Hindawi
International Journal of Energy Research
Volume 2023, Article ID 3525651, 17 pages
https://doi.org/10.1155/2023/3525651

Research Article

Predictive Model for Solar Insolation Using the Deep Learning Technique

Jiwon Park,[1] Sung Hyup Hong,[1] Sang Hun Yeon,[1] Byeong Mo Seo,[2] and Kwang Ho Lee[3]

[1]*Graduate School, Department of Architecture, College of Engineering, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea*
[2]*College of Design, North Carolina State University, 50 Pullen Rd., Campus Box 7707, NC 27695-7701, USA*
[3]*Department of Architecture, College of Engineering, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea*

Correspondence should be addressed to Kwang Ho Lee; kwhlee@korea.ac.kr

Received 12 October 2022; Revised 14 December 2022; Accepted 26 December 2022; Published 3 February 2023

Academic Editor: Saleh N. Al-Saadi

In this study, prediction performances of a regression model and deep learning-based predictive models were comparatively analyzed for the prediction of hourly insolation in regions located at the temperate climate and microthermal climate with high precipitation. Unlike linear regression models, artificial neural networks (ANN) and long short-term memory- (LSTM-) based models achieved reliable predictive performances with CV(RMSE) of 14.0% and 15.8%, respectively. This study proposed the direction of future research by improving the performance of predicting insolation at 1 hour after the current time-step, which has time-dependent characteristics, by utilizing insolation at 24 hours before the current time-step and insolation at the current time-step in addition to the forecasted weather data. In the proposed models, a large error occurred at sunrise and sunset times, suggesting the possibility of improving predictive performance by utilizing variables related to sunrise and sunset in the future. Along with Cheongju, the proposed model could properly predict the hourly insolation in other regions around the world. The results of predicting other regions derived slightly higher prediction errors than Cheongju. However, it is expected that it will be possible to predict the hourly insolation in other regions with better prediction performance if variables related to geographical location are additionally considered in the future.

1. Introduction

1.1. Research Background. Worldwide, many people are paying close attention to climate change caused by greenhouse gas emissions. As of October 2021, 55 countries have declared, documented, or legislated carbon neutrality with a clear target year [1]. Since approximately 76% of greenhouse gas emissions are generated during energy consumption [2], energy consumption must be reduced, and existing fossil fuels that require combustion processes must be replaced by other energy sources, to reduce greenhouse gas emissions.

Since the energy consumed in buildings is nearly 40% of the total energy use [3], reducing the energy used in buildings can significantly reduce the total energy use. Therefore, high-efficiency systems for buildings are being developed to reduce energy consumed and renewable energy systems that can produce energy on-site. As a result, the concept of smart cities emerges, which optimizes the overall city energy system and enables the city to be energy independent. Smart cities are closely connected to the net-zero-energy building that self-produces energy from renewable energy sources, away from the existing method of receiving energy through fossil fuels. Each country has established various policies related to zero-energy buildings that are optimized for energy production and consumption in buildings [4–7].

In order to realize net-zero-energy building, it is essential to reduce energy use and optimize energy usage efficiency. Accordingly, the need for a building management system (BMS) emerged [8]. The BMS includes the building automation system (BAS), which monitors the status of facilities and automatically controls the operation, and the energy management system (EMS), which provides optimized energy management measures while maintaining a comfortable

indoor environment. Using EMS can save about 16% of the annual energy used in buildings [8]. Several preconditions must precede to operate the EMS stably. The critical point is that the EMS has to predict the amount of energy consumption and production to operate energy as necessary [9]. Furthermore, that energy production and consumption prediction can stabilize the operation of the energy grid of the district unit and city [10]. For this reason, many attempts have been made to predict renewable energy production through various data. For example, Balali et al. predicted solar and wind power generation through variables that indirectly affect power generation, such as the human development index [11].

Among various renewable energies, power generation in a building is performed by solar thermal, photovoltaic, geothermal, wind, and fuel cells that can be applied to the building in consideration of the size and system of such power generation facilities. The total amount of renewable energy generation worldwide is 3,063,926 MW, and the most actively applied photovoltaic power generation and solar thermal collection are close to 849,473 MW, which is about 28% of overall renewable energy generation [12]. Solar insolation greatly influences solar power generation and solar heat collection. Since the amount of insolation reaching the ground varies significantly due to sudden changes in weather [13] and particle matters in the atmosphere [14], accurate insolation prediction must precede predicting the amount of photovoltaic power generation and solar thermal collection [15]. To predict building power consumption, it is also necessary to accurately predict insolation [16].

Insolation refers to the total amount of solar radiation reached over a certain period, and thus, it is affected by all the processes it takes to reach the Earth, which makes predicting solar radiation very difficult. Various attempts have been made to develop predictive models of high reliability. The existing prediction models might be classified into physical and statistical models [17]. The physical model is a method of predicting insolation by analyzing actual physical elements such as weather observation information and includes a method of utilizing sky image and a method of using satellite images [18]. The representative physical models include the ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) model and Ahmad and Tiwari model [19]. Ahmad and Tiwari analyzed the solar radiation prediction performance of 50 physical models based on Konya region data. The model that derived the best performance uses nine input variables. The variables are the extraterrestrial radiation, solar declination, relative humidity, ratio of sunshine duration, mean air temperature, soil temperature, cloudiness, precipitation, and evaporation [20].

The statistical model is a model that statistically analyzes the correlation between past weather variables and insolation to predict insolation as a weather variable at a future time [17]. Recently, machine learning techniques such as support vector machine (SVM) and random forest (RF) and artificial intelligence such as artificial neural networks (ANN) and convolutional neural network (CNN) have been most frequently used [21]. In this circumstance, due to the importance and complexity of the accurate prediction of insolation for heating and cooling energy saving in buildings, a variety of studies have been conducted thus far focusing on solar insolation prediction.

1.2. Literature Review. As mentioned above, numerous structured AI-based models have been conducted recently due to the tendency for high prediction performance. Most research utilizes weather data to reflect the sky condition on insolation prediction. Hwang et al. predicted the daily insolation by building rain forest, a generalization acceleration model, and extreme gradient boost (XG Boost) model using weather data and air pollutant data. In this research, the prediction performance of the optimum model is 0.979 of R-square [22]. Ekici constructed a least squares support vector machine (LS-SVM) model to predict the daily insolation using weather data. This research conducted 0.094% of the coefficient of variation of the root mean square error (CV(RMSE)) in daily prediction [23]. Kumari and Toshniwal constructed an extreme gradient boosting forest-deep neural network (XGBF-DNN) model based on weather data to predict daily insolation. In this research, developed models had reduced prediction accuracy during the monsoon. The optimal prediction model has a prediction performance of 51.35 W/m^2 of root mean square error (RMSE) [24]. Chung constructed an ANN-based model based on weather data to predict daily insolation [25]. Husein and Chung constructed a long short-term memory (LSTM) model to predict daily insolation based on weather data [26]. To improve the accuracy of the daily insolation prediction model, many researchers tried to reflect accurate sky condition to the model.

Due to the time series characteristics of solar insolation, several researchers utilize previous insolation data to reflect a short-term meteorological trend. Diez et al. predicted daily insolation by a designated ANN model based on insolation for the past three days. The model that showed the best prediction performance was the model that utilized the amount of insolation and the date of the previous day only [27]. Behrang et al. constructed a multilayer perceptron (MLP) and radial basis function (RBF) models based on weather data and historical insolation data [28]. Cheng et al. constructed a convolutional LSTM model based on weather data and previous insolation data to predict daily insolation. The studies predicted daily insolation by developing an artificial intelligence- (AI-) based prediction model using various weather data and previous insolation data [29].

The meteorological data, measured by a set time-step, includes information about weather changes. Therefore, with the nearly real-time records, the tendency of data could certainly explain the sudden weather change. However, daily meteorological data is measured or generated by summation of 24 hourly data. Because of the tendency variation induced by sudden weather change to vanish during this process, the daily data have the essential disadvantage of predicting sudden weather change. Consequently, the daily prediction models are not sufficient for utilization in EMS.

Numerous researches on hourly prediction have been presented due to the limitation of applying daily prediction

data to EMS. Solmaz and Ozgoren developed an ANN model based on location data and average dry-bulb temperature to predict the hourly insolation. In this study, it was concluded that the ratio of dividing the data set affects the prediction performance [30]. Kuk Yeol et al. predicted hourly insolation by developing a SVM-based model using weather data. In this study, the SVM was derived from having higher predictive performance on the day of weather change than on nonlinear autoregressive (NAR) or ANN-based models [31]. Pang et al. constructed ANN and recurrent neural network (RNN) models based on weather data to predict hourly insolation. In this study, all cases proved that RNN-based models showed higher performance than ANN-based models and that the application of moving window further improved prediction performance [32]. Ji and Chee predicted hourly insolation by developing a model that combines autoregressive moving average (ARMA) and time delay neural network (TDNN). In this study, the RMSE of the hybrid model for every two-day time-step is in the range of about 25 to 270 Wh/m^2 [33]. This study confirmed that the higher the daily insolation, the lower the error rate. Voyant et al. predicted hourly insolation by designing a model that utilizes ANN and ARMA based on weather data measured in the Mediterranean climate. In this study, the proposed model did not significantly improve the predictive performance of the baseline ANN model. The overall average of NRMSE (normalized root mean square error) for optimum model is 16.3% [34]. Bamisile et al. predicted hourly global solar radiation (GSR) and diffused solar radiation (DSR) through models based on machine learning and deep learning algorithms. This study found that deep learning is more suitable than machine learning for GSR and DSR prediction. The best models had the following predictive performance: root mean square error (RMSE) of 82.22 W/m^2 and mean absolute error (MAE) of 36.52 W/m^2 [35].

Diverse prediction models that consider the time series feature have been developed to precisely predict the hourly insolation. Although with the great effort recently, the effect on improving prediction performance is limited. Some of the research with high prediction performance often used data from regions with a clear distinction between dry and humid seasons. Other researchers with high performance used data from regions where the climate was consistent throughout the year with little sudden change in weather conditions.

