IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ARTIFICIAL INTELLIGENCE TECHNOLOGIES FOR ELECTRIC POWER SYSTEMS

Received March 7, 2020, accepted March 14, 2020, date of publication March 18, 2020, date of current version March 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981819

Feature-Selective Ensemble Learning-Based Long-Term Regional PV Generation Forecasting

HANEUL EOM, YONGJU SON, AND SUNGYUN CHOI, (Member, IEEE)
School of Electrical Engineering, Korea University, Seoul 02841, South Korea

Corresponding author: Sungyun Choi (sungyun.choi@ieee.org)

The work was supported in part by the Korea University Grant, and in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) under Grant 2019R1F1A1064164.

ABSTRACT Because of Korea's rapid expansion in photovoltaic (PV) generation, forecasting long-term PV generation is of prime importance for utilities to establish transmission and distribution planning. However, most previous studies focused on long-term PV forecasting have been based on parametric methodologies, and most machine learning-based approaches have focused on short-term forecasting. In addition, many factors can affect local PV production, but proper feature selection is needed to prevent overfitting and multicollinearity. In this study, we perform feature-selective long-term PV power generation predictions based on an ensemble model that combines machine learning methods and traditional time-series predictions. We provide a framework for performing feature selection through correlation analysis and backward elimination, along with an ensemble prediction methodology based on feature selection. Utilities gather predictions from various sources and need to consider them to make accurate forecasts. Our ensemble method can produce accurate predictions using various prediction sources. The model with applied feature selection shows higher predictive power than other models that use arbitrary features, and the proposed feature-selective ensemble model based on a convolutional neural network shows the best predictive power.

INDEX TERMS Ensemble learning, forecasting, long-term forecast, machine learning, power system planning.

I. INTRODUCTION

Korea is pursuing conversion to an energy mix with a photovoltaic (PV) focus and plans to increase PV power generation by about 5.7 times in the coming years, from 5,835 MW in 2017 to 33,530 MW in 2030 [1]. As a result, it is anticipated the large fluctuations in PV power generation will cause difficulties in utilities' grid management [2]; for example, PV curtailment will increase because of steep load ramping. In this sense, utilities' proactive response to long-term PV installation will be of primary importance [3]. A fundamental tool to handle PV fluctuation is accurate forecasting of PV generation, and many researchers have worked on renewable forecasting.

A large portion of PV power generation prediction studies target PVs that have already been installed and their production, and the input variables used for forecasting are generally limited to weather and equipment-related values [4]–[6]. However, from the standpoint of utility planning, existing forecasting methods have limitations in the prediction of future PV installations, which would significantly affect PV fluctuations [7]. Although weather and equipment variables are essential factors in power generation forecasting for installed PVs, various other factors, such as economic indices and policies, determine the prediction of PVs that will be newly installed [8]. The authors of [9] investigated the expansion of new PV facilities. The study attempted to predict PV generation in Shanghai, China, through analysis of Granger causality and logistic functions with policy and economic factors. However, Granger causality analysis does not consider nonlinear relationships, and logistic functions should assume any parametric probability distribution.

Many statistical methods and models have been used for the renewable energy forecasting. Among them, the autoregressive integrated moving average (ARIMA) model depends on time-series data analysis, assuming that prior knowledge or experience influences future trends. In [10], a multi-period prediction method based on 1 hour of solar radiation data is used to determine the best period for the model. A hybrid model combining ARIMA and the wavelet

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

transform technique was used in [11] for day-ahead forecasting, and empirical mode decomposition and the ARIMA model were used in [12]. Work in [13] described a 24-hour PV forecasting model at the aggregated system level by comparing the stability of ARIMA, radial basis function neural network (RBFNN), and least-squares support vector machine (LSSVM) with a persistence model.

The vector autoregression (VAR) model is a general form of a univariate autoregressive model that predicts time-series vectors. Unlike other autoregression (AR) models, VAR focuses on finding correlation among variables using impact analysis focused on how dynamically a change in a variable affects endogenous variables. Furthermore, the relative impact on each endogenous variable that contributes to the overall change can be determined by the variance. The authors of [14] applied VAR in a multienergy system load and found it presented higher accuracy than a single-variable prediction method. In [15], multivariate wind data, including wind direction, wind speed, and wind farm layout, improved forecasting performance when compared with univariate autoregressive prediction models. Researchers in [16] combined the VAR model with the least absolute shrinkage and selection operator (LASSO) framework, thereby presenting better prediction performance than the conventional AR model.

