smaller, but the degree

**TABLE 3.** Performance results with wind speed and rain variables.

| Station | +Wind speed | | +Rain | | +Wind speed +rain | | Base input var. | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| Andong | 0.0805 | 0.1652 | 0.0821 | 0.1731 | 0.0793 | 0.1662 | 0.0835 | 0.1729 |
| Changwon | 0.0842 | 0.1747 | 0.0863 | 0.1841 | 0.0833 | 0.1734 | 0.089 | 0.1857 |
| Cheongju | 0.0904 | 0.1816 | 0.0907 | 0.1859 | 0.09 | 0.1825 | 0.0891 | 0.1817 |
| Chuncheon | 0.0916 | 0.1857 | 0.0907 | 0.1888 | 0.0901 | 0.1854 | 0.0928 | 0.1907 |
| Daegu | 0.0865 | 0.1762 | 0.0881 | 0.1842 | 0.0844 | 0.17 | 0.0889 | 0.186 |
| Daejeon | 0.0814 | 0.1669 | 0.0832 | 0.1716 | 0.0816 | 0.1673 | 0.0835 | 0.1698 |
| Gangneung | 0.0915 | 0.2051 | 0.09 | 0.2049 | 0.0919 | 0.2098 | 0.0906 | 0.2021 |
| Gwangju | 0.09 | 0.1884 | 0.089 | 0.1904 | 0.0892 | 0.1888 | 0.0904 | 0.1912 |
| Heuksan island | 0.1304 | 0.2672 | 0.1286 | 0.2651 | 0.1295 | 0.2683 | 0.1348 | 0.274 |
| Incheon | 0.0945 | 0.1907 | 0.0934 | 0.1901 | 0.0915 | 0.189 | 0.0949 | 0.1927 |
| Jeju island | 0.1026 | 0.2185 | 0.0995 | 0.209 | 0.0983 | 0.2095 | 0.1011 | 0.2138 |
| Jeonju | 0.0874 | 0.1832 | 0.0904 | 0.1928 | 0.0828 | 0.1782 | 0.0893 | 0.1887 |
| Mokpo | 0.0993 | 0.1987 | 0.0991 | 0.1999 | 0.0981 | 0.1963 | 0.0987 | 0.1987 |
| Pohang | 0.1264 | 0.2492 | 0.1241 | 0.2476 | 0.123 | 0.2443 | 0.124 | 0.2489 |
| Seoul | 0.0917 | 0.1843 | 0.0961 | 0.1926 | 0.0924 | 0.1856 | 0.0945 | 0.188 |
| Suwon | 0.0907 | 0.1855 | 0.0949 | 0.1918 | 0.0955 | 0.1912 | 0.0924 | 0.1901 |
| Ulleung Island | 0.1843 | 0.359 | 0.1821 | 0.3574 | 0.1829 | 0.3573 | 0.186 | 0.3634 |
| Yeosu | 0.12 | 0.2389 | 0.1205 | 0.2401 | 0.121 | 0.2407 | 0.1247 | 0.2479 |
| Average | 0.1013 | 0.2066 | 0.1016 | 0.2094 | 0.1003 | 0.2058 | 0.1027 | 0.2104 |

of influence on wind speed and precipitation differed for each point.

However, the model considering wind speed and precipitation performed better at most locations, as the MAE became smaller except in three locations (Gangneung, Suwon, and Cheongju). Gangneung is located on the coast and has higher humidity and more precipitation than other areas. Therefore, it showed an effect when only precipitation was added to the input variables rather than an effect on wind speed. In the case of Suwon, the average wind speed and precipitation were lower than in the other regions. Although Suwon is inland, the difference between when the wind does and does not blow is significant enough to be distinguished. Therefore, the results were improved only when wind speeds with these characteristics were reflected. Cheongju is an area with less precipitation and very low wind speed compared to the other regions, so overfitting seems to have occurred by adding precipitation and wind speed. However, information on wind speed and precipitation does not significantly affect solar radiation, given that the accuracy is not as effectively reflected in the model.

