

Article

Weighted Average Ensemble-Based PV Forecasting in a Limited Environment with Missing Data of PV Power

Dae-Sung Lee  and Sung-Yong Son *

Department of Next Generation Smart Energy System Convergence, Gachon University, Seongnam-si 13120, Republic of Korea; leeds96@gachon.ac.kr

* Correspondence: xtra@gachon.ac.kr

Abstract: Photovoltaic (PV) power is subject to variability, influenced by factors such as meteorological conditions. This variability introduces uncertainties in forecasting, underscoring the necessity for enhanced forecasting models to support the large-scale integration of PV systems. Moreover, the presence of missing data during the model development process significantly impairs model performance. To address this, it is essential to impute missing data from the collected datasets before advancing with model development. Recent advances in imputation methods, including Multi-variate Imputation by Chained Equations (MICEs), K-Nearest Neighbors (KNNs), and Generative Adversarial Imputation Networks (GAINs), have exhibited commendable efficacy. Nonetheless, models derived solely from a single imputation method often exhibit diminished performance under varying weather conditions. Consequently, this study introduces a weighted average ensemble model that combines multiple imputation-based models. This innovative approach adjusts the weights according to “sky status” and evaluates the performance of single-imputation models using criteria such as sky status, root mean square error (RMSE), and mean absolute error (MAE), integrating them into a comprehensive weighted ensemble model. This model demonstrates improved RMSE values, ranging from 74.805 to 74.973, which corresponds to performance enhancements of 3.293–3.799% for KNN and 3.190–4.782% for MICE, thereby affirming its effectiveness in scenarios characterized by missing data.

Keywords: weighted average ensemble; imputation; MICE; KNN; GAIN; CNN-GRU; missing data



Citation: Lee, D.-S.; Son, S.-Y.

Weighted Average Ensemble-Based PV Forecasting in a Limited Environment with Missing Data of PV Power. *Sustainability* **2024**, *16*, 4069. <https://doi.org/10.3390/su16104069>

Academic Editor: Mohamed A. Mohamed

Received: 6 March 2024

Revised: 5 May 2024

Accepted: 9 May 2024

Published: 13 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Global energy demand has escalated significantly due to rapid population growth and technological advancements. This surge in demand has resulted in the heightened use of fossil fuels, which has become a primary contributor to global warming. Consequently, there are concerted global efforts to transition from fossil fuels to renewable energy sources. As of 2023, the global capacity for renewable energy is approximately 3870 GW, reflecting an increase of approximately 13.9% since 2022 [1]. Among renewable energy sources, photovoltaics (PV) are particularly preferred due to their minimal maintenance costs [2] and longer lifespan relative to other renewable options [3]. These benefits have led to PV constituting approximately 37% of the total global renewable energy capacity [1].

However, despite the various advantages of PV systems, difficulties in facility expansion have arisen because of several key problems. PV systems exhibit characteristics that lead to output variability influenced by various factors, including solar radiation, cloud cover, wind speed, module temperature, and ambient temperature [4–6]. The variability in PV output results in forecasting uncertainty and causes an imbalance between the energy supply and demand [5,7]. Therefore, further research is required to enhance PV forecasting for the sustained expansion of PV systems.

Research encompassing approaches, such as ensemble methods, is currently underway to refine PV forecasting. Generally, time-series data such as PV power exhibit irregular variations owing to complex factors such as trends and seasonality over time. These problems

contribute to increased uncertainty and reduced forecasting reliability. Ensemble methods are gaining attention as alternatives to mitigate uncertainty and improve performance by combining individual models that exhibit uncertainty in their forecasting [8]. Akhter, M. N. et al. [9] proposed a hybrid model called slap swarm algorithm-recurrent neural network-long short-term memory (SSA-RNN-LSTM) to predict the output of three different PV systems. The model demonstrated improved performance compared with other hybrid models for polycrystalline, monocrystalline, and thin-film (amorphous silicon) PV systems. Lateko, A. A. et al. [10] introduced a multi-level stacking recurrent neural network (RNN) to enhance PV forecasting models that exhibit varying performances based on weather conditions. The proposed model outperformed the traditional stacking RNN and Random Forest (RF) models under sunny, light-cloudy, cloudy, heavy-cloudy, and rainy weather conditions. AlKandari, M. et al. [11] proposed a machine learning and statistical hybrid model (MLSHM) by combining machine learning (ML) and a theta statistical method to accurately predict the uncertainty in the PV output caused by the variability of weather factors. The MLSHM demonstrated effectiveness compared to traditional single ML models by aggregating ML models using statistical methods to achieve structural diversity. Banik, R. et al. [12] suggested an ensemble approach that combined the results derived from RF and categorical boosting (CatBoost) using Bayesian model averaging (BMA) to accurately predict PV in regions with significant weather variability. Considering the uncertainty of the model through the BMA, this approach showed improved performance compared to traditional ensemble models. Sharma, N. et al. [13] proposed a sequential ensemble model based on the maximal overlap discrete wavelet transform-long short-term memory (MODWT-LSTM), claiming its effectiveness at various resolutions, such as 1 h, 1 d, 10 d, and 1 month, compared with traditional single deep learning (DL) models.

Research related to PV forecasting has generally focused on model refinement. However, in the actual development environment of PV forecasting models, missing measurement data can be easily observed. Variables such as PV power face instances of data omission during the data collection process owing to various factors, such as inverter malfunctions and communication errors [3,14]. Concerning the problem of missing data in PV systems, Zhang, W. et al. [15] claimed that up to 40% of missing data can occur during the PV measurement process. As missing data can hinder the performance of PV forecasting models, the imputation process is crucial prior to model development. Numerous recent studies have addressed the problem of missing data associated with PV.

To impute missing data in PV datasets, Park, S. et al. [16] utilized the K-means clustering method to replace faulty measurements with normal data when anomalies occurred in data collected from small-scale PV systems. The imputation accuracy was enhanced by utilizing measurement data from PV systems installed nearby. Fan, Y. et al. [17] proposed a spatiotemporal denoising graph autoencoder (STD-GAE) framework is proposed to improve the quality of PV time-series data by imputing missing data. The authors claim that their model is robust against the variability of weather factors by considering temporal and spatial coherence. De-Paz-Centeno, I. et al. [18] addressed missing PV data in a constrained environment where external weather factors could not be used. They proposed an encoder-decoder structured artificial neural network (ANN)-based solution and demonstrated superior performance compared with non-parametric imputation methods in situations where predictive parametric models could not be created. Lindig, S. et al. [19] explored various statistical and machine learning-based techniques to address missing data issues in temperature and irradiance metrics, which are essential for PV system performance assessments. They recommended multivariate adaptive regression splines as an effective imputation method when ample training data are available. Kang, M. et al. [20] introduced a cross-modal generative adversarial network (CM-GAN) that combines cross-modal data fusion with a conventional GAN generator, showing enhanced imputation performance in long-term missing data scenarios in PV datasets. Liu, X. et al. [21] proposed a super-resolution perception convolutional neural network (SRPCNN), which combines a naïve SRPCNN with linear imputation to mitigate performance degradation in PV forecasting

models because of missing data. This model has demonstrated superior efficacy over traditional imputation methods under various missing data conditions, such as absent PV output and temperature readings. Bülte, C. et al. [22] proposed a two-stage imputation strategy that accounts for temporal dynamics and cross-dimensional correlations, which was proved effective in learning the distribution of missing time series data in energy contexts such as PV, outperforming conventional methods such as KNN. Guo, G. et al. [23] employed a multivariate least-squares method (MLSM) for managing missing data in real-time control of renewable resources such as PV, addressing data gaps caused by communication network delays through spatial correlations among multiple PV systems.