In addition to statistical forecasting methods, many efforts have been made to use the machine learning approach for renewable forecasting, including deep learning algorithms. Among the algorithms, long short-term memory (LSTM) is a type of recurrent neural network (RNN) algorithm that has a chain structure and thus can deal with sequential data. Although RNN has a vanishing gradient problem that makes it hard to use when solving long-term dependency problems, LSTM, which has acknowledged performance in the processing of series data, is designed to avoid this problem [17]. Therefore, LSTM has been used in diverse research areas, including renewable forecasting. For example, the authors of [18] used LSTM with the k-means method to differentiate cloudy days from sunny days and enhance the forecast accuracy on cloudy days; the accuracy of typical forecasting models is low in cloudy days. The study in [19] proposed an LSTM model combined with principal-component analysis (PCA) to reduce training time for PV forecasting and to prevent overfitting.

This paper is organized as follows: Section II explains the related works of the proposed work, followed by the description of the proposed methodology in Section III. Then, Section IV describes variable selection and data acquisition, and experimental setup is depicted in Section V. Then, Section VI presents simulation results, and the paper is finally concluded in Section VII.

II. RELATED WORKS

Various forecasting methods based on statistics and machine learning have been studied, but any single method has its limitations. The traditional time-series models, such as ARIMA and VAR, are difficult fit into data that include nonlinear correlations. In contrast, the machine learning technique has the advantage of reducing the bias of variable selection by grasping nonlinear relationships among variables, and a deep learning model, like LSTM, can automatically derive characteristics from the data. However, difficulty remains in obtaining enough data, because the data frequency of macroeconomic indicators or policy-related variables is low; therefore, the learning is not great enough to be effective.

To overcome these limitations, this study proposes an ensemble model that combines traditional time-series models and machine learning-based models. The rationale for using the ensemble model is that each prediction model has an essential property. For example, although most persistence models have high accuracy in short-period prediction, the average model presents flat accuracy over the entire prediction horizon. Hence, the proper combination of various models can contain the best accuracy section of each model.

Several studies have focused on the ensemble method [20]–[22]. Ensemble prediction based on the Gaussian process and neural networks for short-term wind power forecasting was proposed in [20], and the authors of [21] proposed a combination model of deep belief network, autoencoder, and LSTM, which improved renewable energy forecasts compared with single usage of physical models or machine learning models. In [22], the authors proposed a framework for hybrid ensemble deep learning that used two LSTM models to perform short-term prediction of PV power generation.

However, these studies mainly focused on short-term forecasting, with only a few considerations related to deep learning methods for PV planning. When it comes to long-term PV forecasting, utilities must deal with diverse factors. To be specific, the ARIMA and VAR models are adequate for forecasting near-future PV generation, and the LSTM model has a substantial advantage in cases with various training data.

In this paper, three representative meta-learner models (ARIMA, VAR, and LSTM) are chosen to form the ensemble. Our goal is to make accurate predictions based on predictions from various sources, because utilities gather and need to consider predictions from various sources to forecast accurately. In addition, we use the simple average method, multilayer perceptron (MLP) method, and convolutional neural network (CNN) method to form an ensemble algorithm. MLP and CNN are neural network-based methodologies and are suitable as ensemble algorithms in that they can identify nonlinear relationships among the predictions of each meta-learner.

Before creating the ensemble model, we performed feature selection to prevent overfitting and multicollinearity. Correlation analysis was performed to solve the problem of multicollinearity among candidate variables. Significant variables were selected through backward elimination for the first selected variables through correlation analysis. We provided a framework for performing feature selection through correlation analysis and backward elimination, along with

an ensemble prediction methodology based on feature selection. To demonstrate the proposed methodology's adequacy, we analyzed the effects of feature selection and compared the predictive power between proposed ensemble models and single models.

The main contributions of this study are as follows:

- Because of Korea's rapid expansion of PV power generation, it is important to forecast mid- and long-term PV generation to establish utilities' transmission and distribution planning. Unlike most previous studies, which mainly focused on short-term predictions, this study attempts to make mid- and long-term predictions that may help utility transmission and distribution planning.
- Many factors can affect local PV production, but proper feature selection is needed to prevent overfitting and multicollinearity. We provide a framework for performing feature selection through correlation analysis and backward elimination, along with an ensemble prediction methodology based on feature selection. When the variables derived through feature selection are used, the proposed methodology shows higher predictive power than other models using the variable that an arbitrary feature is applied.
- Utilities gather predictions from various sources and need to consider them to make accurate forecasts. Our ensemble method can produce more accurate forecasting using various predictions. The proposed CNN-based ensemble method shows better predictive power than other models and is superior to the simple average of each predicted value.

III. METHODOLOGY

A. PROPOSED METHODOLOGY AND FRAMEWORK

The forecasting methodology is based on a feature-selective ensemble learning method. In general, the process of long-term regional PV power generation forecasting is composed of three steps: feature selection based on correlation analysis and backward elimination, meta-learner training based on k-fold validation, and construction of an ensemble model that uses meta-learners' predictions as input data for PV forecasting. The proposed forecasting methodology's framework is shown in Fig. 1.