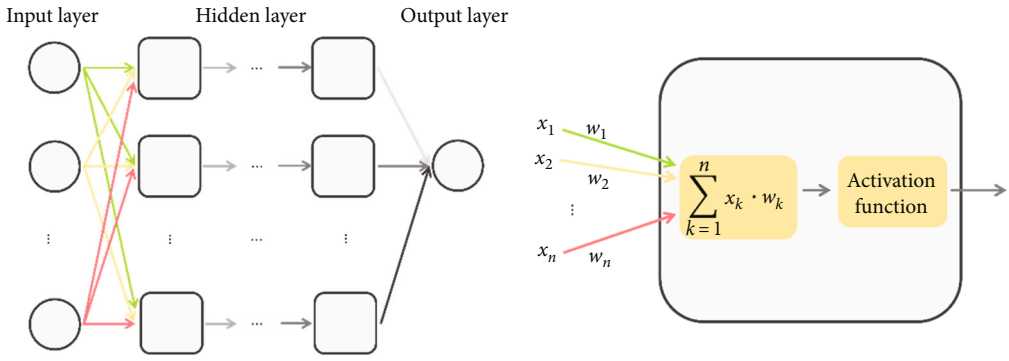
Recently, diverse LSTM-structured models have been presented due to the benefits induced by the structural characteristics of LSTM in processing time series data. Kim et al. predicted the hourly insolation by constructing a LSTM model based on weather data by CV(RMSE) of 26.87% [36], and Jeon et al. built an LSTM model using cloudiness and the previous day's insolation data. In this study, the LSTM model derived a high prediction performance only with other regions' insolation data and the previous day's insolation pattern [37]. Obiora et al. constructed a model of LSTM structure based on weather data to predict the hourly insolation and compared it with SVM-based models. This study's model of the LSTM structure derived a significantly higher prediction rate than other existing methods [38]. Qing and Niu constructed an LSTM model based on weather forecast data to predict the hourly insolation. Compared to other structures, the result was that LSTM showed higher predictive performance with less probability of overfitting. In this study, weather forecast data were used, so there was a limitation that the accuracy of predicting insolation significantly changed depending on the accuracy of the weather forecast [39].

In previous studies, insolation prediction models were constructed with various structures. However, in the daily prediction model, the prediction cycle was longer than an hour, making it difficult to be used in the EMS that receives real-time information for optimal control. In addition, most studies predicting hourly solar radiation tended to find it difficult to detect changes in solar radiation. On the hourly prediction, LSTM-structured models have become a general trend recently. However, most studies have been challenging to apply in practice or examine the model's characteristics to provide a background for further research. In some studies, a wide variety of data that needed additional estimation devices were used as input parameters. In some studies, the subject of study was only a short period when the sun was always up during the year. In some studies, only data from nonrainy periods were the subject of the study. We tried to do research that could be the basis for further research and could be applied immediately in practice. Therefore, we minimized the arbitrary data processing stage with only common weather data provided by the National Meteorological Agency and calculated solar location value. For this reason, we used the whole 365-day and 24-hour data to construct the model.

The studies on temperate or microthermal climates with high summer precipitation and distinct seasons were insufficient. Therefore, we aimed to be a fundamental background for further research and facilitation in these climates. Therefore, in this study, the ANN- and LSTM-based models are developed to accurately predict changes in insolation due to sudden weather changes based on weather data in Cheongju, Republic of Korea, located on the boundary between the temperate climate and the microthermal climate.

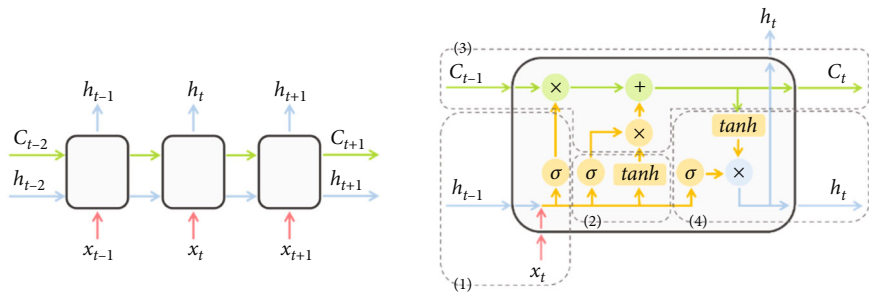
2. Overview of Predictive Models and Their Performance Evaluation Metrics

2.1. Artificial Neural Networks. Artificial neural networks (ANNs) are the statistical learning algorithms inspired by a neural network in biology. ANN-based model refers to the structure of a model in which artificial neurons that form networks by synaptic combinations repeatedly learn, changing the strength of the combination, and thus have problem-solving capabilities. The structure of ANN is shown in Figure 1. ANN consists of an input layer, hidden layer, and output Layer, and the input value is processed into various calculations and presented as a result value. ANN calculates each value of the input layer or the result value of the immediately preceding layer with each weighting factor and then derives the output through the activation function operation. The derived output is used again for the calculation of the next layer [40].



(a) Overall structure of conventional ANN model (b) Interior structure of ANN hidden neuron

Figure 1: Structure of ANN. (a) ANN model consisted with input, hidden, and output layers. The input data propagate to the output layer progressively. (b) In the neuron, propagated data is transformed with designated activation function.



(a) Overall structure of conventional LSTM model (b) Interior structure of LSTM cell

Figure 2: Structure of LSTM. (a) LSTM model consisted with diverse cells which sequentially aligned. (b) In the cell, there are four steps (forget gate, input gate, cell state, and output gate) to propagate data to the next cell. Each of the four steps has extinct role.

ANN produces robust results for several noises or incomplete data and has the advantage of being able to perform well in complex models [41] but has the disadvantages of passing through the hidden layer and transferring initial input data in a transformed state and vanishing gradients in finding the optimal value of parameters [42].

2.2. Recurrent Neural Networks. Recurrent neural networks (RNNs) are artificial neural networks that can utilize inference about past events due to the loop of delivering data and the loop of keeping initial input data intact. The RNN has the advantage of considering the context of the previous time-step when computing the current time-step. Due to this advantage, it is advantageous for repetitive and sequential time series data processing. However, like ANN, it has the disadvantage that Vanishing Gradient occurs in the process of finding the optimal value of parameters.

Therefore, the longer the time-step of the data leads to the long-term dependency problem, making it challenging to connect information and complex to utilize the context of the previous data [43].

2.3. Long Short-Term Memory. Long short-term memory (LSTM) is a type of RNN. The structure of the LSTM is shown in Figure 2. Unlike standard RNN, which has only one activation function in the loop, keeping the initial input data intact, LSTM contains multiple activation functions and operations within four interactive layers. In an LSTM structure, the section through which information is selectively passed is called "Gate", and the path through which the information flows in a straight line along the entire LSTM cell is called "Cell State". Each gate is composed of functions such as sigmoid and tanh and calculations [43].

The first layer of LSTM is the forget gate, which receives the input value x_t of the current time-step and the output value h_{t-1} of the previous time-step and determines information to be discarded through the sigmoid operation. The second layer is the input gate, which receives the input value x_t of the current time-step and the output value h_{t-1} of the previous time-step and determines what information to store through the sigmoid and tanh operations. The third layer is a layer that updates the cell state output value C_{t-1} of the previous time-step to the cell state C_t of the current time-step, and the fourth layer is a layer that determines information to be output based on the cell state C_t of the current time-step [43].

LSTM is designed to solve the long-term dependency problem of RNNs [44]. As mentioned, it aims to derive more accurate output values by properly combining appropriate historical and current information through multiple layers. As LSTM is a type of RNN, it is beneficial for models that use time series data, and when the length of the input data is very long, it usually derives higher predictive performance than RNN [43].

2.4. Performance Evaluation Metrics. In this study, mean absolute error (MAE), coefficient of variation of the root mean square error (CV(RMSE)), and normalized mean bias error (NMBE) are used as performance evaluation metrics to analyze the prediction performance of developed models. The low MAE and CV(RMSE), the low absolute value of NMBE, and high R^2 mean the difference between measured and prediction value is low, indicating the high prediction performance. For the equation of each performance evaluation metric, y_k means the k^{th} measured value and $\widehat{y_k}$ means the k^{th} prediction value.

(1) MAE is an intuitive indicator of the magnitude of the mean error, which utilizes the absolute value of the difference between predicted and measured values [45].

$$\text{MAE} = \frac{\sum_{k=1}^{n} |y_k - \widehat{y_k}|}{n} \tag{1}$$

(2) CV(RMSE) is the performance indicator that utilizes the proportion of square root of square value of the difference between predicted and measured values. According to the ASHRAE Guideline 14, the hourly prediction model has effectiveness with less than 30% of CV(RMSE) [46, 47].

$$\text{CV(RMSE)} = \frac{\sqrt{\left(\sum_{k=1}^{n} (y_k - \widehat{y_k})^2\right)/n}}{\sum_{k=1}^{n} y_k / n} * 100(\%) \tag{2}$$

(3) NMBE is the performance indicator that utilizes the proportion of summation of the difference between predicted and measured values. According to the ASHRAE Guideline 14, the hourly prediction model has effectiveness with the range between -10% and 10% of the NMBE [45, 46].

$$\text{NMBE} = \frac{\sum_{k=1}^{n} (y_k - \widehat{y_k})}{\sum_{k=1}^{n} y_k} * 100(\%) \tag{3}$$

(4) R-square (R^2) is the performance indicator that explains the degree of suitability of the prediction model to the measured value. Therefore, it shows the distance between the regression line of the measured line and the prediction value. R-square ranges between 0 and 1, and the closer it is to 1, the higher the prediction performance. According to the ASHRAE Handbook, R-square higher than 0.75 has effectiveness [48, 49].

$$R^2 = 1 - \frac{\sum_{k=1}^{n} (\widehat{y_k} - y_k)^2}{\sum_{k=1}^{n} (((\sum_{k=1}^{n} y_k)/n) - y_k)^2} \tag{4}$$

3. Data Processing and Predictive Model Development

3.1. Data Preprocessing

3.1.1. Meteorological Data Set. In this study, weather data provided by the Korea Meteorological Administration (KMA) were measured every hour at the meteorological observatory in Cheongju, Republic of Korea (36°38 north and 127°26 east) from 2012 to 2021. Of the 87,672 time-steps during the period, 87,670 data were used, excluding the two time-steps where all data were missing. Among the 87,670 data, in the case of insolation, when the sun did not rise, the corresponding value was assumed to be 0, and then, learning was conducted. For model verification, data from Jeju, Republic of Korea (33°51 north and 126°53 east); Santa Barbara, CA, USA (34°41 north and 119° 88 west); Millbrook, Duchess, NY, USA (41°79 north and 73°74 west) provided by the National Centers for Environmental Information (NCEI) is used. When the sun is below the horizon, the altitude angle is assumed to be 0.