Additional studies have focused on developing predictive models in contexts where PV-related data is absent. Shireen, T. et al. [24] proposed a multi-task learning for time series (MTL-GP-TS) approach was advanced to boost forecasting accuracy in scenarios with missing data, through information sharing among neighboring PV systems. Kim, T. et al. [25] proposed the impact of missing weather information on PV forecasting performance was analyzed by applying imputation methods such as LI and MICE before developing the forecasting model, wherein the KNN-applied model outperformed other imputation-applied models. Girimurugan, R. et al. [26] proposed an adaptive neural imputation module (ANIM) coupled with a recurrent neural network (RNN) and an attention imputed gate recurrent unit (AIGRU), which effectively utilizes the distribution of each weather variable during the imputation process, surpassing traditional methods like KNN in efficacy. Lastly, Lee, D. S. et al. [27] devised a two-stage approach employing linear regression and KNN to address missing weather data in the development of short- and medium-term PV forecasting models, which yielded enhanced performance compared to models that did not incorporate these imputation techniques.

Previous studies have primarily focused on developing forecasting models and evaluating their performance after imputing missing PV data [24–27]. Our review identified that different forecasting performances emerged when single imputation methods were applied to PV data, influenced by the characteristics of the imputation method and weather conditions. Recognizing a critical limitation, the performance of imputation methods in real model development environments remains unpredictable. To address these gaps, this study introduces an ensemble approach that improves the forecasting performance and innovatively adjusts the weights based on weather conditions during the integration of various imputation-applied models into a unified ensemble model.

The remainder of this paper is organized as follows. In Section 2, we describe the four imputation methods used in the study and the convolutional neural network-gated recurrent unit (CNN-GRU) model employed for forecasting. In addition, we discuss an ensemble model that integrates the developed forecasting models. Section 3 evaluates the imputation performance of each imputation method. Section 4 develops forecasting models using imputed data and evaluates their forecasting performance. Finally, we present our conclusions in Section 5.

2. Methods

The purpose of this study is to enhance PV forecasting models in a limited environment with missing PV power data. The research procedure is illustrated in Figure 1. Initially, data from a PV plant without missing information and weather forecast data provided by a meteorological agency were combined. The combined data were then split into training data for learning and testing data for validation. Subsequently, in the development phase of the forecasting model, random missing data were introduced into the existing PV power. The reason for creating missing data in the complete dataset is to evaluate the performance of each imputation method. In addition, forecasting models were developed using the data generated with each imputation method applied. Finally, based on the proposed method, an ensemble model was developed, and its forecasting performance was compared.

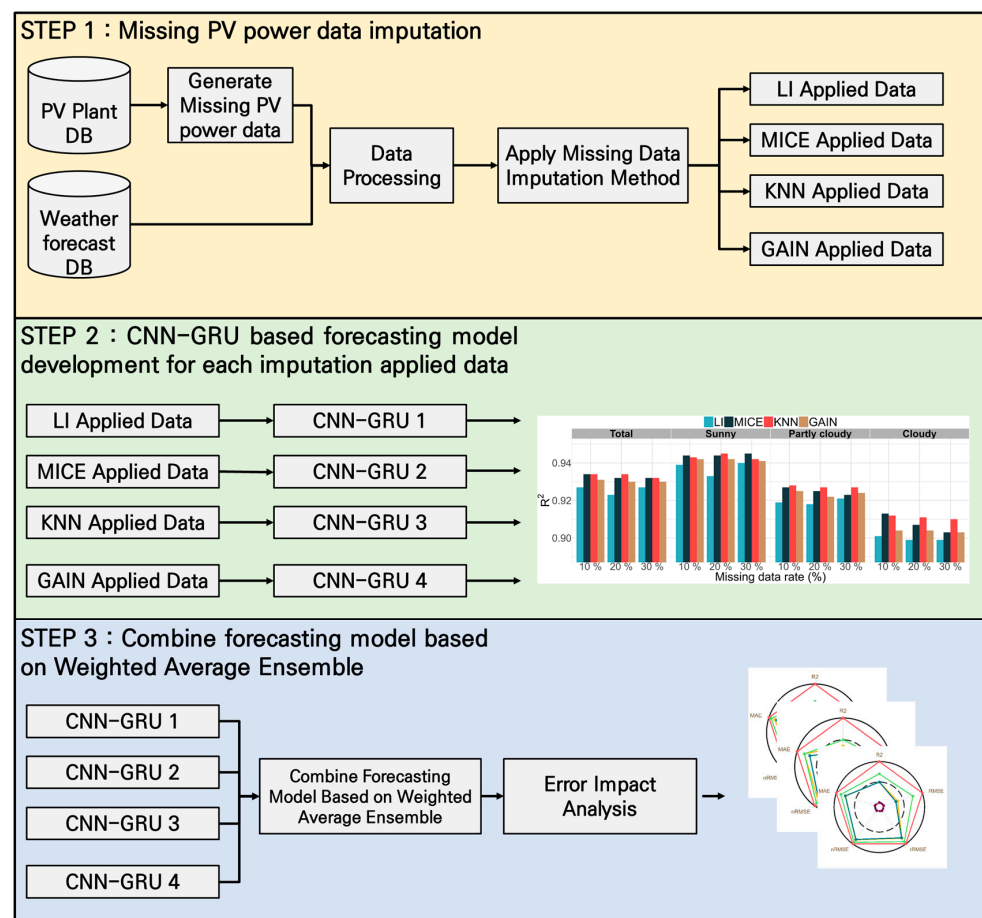


Figure 1. Research methodology.

2.1. Imputation Methods

We have primarily generated multiple datasets by applying imputation methods to enhance the forecasting model in environments with missing PV power. In this study, four methods—linear imputation (LI), multivariate imputation by chained equations (MICE), K-nearest neighbor (KNN), and generative adversarial imputation networks (GAIN)—are employed to impute the missing PV power. Previous research has demonstrated that KNN and MICE are highly effective in addressing missing data issues [22,25,27,28]. GAIN is a method gaining attention in recent imputation-related research, with numerous studies underway to refine GAIN [20,29]. Therefore, to adopt effective methods from existing imputation studies, we have not only used LI, KNN, and GAIN from previous research [28] but also added MICE to increase the number of imputed datasets.

LI is categorized as a deterministic imputation method. When missing data occur, the LI estimates use values in the vicinity of the missing location [30]. LI is widely used owing to its simplicity and speed and requires only two adjacent values for imputation. Although it is commonly used for short-term missing data, a drawback is observed when dealing with long-term missing data, which degrades the imputation performance.

MICE belongs to the category of model-based imputation methods. In contrast to deterministic imputation methods, model-based imputation generates multiple values for missing data and introducing uncertainty [31]. MICE conducts imputation in the following sequence. Initially, temporary imputation is performed for missing data in each variable using methods such as mean values. Subsequently, a regression model is trained by taking one variable as the dependent variable and the remaining variables as independent variables, and imputation was performed. This process was performed iteratively until

convergence was achieved for each variable regression model. MICE offers significant advantages over other imputation methods in terms of model flexibility [32].

KNN, like MICE, is classified as a model-based imputation method, and is one of the most widely used imputation methods across various fields. Unlike MICE, KNN can only impute missing data as univariate variables. KNN operates through the following process. First, the original dataset is divided into a test dataset with missing data and a training dataset without missing data. Subsequently, using distance measurement formulas such as the Euclidean distance and Manhattan distance, the distances between the test and training datasets are calculated. Through this process, the closest neighbors to the missing data can be identified. Scholars can set the parameter ' k ' to determine how many neighbors should be considered for imputation. Model-based imputation methods such as MICE and KNN have the advantage of considering the relationships between various variables during the imputation of missing data [33]. However, their imputation performance depends heavily on the type and characteristics of the data, and when dealing with large datasets, they face limitations owing to the significant runtime cost of applying the model [30].

GAIN is a recently developed generative adversarial network-based imputation method. GAIN comprises three sub-models: generator, discriminator, and hint generator. The GAIN process unfolds as follows. The generator acquires real observed data, imputes missing data, and outputs the completed dataset. The discriminator learns from the completed dataset to discriminate between the real and generated data. The hint generator provides additional information to the discriminator in the form of a hint vector that aids the discrimination process. GAIN has the advantage of being applicable to a broad range of source data types. However, it has the drawback of a complex and challenging learning process that requires substantial time and effort to achieve high accuracy [29].

