IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 December 2023, accepted 4 January 2024, date of publication 10 January 2024, date of current version 26 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3352038

RESEARCH ARTICLE

PV Forecasting Model Development and Impact Assessment via Imputation of Missing PV Power Data

DAE-SUNG LEE[1], (Member, IEEE), AND SUNG-YONG SON[2], (Member, IEEE)
[1]Smart Energy System Convergence Research Institute, Gachon University, Seongnam 13120, South Korea
[2]Department of Electrical Engineering, Gachon University, Seongnam 13120, South Korea

Corresponding author: Sung-Yong Son (xtra@gachon.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under Grant RS-2023-00259004; and in part by the Human Resources Development of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) Grant funded by the Korean Government Ministry of Trade, Industry and Energy under Grant 20214000000060.

ABSTRACT Photovoltaics (PV) have attracted considerable attention owing to their longer lifespans and higher generation potentials compared with other renewable energy sources. However, the intermittent nature of PV systems can degrade the power quality, hindering their widespread adoption. To mitigate the power-quality degradation resulting from the proliferation of PV, high forecasting accuracy is essential. However, missing data during the development of forecasting models can degrade performance. Therefore, appropriate imputation procedures are required. Typically, linear imputation is used. However, there is a tendency for the performance of the forecasting model to decline owing to errors between the actual and imputed values. In this study, we addressed missing PV power data using direct deletion, linear imputation, k-nearest neighbors imputation, and Generative Adversarial Imputation Nets. Subsequently, to assess the impact of weather variability on the imputation performance, we employed the ''sky status'' to categorize the replaced data and analyze whether differences in imputation performance emerged. Finally, we developed a PV forecasting model using the replaced data and evaluated its forecasting performance.

INDEX TERMS CNN-GRU, GAIN, KNN, missing data imputation, PV forecasting model.

I. INTRODUCTION

The importance of renewable energy resources is steadily increasing because of global warming and rapid fossil-fuel depletion [1]. In particular, photovoltaics (PV) have attracted attention compared with other renewable energy sources owing to their advantages of low maintenance and operational costs, as well as long lifespans [9], [10].

However, PV generation is subject to significant variability owing to various weather conditions including cloud cover, solar irradiance, and temperature [21], [22]. The increased variability in PV generation owing to worsening weather conditions makes prediction difficult, thus emphasizing the increasing importance of advanced PV generation forecasting [23].

As the importance of PV forecasting has been emphasized, related research is actively underway. In [38], a stacked ensemble forecast approach was employed combining deterministic models that consider the physical characteristics of PV systems with ensemble models. This method demonstrated a high predictive performance under both sunny and cloudy conditions, outperforming traditional single models. In [39], a probabilistic forecasting approach was adopted using deterministic methods, such as clear-sky models and partial shading detection, to calculate meteorological variables as input variables, with the aim of providing reliable PV forecasts for energy management systems. In [40], a seasonal multi-model based on extreme learning machines was proposed to ensure accurate predictions for grid-connected

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

PV systems and evaluated the distribution of prediction errors across different seasons. In [41], a feed-forward artificial neural network (FF-ANN) prediction model utilizing real-time monitored data logs and predicted solar irradiance was introduced, facilitating the evaluation of the PV system operational efficiency based on predicted PV generation. In [42], a hybrid deep learning approach combining wavelet packet decomposition (WPD) and long short-term memory (LSTM) was proposed to exploit hidden nonlinear relationships in PV power. This approach exhibits superior performance under various seasonal and weather conditions, particularly in scenarios with high variability, such as ''cloudy'' and ''rainy'' conditions. The study in [43] involved analyzing the spatiotemporal correlations among distributed PV systems in the same region, and developed a PV power prediction model that integrated spatial similarity and temporal correlation into a Bayesian network.

The introduction of renewable energy forecasting, including PV forecasting, has improved the stability of power system operations [24], [31]. In [20], the variables commonly considered when developing PV generation forecasting models were investigated, with a focus on PV power, weather, and solar irradiance.

Forecasting research is typically conducted under the assumption of complete information, and the performance of forecasting models is degraded by forecasting models because of data gaps caused by communication errors and other factors [27], [28], [29]. Performance degradation of forecasting models can reduce the stability of power system operations. Therefore, the importance of research on handling missing data has steadily increased [16], [19], [26].

Various studies have been conducted to replace missing values in PV forecasting. In [11], imputation methods were applied to training and test datasets to analyze the impact of missing weather data on the forecasting performance, and a forecasting model was developed and compared with a perfect information-based model. In [12], the irradiance at locations that were not directly measured was estimated using multiple linear regression (MLR)—a statistical-based method—to replace the missing values in irradiance data measured at nine weather stations. In [13], SolarGAN was used to replace missing PV data, and the results indicated its effectiveness across a wide range of missing data rates. In [14], methods were proposed for detecting and replacing errors in small-scale PV systems, which involve the utilization of neighboring PV data when abnormal patterns are identified. In [15], missing data were imputed to analyze the performance and reliability of PV systems, and the authors proposed the use of data inference techniques such as the Sandia module temperature model (SMTM) for imputation when the missing data rate exceeded 10%. In [16], the impact of missing data on the estimation of the long-term degradation rate of PV systems was analyzed, and missing data were imputed using the Sandia PV Array Performance Model (SAPM). In [17], research was conducted to address the issue of missing irradiance data in tropical regions, and suitable replacement methods based on weather types such as ''Sunny'' and ''Intermittent'' were proposed. In [18], the accuracies of replacement methods for missing irradiance data were compared, and appropriate replacement methods were suggested for missing data such as minute and hourly series. In [19], to replace missing values in PV data, an approach combining a Naïve Super-Resolution Perception Convolutional Neural Network (SRPCNN) with linear imputation was proposed, which achieved a imputation performance improvement of over 30% compared to the existing SRPCNN in various missing data rate scenarios.

Previous research in PV-related data imputation primarily focused on replacement studies for situations with missing data such as PV power, weather, and irradiance. However, there has been limited analysis of the impact of missing PV power data on model forecasting performance.

In this study, we analyzed the impact of missing PV power data on the forecasting models. We developed models after applying missing PV power data replacement in various missing data rate scenarios and evaluated the forecasting performance. Additionally, to assess the differences in imputation performance based on weather variability, we divided the data into groups using ''sky status'' data provided by a meteorological agency and calculated the imputation performance. Furthermore, using data with missing values, we developed a forecasting model and analyzed the influence of missing PV power data on the performance of the forecasting model.

The remainder of this paper is organized as follows. Section II introduces the research design and methodologies employed for the imputation and forecasting. Section III presents the evaluation of the replacement performance for different missing data rates. In Section IV, we assess and analyze the forecasting models using imputed data. Finally, Section V presents conclusions.