In the first step, we set categories of variables based on previous research and derived a candidate variable for each category. To prevent multicollinearity of candidate variables, dependent variables and variables with a correlation coefficient of |0.9| or more were removed. Insignificant variables were removed from the derived candidate variables. For the remaining variables, a final variable was selected that had a significant relationship with the dependent variable through backward elimination. The backward elimination algorithms are shown in Fig. 2.

In the second step, meta-learner models were trained using the data after feature selection. Because our goal was to make accurate forecasts based on predictions from various sources, we chose three representative meta-learner models of time-series prediction. We choose the representative models from traditional time-series models such as ARIMA, VAR, and recursive machine learning models like LSTM.

Each model chosen is a commonly used model for time-series prediction. The ensemble model is a methodology that performs forecasting based on the predictions of various meta-learners. We focused on applying the ensemble rather than individual prediction models. To train the ensemble model, meta-learner training and validation had to be done in the training set. To do this, we divided the training set into k folds, trained the meta-learner based on each fold, and computed the predictions from the other folds.

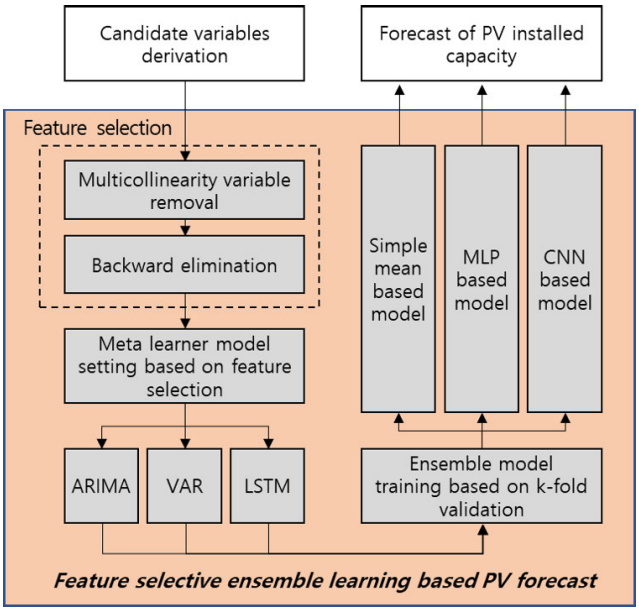


FIGURE 1. Framework of the methodology.

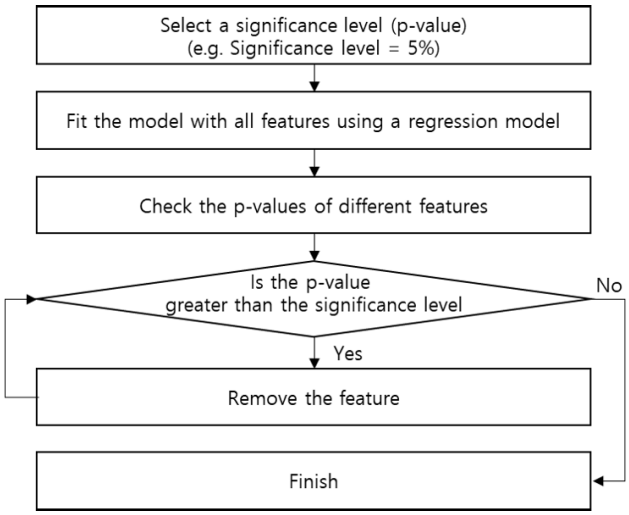


FIGURE 2. Backward elimination algorithms for feature selection.

In the last step, the ensemble model was trained based on the predictions generated by the meta-learner and the actual values. To compare the ensemble models' predictive power, we introduced and compared different ensemble algorithms. In this study, we used the simple average method, MLP method, and CNN method to form an ensemble algorithm. Finally, to find the effects of the predictive power of the ensemble algorithm, the predictive power was compared by classifying whether the ensemble was applied and the ensemble algorithm types. For a fair comparison, individual predictive models that did not apply the ensemble used the entire training set for training.

B. META-LEARNER

1) ARIMA

In general, time-series information has regular patterns and irregular patterns, and the regular pattern can be classified as autocorrelation or moving average. To use these regular patterns for time-series forecasting, Box and Jenkins proposed the ARIMA model, which is based on the autoregressive moving average model (ARMA) and the momentum of past data. The ARIMA model's formula can be expressed as follows:

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + +\varphi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q}, \quad (1)$$

where y_t is the actual value at period t, ϵ_t is the random errors, which are distributed with zero mean and a constant variance, and φ_i for $i = 0, 1, \cdots, p$ and θ_j for $j = 1, 2, \cdots, q$ are model parameters.