3.1.2. Data Feature Analysis. Cheongju, the measurement location of the data used in this study, has an average temperature of 13.72°C from 2012 to 2021 and average annual precipitation of 1,113.66 mm. Cheongju has 17.1 days of precipitation among 31.8 days of the summer rainy season. During this season, the average seasonal precipitation is about 357.93 mm.

The average monthly temperature in January, the coldest month, is -0.86°C, located at the boundary between the temperate and microthermal climates. Cheongju has intermediate climate characteristics between the microthermal climate at the northern end of the Korean Peninsula and the temperate climate at the southern end, which tends to be generally consistent with the average annual average of the Republic of Korea [50]. Cheongju is located at the boundary between the temperate climate and the microthermal climate with a clear four-season climate. In the Köppen-Geiger climate classification [51], Cheongju is classified by microthermal climate. However, the recent climate records have changed slightly due to global warming. As a result, Cheongju has characteristics of both the microthermal climate and the temperate climate.

The range of changes in insolation is extensive due to summer precipitation, which has the most radiation reaching the atmosphere. Therefore, in order to accurately predict insolation in Cheongju, a model that predicts insolation changes caused by sudden changes in weather conditions with high accuracy must be developed.

3.1.3. Research Methodology. Figure 3 demonstrates the overall flow chart of research methodology.

3.1.4. Input Variable Selection through the Pearson Correlation Coefficient Analysis. Before selecting the input variables for the prediction, a correlation analysis was conducted using the Pearson correlation coefficient (PCC) analysis to find the correlation between the output variable and the input variable, which is the goal of the prediction. PCC is an index

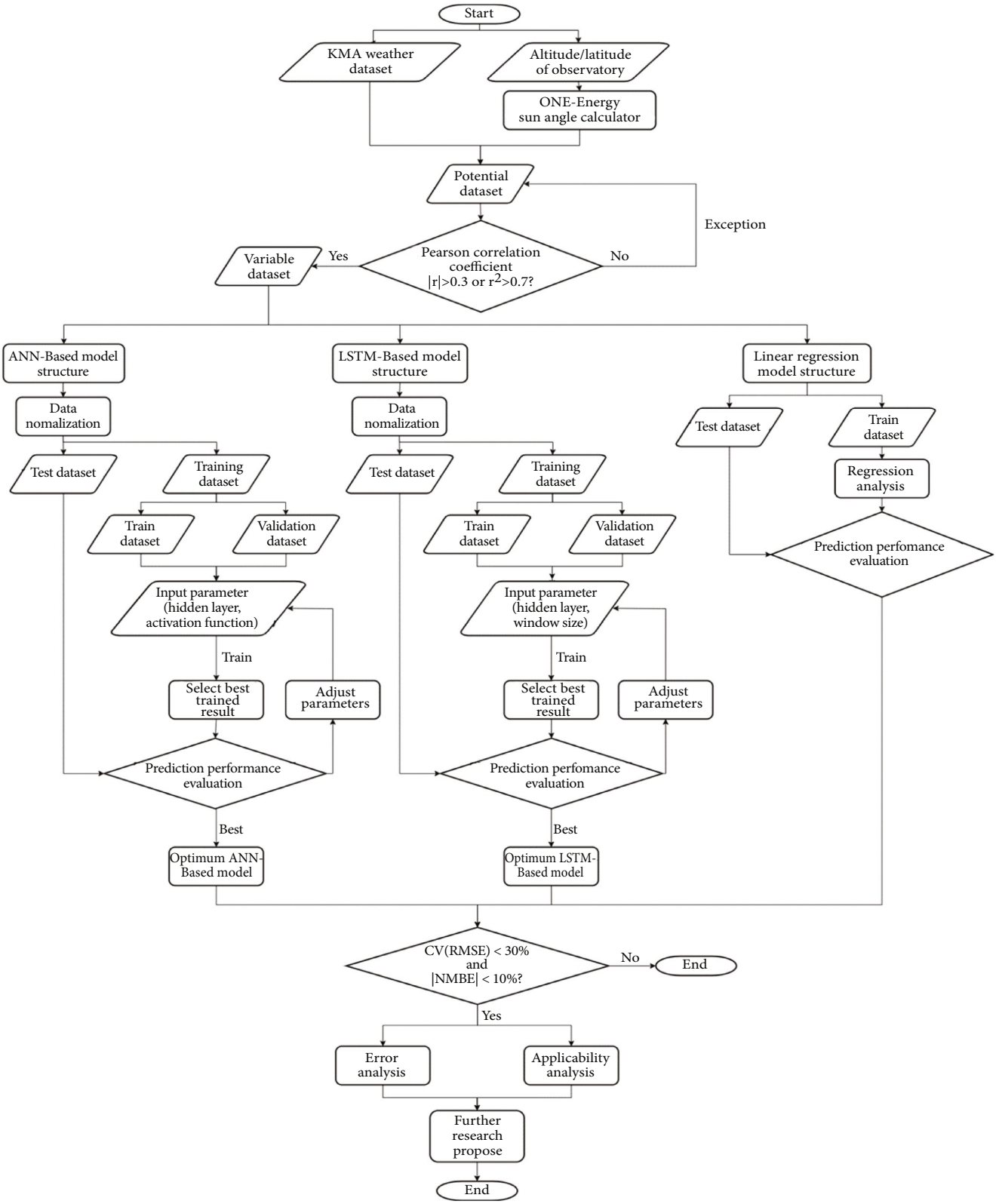


Figure 3: Overall flow chart of research methodology. In this project, ANN-based, LSTM-based, and linear regression models are constructed to derive the optimal model from predicting hourly insolation. After deriving the optimal model, a detailed analysis is processed to verify the result and propose further research.

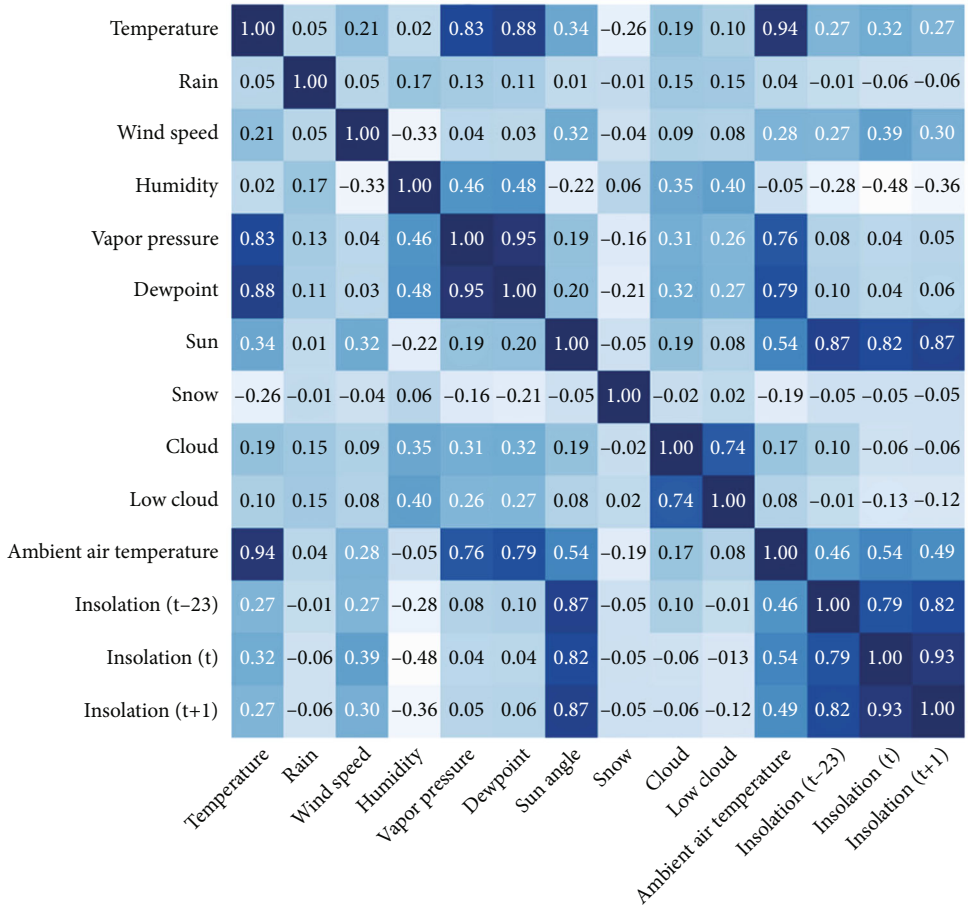


FIGURE 4: PCC analysis among meteorological variables. The overall result of PCC for every available data. A number of each box indicates the PCC between the x-axis variable and the y-axis variable. The darker the color of the box, the higher the Pearson correlation.

that represents a linear correlation between two variables and is the most frequently used correlation measure. The following equation can express PCC [52]:

$$r_{xy} = \frac{\sum(x_k - \bar{x})\sum(y_k - \bar{y})}{\sqrt{\sum(x_k - \bar{x})^2}\sqrt{\sum(y_k - \bar{y})^2}}, \quad (5)$$

where $(\bar{x} = \sum_{k=1}^{n} x_k/n, \bar{y} = \sum_{k=1}^{n} y_k/n.)$, $|r_{xy}| = 1$ (x and y are closely correlated), and $|r_{xy}| = 0$ (the linear correlation between x and y is not obvious).