II. METHODOLOGY

The objective of this study was to analyze the impact of missing PV power data on the performance of the forecasting model, as shown in Fig. 1. Initially, faultless PV and weather forecast data were categorized into training and test datasets. During development of the forecasting model, missing data were identified by intentionally introducing omissions into the dataset with perfect information. The reason for introducing omissions was to evaluate the imputation and forecasting performance. The missing data were replaced, and a forecasting model was developed and used to assess forecasting errors.

A. DATA INGESTION

This study utilized data measured at a PV power plant and weather forecast data provided by the Korea Meteorological Administration. The PV power plant is located in Jeongseon County, Gangwon Province, with solar panels installed at the azimuth of 180°, tilt angle of 20°, and capacity of 1 MW. The data used in the research were collected at one-minute intervals using sensors installed in the inverters at the PV

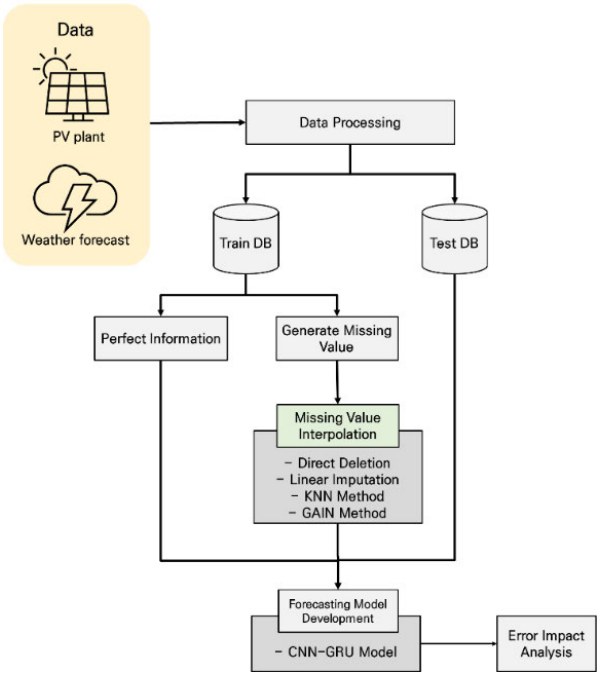


FIGURE 1. Research methodology.

power plant. Five variables from the measured data were used: horizontal solar radiation, inclined solar radiation, module temperature, outside temperature, and PV generation. Horizontal solar radiation represents the direct solar irradiance on the PV panels, whereas inclined solar radiation refers to the solar irradiance perpendicular to the PV power plant. The PV data measured at the solar power plant was utilized at one-minute intervals, and for the development of the prediction model, 15-minute averages were used. After the pre-processing step with 15-min intervals, were reconstructed into a dataset with perfect information, free from missing data, for the period from January 1, 2022, to April 30, 2023.

Weather forecast data were obtained from an open portal operated by the Korea Meteorological Administration. Weather forecast data was obtained from an open meteorological data portal operated by the Korea Meteorological Administration [44]. The short-term forecasts provided by the Korea Meteorological Administration were updated every 3 h. These short-term forecasts cover a time range starting from 6 h after the data update time and extending up to 79 h into the future. In this study, the weather forecast data were based on the forecast data closest to the prediction time. Data on sky status and estimated precipitation for the Jeongseon County area, where the power plant is located, were downloaded in the CSV format. Weather forecast data was obtained from an open meteorological data portal operated by the Korea Meteorological Administration [46]. Weather forecast data were calculated through numerical weather prediction by the Korea Meteorological Administration, and then analyzed using the perfect prog method (PPM) before being made available to the public [47]. The sky status indicates the number of clouds in the sky and ranges from one to four: 1 represents the ''sunny'' state with the least amount of clouds, while 4 represents the ''cloudy'' state with the most clouds.

B. IMPUTATION METHODOLOGY

For handling missing data, we applied four commonly used imputation methods: direct deletion, linear imputation, k-nearest neighbor (KNN) imputation, and Generative Adversarial Imputation Nets (GAIN).

Direct deletion involves removing all rows containing missing values and using only the completely measured data [32]. However, this data processing method can lead to significant information loss when there is a large amount of missing data, adversely affecting subsequent analyses. In this study, direct deletion was used to assess its impact on the predictive performance when the data were removed without replacement.

Linear imputation involves replacing missing values using nearby data points [33]. It is typically used when missing data must be imputed over short intervals, and the imputation performance declines as the length of continuous missing data increases.

In the KNN imputation method, the results of calculations using distance measurement formulas, e.g., Euclidean distance, are employed for replacement. Missing data are replaced using the values of the k-nearest neighbors [4]. The KNN algorithm is generally recognized as suitable for imputing missing values in time-series data, such as power consumption data [5].

GAIN is a modified version of the generative adversarial network (GAN) model [6]. A hint generator is added to the generator and discriminator components used in GAN. The generator is responsible for replacing missing values, whereas the discriminator determines whether the input data are real or generated by the generator [7]. Finally, the hint generator provides a hint matrix to the discriminator, assisting in discriminating the generated data [8].

C. FORECASTING MODEL

In this study, a convolutional neural network–gated recurrent unit (CNN-GRU) model was employed as the forecasting model. The convolutional neural network (CNN) model was designed to remove noise from the input data and extract important features [34]. The structure of a CNN includes convolution, pooling, and fully connected layers. The convolution layer processes the input data through convolution operations to remove the noise present in the input data and produces activation maps. The output activation maps undergo feature extraction and dimension reduction in the pooling layer [44]. Generally, methods such as max pooling and average pooling are considered; in this study, the max pooling method was adopted. Finally, the dimension-reduced activation maps are passed through the fully connected layer to produce the ultimate output.

The gated recurrent unit (GRU) is a variation of the long short-term memory (LSTM) network. A GRU comprises the reset gate layers, update gate layers, and candidate layers. The reset gate layer determines how much previous information should be forgotten, while the update gate layer determines how much previous information should be retained [45]. The update gate layer simultaneously performs the roles of the input and forget gate layers of the LSTM. It has a higher learning speed than LSTM because the forget and input gates are combined into a single gate [35]. The CNN-GRU model is a hybrid model that combines the CNN and GRU. During training, the input data processed using the CNN model undergo convolution operations to remove noise [2]. After feature extraction, the data undergo dimension reduction through pooling layers and are then processed through GRU layers to model the temporal features [3]. The output layer was used to derive the prediction results through a connected layer. Fig. 2 illustrates the architecture of the CNN-GRU model used in this study.

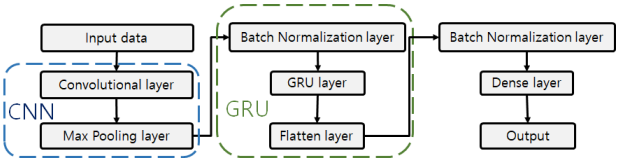


FIGURE 2. CNN-GRU hybrid model architecture.