2) LSTM

LSTM is an artificial RNN that has feedforward neural networks and feedback connections, processing single data points and entire sequences of data. The LSTM model prevents gradient disappearance and explosion by using gates and memory cells, which are suitable for learning long-term dependencies. An LSTM unit has three gates (i.e., input, output, and forget) and a memory cell that can hold data for a certain period. In addition, the three gates can control the cell input and output. The LSTM structure is shown in Fig. 3 and is expressed as follows:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f), \quad (2)$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i), \quad (3)$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o), \quad (4)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c), \quad (5)$$
$$h_t = o_t \circ \sigma_h(c_t), \quad (6)$$

where x_t is the input vector of the LSTM unit, h_t is the output vector of the LSTM unit, f_t is the activation vector of the forget gate, i_t is the activation vector of the input gate, o_t is the activation vector of the output gate, c_t is the cell state vector, σ is the sigmoid function, and the subscript t denotes the time step. W, U, and b are weight matrices and bias vector parameters that need to be updated during training.

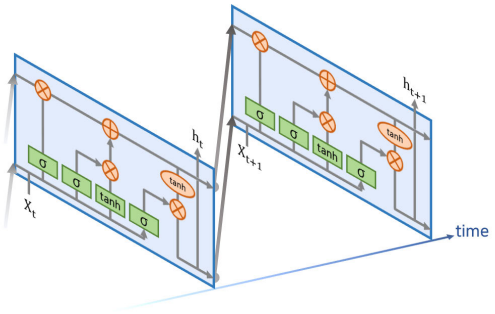


FIGURE 3. LSTM architecture.

3) VAR

The traditional regression model is based on variable correlation to calculate the dependent variable Y from several explanatory variables X_t, but the model cannot reflect time variation. Although the ARIMA method can deal with the time-series information, it ignores the interaction among variables. To overcome the drawbacks of the regression model and ARIMA, the VAR model, which combines both regression and time-series analysis, has been applied for forecasting. The VAR model X_t has N multivariate stationary time series: $X_t = (X_{1,t}, X_{2,t}, \cdots, X_{N,t})$. The VAR model can be expressed as follows:

$$X_t = C + \theta_1 X_{t-1} + \theta_2 X_{t-2} + \cdots + \theta_p X_{t-p} + \epsilon_t, \quad (7)$$

where C is the vector of constants, θ_i is the matrix of time-invariant coefficients, and ϵ_t is the white noise vector.

C. ENSEMBLE MODEL

No single model can perform well in all situations. The ensemble model (or stacking ensemble model) used in this study produced improved performance by combining different models. Stacking models can be constructed by combining various algorithms, and such combination makes up for the weaknesses of each algorithm. In this study, the stacking ensemble model was constructed using simple average, multilayer perceptron, and CNN. To compare the ensemble models' predictive power, three ensemble models were constructed.

1) SIMPLE MEAN

An ensemble model combines meta-learner model predictions with certain rules to produce new predictions. The simplest ensemble model is simply averaging each prediction. Despite the simplicity of this model, it is possible to attain good performance if enough meta-learner models are available.

2) MLP

MLP is a neural network model composed of an input layer, a hidden layer, and an output layer as shown in Fig. 4. This algorithm is more advanced than a single-layer neural network, which is composed of one input layer and a hidden layer. The MLP methodology constructs and analyzes many

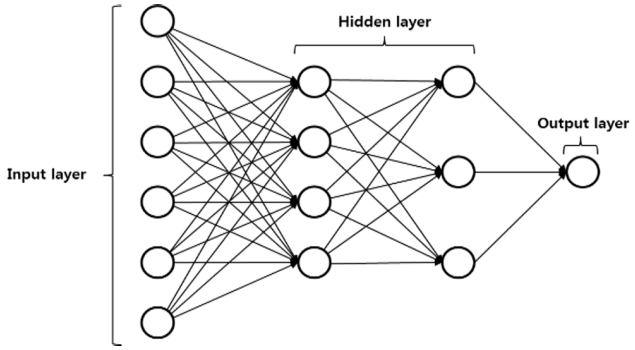


FIGURE 4. MLP conceptual diagram.

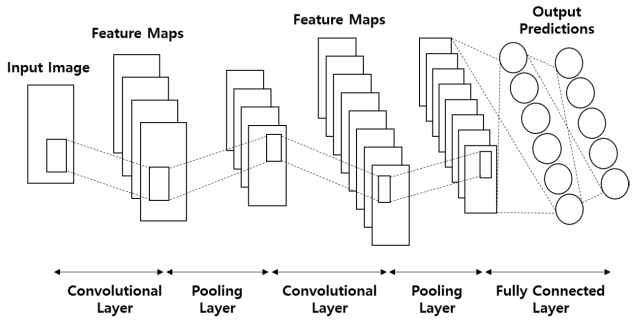


FIGURE 5. CNN conceptual diagram.