The closer the r value indicating the correlation between variables is to 0, the lower the correlation between the two variables is, and the closer the 1 is, the higher the correlation between the two variables is. In this study, among Cohen's detection, $r^2 > 0.7$, the criterion for determining the suitability of variables in the engineering field was selected as the primary criterion [53]. If there is no variable suitable for this criterion, $|r| > 0.3$ was selected as the secondary criterion according to previous studies to determine and select the suitability of the input variable [54]. IBM Statistics SPSS 26 was used as software for PCC analysis.

Figure 4 shows the Pearson correlation coefficient r value derived through PCC analysis using longitudinal meteorological observation data provided by the KMA. According to the criteria for determining the suitability of the input variable, the selected input variable for predicting the insolation$(t + 1)$ includes wind speed, humidity, solar altitude angle, ambient air temperature, insolation$(t - 23)$, and insolation(t). Hereafter, insolation$(t + 1)$, insolation$(t - 23)$, and insolation(t) indicate the insolation at one hour after the current time-step, insolation at 23 hours before the current time-step, and insolation at the current time-step, respectively.

Pearson correlation coefficients r and r^2 of the selected input variables are shown in Table 1. In this study, in addition to the input variable selected by PCC, the model was constructed by adding the classification data, hour, to the variable to improve the model's accuracy.

3.1.5. Data Scaling. The most frequently used normalization function for regression analysis is MinMax Scaler, which transforms the entire data by transforming the minimum value of data to zero and the maximum value to 1. Therefore, there is a disadvantage that the inclusion of data outliers that are too large or too small in size affects the entire data set [55]. The distribution of each input variable does not change even if the maximum and minimum values of the input variables are converted equally through normalization, and it is

Table 1: PCC analysis among selected input variables.

	Wind speed	Humidity	Solar altitude angle	Ambient air temperature	Insolation($t-23$)	Insolation(t)
r	0.302	-0.361	0.868	0.495	0.823	0.927
r^2	0.091	0.130	0.753	0.245	0.677	0.859

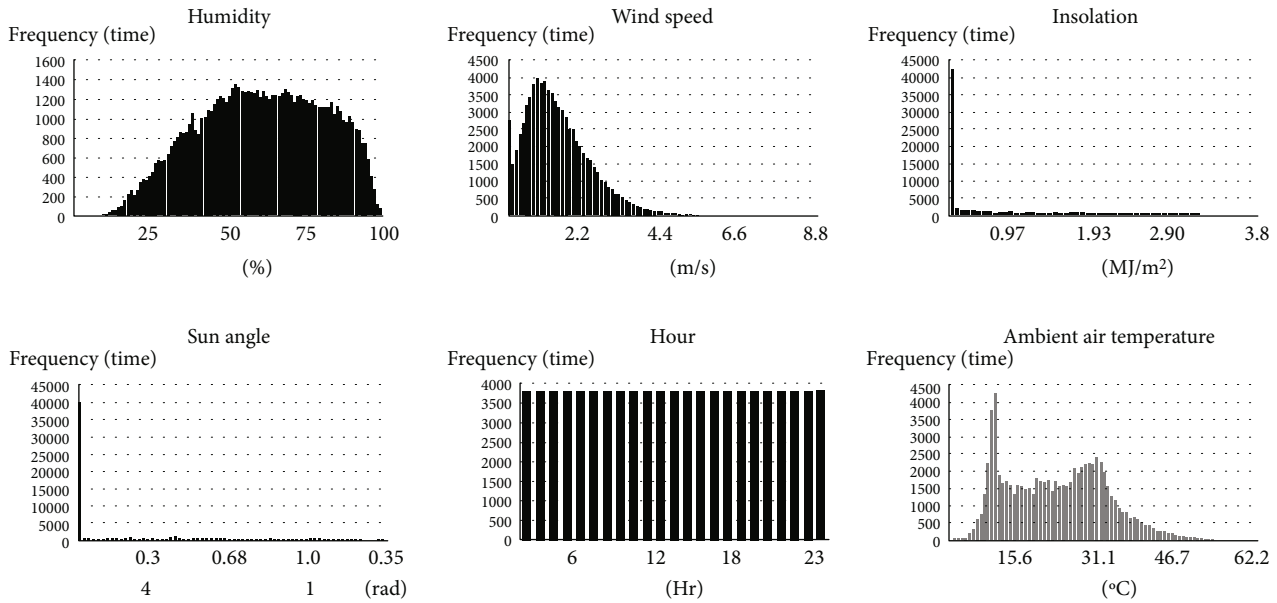


Figure 5: Distribution of input variables. The selected input variables have distinct distributions in frequency. The insolation and sun angle has a very skewed distribution, and the hour has an even distribution.

most appropriate to use quantile transformer in terms of computational speed, variance, and bias when using input variables with various distributions [56].

The input variables used in this study have various distributions as shown in Figure 5. The x-axis represents the absolute value of the input variable, and the y-axis represents the frequency of occurrence. In particular, the insolation(t) and solar altitude angle, which were the most correlated with insolation($t+1$) in the PCC analysis described in Section 3.1 (3), showed that about 50% of the data was zero, indicating that the data is concentrated to one side. Therefore, if normalization is carried out using a normalization function that uses the size of data values, such as the minimum and maximum values of variables, the weight for insolation(t) and solar altitude angles may be distorted. Therefore, normalization was carried out using quantile transformer, which converts diverse data sets nonlinearly [55] using the quantile among the normalization functions [57].

Table 2: Prediction performance of regression analysis.

Case	MAE	NMBE	CV(RMSE)	R^2
R1	0.1964	6.7648	48.3285	0.8831

Table 3: Prediction performance of ANN-based models.

Case	Layer	Activation function	MAE	NMBE	CV(RMSE)	R^2
A1	1	ReLu	0.0274	-0.0735	14.0353	0.9798
A2	1	Sigmoid	0.0286	0.0675	14.1840	0.9794
A3	2	ReLu	0.0263	-0.0396	14.2471	0.9792
A4	2	Sigmoid	0.0267	0.3935	14.0443	0.9799
A5	3	ReLu	0.0255	0.4514	14.2509	0.9794
A6	3	Sigmoid	0.0276	0.8078	14.1727	0.9797

3.2. Predictive Model Development and Performance Evaluation

3.2.1. Regression Analysis. Regression analysis was conducted using IBM Statistics SPSS 26, a statistical analysis software. To make the regression equations, 78,910 time-step data sets from 2012 to 2020 were used among the 87,670 data sets. To verify the regression equation, 8,760 time-step data sets corresponding to 2021 were used as the test set. Regression analysis was conducted using variables selected through PCC analysis. According to the regression analysis results, the most effective regression equation was created when all variables selected by PCC analysis were included, and the resulting regression equation is as follows: $\text{Insolation}_{t+1} = -0.047x_1 + 0.001x_2 - 0.004x_3 + 0.719x_4 + 0.579x_5 - 0.006x_6 + 0.090x_7 + 0.050$ (x_1 = wind speed, x_2 = humidity, x_3 = ambient air temperature, x_4 = insolation(t), x_5 = sun angle, x_6 = hour, and x_7 = insolation($t-23$)).

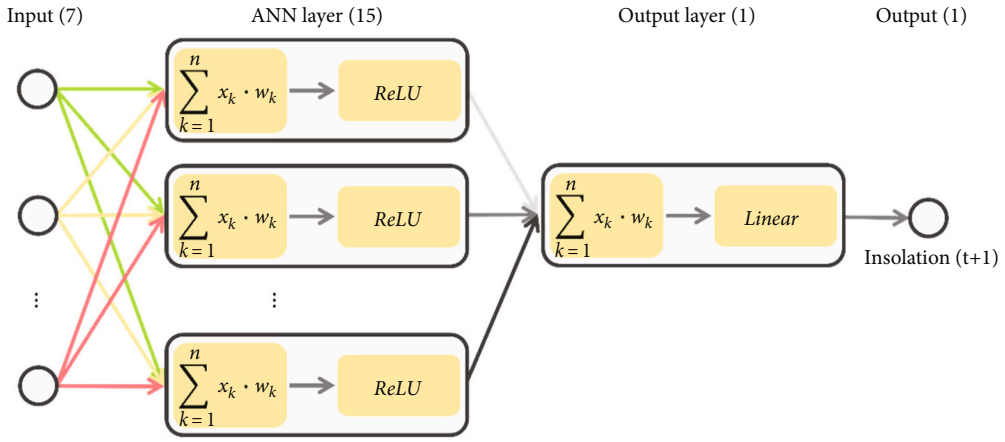


Figure 6: Structure of A1 model. Each hidden neuron of the A1 model has ReLu as an activation function. The output layer consists of linear function.

Table 2 shows the results of predicting 2021 data through the derived regression equation.

3.2.2. Deep Learning-Based Predictive Model Development. Of the 87,670 total data used in the study, 8,760 data sets corresponding to 2021 were used as test sets, 70,975 data sets corresponding to 90% of the 2012-2020 data sets were used as train sets, and 7,887 data corresponding to 10% were used as validation sets. In this study, LSTM and ANN algorithms were developed on Intel core i7 at 3.00 GHz, and 16 GB memory computers and Google Collaboration based on python were utilized to use various data analysis packages. To construct the LSTM- and ANN-based models, Keras Library, which provides various tools to simplify algorithms, was used. Based on the PCC analysis in Section 3.1 (2), a model was designated to derive optimal predictive performance by adjusting parameters such as window size and hidden layer using the selected input variables. Mean square error was used as a loss function, Adam was used as an optimizer, and the number of hidden neurons was selected according to previous studies [58].