D. EVALUATION INDEX

To evaluate the imputation performance, the coefficient of determination ($\boldsymbol{R^2}$) was used. Various indices were employed to assess the forecasting performance including $\boldsymbol{R^2}$ and the root mean square error (RMSE), relative root mean squared error (rRMSE), normalized root mean squared error (nRMSE), mean absolute error (MAE), mean relative deviation (MRD), and root mean squared deviation (RMSD), as defined by (1)–(7), respectively. Here, $\boldsymbol{y_t}$ represents the actual PV power, $\hat{\boldsymbol{y}}_{\boldsymbol{t}}$ represents the forecasted PV power, $\bar{\boldsymbol{y}}_{\boldsymbol{t}}$ represents the mean of the actual PV power, n represents the number of forecasted data points, $\boldsymbol{P_t}$ represents the forecasted PV power of the reference model, $\hat{\boldsymbol{P}}_{\boldsymbol{t}}$ represents the predicted generation of the imputed-data-based forecasting model, and $\boldsymbol{P_{total}}$ represents the PV installation capacity.

$$R^2 \quad 1-\left(\sum_{t=1}^{n}\left(y_t-\hat{y}_t\right)^2\right)/\left(\sum_{t=1}^{n}\left(y_t-\bar{y}_t\right)^2\right) \tag{1}$$

$$\text{RMSE} \quad \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t-\hat{y}_t\right)^2} \tag{2}$$

$$rRMSE \quad \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t-\hat{y}_t\right)^2}}{\bar{y}} \tag{3}$$

$$n\text{RMSE} \quad \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t-\hat{y}_t\right)^2}}{y_{max}-y_{min}} \tag{4}$$

$$MAE \quad \frac{1}{n}\sum_{t=1}^{n}\left|y_t-\hat{y}_t\right| \tag{5}$$

$$MRD \quad \frac{1}{n}\sum_{t=1}^{n}\frac{\left|P_t-\hat{P}_t\right|}{P_{total}} * 100 \tag{6}$$

$$RMSD \quad \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(P_t-\hat{P}_t\right)^2} \tag{7}$$

R^2 ranges from 0 to 1, with values closer to 1 indicating a higher accuracy. The RMSE represents the standard deviation of the estimation errors [36]. The rRMSE is obtained by dividing the RMSE by the mean [37]. The nRMSE is calculated by dividing the RMSE by the difference between the maximum and minimum values of the measured data [25]. The MAE reflects the difference between the measured and predicted values and is less sensitive to outliers than the RMSE [30]. The MRD and RMSD were used to assess the differences between the reference model and the imputation-based model [11].

III. IMPUTATION OF MISSING DATA

A. DATA INTRODUCTION

In this study, experiments were conducted using data from a 1-MW-capacity PV power plant in Jeongseon-gun, Gangwon Province, South Korea. To compare the performance of the forecasting models based on different data imputation methods, five variables measured at the PV power plant were used as input variables: direct radiation, diffuse radiation, module temperature, outside temperature, and PV power. Additionally, two variables provided by a meteorological agency—estimated precipitation and sky status—were used as input variables. The estimated precipitation represents the amount of precipitation within 1 h. The sky status indicates the number of clouds in the sky and ranges from 1 to 4, with 1 representing the least cloudy conditions (clear sky) and 4 representing the cloudiest conditions.

The data used in this study were collected from January 1, 2022 to April 30, 2023. The forecasting model was developed using data collected in 2022, and to evaluate its performance, data from January 1, 2023 to April 30, 2023 were used for testing.

B. MISSING DATA GENERATION

To create missing data conditions, missing data with rates of 10%, 20%, and 30% were generated using perfect PV data. An example of the missing PV power data is shown in Fig. 3. Missing data were randomly generated in 15-min intervals from the normally measured PV power data. The black solid line and blue dots represent the measured actual PV power and artificially missing PV power, respectively. The missing values were randomly generated throughout the entire 2022 dataset used in the forecasting model. To prevent bias in the results, missing data were generated five times, and the imputation and forecasting results were examined.

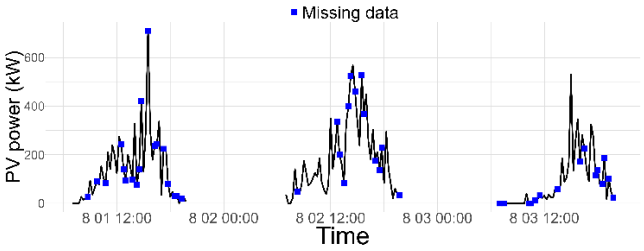


FIGURE 3. Example of missing PV power data: missing data rate of 30%.

C. IMPUTATION OF MISSING DATA

Missing data rates of 10%, 20%, and 30% were considered, and the linear, KNN, and GAIN imputation methods were applied.

1) APPLYING IMPUTATION METHOD BASED ON MISSING DATA

Linear imputation was performed using the na.approx() function provided by the Zoo package in R. Fig. 4 illustrates an example of missing data, where red and blue represent missing and imputed data, respectively. After five experiments, the average imputation performance for missing data rates of 10%, 20%, and 30% was calculated, resulting in R^2 values of 0.871, 0.866, and 0.865, respectively. Examining the imputed results for the period from 2022-08-01 12:30 to 2022-08-01 13:00, within which there were continuous missing data, revealed differences between the measured PV power and imputed PV power. These results indicate that the use of linear imputation for long-term missing data in periods of high PV power volatility may be unsuitable.

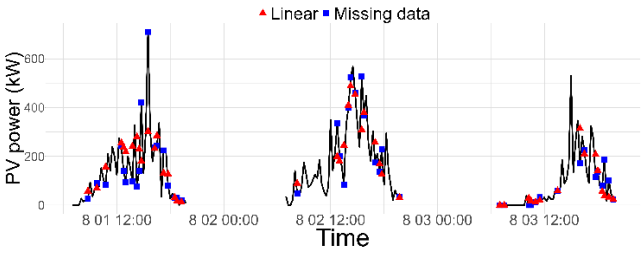


FIGURE 4. Linear imputation trial results: missing data rate of 30%.

KNN imputation was performed using the "knn.reg()" function from the "caret" package in R. The parameter "k" was varied from 1 to 9 to determine the optimal k value, which was found to be 7. An example of the results obtained with k set as 7 is shown in Fig. 5. Similar to Fig. 4, the red and blue points represent the missing and imputed data, respectively. The average imputation performance for the five experiments exhibited R^2 values of 0.975, 0.977, and 0.976 for the missing data rates of 10%, 20%, and 30%, respectively. This represents an average performance improvement of 0.109 compared with the commonly used linear imputation method.

GAIN was implemented using the "keras" library in Python to construct the generator, discriminator, and hint generator. The parameter settings are presented in Table 1.

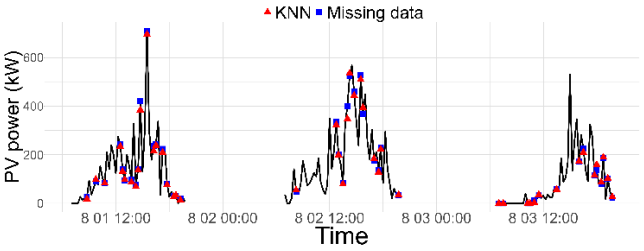


FIGURE 5. KNN imputation trial results: missing data rate of 30%.