hidden layers in an artificial neural network. This analytical methodology is widely used for pattern classification, recognition, and prediction and is extended to more advanced neural network analysis depending on the hidden layer's shape and function. MLP can be written as shown in (8). v_i is the input layer or previous hidden layer signal, and b_j and b_k represent the bias of the hidden layer and the output layer, respectively. w_{ij} and w_{jk} denote coefficient values of the hidden layer and the output layer, respectively:

$$Y_k = \sum_{j=1}^{m} f\left(\sum_{i=1}^{n} v_i w_{ij} + b_j\right) w_{jk} + b_k, \tag{8}$$

where f represents an activation function, and sigmoid and rectified linear unit functions are commonly used. The resultant value Y of the output layer can be obtained through the path of (8).

3) CNN

CNN is a regularized version of multilayer perceptrons. Multilayer perceptrons usually mean fully connected networks; that is, each neuron in one layer is connected to all neurons in the next layer. Typical methods of regularization include adding some form of magnitude measurement of weights to the loss function.

CNNs take a different approach to regularization: they take advantage of the data's hierarchical pattern and assemble complex patterns using smaller and simpler patterns. Therefore, on the scale of connectedness and complexity, CNNs are on the lower extreme. Composite product neural networks like CNNs have shown significant performance in the field of imaging. As shown in Fig. 5, a CNN is composed of one input layer and one output layer, one or more convolutional layers, and a pooling layer. Data are input through the input layer and filtered through a convolutional layer to extract the appropriate features. At this time, the number of feature maps is determined according to the number of filters.

IV. VARIABLES SELECTION AND DATA ACQUISITION

A. DATA DESCRIPTION

As the amount of PV generation continues to increase, the difficulty utilities have in managing loads increases, because PV power generation is concentrated during the daytime and the volatility of PV generation is high. In particular, PV power generation has a large variation in scale from region to region, and regional distribution networks and transformer capacities are different, so regional prediction of PV installation is important. Korea is among the countries most rapidly pursuing a PV-focused energy mix, so forecasting for PV installation is highly important. PV generation in Korea is the highest in Jeollanam-do and Gwangju, so we tried to forecast local PV installation in Jeollanam-do and Gwangju. The main factors that affect the amount of PV already installed are weather and equipment-related variables, but new PV installation relies on the profitability of the PV investment, the mandatory PV generation ratio set by the government, the climate and demographic characteristics of the region, etc. Therefore, this study selected candidate variables for economic, policy, and environmental factors that affect existing and new PV installations and created forecasts based on them. Table 1 and Table 2 displays symbols for the variables and their descriptive statistical results, respectively. We used monthly data for forecasting.

1) ECONOMIC FACTORS

Economic factors that can affect PV installations include the profits and costs of PV power generation and the relative profitability of other investments, such as stock prices and Treasury yield. System marginal cost (SMP) was selected as a candidate variable to judge the profit of PV generation, and the PV installation price announced by Bloomberg was selected as a variable representing the cost. In addition, Korea's consumer prices and GDP should be considered to determine the relative profitability and purchasing power by year. We also considered the Korea Composite Stock Price Index (KOSPI) and Treasury yield (3 year) for comparison with alternative investments.

2) POLICY FACTORS

PV installations have traditionally been heavily influenced by government policy decisions. The government sets targets for PV power generation from an energy-mix perspective. To determine these policy factors, we selected the renewable portfolio standard (RPS) target as a candidate variable.

TABLE 1. Symbols for regional PV capacity forecasting.

Symbol	Variables
Y_0	PV generation in Jeollanam-do and Gwangju (MWh)
V_1	SMP (KRW/kWh)
V_2	PV installation price ($/W)
V_3	Korea's consumer prices (2015 index = 100)
V_4	Korea GDP (one million KRW)
V_5	KOSPI (index)
V_6	Treasury yield (3 year, %)
V_7	RPS (%)
V_8	Power consumption (TWh)
V_9	Maximum power (MW)
V_{10}	Supply reserves (%)
V_{11}	Regional transformer capacity (kVA)
V_{12}	Local mean temperature ($^\circ C$)
V_{13}	Local average sunshine (hours)
V_{14}	Local population density (person/m^2)

TABLE 2. Descriptive statistical results.

Symbol	Mean	Standard deviation	Min	Max
Y_0	348.5	99.4	186.5	533.7
V_1	117.7	32.6	65.2	163
V_2	1.4	0.2	1.1	1.6
V_3	99.4	1.4	97.5	101.2
V_4	1,615,637	92,371	1,500,819	1,740,780
V_5	1,985.10	53.6	1,872	2,114.90
V_6	2.2	0.6	1.2	2.9
V_7	3.3	0.6	2.5	4
V_8	483.2	8.7	474.7	497
V_9	1,995,170	147,896	1,785,518	2,324,888
V_{10}	22.1	6	11.8	35.4
V_{11}	18,500,000	185,502	18,200,000	18,700,000
V_{12}	13.8	8.6	0.4	27.7
V_{13}	5.6	1.5	2.9	9
V_{14}	145.8	0.4	145	146

In addition, power consumption, maximum power, supply reserves, and regional (Jeollanam-do and Gwangju) transformer capacity were considered.