3.2.3. ANN-Based Model Optimization. The prediction performance of the ANN-based model is shown in Table 3. In the case of ANN-based models, as shown in Table 3, the type of activation function and prediction performance was not significantly related, and it was confirmed that NMBE increased as the number of layers increased. Since the time required for computation increases as the number of layers increases, the A1 model with acceptable prediction performance, less bias, and faster computational speed was determined as the optimized ANN-based model.

The structure of the optimized ANN model is presented in Figure 6. The optimized A1 model has 2 layers. The ANN layer of the A1 model is processed by the activation function of ReLu with 15 hidden neurons.

The output layer of the A1 model integrates all neuron's results and is converted to the output value, which is the objective, insolation$(t+1)$ with linear activation function.

Table 4: Prediction performance of LSTM-based models.

Case	Layer	Window size	MAE	NMBE	CV(RMSE)	R^2
L1	1	12	0.0359	1.5882	19.0363	0.9640
L2	1	24	0.0312	0.2429	15.8019	0.9745
L3	1	36	0.0909	0.4617	15.9080	0.9742
L4	2	12	0.0343	1.2526	18.5190	0.9657
L5	2	24	0.0296	0.7412	15.2855	0.9764
L6	2	36	0.0307	0.9032	15.5197	0.9757
L7	3	12	0.0344	1.6601	18.3363	0.9666
L8	3	24	0.0301	0.3170	15.5485	0.9753
L9	3	36	0.0300	-0.3293	15.4208	0.9754

The model has a learning process with an Adam optimizer based on a 0.001 initial learning rate.

3.2.4. LSTM-Based Model Optimization. Table 4 shows the prediction performance of the LSTM-based model. Since a low CV(RMSE) was derived when the window size was 24 or 36 and a low NMBE was derived when the window size was 24, it can be interpreted that the model's error rate was low and the degree of deflection was the smallest when the window size was 24. Therefore, the optimal LSTM model was selected as L2, considering the performance metrics in Table 4 and the computational time.

The structure of the optimized LSTM model is presented in Figure 7. The optimized L2 model has 1 LSTM layer with 15 hidden neurons, learning with Adam optimizer based on a 0.001 initial learning rate. The result of the LSTM layer is collected and converted into the final output, which is our objective insolation$(t+1)$ at the output layerwith the activation function of linear.

The LSTM layer is constructed with LSTM cells every time-step. The cell structure of the optimized L2 model is presented in Figure 8. Each cell of LSTM utilizes the concatenated result by 7 input data and h_{t-1}, the output value of prior time-step, for the prediction. In addition, C_{t-1}, the cell state value of prior time-step, is also used to predict the output value h_t.

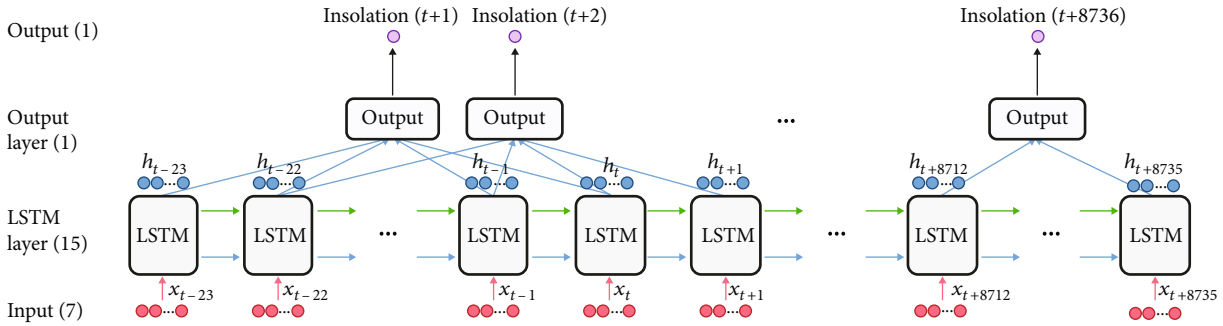


Figure 7: Structure of L2 model. The L2 model has 1 LSTM layer with a window size of 24 and 1 output layer.

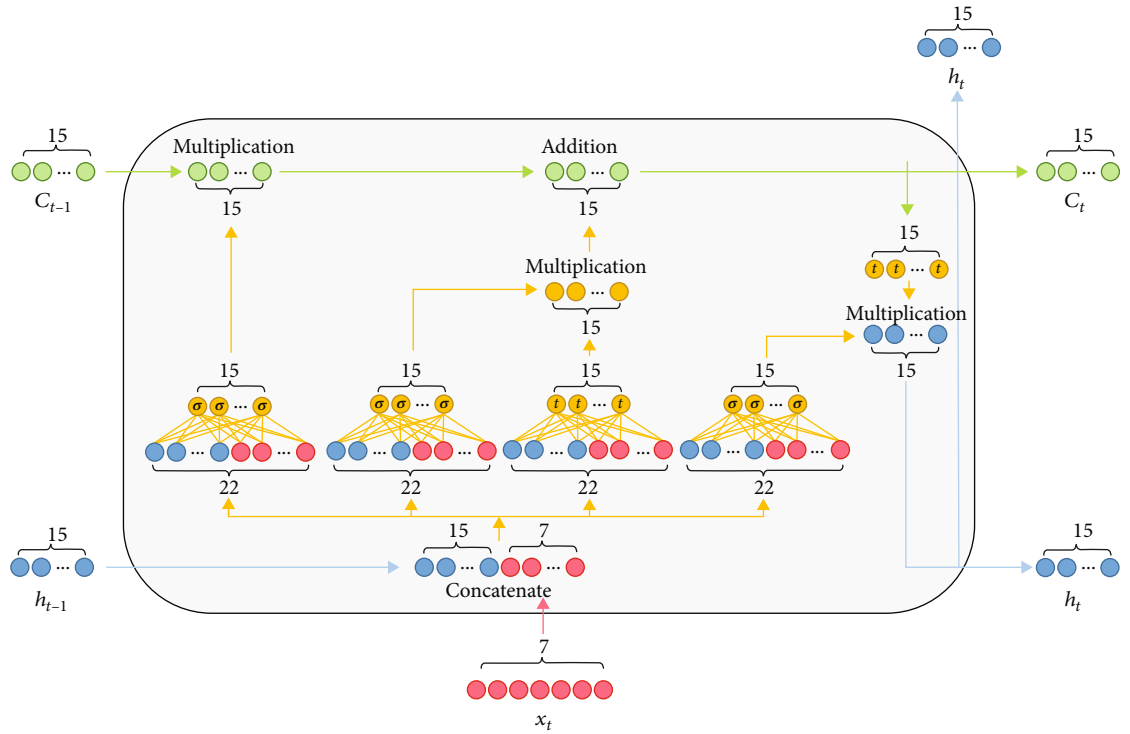


Figure 8: Cell structure of L2 model. In the LSTM layer, there are sequential LSTM cells exist. In each LSTM cell, input data and propagated data are transformed through the concatenate, multiplication, addition, and functionalized process.

4. Prediction Results Analysis and Discussion

Unlike previous papers that predicted daily insolation, this study has a high utilization in EMS by establishing a model for predicting hourly insolation [24–29]. It also has the advantage of being the basis for future studies or being applied in practice using only the weather observation data generally provided by the Meteorological Agency and calculated value. In addition, unlike previous studies, one model using data in all time zones was derived so that it can be used all year round without additional work [37–39]. Significantly, a model that can be used in other regions in the climate zone has been constructed using regional data located at the boundary between temperature and microthermal climate, which are rare in previous studies [23, 24, 33, 34].

4.1. Prediction Accuracy Comparison. In this study, the prediction accuracy results of the regression analysis, ANN-based model, and LSTM-based model were compared, as shown in Table 5. Hereafter, in cases of regression analysis, ANN and LSTM will be designated as R1, A1, and L2, respectively. According to ASHRAE Guideline 14, if predictive performance is derived within CV(RMSE) < 30% and −10% < NMBE < 10% based on hourly data, it can be interpreted as a valid model. Based on this guideline and the prediction performance of each model analyzed earlier, it can be confirmed that the deep learning-based A1 and L2 models,

Table 5: Prediction result of optimum model.

Case	MAE	NMBE	CV(RMSE)	R^2
R1	0.1964	6.7648	48.3285	0.8831
A1	0.0274	-0.0735	14.0353	0.9798
L2	0.0312	0.2429	15.8019	0.9745

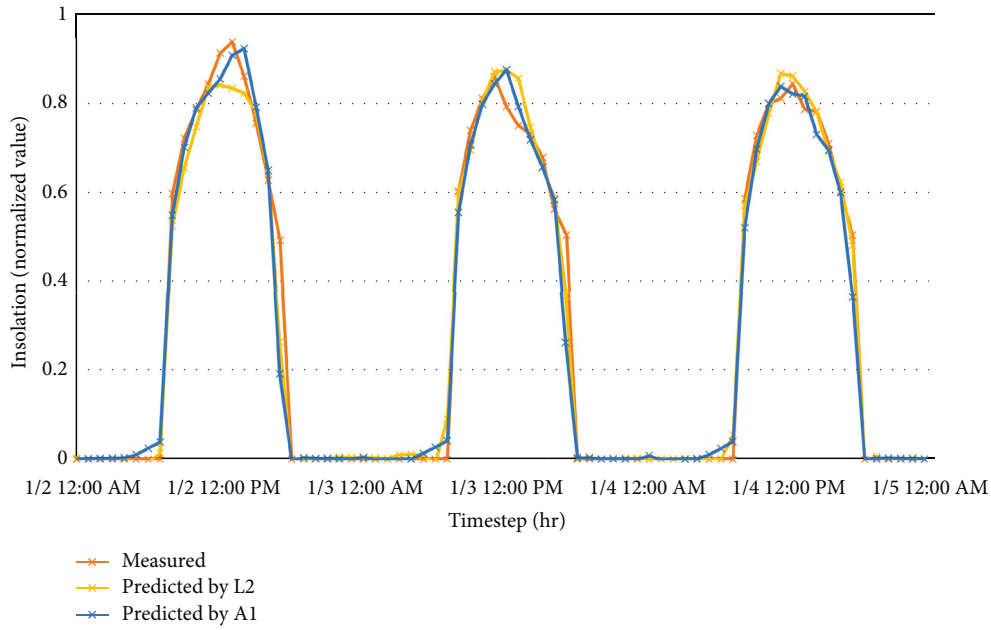


Figure 9: Line graph comparing prediction and measured values. Comparison between the measured insolation value and predicted value by A1 and L2 models.

excluding regression analysis-based model R1, derive appropriate prediction performance. Therefore, the result analysis of the prediction model was conducted in further detail based on the A1 and L2 models that derived reliable prediction performance.