TABLE 1. Generator and discriminator model configurations.

Model layer		Generator	Discriminator
Layer (1)	Unit	32	32E
	activation	ELU	softmax
Layer (2)	Unit	64	64
	activation	ELU	softmax
Layer (3)	Unit	9	9
	activation	ELU	softmax

An example illustrating the imputation performance is shown in Fig. 6. The average imputation performance for each case was calculated, and R^2 values of 0.924, 0.916, and 0.941 were obtained for missing data rates of 10%, 20%, and 30%, respectively. The imputation performance of GAIN was approximately 0.060 times higher than that of linear imputation. Analyzing the imputation results for the data corresponding to August 1, 2022, as shown in Fig. 6, revealed cases in which the replaced PV power was higher or lower than the actual PV power. Furthermore, compared with KNN and GAIN, linear imputation exhibited higher variability in imputation performance. The variability was calculated as the difference between the highest and lowest imputation performance values obtained from the experiments. The imputation performance variability for KNN imputation was 0.007, 0.005, and 0.009 sequentially for missing data rates of 10%, 20%, and 30%, respectively, whereas for GAIN imputation, it was 0.021, 0.045, and 0.027, respectively.

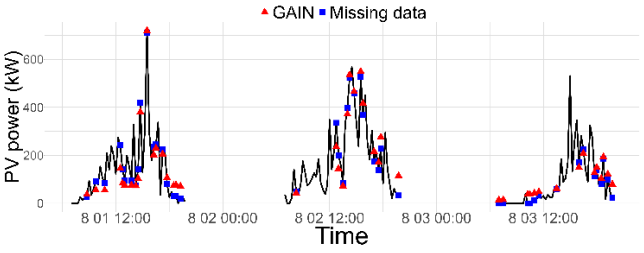


FIGURE 6. GAIN trial results: missing data rate of 30%.

2) IMPUTATION CONSIDERING SKY STATUS

We analyzed the impact of cloud cover on the imputation performance. The sky status variable was categorized into three groups: sky status 1 was labeled as "Sunny," status 4 was labeled as "Cloudy," and other values were labeled as "Partly cloudy."

Table 2 presents the imputation results for each sky-state category. Linear imputation consistently exhibited lower performance than the other imputation methods across all sky-state categories. In particular, for the "Cloudy" category, the imputation performance exhibited a sharp decline. Therefore, linear imputation is not suitable for imputation in situations with high variability in PV power.

KNN imputation consistently outperformed the other imputation methods—particularly when the sky state was "Partly cloudy" or "Cloudy." At a missing data rate of 30%, when the sky state was "Partly cloudy," the R^2 values were 0.861, 0.977, and 0.944 for linear, KNN, and GAIN, respectively. For the "Cloudy" category, the values were 0.816, 0.987, and 0.934, respectively. These results suggest that KNN-based imputation is appropriate for situations with significant weather variability.

TABLE 2. Comparison of imputation performance according to sky status classification (R^2).

Method	Case	Missing data rate (%)	Sky status			
			Total	Sunny	Partly cloudy	Cloudy
Linear	1	10	0.880	0.899	0.857	0.814
		20	0.870	0.880	0.861	0.828
		30	0.870	0.875	0.862	0.821
	2	10	0.873	0.899	0.857	0.814
		20	0.869	0.880	0.861	0.828
		30	0.865	0.875	0.862	0.821
	3	10	0.864	0.876	0.845	0.835
		20	0.859	0.877	0.843	0.832
		30	0.865	0.885	0.858	0.821
	4	10	0.870	0.892	0.865	0.812
		20	0.868	0.876	0.857	0.840
		30	0.863	0.873	0.860	0.820
	5	10	0.867	0.893	0.841	0.815
		20	0.864	0.892	0.861	0.801
		30	0.860	0.883	0.862	0.798
	Average	10	0.871	0.892	0.853	0.818
		20	0.866	0.881	0.857	0.826
		30	0.865	0.878	0.861	0.816
KNN	1	10	0.980	0.962	0.977	0.983
		20	0.980	0.969	0.982	0.985
		30	0.970	0.967	0.978	0.987
	2	10	0.973	0.962	0.977	0.983
		20	0.978	0.969	0.982	0.985
		30	0.977	0.967	0.978	0.987
	3	10	0.973	0.956	0.978	0.991
		20	0.976	0.963	0.980	0.989
		30	0.977	0.968	0.977	0.987
	4	10	0.977	0.964	0.982	0.987
		20	0.975	0.961	0.978	0.989
		30	0.976	0.963	0.976	0.988
	5	10	0.973	0.965	0.971	0.982
		20	0.978	0.971	0.974	0.987
		30	0.979	0.971	0.977	0.987
	Average	10	0.975	0.962	0.977	0.985
		20	0.977	0.967	0.979	0.987
		30	0.976	0.967	0.977	0.987
GAIN	1	10	0.940	0.933	0.951	0.926
		20	0.935	0.938	0.955	0.950
		30	0.941	0.931	0.945	0.941
	2	10	0.922	0.910	0.946	0.932
		20	0.913	0.922	0.896	0.888
		30	0.939	0.934	0.942	0.936
	3	10	0.919	0.895	0.944	0.937
		20	0.890	0.888	0.909	0.878
		30	0.924	0.928	0.916	0.911
	4	10	0.920	0.919	0.918	0.904
		20	0.911	0.917	0.925	0.905
		30	0.951	0.946	0.955	0.946
	5	10	0.921	0.928	0.950	0.936
		20	0.929	0.944	0.925	0.919
		30	0.951	0.954	0.964	0.939
	Average	10	0.924	0.917	0.942	0.927
		20	0.916	0.922	0.922	0.908
		30	0.941	0.939	0.944	0.935

IV. RESULTS OF PV FORECASTING MODEL

A. RESULTS OF PV FORECASTING MODEL DEVELOPED USING PERFECT INFORMATION

To analyze the impact of missing data on the forecasting models, we developed a forecasting model using perfect information. In the development of the forecasting model, the data points before 7 AM and after 8 PM were excluded because the PV power during these periods tended to be close to or equal to zero.

The forecasting model was based on the CNN-GRU model, predicting PV power 1 h ahead at 15-min intervals. We used the Adam optimizer with the MAE as the loss function. The batch size was set as 32. We set the number of epochs as 100 but employed an "early stop" mechanism, which halted training if the model's error did not decrease. The key hyper-parameter sets for each layer are as follows: for the CNN, the padding was set to 'same' to ensure identical number of input and output data. Setting 'same' prevents information loss caused by data reduction. For the GRU, both the kernel initializer and recurrent initializer were set to 'he-normal' to initialize the weights. The reason for choosing 'he normal' for weight initialization is the better prediction performance compared with other weight initialization methods. In particular, when using the 'glorot' method for weight initialization, the prediction results tended to converge to 0. The parameters of the forecasting model are presented in Table 3.