2.2. Forecasting Methods

2.2.1. CNN-GRU

Recently, research on hybrid models that combine multiple DL models to extract features from data and perform forecasting has been ongoing. The CNN-GRU used in this study is a hybrid model that combines an upper-layer convolutional neural network (CNN) and a lower-layer gated recurrent unit (GRU) into a single model, as illustrated in Figure 2. When a GRU is used alone, there are limitations to properly reflecting the characteristics of the data. Therefore, a hybrid model approach combining a CNN and GRU has recently been adopted [34]. The characteristics of each model are as follows.

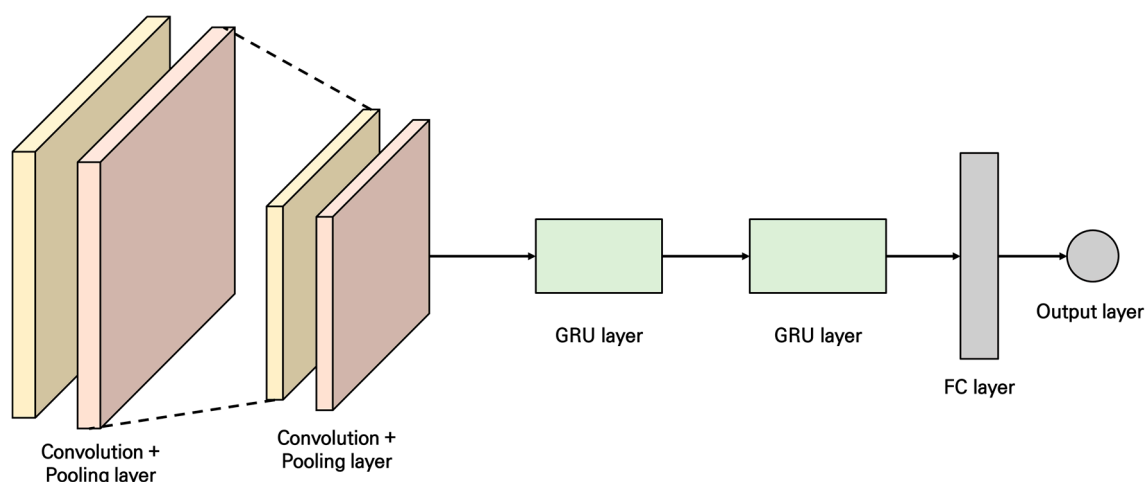


Figure 2. CNN-GRU structure.

CNNs are primarily utilized to extract local features from images [35] and have been increasingly applied to time-series data in recent years. The CNN structure can be classified

into convolutional and pooling layers. The convolution layer employs the convolution operation to filter noise, represented by Equation (1):

$$y_{\text{Conv}} = W * X + b \quad (1)$$

Here, W , $*$, and b , respectively, represent the weight of the convolution layer, the convolution operation, and the bias of the convolution layer.

The extracted features are output in the form of activation maps and utilized as input data for the pooling layer [36]. Although pooling layers exist in various forms, such as max, sum, and average, the max-pooling layers were utilized herein. Thereafter, the input activation map was used to extract the features in the pooling layer, which served as the input variable for the GRU. Equation (2) expresses the formula used for the max pooling layer in the development of the forecasting model.

$$y_{\text{Pool}} = \max(x_{m,n}) \quad (2)$$

A GRU is a model proposed to address the gradient vanishing problem that occurs in a RNN. It stores and learns temporal information regarding important features extracted from the CNN [37]. The structure of a GRU is broadly divided into a reset gate and an update gate, which are controlled through a single-gate controller. The reset gate determines the method for combining new and previous information, while the update gate maintains past information and calculates the new state [38].

The operational dynamics within the GRU layer are expressed in the following equations:

$$z_t = \sigma(w_z[h_{t-1}, x_t]) \quad (3)$$

$$r_t = \sigma(w_r[h_{t-1}, x_t]) \quad (4)$$

$$\tilde{h}_t = \tanh(r_t \times h_{t-1}, x_t) \quad (5)$$

$$h_t = ([1 - z_t] \times h_{t-1}) + z_t \times \tilde{h}_t \quad (6)$$

The information extracted from the GRU layer ultimately passes through a fully connected (FC) layer, deriving the forecasting results. Equations (3)–(6) represent the formulas that compose the GRU layer.

Where z_t and r_t represent the update gate and reset gate, respectively, and w_z and w_r denote the weights of the update gate and reset gate. Additionally, $\tilde{h}_t$ and h_t indicate the output before and the final output from the layer, respectively.

2.2.2. Weighted Average Ensemble

The Ensemble method combines multiple base models into a single model to construct an improved model compared to a single model [39]. Various methods exist within ensemble methods, including RF, bagging, stacking, and boosting. In this study, the weighted average ensemble method, which is a type of bagging method, is utilized. This specific method operates on the principle that not all sub-models contribute equally to forecasting accuracy. To address disparities in performance among sub-models, weights are assigned based on their relative accuracy, thereby enhancing the overall predictive capability of the ensemble.

This approach enables the mitigation of errors from less accurate sub-models through compensation by more accurate ones, generally yielding better overall performance than any individual model could achieve [40]. Additionally, the aggregation inherent in the ensemble approach helps reduce forecasting bias [41]. However, a notable drawback of this method is the potential degradation of overall performance if the constituent sub-models are weak or if the weighting mechanism is not optimally calibrated for combining the sub-models [42].

The proposed weighted average ensemble model is designed to combine each imputation-applied forecasting model. Previous studies on weighted average ensemble models only calculated one weight for each model [12,13]. However, each imputation-applied model yields different performance based on the weather conditions [28]. Therefore, it is necessary to calculate the appropriate weights for each weather condition (Sunny, Partly cloudy, or Cloudy). The formula for calculating the weights by considering the characteristics of these models is shown in Equation (7). Here, N represents the number of evaluation indices used to assess the performance of the forecasting models, and $S_{i,j,k}$ represents the score of the i -th single imputation-applied model evaluated based on the j -th evaluation index under a specific weather condition (k). The score indicates the value based on the ranking of single imputation-applied models evaluated based on a specific evaluation index, where the most superior model has a score of N , and the lowest-performing model has a score of 1.

$$W_{i,k} = \sum_{j=1}^N S_{i,j,k} \quad (7)$$

The predicted results of the integrated ensemble model utilizing the calculated weights are calculated using Equation (8). Here, M represents the number of integrated models, and $\hat{y}_{i,k}$ represents the forecasting result of the i -th model.

$$\hat{y}_{\text{ensemble},k} = \frac{\sum_{i=1}^M \hat{y}_{i,k} \times W_{i,k}}{\sum_{i=1}^M W_{i,k}} \quad (8)$$

2.3. Performance Evaluation Index

In this section, we explain the indices used to evaluate the imputation and forecasting performances. Initially, the coefficient of determination (R^2) was used as an indicator for evaluating imputation performance. Various indices commonly used for evaluating PV forecasting models, such as R^2 , root mean square error (RMSE), relative root mean squared error (rRMSE), normalized root mean squared error (nRMSE), and mean absolute error (MAE), were utilized for the evaluation, as shown in Equations (9)–(13): Here, y_t , $\hat{y}_t$, $\bar{y}_t$, and n represent the measured PV power, predicted PV power, average measured PV power, and number of data points, respectively.

$$R^2 = 1 - \left(\frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y}_t)^2} \right) \quad (9)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \quad (10)$$

$$\text{rRMSE} = \frac{\sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}}{\bar{y}} \quad (11)$$

$$\text{nRMSE} = \frac{\sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}}{y_{\max} - y_{\min}} \quad (12)$$

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (13)$$

3. Imputation of Missing PV Power Data

3.1. Data Description

This study was conducted using data collected from a 1 MW capacity PV plant located in Jeongseon-gun, Gangwon Province, Republic of Korea, spanning from 1 January 2022, to 30 April 2023. The measured variables include the PV power, inclined solar radiation, horizontal solar radiation, module temperature, and outside temperature. To develop a

model that predicts one hour ahead with a 15-min interval, the data were collected every minute and stored in a database, with processing conducted at 15-min intervals. Additionally, the periods before 7 a.m. and after 8 p.m. were excluded from the experiment because the amount of generation was close to or equal to zero during these times. Furthermore, weather forecast data for the location of the PV plant provided by the Korean Meteorological Administration were utilized, focusing solely on the sky status. Sky status is a categorical variable that ranges from 1 to 4, indicating an increasing number of clouds. In this study, the values were transformed, with 1 representing “Sunny”, 4 representing “Cloudy”, and other numbers converted to “Partly cloudy” for subsequent use.