## V. CONCLUSION

In this paper, we aimed to predict the seasonal wind power generation in Gangwon-do using the wind direction and wind speed variables that affect each season as exogenous variables. The raw data were transformed using log transformation, and the existing time-series models (ARIMA and ARIMAX) and the machine-learning regression models (support vector machine, RF, and XGBoost) were used. When comparing the evaluation indicators MAE and MAPE, the machine-learning model displayed the best predictive performance compared to the time-series models. Among the machine-learning models, the RF model had the best predictive performance, followed by It was followed by the XGBoost model and the SVR models. In this study, prediction was attempted using log transformation and a single machine-learning model, but in future studies, other data transformation techniques, the addition of new weather variables, or complex machine-learning models may be evaluated. As interest in solar power prediction increases, the importance of solar radiation prediction is increasing. Therefore, this study was conducted to improve the accuracy of predicting solar radiation.

The solar radiation was predicted using three weather variables, temperature, humidity, and total cloud volume, and the results were compared using various models. The boosting (XGBoost and CatBoost) and RNN (LSTM and GRU) models were suitable for determining the optimal hyperparameters for each point. The point-by-point average MAE was 0.1251 for XGBoost, 0.1157 for CatBoost, 0.1599 for LSTM, and 0.1515 for GRU. Thus, the CatBoost model was the best. Additionally, the CNN (first layer) model was fitted; thus, the average MAE was 0.1104, slightly improving the performance. Subsequently, the hybrid model was selected as the final hybrid model, combining the CNN and CatBoost models and lowering the average MAE of the solar radiation prediction accuracy from 0.1104 to 0.1027.

In addition, the influence of wind speed and precipitation was identified by considering these parameters in addition to the temperature, humidity, and cloud volume that have frequently been used. Wind speed and precipitation were added to improve accuracy, resulting in an average MAE of 0.1003. The model that considers wind speed and precipitation improved at all locations except three (Gangneung, Suwon, and Cheongju). However, considering that the MAE value changed little, wind speed and precipitation do not significantly affect solar radiation.

The analysis focused on comparing the model based on weather variables. A deeper study of the relationship between weather variables and solar radiation is needed. A diverse approach should be used, such as examining the relationship between weather variables through principal component analysis by assessing multicollinearity and extracting appropriate characteristics for solar radiation prediction using various methods. The accuracy of solar radiation prediction can be improved through the proposed future research to establish future renewable energy generation plans based on improved prediction accuracy.

## REFERENCES

[1] Z. Zhongming, L. Linong, Y. Xiaona, Z. Wangqiang, and L. Wei, ''World adds record new renewable energy capacity in 2020,'' Global S&T Develop. Trend Anal. Platform Resour. Environ., 2021. [Online]. Available: http://119.78.100.173/C666/handle/2XK7JSWQ/321399

[2] International Renewable Energy Agency (RENA). (Apr. 2021). *Renewable Capacity Statistics 2021*. Distributed by RENA. [Online]. Available: https://www.irena.org/publications/2021/rch/Renewable-Capacity-Statistics-2021

[3] A. Alzahrani, P. Shamsi, C. Dagli, and M. Ferdowsi, ''Solar irradiance forecasting using deep neural networks,'' *Proc. Comput. Sci.*, vol. 114, pp. 304–313, Jan. 2017.

[4] B. M. Alluhaidah, S. H. Shehadeh, and M. E. El-Hawary, ''Most influential variables for solar radiation forecasting using artificial neural networks,'' in *Proc. IEEE Electr. Power Energy Conf.*, Nov. 2014, pp. 71–75.

[5] Y. Kwon, A. Kwasinski, and A. Kwasinski, ''Solar irradiance forecast using Naïve Bayes classifier based on publicly available weather forecasting variables,'' *Energies*, vol. 12, no. 8, p. 1529, Apr. 2019.