TABLE 3. Forecasting model parameters.

Conv1D	**filters**	**32**
	activation	ELU
Conv1D	filters	64
	activation	ELU
Conv1D	filters	128
	activation	ELU
Conv1D	filters	256
	activation	ELU
Max Pooling	Pool size	1
GRU	Unit	128
	activation	ELU
GRU	Unit	64
	activation	ELU
Flatten Layer		
Batch Normalization Layer		
Output	activation	ELU

The forecasting results of the prpposed model are presented in Fig. 7. The forecasting model yielded an R^2 value

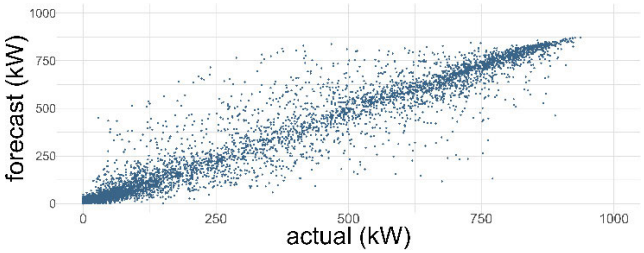


FIGURE 7. Forecasting results before the generation of missing data.

of 0.933. For the "Sunny," "Partly cloudy," and "Cloudy" sky statuses, the, R^2 scores were calculated as 0.946, 0.922, and 0.910, respectively. The forecasting performance for "Partly cloudy" and "Cloudy" was lower than that for "Sunny," with differences of 0.024 and 0.036, respectively, in the R^2 scores. The performance results for each index are presented in Table 4.

TABLE 4. Reference model results.

Evaluation index	Sky status			
	Total	Sunny	Partly cloudy	Cloudy
R^2	0.933	0.946	0.922	0.910
RMSE	19.45	18.30	20.71	19.77
rRMSE	0.278	0.226	0.284	0.399
nRMSE	0.083	0.078	0.089	0.086
MAE	10.71	9.56	11.83	11.47

B. RESULTS OF PV FORECASTING MODEL DEVELOPED USING IMPUTED DATA

We developed forecasting models for each missing data case to evaluate the forecasting performance. The parameters for the model development are consistent with those presented in Table 3. We developed forecasting models for each imputation method using the data generated from the five experiments. Subsequently, we categorized the forecasting results for each imputation method and the missing data rate and calculated the average forecast PV power. The forecasting performance was evaluated, and the results are shown in Fig. 8.

As shown in Fig. 8(a), regarding R^2, the KNN and GAIN models outperformed the existing linear and direct deletion methods. When the missing data rate was 30%, the R^2 values of direct deletion, linear, KNN, and GAIN for the all-sky status were 0.930, 0.930, 0.935, and 0.934, respectively. In the missing data rate range of 10–30%, the KNN-based model exhibited the best performance among the models examined, whereas the linear-based model exhibited the lowest performance.

Regarding the RMSE, rRMSE, and nRMSE, as shown in Figs. 8(b)–(d), the performances of the KNN- and GAIN-based models were lower than that of the direct deletion-based model at the missing data rate of 10%; however, it improved as the missing data rate increased. When the sky status was "cloudy" and the missing data rate was 30%, KNN outperformed GAIN.

As shown in Fig. 8(e), the KNN-based model exhibits the best performance at missing data rates of 10% and 20%, whereas the GAIN-based model exhibits the best performance at 30%. The performance of the direct deletion-based model decreased as the rate of missing data increased. When the four indices (RMSE, rRMSE, nRMSE, and MAE) were compared, differences were observed in the results of the GAIN-based model. For the "cloudy" sky status, the performance of the GAIN-based model declined as the missing data rate increased; however, this model exhibited the best performance at 30%. The MAE is less sensitive to outliers, but the other three indicators exhibit a characteristic of being sensitive to outliers. More specifically, it is suspected that the performance of the GAIN model declined for all three indicators, owing to the presence of outliers in the forecasting results.

Figs. 8(f) and (g) present the forecasting performance results in terms of RMSD and MRD. These two indices are associated with the deviation between the reference and imputation-applied models, reflecting the similarity of the compared models. For the all-sky status, the RMSD at the missing data rate of 10% was 37.60 for direct deletion and 40.80 for KNN, whereas in the 30% interval, it was 56.96 and 35.88, respectively. For MRD, with 10% missing data, it was 6.17 for direct deletion and 6.55 for KNN, whereas with 30% missing data, it was 10.81 and 5.93, respectively. As the missing data rate increased, direct deletion exhibited an increasing deviation from the reference model, whereas KNN exhibited the opposite trend, with the deviation decreasing as the rate increased.

Table 5 summarizes the forecasting results and presents the preferred imputation models for each case. The preferred method is defined as the imputation model that exhibits the best performance within each column or row.

C. ANALYSIS OF FORECASTING ERROR WITH RESPECT TO SKY STATUS

Finally, we compared and analyzed the time-specific forecasting errors of the reference model developed using perfect information with those of each imputation-applied model. The forecasting errors were calculated by subtracting the PV power forecasted by the imputation-applied model from that forecasted by the reference model. A positive error indicated that the reference model forecasted a higher value than the imputation model.

The time-specific forecasting errors are shown in Fig. 9. The black line inside the boxes represents the median of the errors, whereas the edges of the boxes represent the 25th and 75th percentiles.

Compared with the reference model in Fig. 9, for all the sky statuses, the linear imputation exhibited the smallest error at a missing data rate of 10%, whereas KNN and GAIN exhibited the smallest errors at 30%. At 10%, KNN and GAIN had the largest errors, whereas at 30%, direct deletion had the

TABLE 5. Preferred imputation methods for different missing data rates and forecasting error calculation methods.

Index	Total			Sunny			Partly cloudy			Cloudy			
Missing data rate (%)	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%	Preferred method
R^2	KNN	KNN	KNN	KNN	KNN	GAIN	KNN	GAIN	KNN	KNN	GAIN	KNN	KNN
RMSE	Direct	GAIN	KNN	Direct	KNN	KNN	Direct	GAIN	KNN	Direct	GAIN	KNN	KNN
rRMSE	Direct	GAIN	KNN	Direct	KNN	GAIN	Direct	GAIN	KNN	Direct	GAIN	KNN	KNN / GAIN
nRMSE	Direct	GAIN	KNN	Direct	KNN	GAIN	Direct	GAIN	KNN	Direct	GAIN	KNN	KNN/ GAIN
MAE	KNN	KNN	GAIN	Direct	KNN	GAIN	KNN	KNN	GAIN	KNN	Direct	GAIN	KNN
RMSD	Direct	GAIN	KNN	Linear	GAIN	KNN	Linear	GAIN	KNN	Linear	GAIN	KNN	KNN/ GAIN
MRD	Linear	GAIN	GAIN	Direct	Direct	GAIN	Linear	GAIN	GAIN	Linear	GAIN	KNN	GAIN
Preferred method	Direct	GAIN	KNN	Direct	KNN	GAIN	Direct	GAIN	KNN	Direct	GAIN	KNN	KNN