3.2. Generate Missing PV Power Data

To evaluate the performance after imputing the missing PV power data, randomly generated missing data in proportions of 10%, 20%, and 30% were introduced into the normally measured PV power data. This is because evaluating imputation performance becomes infeasible when conducting research using pre-existing missing data. An example of the generated missing PV power is illustrated in Figure 3, where the solid black line represents the measured PV power and the red dots represent the generated missing data. To generalize the experimental results, missing data were generated 5 times for each missing data rate, resulting in 15 datasets for the research.

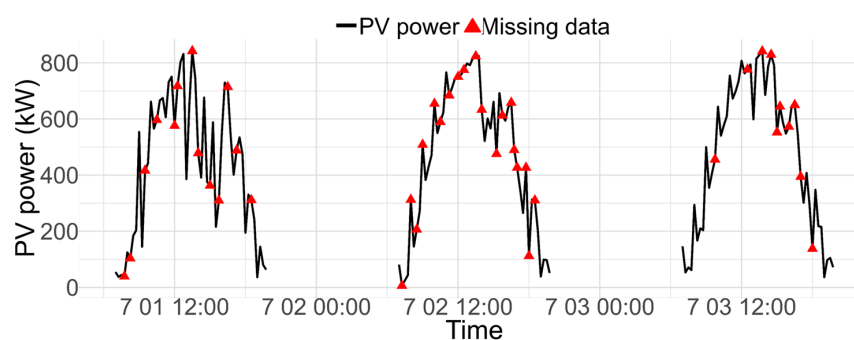


Figure 3. Example of generating missing PV power data: missing data rate of 30%.

Figure 4 illustrates an example visualizing the frequency of occurrence of missing data generated. Frequency analysis was conducted to represent one of the generated cases mentioned above. Missing data that occur at one point indicates a duration of 15 min, and the result was calculated by multiplying 15 min by the number of consecutive occurrences of missing data. In total, 3326 instances of missing data occurred over the course of a year, and multiple points where missing data occurred 2 or more times consecutively accounted for 29.3% of the total. The longest duration of missing data was 105 min, which occurred continuously in 6 instances.

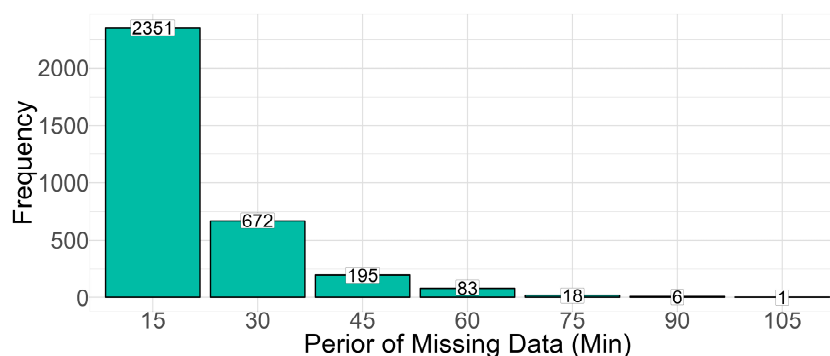


Figure 4. Frequency of occurrence of missing PV power: missing data rate 30%.

3.3. Impute Missing PV Power Data

The imputation methods LI, MICE, KNN, and GAIN were each applied to the simulated missing data. While LI, MICE, and KNN utilized R for implementation, GAIN was executed using Python. Figure 5 displays the imputation process during a period characterized by significant variability in PV power from 1 July 2022 to 3 July 2022. For all methods except LI, additional weather-related variables such as inclined and horizontal solar radiation, module and ambient temperatures, and sky status were employed. These variables were consistently utilized across all imputation methods to ensure coherence in the development of the forecasting model. The measured PV power and the results of the imputation are depicted as a solid line and dots, respectively.

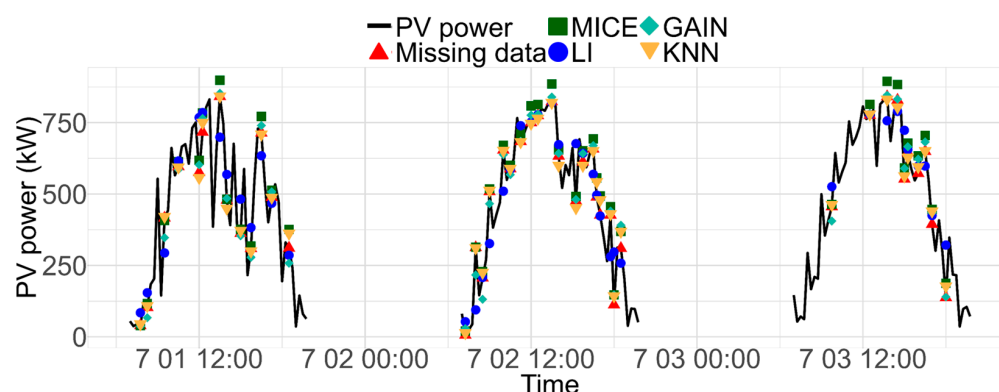


Figure 5. Test results for applying each imputation method: missing data rate 30%.

The evaluation results of the imputation of missing PV power based on the R^2 criterion are presented in Table 1. The evaluation results represent the average imputation performance for each generated case. First, when comparing and analyzing the overall results without distinguishing weather conditions, the LI exhibits a decrease in imputation performance as the missing data rate increases. The LI is effective when short-term missing data occur; however, it exhibits a sharp decline in performance in the cases of long-term missing data. As the missing data rate increases, the probability of long-term missing data also increases, which is presumed to influence the imputation method. Model-based imputation methods, such as MICEs and KNNs, exhibit improved performance compared to other imputation methods. At the highest missing data rate of 30%, LI was 0.865, while the MICE and KNN demonstrated superior performance, exhibiting values of 0.967 and 0.976, respectively. Although the GAIN exhibits better performance than the LI, its performance is lower than that of model-based imputation methods.

Subsequently, the performance of each imputation method based on sky status was compared and analyzed. First, for the commonly used LI, the performance improved by 0.021, 0.015, and 0.013 for “Sunny” compared to “Total” at 10–30%, but it showed a decrease in performance for “Partly cloudy” and “Cloudy.” This phenomenon is presumed to arise from the characteristics of LI. The LI assumes a linear relationship in the data and, therefore, exhibits low performance in conditions with high variability in PV power, such as “Partly cloudy” or “Cloudy.” Other imputation methods, such as KNN, show opposite trends. When comparing KNN, in the 10–30% range, a performance decrease of 0.013, 0.010, and 0.009 for “Sunny” compared to “Total” was observed. Conversely, “Partly cloudy” showed a performance improvement of 0.002, 0.002, and 0.001, and “Cloudy” demonstrated an improvement of 0.010, 0.011, and 0.003. In some cases, the MICE and GAIN exhibited results that contradict those of the KNN but generally showed similar trends. These results suggest that the KNN is not only the superior method compared to other imputation methods, but it is also a robust model against the variability in PV power based on weather conditions.

Table 1. Comparison of imputation performance according to sky status classification: R^2 .