Figure 8 compares the predicted and measured values by the A1 and L2 models for three days from January 2 to 4, 2021. Through Figure 9, it can be confirmed that both the A1 model and the L2 model derive similar prediction results to the actual measured values.

4.2. Detailed Analysis of Prediction Error. As shown in Figure 10, it can be seen that the insolation data is distributed only from 0.5 to 1. This pattern is derived from the normalization process of the insolation data, where about 50% of data appears to be zero because the insolation after sunset is all zero. Quantile transformer, which utilizes quantiles, is a normalization method that sorts the entire data set in order of size and then derives transformed values according to the quantiles.

As shown in Figure 5, unlike other input variables with various distributions, the insolation data have a very biased distribution, i.e., most values corresponding to the 0-50% quantile after normalization are zero. Since the insolation data corresponds to the 50-100% quantile and the same quantile of other data goes through a normalization process, the insolation value after the normalization of the data corresponds to the 50% quantile is considerably different from 0. Therefore, after normalization, there is no insolation value between 0 and 0.5.

Significant prediction errors have occurred with this biased tendency of insolation data distribution. Based on the investigation of the time-steps showing the top 1% significant difference between the measured values and the predicted values, it was found that about 85% error of the A1 model and 83% error of the L2 model were distributed from 5: 00 a.m. to 7: 00 a.m. and from 5:00 p.m. to 7:00 p.m., which are the sunrise and sunset times, respectively. Sunrise time is a value from zero to nonzero for insolation data, and sunset time is a time zone in which insolation data changes from a value of nonzero to zero. Since quantile transformer

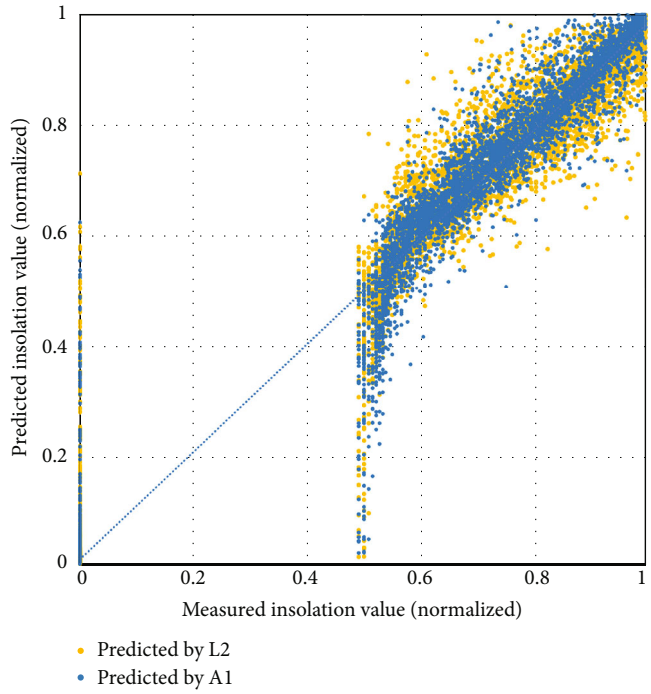
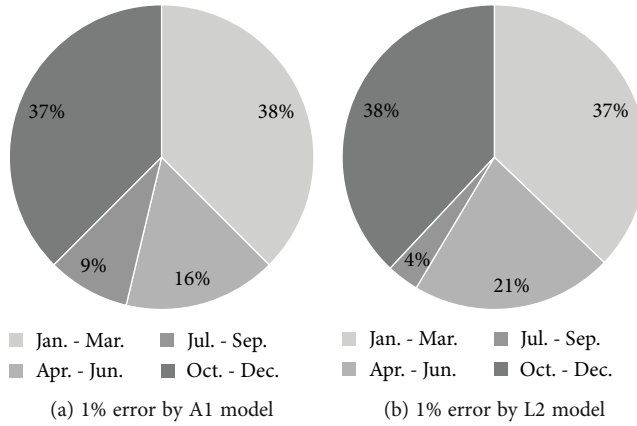


Figure 10: Scatter graph comparing prediction and measured values. Normalized insolation value has a very skewed distribution from a normalized value of 0.5 to 1.



(a) 1% error by A1 model (b) 1% error by L2 model

Figure 11: Seasonal distribution of prediction error. The largest 1% error by A1 and L2 models is concentrated in the fall and winter seasons in the Cheongju.

was used in this study, in the case of data with measured insolation slightly larger than 0, the difference from 0 significantly increases to about 0.5 after normalization. In this process, it is deemed that the model is not predicting the trends related to sunset and sunrise.

Figure 11 shows that about 75% of these top 1% errors are concentrated in the fall and winter seasons corresponding to January-March and October-December. Despite an extensive range of insolation changes induced by severe precipitation in the summer rainy season in Cheongju, only a limited number of errors occur in the summer season. Instead, the significant errors are intensively distributed in the winter, when the sunlight hour is short.

October-December and January-March, when 75% of the top 1% errors occurred, have a small amount of insolation and short sunlight hours with late sunrise and early sunset. Table 6 shows each month's section's sunrise and sunset times in 2021 [59]. Through the results of the intensive distribution of upper errors in October-March, it can be confirmed that the prediction model constructed in this study interprets the sunlight time for a long time. These results suggest the possibility of improving the prediction model's performance by using variables that imply sunrise and sunset times, such as sunlight time, in the future.

As noted in LSTM cell structure in Section 4.1, LSTM is advantageous in processing time series data because it utilizes the data from the previous time-step [43]. This is evidenced by the fact that in most cases, LSTM-based models derive better predictive performance than ANN-based models in existing insolation prediction studies, which compared ANN-based models and LSTM-based models [26, 33, 36, 39]. However, in this work, ANN-based models derive CV(RMSE) approximately 1-4% lower than LSTM-based models. Unlike previous studies, to analyze the cause of ANN-based models' higher prediction performance, a model that removed variables related to time series performance was constructed and predicted. In the corresponding model, the weather variables except for insolation($t-23$) and insolation(t) with time series properties were used as input variables, and the structure of the model was constructed the same as the ANN-based A1 model and LSTM-based L2 model with the selected optimal structure. As shown in Table 7, the A1 and L2 models were found to derive effective predictive performance even if only the rest of the weather variables except insolation(t) and insolation($t-23$) were used. Therefore, the prediction performance of the model is shown in Table 7.

Table 6: Sunrise and sunset times of each month interval.

	Jan.–Mar.	Apr.–Jun.	Jul.–Sep.	Oct.–Dec.
Sunrise time	7:12	5:29	5:48	7:07
Sunset time	18:08	19:28	19:14	17:29

Table 7: Prediction performance of selected models excluding the past insolation data as input variables.

	MAE	NMBE	CV(RMSE)	R^2
A1_excluded	0.0515	2.7754	24.3520	0.9425
L2_excluded	0.0401	1.9806	19.6182	0.9620

Figure 12 compares the predictive performance of the L2 and A1 models with insolation(t) and insolation($t-23$) usage differences. A1 model increases 2.7% of NMBE and 10% of CV(RMSE) without insolation(t) and insolation($t-23$) usage. In L2 model, 1.7% of NMBE and 4% of CV(RMSE) increase without insolation(t) and insolation($t-23$) usage. Therefore, when insolation(t) and insolation($t-23$) are used as input variables, it is found that both ANN and LSTM models improve performance more than when they are not utilized. On the other hand, when insolation(t) and insolation($t-23$) were not used as input variables, LSTM derived better prediction performance, but ANN derived better prediction performance when used.

The insolation data used in this study is the amount of insolation measured from the ground. Change in the insolation is derived when the sun is covered or scattered by particles in the atmosphere or clouds and physical obstacles. An insolation meter is installed in a space without trees. In detail, the distance to the nearest building is bigger than the height of the building [60, 61]. Therefore, the effect of

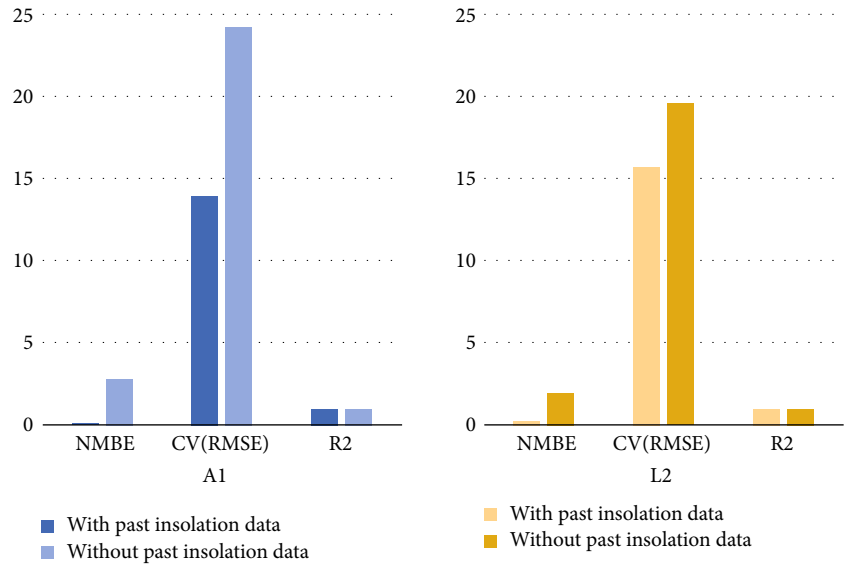


Figure 12: Comparison of prediction performance with and without the past insolation data. The A1 and L2 models have improved with past insolation data usage. Especially, the improvement of the A1 model is about 10%.

blocking or scattering the sun by physical elements installed on the ground is insignificant.