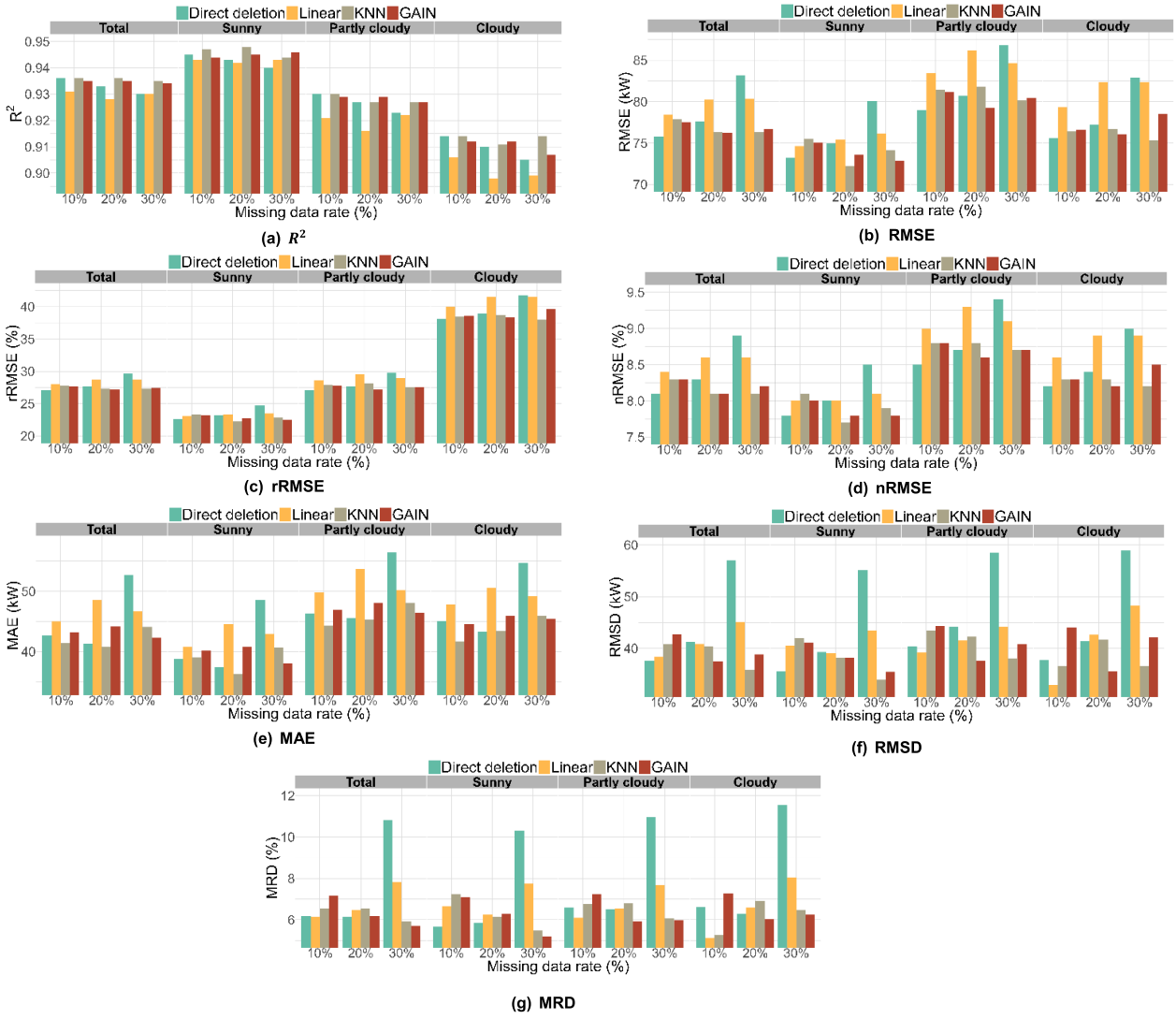


FIGURE 8. Comparison of power forecasting errors for different imputation methods.

largest error. When the sky status was "Sunny," the average errors at 10% were –2.673, –1.258, –5.167, –2.934 for direct deletion, linear, KNN, and GAIN, respectively. At 30%, they were –8.369, –5.051, –2.487, and –2.150, respectively. These results are consistent with the MRD findings presented in Fig. 8(e).

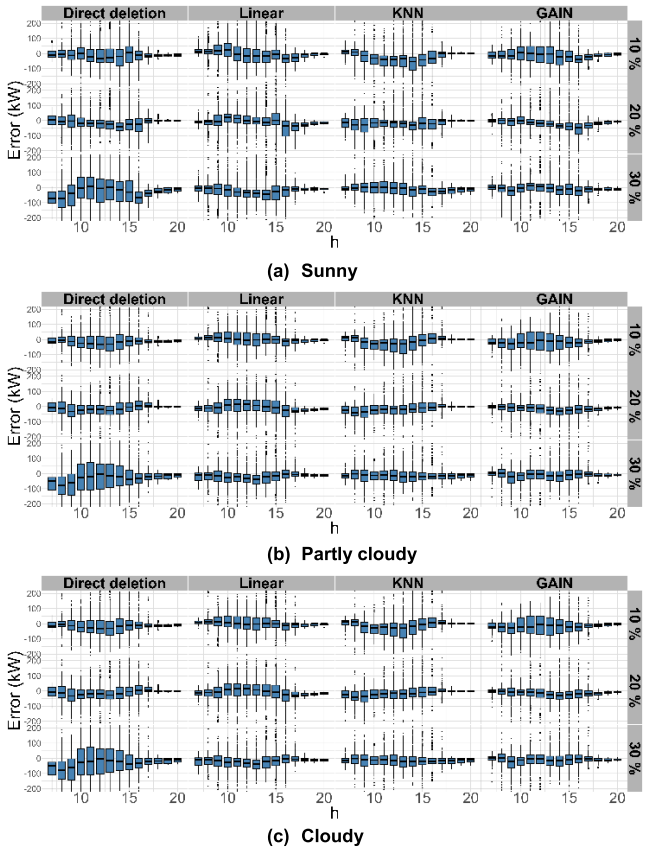


FIGURE 9. Comparison of the forecasting errors for different imputation methods.

V. CONCLUSION

We analyzed the impact of replacing missing PV power data on the results of forecasting models. Missing data were randomly generated in PV data with perfect information and were subsequently imputed using linear, KNN, and GAIN methods. Forecasting models were developed using the imputed data, and their performance was evaluated using seven indices. Regarding the imputation performance, KNN exhibited the highest R^2 at missing data rates of 10%, 20%, and 30%, with values of 0.975, 0.977, and 0.976, respectively. However, when forecasting models were developed using the imputed data and compared, direct deletion performed the best at 10%, whereas the KNN and GAIN models performed better at 20% and 30%. The results for different missing data rates and evaluation indices indicate that KNN is generally a good imputation method but also highlight the importance of selecting the appropriate imputation method depending on the situation.