[6] O. Kisi, M. Alizamir, S. Trajkovic, J. Shiri, and S. Kim, ''Solar radiation estimation in Mediterranean climate by weather variables using a novel Bayesian model averaging and machine learning methods,'' *Neural Process. Lett.*, vol. 52, no. 3, pp. 2297–2318, Dec. 2020.

[7] T. C. McCandless, S. E. Haupt, and G. S. Young, ''A regime-dependent artificial neural network technique for short-range solar irradiance forecasting,'' *Renew. Energy*, vol. 89, pp. 351–359, Apr. 2016.

[8] J. Wojtkiewicz, M. Hosseini, R. Gottumukkala, and T. L. Chambers, ''Hour-ahead solar irradiance forecasting using multivariate gated recurrent units,'' *Energies*, vol. 12, no. 21, p. 4055, Oct. 2019.

[9] X. Qing and Y. Niu, ''Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM,'' *Energy*, vol. 148, pp. 461–468, Apr. 2018.

[10] A. P. Yadav and L. Behera, ''Solar radiation forecasting using neural networks and wavelet transform,'' *IFAC Proc. Volumes*, vol. 47, no. 1, pp. 890–896, 2014.

[11] S. Kim, ''A study on solar irradiance forecasting with weather variables,'' *Korean J. Appl. Statist.*, vol. 30, no. 6, pp. 1005–1013, 2017.

[12] J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, and Y. Xiang, ''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 Convers. Manag.*, vol. 164, pp. 102–111, May 2018.

[13] Z. Pang, F. Niu, and Z. O'Neill, ''Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons,'' *Renew. Energy*, vol. 156, pp. 279–289, Aug. 2020.

[14] S. Ghimire, R. C. Deo, N. Raj, and J. Mi, ''Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms,'' *Appl. Energy*, vol. 253, Nov. 2019, Art. no. 113541.

[15] A. Agga, A. Abbou, M. Labbadi, and Y. El Houm, ''Short-term self consumption PV plant power production forecasts based on hybrid CNN-LSTM, ConvLSTM models,'' *Renew. Energy*, vol. 177, pp. 101–112, Nov. 2021.

[16] C. S. Lai, C. Zhong, K. Pan, W. W. Y. Ng, and L. L. Lai, ''A deep learning based hybrid method for hourly solar radiation forecasting,'' *Expert Syst. Appl.*, vol. 177, Sep. 2021, Art. no. 114941.

[17] Y. Gala, Á. Fernández, J. Díaz, and J. R. Dorronsoro, ''Hybrid machine learning forecasting of solar radiation values,'' *Neurocomputing*, vol. 176, pp. 48–59, Feb. 2016.

[18] H. Kim, S. Park, and S. Kim, ''Solar radiation forecasting using boosting decision tree and recurrent neural networks,'' *Commun. Stat. Appl. Methods*, vol. 29, no. 6, pp. 709–719, Nov. 2022.

[19] A. V. Dorogush, V. Ershov, and A. Gulin, ''CatBoost: Gradient boosting with categorical features support,'' 2018, *arXiv:1810.11363*.

[20] L. Wu, G. Huang, J. Fan, F. Zhang, X. Wang, and W. Zeng, ''Potential of kernel-based nonlinear extension of Arps decline model and gradient boosting with categorical features support for predicting daily global solar radiation in humid regions,'' *Energy Convers. Manage.*, vol. 183, pp. 280–295, Mar. 2019.

[21] R. Kohavi and C. H. Li, ''Oblivious decision trees, graphs, and top-down pruning,'' in *Proc. IJCAI*, Aug. 1995, pp. 1071–1079.

[22] M. Massaoudi, S. S. Refaat, H. Abu-Rub, I. Chihi, and F. S. Wesleti, ''A hybrid Bayesian ridge regression-CWT-Catboost model for PV power forecasting,'' in *Proc. IEEE Kansas Power Energy Conf. (KPEC)*, Jul. 2020, pp. 1–5.