Therefore, the amount of insolation is affected by the atmosphere, which is reduced by the sun being covered by clouds or fog [62] or by the concentration of fine dust in the atmosphere [63]. Although the trend can be interpreted through data from the previous time-step if the sun is obscured throughout the day, it is impossible to interpret the trend through the previous time-step data if there is a sudden change in the cloud covering the sun only at a particular time.

The L2 model uses two time-step data, i.e., insolation(t) and insolation($t-23$), designated as input variables, and the other selectively important information within the past time-step data corresponding to the window size of 24. In comparison, A1 model directly utilizes only the two time-step data, i.e., insolation(t) and insolation($t-23$), as input variables. Therefore, the reason why the A1 model's performance is significantly improved by about 10% when using insolation(t) and insolation(t-23) in this study is expected to be due to the fact that the A1 model examines the tendency of shorter time-step than the L1 model and thus can appropriately capture sudden weather changes.

4.3. Applicability Verification of Prediction Models. In order to determine the applicability of the model developed in this study, verification was conducted using 2021 data from Jeju in the Republic of Korea, Santa Barbara in the United States, and Millbrook in the United States as a test set. A1 model and the L2 model were utilized to evaluate. About 0.06% of the meteorological data measured in Santa Barbara and Millbrook differed by more than 500% from the data before and after 1 hour, and in this case, the data at that time zone were considered an error and replaced with the average value of the data after 1 hour.

Tables 8 and 9 show the results of predicting insolation from Jeju, Santa Barbara, and Millbrook through ANN- and LSTM-based models built using Cheongju weather data in this study. Since ASHRAE Guideline 14-2014 considers a valid model when $-10\% < \text{NMBE} < 10\%$ and $\text{CV(RMSE)} < 30\%$, it is judged that both ANN-based and LSTM-based models can be used to predict hourly insolation outside Cheongju. Meanwhile, Jeju, Santa Barbara, and Millbrook all derived slightly lower predictive performance than Cheongju's predictive performance, and in particular, Santa Barbara and Millbrook derived lower predictive performance than Jeju. It was estimated that this was caused by various causes, such as the distribution characteristics of the data and the difference in the measurement process. Significantly, the following three possibilities are proposed in this study.

Table 8: Prediction performance of selected ANN model based on data from other regions.

A1	Jeju	Santa Barbara	Millbrook
MAE	0.0364	0.0360	0.0388
NMBE	-1.8333	-1.7962	-3.0687
CV(RMSE)	19.85	23.9967	23.5770
R^2	0.9628	0.9430	0.9446

Table 9: Prediction performance of selected LSTM model based on data from other regions.

L1	Jeju	Santa Barbara	Millbrook
MAE	0.0402	0.0405	0.0442
NMBE	0.5096	-0.1945	-1.4579
CV(RMSE)	21.8708	24.6502	25.1121
R^2	0.9569	0.9417	0.9390

First, Millbrook had the lowest ambient air temperature of -23.9°C, which was significantly lower than -10.9°C in Cheongju. Santa Barbara and Jeju had the highest wind speeds of 14.8 m/s and 13.7 m/s, respectively, significantly higher than 8.7 m/s in Cheongju. Therefore, a difference in prediction performance may have occurred in testing with data beyond the range of learning data.

Second, in the case of insolation data used in this study, the amount of global radiation throughout 1-hour data is composed of diffuse and direct radiation. Since diffuse insolation varies depending on the environment around the insolation meter, some differences may occur from region to region. In addition, Jeju data is provided by the KMA, while NCEI provides Santa Barbara and Millbrook data. The KMA measures wind speed at 10 m height [60], but NCEI estimates wind speed at 1.5 m [61]. Therefore, data measurement and estimation methods may have caused differences in prediction performance.

Third, this study used solar altitude angle data using latitude as input variables, but variables directly related to the geographical location such as latitude and longitude were not fully integrated in the predictive model. Therefore, the prediction performance may have deteriorated due to the difference in geographical location.

5. Conclusion and Future Work

This study evaluated prediction performance by constructing a regression model and deep learning-based predictive models to predict hourly insolation in areas located in temperate and microthermal climates with high precipitation. This study developed six models based on ANN, nine models based on LSTM, and 1 model based on linear regression, making up a total of 16 models. In this study, input variables were selected through PCC analysis. As an input variable, wind speed, humidity, solar altitude angle, ambient air temperature, insolation($t-23$), and insolation(t) were selected. This study is based on the hourly time-step and was trained based on data from the years of 2012 to 2020 to test 8,760 hours data corresponding to the year of 2021. The outstanding points of this study are highlighted as follows:

(i) The linear regression model was found to be inappropriate for predicting hourly insolation according to ASHRAE Guideline 14, whereas ANN- and LSTM-based models achieved reliable predictive performance

(ii) All ANN and LSTM-based models developed in this study were evaluated as models suitable according to the ASHRAE criteria. The A1 model, which showed the best predictive performance, derived CV(RMSE) of 14.04, NMBE of -0.0735, R-square of 0.9798, and MAE of 0.0274

(iii) This study proposed the direction of future research for improving the performance of predicting insolation at 1 hour after the current time-step, which has time-dependent characteristics, by utilizing insolation 24 hours before the current time-step and insolation at the current time-step in addition to weather data

(iv) Unlike other studies, the ANN-based model derived better performance than the LSTM-based model. This was considerd to be due to the fact that ANN performance was significantly improved using two past (t and $t-23$) insolation data

(v) In the optimal models, a significant error occurred at sunrise and sunset times, suggesting the possibility of further improving predictive performance by utilizing variables related to sunrise and sunset in the future

(vi) Along with Cheongju, the proposed model could adequately predict hourly insolation in other regions around the world such as Jeju, Santa Barbara, and Millbrook. The results of predicting other regions derived slightly higher prediction errors than Cheongju. However, it is expected that it will be possible to predict the hourly insolation with better prediction performance if variables related to the geographical location, such as latitude and longitude, can be fully integrated with the predictive model in the future

Abbreviations

ANN:	Artificial neural network
LSTM:	Long short-term memory
CV(RMSE):	Coefficient of variation of the root mean squared error
LS-SVM:	Least squares support vector machine
BMS:	Building automation system
BAS:	Building automation system
EMS:	Energy management system
ASHRAE:	American Society of Heating, Refrigerating and Air-Conditioning Engineers
SVM:	Support vector machine
RF:	Random forest
CNN:	Convolutional neural network
XG Boost:	Extreme gradient boost
XGBF-DNN:	Extreme gradient boosting forest-deep neural network
MLP:	Multilayer perceptron
RBF:	Radial basis function
AI:	Artificial intelligence
NAR:	Nonlinear autoregressive
RNN:	Recurrent neural network
ARMA:	Autoregressive moving average
TDNN:	Time delay neural network
NRMSE:	Normalized root mean square error
GSR:	Global solar radiation
DSR:	Diffused solar radiation
RMSE:	Root mean square error
MAE:	Mean absolute error
NMBE:	Normalized mean bias error
R^2:	R-square

NCEI: National Centers for Environmental Information
KMA: Korea Meteorological Administration
PCC: Pearson correlation coefficient.

Data Availability

Some or all data, models, or code generated or used during the study are available from the corresponding author by request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C2006469). This work was also supported by a Korea University grant (No. K2121761).

References

[1] Korea Energy Economics Institute [KEEI], *World Energy Market Insight*, KEEI, 2021.

[2] Historical GHG Emissions, "Climatewatch," 2021, https://www.climatewatchdata.org/ghg-emissions?source=CAIT.

[3] Y. Wu, Y. Wu, J. M. Guerrero, J. C. Vasquez, E. J. Palacios-Garcia, and Y. Guan, "IoT-enabled microgrid for intelligent energy-aware buildings: a novel hierarchical self-consumption scheme with renewables," *Electronics*, vol. 9, no. 4, p. 550, 2020.

[4] OECD/IPEEC, "Zero energy building definitions and policy activity: an international review," *IPEEC Building Energy Efficiency Taskgroup*, 2018, https://globalabc.org/resources/publications/zero-energy-building-definitions-and-policy-activity-international-review.

[5] The National Institute of Building Sciences, *A common definition for zero energy buildings.(EE-1247)*, U.S. Department of Energy, 2015, https://www.energy.gov/eere/buildings/downloads/common-definition-zero-energy-buildings.

[6] Directive, "2010/31/EU On the energy performance of buildings, CELEX: 32010L0031 E *U §*," vol. 12, 2010https://eur-lex.europa.eu/eli/dir/2010/31/oj.

[7] Zero Energy Building Certification System, *Zero Energy Building Certification Guideline*, Ministry of Land, Infrastructure and Transport, 2020, https://zeb.energy.or.kr/BC/BC03/BC03_05_003.do.

[8] D. Lee and C.-C. Cheng, "Energy savings by energy management systems: a review," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 760–777, 2016.

[9] J. Y. Kim and S. B. Cho, "Electric energy consumption prediction by deep learning with state explainable autoencoder," *Energies*, vol. 12, no. 4, p. 739, 2019.

[10] N. Shirzadi, F. Nasiri, C. El-Bayeh, and U. Eicker, "Optimal dispatching of renewable energy-based urban microgrids using a deep learning approach for electrical load and wind power forecasting," *International Journal of Energy Research*, vol. 46, no. 3, pp. 3173–3188, 2022.