In addition, the errors between the reference and imputation-applied models were consistent with the results of the MRD analysis. As the missing data rate increased, the average errors of KNN and GAIN, which excelled in PV power imputation, decreased. This reflects the impact of missing PV power data on the performance of the forecasting models; i.e., the importance of the imputation performance increases with the missing data rate.

Our study has the following limitations:

First, since this study targeted only one PV plant, the generalizability of the research results is limited. In future studies, we plan to apply interpolation methods to a large number of PV plants to generalize an appropriate imputation method for PV plants.

Second, since this study primarily focused on analyzing the impact of missing PV power data on the forecasting model, we did not consider the issues of missing data and diversity in weather forecasts. However, PV power-forecasting models require various input variables, including weather forecasts, PV power, and solar irradiance. Therefore, we anticipate that additional research will simultaneously address the missing data issues of the key variables used in PV power forecasting models.

Third, there is a limitation in using only a short-term PV dataset. The data used in this study covers the period from January 1, 2022, to April 30, 2023. The period used for model validation is limited to January 1, 2023, to April 30, 2023. Due to the short duration of analysis, there is a limitation in not considering the seasonal characteristics of PV power. In future work, we plan to conduct additional research that takes into account the seasonal characteristics of PV power.

In future research, we plan to improve the forecasting performance through ensemble methods that combine various imputation methods for missing PV data.

REFERENCES

[1] W. U. Rehman, A. R. Bhatti, A. B. Awan, I. A. Sajjad, A. A. Khan, R. Bo, S. S. Haroon, S. Amin, I. Tlili, and O. Oboreh-Snapps, "The penetration of renewable and sustainable energy in Asia: A state-of-the-art review on net-metering," *IEEE Access*, vol. 8, pp. 170364–170388, 2020.

[2] M. Suresh, M. S. Anbarasi, J. Divyabharathi, D. Harshavardeni, and S. Meena, "Household electricity power consumption prediction using CNN-GRU techniques," in *Proc. Int. Conf. Syst., Comput., Autom. Netw. (ICSCAN)*, Jul. 2021, pp. 1–6.

[3] Q. Li, X. Zhang, T. Ma, D. Liu, H. Wang, and W. Hu, "A multi-step ahead photovoltaic power forecasting model based on TimeGAN, soft DTW-based K-medoids clustering, and a CNN-GRU hybrid neural network," *Energy Rep.*, vol. 8, pp. 10346–10362, Nov. 2022.

[4] L. Erhan, M. Di Mauro, A. Anjum, O. Bagdasar, W. Song, and A. Liotta, "Embedded data imputation for environmental intelligent sensing: A case study," *Sensors*, vol. 21, no. 23, p. 7774, Nov. 2021.

[5] C. Bülte, M. Kleinebrahm, H. Ü. Yilmaz, and J. Gómez-Romero, "Multivariate time series imputation for energy data using neural networks," *Energy AI*, vol. 13, Jul. 2023, Art. no. 100239.

[6] Y. Sun, J. Li, Y. Xu, T. Zhang, and X. Wang, "Deep learning versus conventional methods for missing data imputation: A review and comparative study," *Expert Syst. Appl.*, vol. 227, Oct. 2023, Art. no. 120201.

[7] J. Yoon, J. Jordon, and M. Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. ICML*, 2018, pp. 5689–5698.

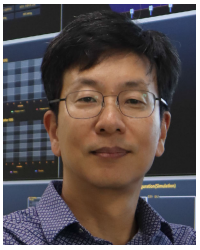
[8] Y. Zhang, R. Zhang, and B. Zhao, "A systematic review of generative adversarial imputation network in missing data imputation," *Neural Comput. Appl.*, vol. 35, no. 27, pp. 19685–19705, Sep. 2023.

[9] U. K. Das, K. S. Tey, M. Seyedmahmoudian, S. Mekhilef, M. Y. I. Idris, W. van Deventer, B. Horan, and A. Stojcevski, "Forecasting of photovoltaic power generation and model optimization: A review," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 912–928, Jan. 2018.

[10] M. N. Akhter, S. Mekhilef, H. Mokhlis, Z. M. Almohaimeed, M. A. Muhammad, A. S. M. Khairuddin, R. Akram, and M. M. Hussain, "An hour-ahead PV power forecasting method based on an RNN-LSTM model for three different PV plants," *Energies*, vol. 15, no. 6, p. 2243, Mar. 2022.