3) ENVIRONMENTAL FACTORS

In addition to economic and policy factors, the main environmental factors affecting PV installation are weather and demographics in the region. In this study, we selected local mean temperature and local average sunshine as weather candidates and used local population density representing population per unit area as the demographic factor.

B. FEATURE SELECTION

1) MULTICOLLINEARITY VARIABLE ELIMINATION

The result of correlation analysis with all features (variables) is shown in Fig. 6. The variables with high correlation coefficients ($|\rho| > 0.9$) that showed multicollinearity were PV

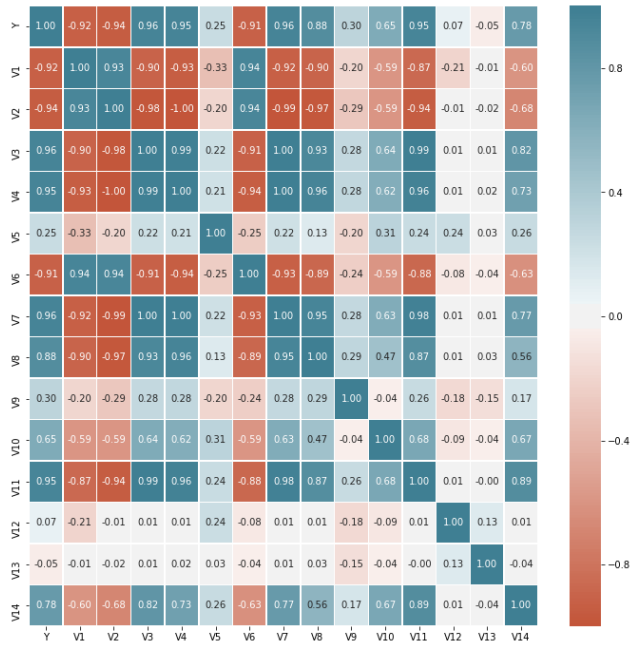


FIGURE 6. Correlation analysis results for all candidate variables.

TABLE 3. Model fit results (1st round).

Feature	Coef	Std. Err.	t	$p > \|t\|$
V_1	-2.5197	0.213	-11.807	0.000
V_5	-0.1282	0.103	-1.239	0.222
V_9	7.252×10^{-5}	3.79×10^{-5}	1.914	0.063
V_{10}	2.6646	1.123	2.373	0.022
V_{12}	-0.5078	0.645	-0.788	0.435
V_{13}	-2.3940	3.375	-0.709	0.482
V_{14}	4.9163	1.697	2.896	0.006

Adj. R squared, 0.991; F statistic, 787.7; AIC, 480.2; BIC, 493.3

installation price, Korea's consumer prices, Korean GDP, Treasury yield (3 year), RPS, power consumption, and local mean temperature. Therefore, we excluded those variables before proceeding to the next step.

2) BACKWARD ELIMINATION

Backward elimination was performed based on the assumption that the multicollinearity issue was resolved. Table 3 shows the result of constructing the regression model with each feature as an independent variable and the target variable as a dependent variable. The variable with the highest p-value was removed, and the regression model was rebuilt using the remaining variables. V_{13}, V_{12}, and V_5 was removed from Table 3, Table 4, and Table 5, respectively. In Table 6, we can see that the p-values of all features were lower than the threshold.

V. EXPERIMENTAL SETUP

We used monthly data up to time t to predict $t + 1$ years. This means that the forecast horizon was 1 year. Because the

TABLE 4. Model fit results (2nd round).

Feature	Coef	Std. Err.	t	$p > \|t\|$
V_1	-2.5117	0.212	-11.856	0.000
V_5	-0.1268	0.103	-1.233	0.225
V_9	7.631×10^{-5}	3.73×10^{-5}	2.047	0.047
V_{10}	2.7179	1.114	2.441	0.019
V_{12}	-0.5405	0.639	-0.845	0.403
V_{14}	4.7412	1.669	2.84	0.007

Adj. R squared, 0.991; F statistic, 929.9; AIC, 478.8; BIC, 490.0

TABLE 5. Model fit results (3rd round).

Feature	Coef	Std. Err.	t	$p > \|t\|$
V_1	-2.4468	0.197	-12.432	0.000
V_5	±0.1414	0.101	-1.400	0.169
V_9	8.438×10^{-5}	3.59×10^{-5}	2.349	0.023
V_{10}	3.0521	1.037	2.942	0.005
V_{14}	4.6763	1.662	2.814	0.007

Adj. R squared, 0.992; F statistic, 1123.0; AIC, 477.6; BIC, 487.0

TABLE 6. Model fit results (4th round).