Imputation Method	Missing Data Rate (%)	Sky Status			
		Total	Sunny	Partly Cloudy	Cloudy
LI	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
MICE	10	0.976	0.969	0.981	0.980
	20	0.972	0.969	0.971	0.970
	30	0.967	0.961	0.965	0.972
KNN	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	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

Figure 6 illustrates the variation in imputation performance as the duration of missing data extends. At a missing data rate of 30% and using an R^2 metric, we observed a performance degradation in LI from 0.896 to 0.839 as missing intervals lengthened. Conversely, KNN and the MICE exhibited more stable performances, with KNN ranging from 0.976 to 0.971 and the MICE from 0.960 to 0.964, suggesting no significant declining trend. Thus, in this case, KNN and the MICE remained minimally affected by the duration of missing data.

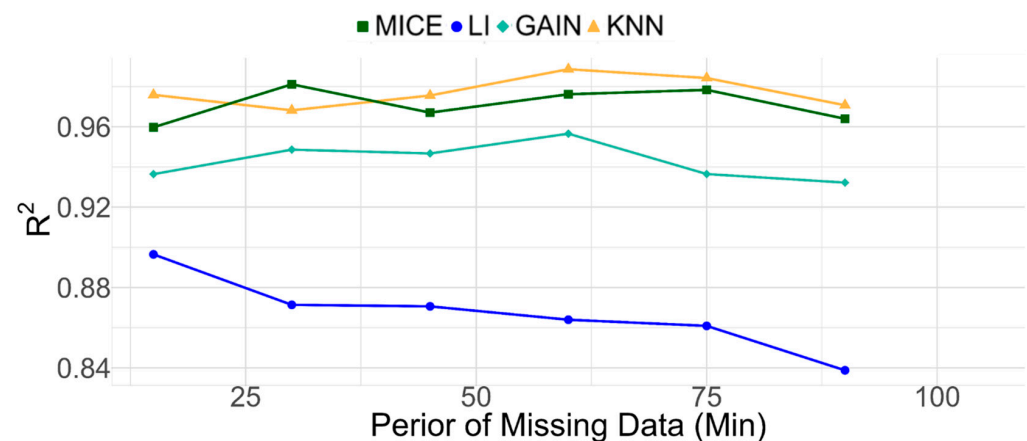
**Figure 6.** Trend in imputation performance as period of missing data increases: missing data rate 30%, R^2 .

Table 2 shows the time required to apply each imputation method. The average time consumed for LI, MICE, KNN, and GAIN are 0.02, 1.16, 0.03, and 26.65 s, respectively. As mentioned in Section 2.1, LI demonstrates the shortest consumption time among the imputation methods. When comparing the model-based methods KNN and MICE, KNN requires about 1.13 s less. The MICE, executing until the regression model converges iteratively, takes more time than the KNN. The GAIN consumes significantly more time than the other methods because it involves repeated interactions between the generator and discriminator to generate suitable imputed data. The ensemble model generally requires the sum of individual methods, but the time consumed for the GAIN was dominant in the case study.

Table 2. Time taken to impute missing PV data: s.

Missing Data Rate (%)	Imputation Method			
	LI	MICE	KNN	GAIN
10	0.02	1.13	0.03	30.88
20	0.01	1.40	0.03	23.76
30	0.02	0.94	0.04	25.30
Average	0.02	1.16	0.03	26.65

4. Developing a PV Forecasting Model Using Imputation Method Applied Data

4.1. Developed a PV Power Forecasting Model Applying a Single Imputation

We developed a model to predict the PV power for the next hour at 15-min intervals using data processed with the imputation method. The model was trained using data measured from 1 January 2022, to 31 December 2022, and evaluated using data from 1 January 2023, to 30 April 2023. Additionally, after applying min–max scaling to all the variables, they were input into the model. Table 3 presents the structures of the CNN-GRU-based PV forecasting model. The convolutional layer extracts the temporal features of multivariate input variables and passes them to the GRU layer. The transferred features are modeled in the GRU layer, and the PV power is predicted through the output layer. Furthermore, MAE and Adam were adopted as the loss function and optimizer, respectively. To compare and analyze all forecasting models in the same environment, layers were consistently applied to all datasets during development.

Table 3. Layers configuration of proposed forecasting model.

Proposed Model		
Conv1D	Filters	32
	Activation	ELU
Conv1D	Filters	64
	Activation	ELU
Conv1D	Filters	128
	Activation	ELU
Conv1D	Filters	256
	Activation	ELU
GRU	Hidden node	128
	Activation	ELU
GRU	Hidden node	64
	Activation	ELU
Output	Hidden node	1
	Activation	ELU

We developed a forecasting model using datasets from each imputation method. The models were developed and evaluated for each case, and the average forecasting performance for each case was calculated. The performances of the developed models were compared using the performance evaluation indices described in Section 2.

Figure 7a presents a visualization of the results based on R^2 . When comparing situations with a 30% missing data rate, the models with LI, MICE, KNN, and GAIN applied show R^2 values of 0.927, 0.932, 0.932, and 0.930, respectively, based on the “Total” criterion. Models with high imputation performances, particularly those with KNN and MICE, demonstrated improved performances. However, when comparing and analyzing

weather conditions in detail, models with MICE, KNN, and KNN applied showed higher performance in “Sunny”, “Partly cloudy”, and “Cloudy”, respectively.

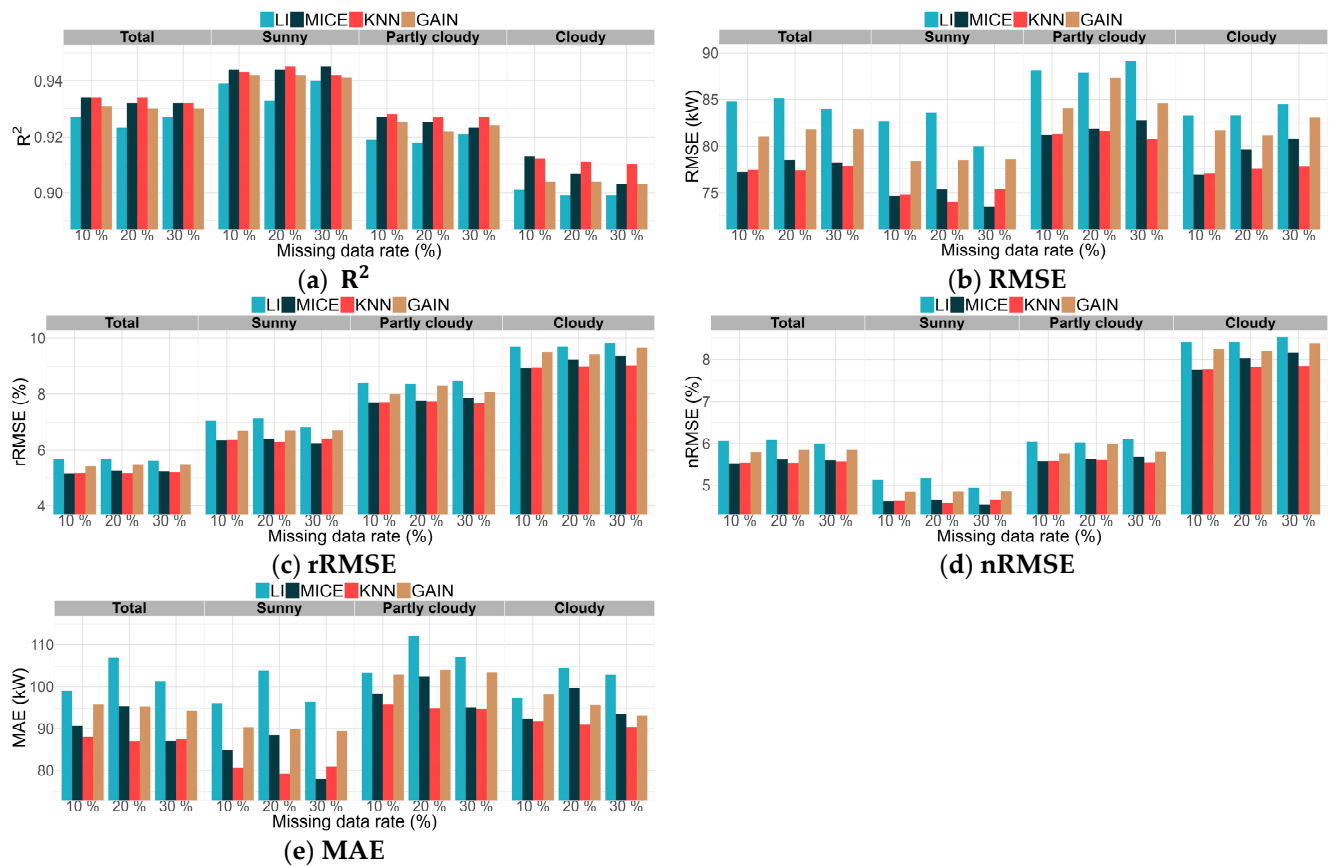


Figure 7. Comparison of power forecasting performance for different imputation methods.