[11] T. Kim, W. Ko, and J. Kim, "Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting," *Appl. Sci.*, vol. 9, no. 1, p. 204, Jan. 2019.
[12] C. Turrado, M. López, F. Lasheras, B. Gómez, J. Rollé, and F. Juez, "Missing data imputation of solar radiation data under different atmospheric conditions," *Sensors*, vol. 14, no. 11, pp. 20382–20399, Oct. 2014.
[13] W. Zhang, Y. Luo, Y. Zhang, and D. Srinivasan, "SolarGAN: Multivariate solar data imputation using generative adversarial network," *IEEE Trans. Sustain. Energy*, vol. 12, no. 1, pp. 743–746, Jun. 2021.
[14] S. Park, S. Park, M. Kim, and E. Hwang, "Clustering-based self-imputation of unlabeled fault data in a fleet of photovoltaic generation systems," *Energies*, vol. 13, no. 3, p. 737, Feb. 2020.
[15] A. Livera, M. Theristis, E. Koumpli, S. Theocharides, G. Makrides, J. Sutterlueti, J. S. Stein, and G. E. Georghiou, "Data processing and quality verification for improved photovoltaic performance and reliability analytics," *Prog. Photovoltaics, Res. Appl.*, vol. 29, no. 2, pp. 143–158, Feb. 2021.
[16] I. Romero-Fiances, A. Livera, M. Theristis, G. Makrides, J. S. Stein, G. Nofuentes, J. de la Casa, and G. E. Georghiou, "Impact of duration and missing data on the long-term photovoltaic degradation rate estimation," *Renew. Energy*, vol. 181, pp. 738–748, Jan. 2022.
[17] N. B. Mohamad, A.-C. Lai, and B.-H. Lim, "A case study in the tropical region to evaluate univariate imputation methods for solar irradiance data with different weather types," *Sustain. Energy Technol. Assessments*, vol. 50, Mar. 2022, Art. no. 101764.
[18] H. Demirhan and Z. Renwick, "Missing value imputation for short to mid-term horizontal solar irradiance data," *Appl. Energy*, vol. 225, pp. 998–1012, Sep. 2018.
[19] X. Liu, C. Huang, L. Wang, and X. Luo, "Improved super-resolution perception convolutional neural network for photovoltaics missing data recovery," *Energy Rep.*, vol. 9, pp. 388–395, Sep. 2023.
[20] P. Gupta and R. Singh, "PV power forecasting based on data-driven models: A review," *Int. J. Sustain. Eng.*, vol. 14, no. 6, pp. 1733–1755, Nov. 2021.
[21] F. Katiraei and J. R. Agüero, "Solar PV integration challenges," *IEEE Power Energy Mag.*, vol. 9, no. 3, pp. 62–71, May 2011.
[22] L. Bird, M. Milligan, and D. Lew, "Integrating variable renewable energy: Challenges and solutions," Nat. Renew. Energy Lab. (NREL), Golden, CO, USA, Tech. Rep. NREL/TP-6A20-60451, 2013, doi: 10.2172/1097911.
[23] B. Li and J. Zhang, "A review on the integration of probabilistic solar forecasting in power systems," *Sol. Energy*, vol. 210, pp. 68–86, Nov. 2020.
[24] R. Widiss and K. Porter, "Review of variable generation forecasting in the west: July 2013-March 2014," Nat. Renew. Energy Lab. (NREL), Golden, Co, USA, Tech. Rep. NREL/SR-6A20-61035, 2014, doi: 10.2172/1126838.
[25] R. L. D. C. Costa, "Convolutional-LSTM networks and generalization in forecasting of household photovoltaic generation," *Eng. Appl. Artif. Intell.*, vol. 116, Nov. 2022, Art. no. 105458.
[26] B. Jing, Y. Pei, Z. Qian, A. Wang, S. Zhu, and J. An, "Missing wind speed data reconstruction with improved context encoder network," *Energy Rep.*, vol. 8, pp. 3386–3394, Nov. 2022.
[27] R. Tawn, J. Browell, and I. Dinwoodie, "Missing data in wind farm time series: Properties and effect on forecasts," *Electr. Power Syst. Res.*, vol. 189, Dec. 2020, Art. no. 106640.
[28] P. D. Rosero-Montalvo, P. Tözün, and W. Hernandez, "Time-series forecasting to fill missing data in IoT sensor data," *IEEE Sensors Lett.*, vol. 7, no. 9, pp. 1–4, Sep. 2023, doi: 10.1109/LSENS.2023.3307072.
[29] W. Liu, C. Ren, and Y. Xu, "PV generation forecasting with missing input data: A super-resolution perception approach," *IEEE Trans. Sustain. Energy*, vol. 12, no. 2, pp. 1493–1496, Apr. 2021.
[30] W.-C. Kuo, C.-H. Chen, S.-H. Hua, and C.-C. Wang, "Assessment of different deep learning methods of power generation forecasting for solar PV system," *Appl. Sci.*, vol. 12, no. 15, p. 7529, Jul. 2022.
[31] E. Ibanez, I. Krad, B.-M. Hodge, and E. Ela, "Impacts of short-term solar power forecasts in system operations," in *Proc. IEEE/PES Transmiss. Distribution Conf. Expo. (T&D)*, May 2016, pp. 1–5.
[32] X. Xu, L. Xia, Q. Zhang, S. Wu, M. Wu, and H. Liu, "The ability of different imputation methods for missing values in mental measurement questionnaires," *BMC Med. Res. Methodol.*, vol. 20, no. 1, pp. 1–9, Dec. 2020.
[33] Y. Zhang and P. J. Thorburn, "Handling missing data in near real-time environmental monitoring: A system and a review of selected methods," *Future Gener. Comput. Syst.*, vol. 128, pp. 63–72, Mar. 2022.
[34] M. Alhussein, K. Aurangzeb, and S. I. Haider, "Hybrid CNN-LSTM model for short-term individual household load forecasting," *IEEE Access*, vol. 8, pp. 180544–180557, 2020.
[35] Z. Ma, S. Guo, G. Xu, and S. Aziz, "Meta learning-based hybrid ensemble approach for short-term wind speed forecasting," *IEEE Access*, vol. 8, pp. 172859–172868, 2020.
[36] M. Elsaraiti and A. Merabet, "Solar power forecasting using deep learning techniques," *IEEE Access*, vol. 10, pp. 31692–31698, 2022.
[37] D. Lauria, F. Mottola, and D. Proto, "Caputo derivative applied to very short time photovoltaic power forecasting," *Appl. Energy*, vol. 309, Mar. 2022, Art. no. 118452.
[38] S.-V. Oprea and A. Bâra, "A stacked ensemble forecast for photovoltaic power plants combining deterministic and stochastic methods," *Appl. Soft Comput.*, vol. 147, Nov. 2023, Art. no. 110781.
[39] W. El-Baz, P. Tzscheutschler, and U. Wagner, "Day-ahead probabilistic PV generation forecast for buildings energy management systems," *Sol. Energy*, vol. 171, pp. 478–490, Sep. 2018.
[40] Y. Han, N. Wang, M. Ma, H. Zhou, S. Dai, and H. Zhu, "A PV power interval forecasting based on seasonal model and nonparametric estimation algorithm," *Sol. Energy*, vol. 184, pp. 515–526, May 2019.
[41] S.-V. Oprea and A. Bâra, "Ultra-short-term forecasting for photovoltaic power plants and real-time key performance indicators analysis with big data solutions. Two case studies-PV Agigea and PV Giurgiu located in Romania," *Comput. Ind.*, vol. 120, Sep. 2020, Art. no. 103230.
[42] A. Mellit, A. M. Pavan, and V. Lughi, "Deep learning neural networks for short-term photovoltaic power forecasting," *Renew. Energy*, vol. 172, pp. 276–288, Jul. 2021.
[43] R. Zhang, H. Ma, W. Hua, T. K. Saha, and X. Zhou, "Data-driven photovoltaic generation forecasting based on a Bayesian network with spatial–temporal correlation analysis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1635–1644, Mar. 2020.
[44] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021.
[45] H. M. Lynn, S. B. Pan, and P. Kim, "A deep bidirectional GRU network model for biometric electrocardiogram classification based on recurrent neural networks," *IEEE Access*, vol. 7, pp. 145395–145405, 2019.
[46] *Open Meteorological Data Portal*. Accessed: Oct. 26, 2023. [Online]. Available: https://data.kma.go.kr/cmmn/main.do
[47] Korea Meteorological Administration. *KMA—Beginner's Forecast Training Textbook*. Accessed: Oct. 26, 2023. [Online]. Available: https://www.kma.go.kr/down/e-learning/beginning/beginning_04.pdf

DAE-SUNG LEE (Member, IEEE) received the B.S. degree from the Department of Applied Statistics, Gachon University, South Korea, in 2022, where he is currently pursuing the M.S. degree in smart energy system engineering with the Smart Energy System Research Institute (SERI). His research interests include energy big data and forecasting.



SUNG-YONG SON (Member, IEEE) received the B.S. and M.S. degrees from the Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 1990 and 1992, respectively, and the Ph.D. degree in mechanical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2000. From 2000 to 2005, he was with 4DHomeNet and Icross-Technology. He was a Visiting Scholar with the Lawrence Berkley National Laboratory (LBNL), in 2014. He is currently a Professor with the Department of Electrical Engineering, Gachon University, South Korea. His main research interests include smart grids, smart homes, and smart cities.