Feature	Coef	Std. Err.	t	$p > \|t\|$
V_1	-2.3804	0.193	-12.328	0.000
V_9	9.838×10^{-5}	3.49×10^{-5}	2.821	0.007
V_{10}	2.9932	1.048	2.856	0.007
V_{14}	2.5135	0.619	4.058	0.000

Adj. R squared, 0.991; F statistic, 1373.0; AIC, 477.8; BIC, 485.2

ensemble model was trained based on meta-learner predicted values, the predicted value of each meta-learner should be calculated in the training set. Therefore, it was necessary to divide the training set to train the meta-learner and to derive the predicted value. In this study, training set was divided in two, as shown in Fig. 7.

The training set of Fold 1 consisted of feature data from January 2013 to December 2014 and target data from January 2014 to December 2015. The validation set of Fold 1 consisted of feature data from January 2015 to December 2015 and target data from January 2015 to December 2016. Similarly, the training set of Fold 2 consisted of feature data from January 2014 to December 2015 and target data from January 2015 to December 2016. The validation set of Fold 2 consisted of feature data from January 2016 to December 2016 and target data from January 2017 to December 2017. Predicted values of the meta-learners (LSTM, ARIMA, and VAR) in folds and actual PV capacity were used to train the ensemble model (simple mean, MLP, and CNN). Each model predicted PV generation for up to 1 year.

A. META-LEARNER MODEL DESCRIPTION

1) LSTM

To make the LSTM model predict PV power generation after 1 year, the data set was set up as follows: The target was the regional PV generation at $t + 1$ year, and the input data were the other variables at t years. Training sets of Fold 1 and Fold 2 were used as training data, and the predictive power was tested through each validation set. Hyperparameters for LSTM model implication are shown in Table 7.

Feature(t)	**Target**(t+1year)	**Fold Composition**	
V_1, V_9, V_{10}, V_{14}	Y	Fold 1	Fold 2
Jan 2013	Jan 2014	Train set	
...	...		
Jan 2014	Jan 2015		Train set
...	...		
Dec 2014	Dec 2015		
Jan 2015	Jan 2016	Validation set	
...	...		
Dec 2015	Dec 2016		
Jan 2016	Jan 2017		Validation set
...	...		
Dec 2016	Dec 2017		
Jan 2017	Jan 2018	Test set	
...	...		
Dec 2017	Dec 2018		

FIGURE 7. Fold composition for ensemble learning.

TABLE 7. Hyperparameter setting value of the LSTM model.

Hyperparameter	Value	Hyperparameter	Value
Input variables	4	Epochs	30
Hidden layer	1	Batch Size	1
Hidden node	20	Optimizer	ADAM

TABLE 8. ARIMA model parameter setting results.

Parameter	Value	Statistics	Value
Order(p,d,q)	(1,1,1)	AIC	-36.035
Seasonal order	(0,1,0,12)	BIC	-34.444

2) ARIMA

We used the seasonal-ARIMA model to consider the data's seasonality. Model identification and order decision were made based on Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). The selected model and order are shown in Table 8. For the analysis, training sets of Fold 1 and Fold 2 were used as input data.

3) VAR

To ensure the time series remains stationary, the first difference was made. The order was chosen based on the data provided. The calculated orders, AIC, and BIC are shown

TABLE 9. VAR model parameter setting results.

Parameter	Value
AR lag order	5
AIC	-29.6156
BIC	-24.2371

TABLE 10. Hyperparameter setting value of the MLP model.

Hyperparameter	Value	Hyperparameter	Value
Input variables	3	Epochs	300
Hidden layer	2	Batch Size	1
Hidden node	100	Optimizer	ADAM

TABLE 11. Hyperparameter setting value of the CNN model.

Hyperparameter	Value	Hyperparameter	Value
Input variables	3	Pool size	2
Dimension	1	Epoch	300
Filter size	128	Loss function	MSE
Kerner size	2	Optimizer	ADAM

in Table 9. As with other models, training sets of Fold 1 and Fold 2 were used as input data.

B. ENSEMBLE MODEL DESCRIPTION

1) SIMPLE MEAN

The simple mean method simply averaged the predictions of each meta-learner (LSTM, ARIMA, and VAR). This method was used as a criterion to compare the predictive power of other ensemble models.

2) MLP

The MLP model's input data were the predicted value of each meta-learner model and the actual target value. The trained MLP model predicted PV power generation using the test set. Hyperparameters for MLP model implication are shown in Table 10.