[23] K. Wu, J. Wu, L. Feng, B. Yang, R. Liang, S. Yang, and R. Zhao, ''An attention-based CNN-LSTM-BiLSTM model for short-term electric load forecasting in integrated energy system,'' *Int. Trans. Electr. Energy Syst.*, vol. 31, no. 1, 2021, Art. no. e12637.

[24] H. Raju and S. Das, ''CNN-based deep learning model for solar wind forecasting,'' *Sol. Phys.*, vol. 296, no. 9, pp. 1–25, Sep. 2021.

[25] B. Gao, X. Huang, J. Shi, Y. Tai, and J. Zhang, ''Hourly forecasting of solar irradiance based on CEEMDAN and multi-strategy CNN-LSTM neural networks,'' *Renew. Energy*, vol. 162, pp. 1665–1683, Dec. 2020.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[27] R. Yan, J. Liao, J. Yang, W. Sun, M. Nong, and F. Li, ''Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering,'' *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114513.

[28] P. Kumari and D. Toshniwal, ''Long short term memory-convolutional neural network based deep hybrid approach for solar irradiance forecasting,'' *Appl. Energy*, vol. 295, Aug. 2021, Art. no. 117061.

[29] X. Hei, H. Zhang, W. Ji, Y. Wang, L. Zhu, and Y. Qiu, ''ConvCatb: An attention-based CNN-CATBOOST risk prediction model for driving safety,'' in *Proc. Int. Conf. Netw. Netw. Appl. (NaNA)*, Oct. 2021, pp. 513–519.

[30] S. Kim, ''A study on solar irradiance forecasting with weather variables,'' *Korean J. Appl. Statist.*, vol. 30, no. 6, pp. 1005–1013, 2017.

**HYOJEOUNG KIM** received the B.S. degree in industrial engineering from the Seoul National University of Science and Technology and the M.S. degree in statistics from Chung-Ang University, Dongjak, Seoul, South Korea, where she is currently pursuing the Ph.D. degree in statistics. From 2018 to 2022, she was a Researcher developing meteorological convergence services at the Korea Meteorological Administration Big Data Application Division. Her main research interests include time-series analysis and renewable energy demand forecasting.

**SUJIN PARK** was born in Seoul, South Korea. She received the B.S. degree in information statistics from Kangwon National University and the M.S. degree in statistics from Chung-Ang University, Dongjak, Seoul, where she is currently pursuing the Ph.D. degree in statistics. Her main research interests include time-series analysis and renewable energy.

**HEE-JUN PARK** received the B.S. degree from the Department of Astronomy and Space Science, Chungbuk National University, Cheongju, South Korea. Since 2012, he has been a Developer of energy management system (EMS) and an Operator of transmission grid with Korea Power Exchange. His main research interests include renewable energy demand forecasting and short-trem electricity load forecasting for stable power systems.

**HEUNG-GU SON** received the B.S., M.S., and Ph.D. degrees in statistics from Chung-Ang University, Seoul, South Korea. From 2017 to 2018, he was a Postdoctoral Researcher at the Korea Transport Institute. Since 2019, he has been a Electricity Load Forecaster with Korea Power Exchange. His main research interests include time-series analysis and short & long term electricity load forecasting, renewable energy demand forecasting, and seasonal demand forecasting.

**SAHM KIM** (Member, IEEE) received the B.S. and M.S. degrees from the Department of Computational Statistics, Seoul National University, Seoul, South Korea, and the Ph.D. degree in statistics from the University of Georgia, Athens, USA. His final thesis title is "Estimating Functions for a Class for Non-linear Time Series Model." From 1985 to 2000, he was a Researcher investigating traffic prediction at the KT Technology Research Center. Since 2000, he has been a Professor of applied statistics with Chung-Ang University. Since 2019, he has been the Vice President of the Korean Statistical Society. In 2001, he received the Presidential Citation from the 51st International Statistical Society and the New Statistician Award from the Korean Statistical Society, in 2002. He achieved the Academic Promotion Award, in 2011.