Figure 7b–d illustrates the forecasting models using the RMSE, rRMSE, and nRMSE. In examining the three indicators, MICE, KNN, and the model with KNN demonstrated superior performances in 10–30%. When comparing at the 30% criterion, the overall performance favors the KNN-applied model; however, restricting the weather condition to “Sunny” surprisingly showed that the MICE-applied model exhibited superior performance. As an example, comparing two imputation-applied models based on RMSE, for the “Total” case, KNN and MICE show values of 78.218 and 77.821, respectively, while for the “Sunny” case, the values are 73.460 and 75.369.

Figure 7e illustrates the results visualized based on MAE. In the 10–30%, KNN, KNN, and MICE demonstrated improved performance. When comparing the MICE-applied model and the KNN-applied model, which showed superior performance in other indices, at the 30% criterion, in “Sunny”, “Partly cloudy”, and “Cloudy”, MICE exhibited values of 77.988, 95.173, and 93.575, respectively, while KNN showed values of 80.914, 94.760, and 90.343, respectively. In situations with low variability in the PV power, the MICE-applied models performed better, whereas in situations with high variability, the KNN-applied models exhibited superior performance. This implies that, even if the overall forecasting performance of the imputation-applied model is excellent, it may not be optimal when considering weather conditions.

Following a comprehensive analysis of multiple indices, the forecasting model with KNN, which showed improved imputation performance, generally demonstrated superior performance. However, despite the high overall forecasting performance, cases in which forecasting models with other imputation methods exhibited superior performance are frequent when evaluating the model considering weather conditions. The result suggests

that the deviation caused by the imputation between the imputed and measured data may have influenced the model learning. In other words, developing a forecasting model using a single imputation method for missing data may not be optimal.

4.2. Combine PV Forecasting Model Based on Weighted Average Ensemble

Using the forecasting performance results of the single imputation-applied models, weights were calculated based on the sky status and the model and integrated into the proposed Ensemble model. Subsequently, a general weighted average ensemble model was developed to evaluate the performance of the proposed model, and a Diebold and Mariano (DM) test [43] was conducted to compare the two developed ensemble models. The null hypothesis of the DM test indicates that there is no statistically significant difference in performance between the proposed model and the general ensemble model, whereas the alternative hypothesis suggests that the proposed model exhibits statistically superior performance compared to the general ensemble model. The interpretation of the results according to the test is as follows: if the p -value is greater than 0.05, the null hypothesis is accepted, whereas if the p -value is less than 0.05, the null hypothesis is rejected in favor of the alternative hypothesis. The DM test results are listed in Table 4. The p -values at 10–30% are 0.001, 0.778, and 0.003, respectively. At 10% and 30%, the proposed model demonstrates superior performance compared with the conventional ensemble model, whereas at 20%, there was no statistically significant performance difference between the two models. This test suggests that the proposed model exhibits similar or improved predictive performance compared to existing methods.

Table 4. DM test results.

Missing Data Rate (%)	Model 1	Model 2	p -Value
10	Proposed	Ensemble	0.001
20	Proposed	Ensemble	0.778
30	Proposed	Ensemble	0.003

Figure 8 exemplifies the comparative effectiveness of the proposed model against established KNN and MICE models, particularly under a scenario with a 30% missing data rate. This visualization highlights how the proposed model more accurately approximates actual PV power outputs during peak hours (12–13 h) and under challenging weather conditions, such as those on 7 February 2023. Notably, the proposed model significantly reduces over-predictions commonly observed in traditional models, showcasing its robustness and reliability in handling missing data efficiently.

Figure 9 showcases a comparative analysis between conventional single imputation models and our proposed Ensemble model. To evaluate the comparative performance across different forecasting models, five statistical indices including R^2 were employed. Given the disparate scales of these indices, min–max scaling was applied to each for normalization prior to analysis. The results illustrate that the Ensemble model consistently outperforms others, indicated by its smaller area on the comparison graph, suggesting superior model efficacy. Specifically, the LI-applied model consistently showed the weakest performance. Detailed RMSE comparisons at missing data rates of 10%, 20%, and 30% further validate the robustness of our model, with respective improvements of 3.293–3.799% for KNN and 3.190–4.782% for MICE, showcasing significant reductions in forecasting error. Therefore, our proposed method is effective in enhancing the performance of PV forecasting models in scenarios with missing data.

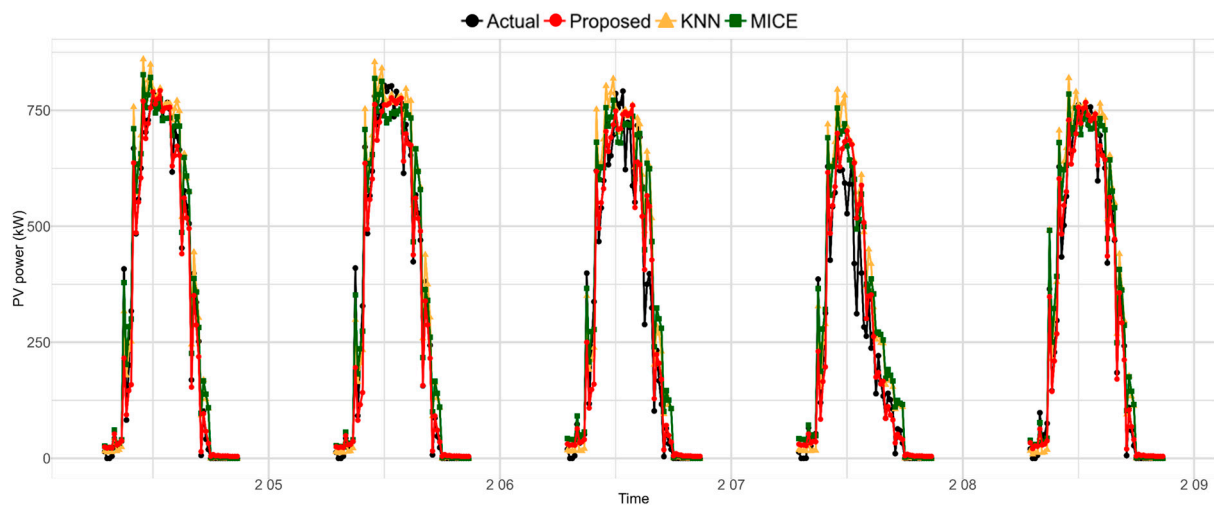


Figure 8. PV power forecasting result example. missing data rate: 30%.