[11] M. H. Balali, N. Nouri, M. Rashidi, A. Nasiri, and W. Otieno, "A multi-predictor model to estimate solar and wind energy generations," *International Journal of Energy Research*, vol. 42, no. 2, pp. 696–706, 2018.

[12] International Renewable Energy Agency, *Renewable Energy Capacity Statistics 2022*, IRENA, 2022, https://www.irena.org/publications/2022/Apr/Renewable-Capacity-Statistics-2022.

[13] Y. El Mghouchi, E. Chham, E. M. Zemmouri, and A. El Bouardi, "Assessment of different combinations of meteorological parameters for predicting daily global solar radiation using artificial neural networks," *Building and Environment*, vol. 149, pp. 607–622, 2019.

[14] A. Masoom, P. Kosmopoulos, A. Bansal et al., "Forecasting dust impact on solar energy using remote sensing and modeling techniques," *Solar Energy*, vol. 228, pp. 317–332, 2021.

[15] Y. Ahn, Y.-J. Lee, E. J. Oh, and B. S. Kim, "Prediction of building power consumption in the short-term through solar radiation calculation based on ultra-short weather forecast data," *Korean Journal of Air-Conditioning and Refrigeration Engineering*, vol. 33, no. 3, pp. 113–121, 2021.

[16] S. S. A. K. Javeed Nizami and A. Z. Al-Garni, "Forecasting electric energy consumption using neural networks," *Energy Policy*, vol. 23, no. 12, pp. 1097–1104, 1995.

[17] I. Sansa, Z. Boussaada, and N. M. Bellaaj, "Solar radiation prediction using a novel hybrid model of ARMA and NARX," *Energies*, vol. 14, no. 21, p. 6920, 2021.

[18] L. F. Zarzalejo, J. Polo, L. Martín, L. Ramírez, and B. Espinar, "A new statistical approach for deriving global solar radiation from satellite images," *Solar Energy*, vol. 83, no. 4, pp. 480–484, 2009.

[19] M. J. Ahmad and G. N. Tiwari, "Evaluation and comparison of hourly solar radiation models," *International Journal of Energy Research*, vol. 33, no. 5, pp. 538–552, 2009.

[20] M. J. Ahmad and G. N. Tiwari, "Solar radiation models—a review," *International Journal of Energy Research*, vol. 35, no. 4, pp. 271–290, 2011.

[21] Y. Zhou, Y. Liu, D. Wang, X. Liu, and Y. Wang, "A review on global solar radiation prediction with machine learning models in a comprehensive perspective," *Energy Conversion and Management*, vol. 235, p. 113960, 2021.

[22] Y. Hwang, D. Kang, M. Na, and S. Yoon, "Insolation prediction using air pollutants and meteorological variables," *Journal of the Korean Data Information Science Society*, vol. 32, no. 5, pp. 997–1005, 2021.

[23] B. B. Ekici, "A least squares support vector machine model for prediction of the next day solar insolation for effective use of PV systems," *Measurement*, vol. 50, pp. 255–262, 2014.

[24] P. Kumari and D. Toshniwal, "Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance," *Journal of Cleaner Production*, vol. 279, p. 123285, 2021.

[25] M. H. Chung, "Estimating solar insolation and power generation of photovoltaic systems using previous day weather data," *Advances in Civil Engineering*, vol. 2020, Article ID 8701368, 13 pages, 2020.

[26] M. Husein and I. Y. Chung, "Day-ahead solar irradiance forecasting for microgrids using a long short-term memory recurrent neural network: a deep learning approach," *Energies*, vol. 12, no. 10, p. 1856, 2019.

[27] F. J. Diez, L. M. Navas-Gracia, A. Correa-Guimaraes, A. Martínez-Rodríguez, and L. Chico-Santamarta, "Prediction of horizontal daily global solar irradiation using artificial neural networks (ANNs) in the Castile and León region, Spain," *Agronomy*, vol. 10, no. 1, p. 96, 2020.

[28] M. A. Behrang, E. Assareh, A. Ghanbarzadeh, and A. R. Noghrehabadi, "The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data," *Solar Energy*, vol. 84, no. 8, pp. 1468–1480, 2010.

[29] H.-Y. Cheng, C.-C. Yu, and C.-L. Lin, "Day-ahead to week-ahead solar irradiance prediction using convolutional long short-term memory networks," *Renewable Energy*, vol. 179, pp. 2300–2308, 2021.

[30] O. Solmaz and M. Ozgoren, "Prediction of hourly solar radiation in six provinces in Turkey by artificial neural networks," *Journal of Energy Engineering*, vol. 138, no. 4, pp. 194–204, 2012.

[31] B. Kuk Yeol, J. Han Seung, and S. Dan Keun, "Hourly solar irradiance prediction based on support vector machine and its error analysis," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 935–945, 2017.

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

[33] W. Ji and K. C. Chee, "Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN," *Solar Energy*, vol. 85, no. 5, pp. 808–817, 2011.

[34] C. Voyant, M. Muselli, C. Paoli, and M. L. Nivet, "Hybrid methodology for hourly global radiation forecasting in Mediterranean area," *Renewable Energy*, vol. 53, pp. 1–11, 2013.

[35] O. Bamisile, A. Oluwasanmi, C. Ejiyi, N. Yimen, S. Obiora, and Q. Huang, "Comparison of machine learning and deep learning algorithms for hourly global/diffuse solar radiation predictions," *International Journal of Energy Research*, vol. 46, no. 8, pp. 10052–10073, 2022.

[36] M. Kim, S. Jung, J. Kim, H. Lee, and S. Kim, "A study on solar radiation forecasting based on long short-term memory considering hourly weather changes," *Journal of Korean Institute of Intelligent Systems*, vol. 31, no. 1, pp. 88–94, 2021.

[37] B.-K. Jeon, K.-H. Lee, and E.-J. Kim, "Development of a prediction model of solar irradiances using LSTM for use in building predictive control," *Journal of the Korean Solar Energy Society*, vol. 39, no. 5, pp. 41–52, 2019.

[38] C. N. Obiora, A. Ali, and A. N. Hasan, "Forecasting hourly solar irradiance using long short-term memory (LSTM) network," in *2020 11th International Renewable Energy Congress (IREC)*, Hammamet, Tunisia, 2020.

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

[40] K. L. Priddy and P. E. Keller, *Artificial Neural Networks: An Introduction*, vol. 68, SPIE press, 2005.

[41] "[ADP] Neural Network Model by R," 2020, SH's learning note. https://todayisbetterthanyesterday.tistory.com/.

[42] S. Hochreiter, "Untersuchungen ZuDynamischenNeuronalen-Netzen [Diploma]," *Technische Universität München*, vol. 91, no. 1, 1991.

[43] C. Olah, "Understanding LSTM networks," 2015, https://colah.github.io/.

[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[45] P. Eguía-Oller, S. Martínez-Mariño, E. Granada-Álvarez, and L. Febrero-Garrido, "Empirical validation of a multizone building model coupled with an air flow network under complex realistic situations," *Energy & Buildings*, vol. 249, article 111197, 2021.

[46] American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE), *Guideline 14-2014, Measurement of Energy and Demand Savings*, Technical Report; American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA, USA, 2014.

[47] J. M. Lee, S. H. Hong, B. M. Seo, and K. H. Lee, "Application of artificial neural networks for optimized AHU discharge air temperature set-point and minimized cooling energy in VAV system," *Applied Thermal Engineering*, vol. 153, pp. 726–738, 2019.

[48] American Society of Heating, Refrigerating and Air-Conditioning Engineers [ASHRAE], *Handbook Fundamentals; American Society of Heating, Refrigerating and Air-Conditioning Engineers*, American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Atlanta, GA, USA, 2021.

[49] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, pp. e623–e623, 2021.

[50] Climate Statistics Analysis, "Korea Meteorological Administration [KMA]," 2022, https://data.kma.go.kr/.

[51] H. Beck, N. Zimmermann, T. McVicar, N. Vergopolan, A. Berg, and E. F. Wood, "Present and future Koppen-Geiger climate classification maps at 1-km resolution," *Scientific Data*, vol. 5, no. 1, article 180214, 2018.

[52] "SPSS Tutorials: Pearson Correlation," 2022, Kent State Library. https://libguides.library.kent.edu/.

[53] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation," *Anesthesia and Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[54] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, L. Erlbaum Associates, 2nd edition, 1988.

[55] "Compare the effect of different scalers on data with outliers," 2022, Scikit-learn. https://scikit-learn.org/.

[56] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[57] J. Lee, S. Kang, J. Jeong, and G. Chun, "Development of groundwater level monitoring and forecasting technique for drought analysis (I) - groundwater drought monitoring using standardized groundwater level index (SGI)," *Journal of Korea Water Resources Association*, vol. 51, no. 11, pp. 1011–1020, 2018.

[58] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *International 1989 Joint Conference on Neural Networks*, Washington, DC, USA, 1989.

[59] Monthly Sun and Moon Appearance Time, "Korea Astronomy and Space Science Institute," 2022, https://astro.kasi.re.kr.

[60] Korea Meteorological Administration [KMA], *Guidelines for the Terrestrial Weather Observation*, KMA, 2016.

[61] H. J. Diamond, T. R. Karl, M. A. Palecki et al., "U.S. climate reference network after one decade of operations: status and

assessment," *Bulletin of the American Meteorological Society*, vol. 94, no. 4, pp. 485–498, 2013.

[62] D.-K. Jo, C.-Y. Yun, K.-D. Kim, and Y.-H. Kang, "A study on the estimating solar radiation in Korea using cloud cover and hours of bright sunshine," *Journal of the Korean Solar Energy Society*, vol. 32, no. 2, pp. 28–34, 2012.

[63] J. Shim and D.-S. Song, "Effect of atmospheric particulate matter concentration on solar insolation," in *Proceedings of the SAREK Summer Conference 2019*, pp. 167–170, Yongpyeong, Gangwon-do, Republic of Korea, 2019.