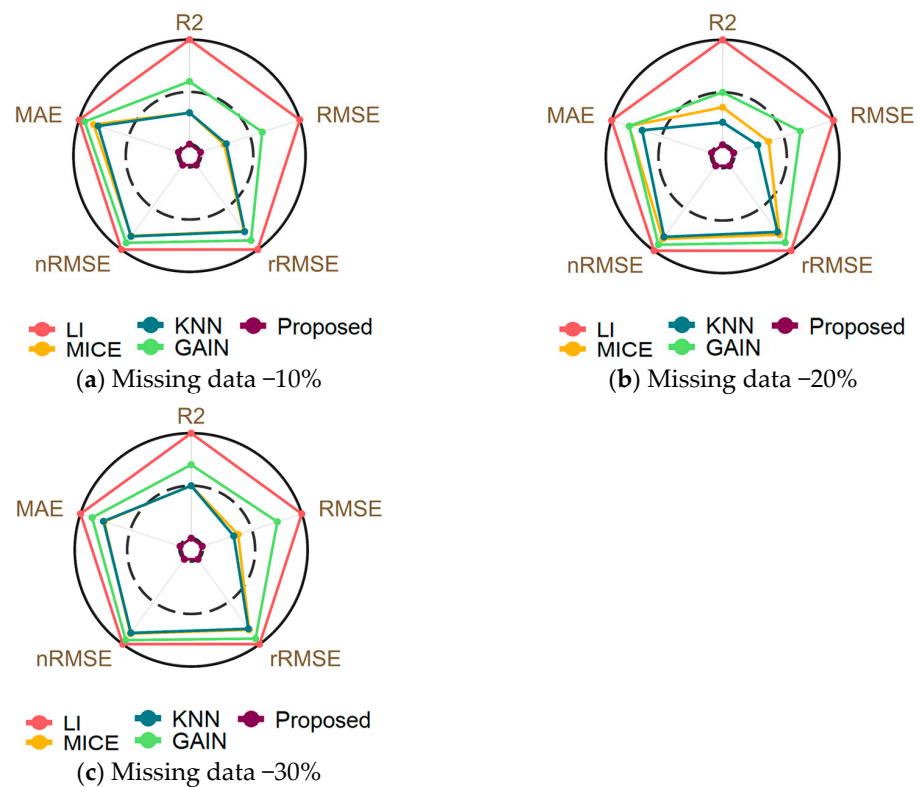


Figure 9. Comparison of forecasting errors for each imputation-applied forecasting model.

Table 5 presents the results of a detailed comparison based on the sky status, showing the models that demonstrate superior performance for each sky status and index. Generally, the ensemble model exhibits improved performance compared with conventional single-imputation models. These results suggest that the proposed model may be effective for environments with missing data.

Table 5. Preferred imputation method for different missing data rates and forecasting performance.

Index	Sunny			Partly Cloudy			Cloudy			Preferred Method
Missing Data Rate	10%	20%	30%	10%	20%	30%	10%	20%	30%	
R^2	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed
RMSE	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed
rRMSE	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed
nRMSE	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed
MAE	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed
Preferred method	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed	Proposed

5. Conclusions

This study proposed a method to enhance the prediction performance in environments with limited PV power data owing to missing data. Imputations for the missing PV power data were conducted using LI, MICE, KNN, and GAIN. A comparative analysis of the imputation results revealed that model-based imputation methods, specifically MICE and KNN, exhibited superior performance compared with other imputation methods. However, GAIN, which has gained attention in recent imputation-related research, showed unsatisfactory performance in imputing the PV power. These results suggest that model-based methods are more suitable for imputing time-series data. Additionally, when developing a PV power forecasting model using a single imputation method, there were instances where a lower forecasting performance than expected was observed for certain sky statuses, based on the characteristics and performance of the imputation. To address the limitations of single-imputation models, we performed model integration using a weighted average ensemble. The proposed ensemble model demonstrated favorable evaluations across most indices in the 10–30%.

The proposed model shows improved performance relative to single-imputation models and allows for easy integration with other forecasting models by accommodating the performance ranking of each index based on sky status. This model is advantageous in scenarios with missing data and can be updated seamlessly by recalculating weights as the performance of sub-models varies. When comparing the time required for imputation, GAIN consumed significantly more time than other imputation methods. However, the time taken to output forecasting results using the proposed model was less than 0.01 s. These results signify that the computational time for the proposed model has minimal impact on practical implementations.

Our study has the following limitations:

First, the dependence on weather forecasts, which may contain missing data, poses a challenge in evaluating these weights. Consequently, in PV forecasting, the absence of key weather data can significantly impair model performance.

Second, this study addressed the missing data phenomena during the forecasting model development phase. However, in practical implementation, problems can arise from missing data present at the time of forecasting beyond the development of the forecasting model itself.

Third, our research adopted an approach that combines all sub-models. However, the performance of the ensemble model varies depending on the combination of sub-models. Therefore, identifying the combination that delivers the best performance among all sub-models would be one of the crucial steps in developing an ensemble model.

Further research is thus warranted to strengthen the robustness of forecasting models against missing data in essential inputs such as weather forecasts and PV power.

Author Contributions: Conceptualization, S.-Y.S.; methodology, S.-Y.S. and D.-S.L.; software, D.-S.L.; validation, D.-S.L.; investigation, D.-S.L.; data curation, D.-S.L.; writing—original draft preparation,

D.-S.L.; writing—review and editing, S.-Y.S.; supervision, S.-Y.S.; funding acquisition, S.-Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the MSIT (Ministry of Science and ICT), Republic of Korea, under the ITRC (Information Technology Research Center) support program (RS-2023-00259004) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and by the Ministry of Trade, Industry and Energy, under the STEP (Strategy for Transition to next generation Energy Portfolio) support program (No.20214000000060) supervised by the KETEP (Korea Institute of Energy Technology Evaluation and Planning).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset available on request from the authors.

Acknowledgments: The authors gratefully acknowledge the financial support of the project Data-driven Energy System Innovation Research Center and Department of Next Generation Energy System Convergence based on Techno Economics.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- IRENA. Renewable Capacity Statistics 2024. Available online: <https://www.irena.org/Publications/2024/Mar/Renewable-capacity-statistics-2024> (accessed on 9 April 2024).
- Iheanetu, K.J. Solar Photovoltaic Power Forecasting: A Review. *Sustainability* **2022**, *14*, 17005. [CrossRef]
- Akhter, M.N.; Mekhilef, S.; Mokhlis, H.; Almohaimeed, Z.M.; Muhammad, M.A.; Khairuddin, A.S.M.; Akram, R.; Hussain, M.M. An Hour-Ahead PV Power Forecasting Method Based on an RNN-LSTM Model for Three Different PV Plants. *Energies* **2022**, *15*, 2243. [CrossRef]
- Mohamad Radzi, P.N.L.; Akhter, M.N.; Mekhilef, S.; Mohamed Shah, N. Review on the Application of Photovoltaic Forecasting Using Machine Learning for Very Short-to Long-Term Forecasting. *Sustainability* **2023**, *15*, 2942. [CrossRef]
- Zhao, J.; Guo, Z.H.; Su, Z.Y.; Zhao, Z.Y.; Xiao, X.; Liu, F. An improved multi-step forecasting model based on WRF ensembles and creative fuzzy systems for wind speed. *Appl. Energy* **2016**, *162*, 808–826. [CrossRef]
- Lee, C.H.; Yang, H.C.; Ye, G.B. Predicting the performance of solar power generation using deep learning methods. *Appl. Sci.* **2021**, *11*, 6887. [CrossRef]
- Udawalpola, R.; Masuta, T.; Yoshioka, T.; Takahashi, K.; Ohtake, H. Reduction of power imbalances using battery energy storage system in a bulk power system with extremely large photovoltaics interactions. *Energies* **2021**, *14*, 522. [CrossRef]
- Wang, X.; Hyndman, R.J.; Li, F.; Kang, Y. Forecast combinations: An over 50-year review. *Int. J. Forecast.* **2023**, *39*, 1518–1547. [CrossRef]
- Akhter, M.N.; Mekhilef, S.; Mokhlis, H.; Ali, R.; Usama, M.; Muhammad, M.A.; Khairuddin, A.S.M. A hybrid deep learning method for an hour ahead power output forecasting of three different photovoltaic systems. *Appl. Energy* **2022**, *307*, 118185. [CrossRef]
- Lateko, A.A.; Yang, H.T.; Huang, C.M. Short-term PV power forecasting using a regression-based ensemble method. *Energies* **2022**, *15*, 4171. [CrossRef]
- AlKandari, M.; Ahmad, I. Solar power generation forecasting using ensemble approach based on deep learning and statistical methods. *Appl. Comput. Inform.* **2020**. [CrossRef]
- Banik, R.; Biswas, A. Improving Solar PV Prediction Performance with RF-CatBoost Ensemble: A Robust and Complementary Approach. *Renew. Energy Focus* **2023**, *46*, 207–221. [CrossRef]
- Sharma, N.; Mangla, M.; Yadav, S.; Goyal, N.; Singh, A.; Verma, S.; Saber, T. A sequential ensemble model for photovoltaic power forecasting. *Comput. Electr. Eng.* **2021**, *96*, 107484. [CrossRef]
- Choi, J.; Lee, I.W.; Cha, S.W. Analysis of data errors in the solar photovoltaic monitoring system database: An overview of nationwide power plants in Korea. *Renew. Sustain. Energy Rev.* **2022**, *156*, 112007. [CrossRef]
- Zhang, W.; Luo, Y.; Zhang, Y.; Srinivasan, D. SolarGAN: Multivariate solar data imputation using generative adversarial network. *IEEE Trans. Sustain. Energy* **2020**, *12*, 743–746. [CrossRef]
- Park, S.; Park, S.; Kim, M.; Hwang, E. Clustering-based self-imputation of unlabeled fault data in a fleet of photovoltaic generation systems. *Energies* **2020**, *13*, 737. [CrossRef]
- Fan, Y.; Yu, X.; Wieser, R.; Meakin, D.; Shaton, A.; Jaubert, J.N.; Flottesmesch, R.; Howell, M.; Braid, J.; Bruckman, L.S.; et al. Spatio-Temporal Denoising Graph Autoencoders with Data Augmentation for Photovoltaic Timeseries Data Imputation. *arXiv* **2023**, arXiv:2302.10860.
- de-Paz-Centeno, I.; García-Ordás, M.T.; García-Olalla, Ó.; Alaiz-Moretón, H. Imputation of missing measurements in PV production data within constrained environments. *Expert Syst. Appl.* **2023**, *217*, 119510. [CrossRef]

19. Lindig, S.; Louwen, A.; Moser, D.; Topic, M. Outdoor PV system monitoring—Input data quality, data imputation and filtering approaches. *Energies* **2020**, *13*, 5099. [\[CrossRef\]](#)
20. Kang, M.; Zhu, R.; Chen, D.; Liu, X.; Yu, W. CM-GAN: A Cross-Modal Generative Adversarial Network for Imputing Completely Missing Data in Digital Industry. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**. [\[CrossRef\]](#)
21. Liu, X.; Huang, C.; Wang, L.; Luo, X. Improved super-resolution perception convolutional neural network for photovoltaics missing data recovery. *Energy Rep.* **2023**, *9*, 388–395. [\[CrossRef\]](#)
22. Bülte, C.; Kleinebrahm, M.; Yilmaz, H.Ü.; Gómez-Romero, J. Multivariate time series imputation for energy data using neural networks. *Energy AI* **2023**, *13*, 100239. [\[CrossRef\]](#)
23. Guo, G.; Zhang, M.; Gong, Y.; Xu, Q. Safe multi-agent deep reinforcement learning for real-time decentralized control of inverter based renewable energy resources considering communication delay. *Appl. Energy* **2023**, *349*, 121648. [\[CrossRef\]](#)
24. Shireen, T.; Shao, C.; Wang, H.; Li, J.; Zhang, X.; Li, M. Iterative multi-task learning for time-series modeling of solar panel PV outputs. *Appl. Energy* **2018**, *212*, 654–662. [\[CrossRef\]](#)
25. Kim, T.; Ko, W.; Kim, J. Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting. *Appl. Sci.* **2019**, *9*, 204. [\[CrossRef\]](#)
26. Girimurugan, R.; Selvaraju, P.; Jeevanandam, P.; Vadivukarassi, M.; Subhashini, S.; Selvam, N.; Hasane, A.S.K.; Mayakannan, S.; Vaithilingam, S.K. Application of Deep Learning to the Prediction of Solar Irradiance through Missing Data. *Int. J. Photoenergy* **2023**, *2023*, 4717110. [\[CrossRef\]](#)
27. Lee, D.S.; Lai, C.W.; Fu, S.K. A Short-and Medium-Term Forecasting Model for Roof PV Systems with Data Pre-Processing. *Heliyon* **2024**, *10*, e27752. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Lee, D.S.; Son, S.Y. PV forecasting model development and impact assessment via imputation of missing PV power data. *IEEE Access* **2024**, *12*, 12843–12852. [\[CrossRef\]](#)
29. Zhang, Y.; Zhang, R.; Zhao, B. A systematic review of generative adversarial imputation network in missing data imputation. *Neural Comput. Appl.* **2023**, *35*, 19685–19705. [\[CrossRef\]](#)
30. Zhang, Y.; Thorburn, P.J. Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. *Future Gener. Comput. Syst.* **2022**, *128*, 63–72. [\[CrossRef\]](#)
31. Aleryani, A.; Bostrom, A.; Wang, W.; Iglesia, B. Multiple Imputation Ensembles for Time Series (MIE-TS). *ACM Trans. Knowl. Discov. Data* **2023**, *17*, 1–28. [\[CrossRef\]](#)
32. Azur, M.J.; Stuart, E.A.; Frangakis, C.; Leaf, P.J. Multiple imputation by chained equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res.* **2011**, *20*, 40–49. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Žliobaitė, I.; Hollmén, J. Optimizing regression models for data streams with missing values. *Mach. Learn.* **2015**, *99*, 47–73. [\[CrossRef\]](#)
34. Chiu, M.C.; Hsu, H.W.; Chen, K.S.; Wen, C.Y. A hybrid CNN-GRU based probabilistic model for load forecasting from individual household to commercial building. *Energy Rep.* **2023**, *9*, 94–105. [\[CrossRef\]](#)
35. Tovar, M.; Robles, M.; Rashid, F. PV power prediction, using CNN-LSTM hybrid neural network model. Case of study: Temixco-Morelos, México. *Energies* **2020**, *13*, 6512. [\[CrossRef\]](#)
36. Alhussein, M.; Aurangzeb, K.; Haider, S.I. Hybrid CNN-LSTM model for short-term individual household load forecasting. *IEEE Access* **2020**, *8*, 180544–180557. [\[CrossRef\]](#)
37. Sajjad, M.; Khan, Z.A.; Ullah, A.; Hussain, T.; Ullah, W.; Lee, M.Y.; Baik, S.W. A novel CNN-GRU-based hybrid approach for short-term residential load forecasting. *IEEE Access* **2020**, *8*, 143759–143768. [\[CrossRef\]](#)
38. Lynn, H.M.; Pan, S.B.; Kim, P. A deep bidirectional GRU network model for biometric electrocardiogram classification based on recurrent neural networks. *IEEE Access* **2019**, *7*, 145395–145405. [\[CrossRef\]](#)
39. Kumar, V.; Ayday, P.S.S.; Minz, S. Multi-view ensemble learning using multi-objective particle swarm optimization for high dimensional data classification. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 8523–8537. [\[CrossRef\]](#)
40. Anand, V.; Gupta, S.; Gupta, D.; Gulzar, Y.; Xin, Q.; Juneja, S.; Shah, A.; Shaikh, A. Weighted average ensemble deep learning model for stratification of brain tumor in MRI images. *Diagnostics* **2023**, *13*, 1320. [\[CrossRef\]](#)
41. Shahhosseini, M.; Hu, G.; Pham, H. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Mach. Learn. Appl.* **2022**, *7*, 100251. [\[CrossRef\]](#)
42. Baradaran, R.; Amirkhani, H. Zero-shot estimation of base models' weights in ensemble of machine reading comprehension systems for robust generalization. In Proceedings of the 26th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran, 3–4 March 2021; pp. 1–5.
43. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.