3) CNN

The CNN model's input data were the predicted value of each meta-learner model and the actual target value, like the MLP model shown earlier. Trained CNN models used test sets to predict PV generation. Hyperparameters for CNN model implication are shown in Table 11.

VI. SIMULATION RESULTS

A. EACH META-LEARNER

To calculate the prediction value for each meta-learner model to be used as the ensemble model's input data, we trained and verified the meta-learner model for each fold. Fig. 8 shows the result of predicting PV power generation in 2016 using training data of Fold 1 up to December 2015

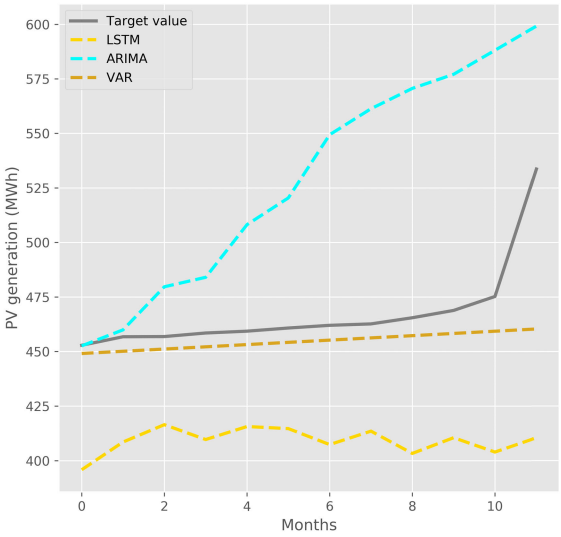


FIGURE 8. Meta-learner's forecasting result of PV generation (Fold 1).

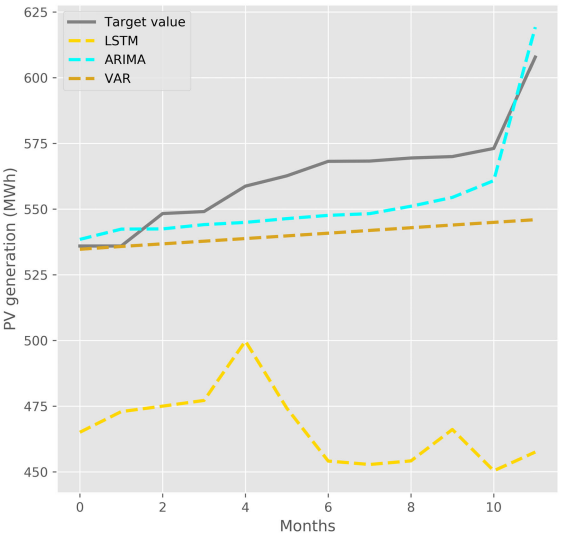


FIGURE 9. Meta-learner's forecasting result of PV generation (Fold 2).

from January 2014, and Fig. 9 shows the result of predicting PV power generation in 2017 using training data of Fold 2 up to December 2016 from January 2015. In the case of Fold 1, ARIMA tended to predict higher PV generation than actual PV generation over time, and LSTM showed a repetitive rise and fall. In the case of Fold 2, ARIMA showed higher predictive power than other models, and LSTM showed significantly lower predictive power than other models. VAR showed more stable predictive power than other models.

Figs. 10 to 12 show the prediction error results of the calculated meta-leaner models. In these cases, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) are calculated as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_t|, \tag{9}$$

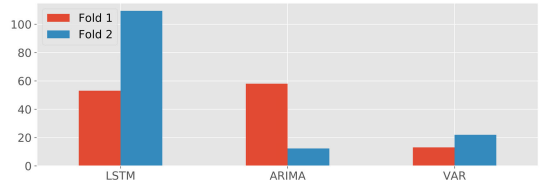


FIGURE 10. Prediction errors of meta-learner models (MAE).

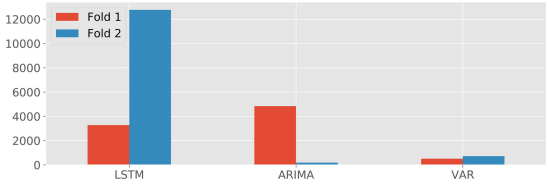


FIGURE 11. Prediction errors of meta-learner models (MSE).

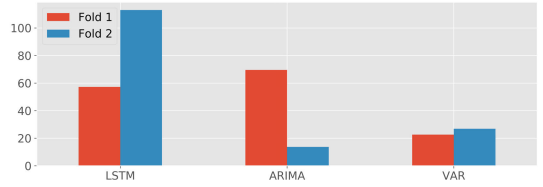


FIGURE 12. Prediction errors of meta-learner models (RMSE).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_t^2, \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_t^2}. \tag{11}$$

In Fold 1, the VAR model showed the highest accuracy, whereas in Fold 2, the ARIMA model showed the highest accuracy. The ensemble model was trained based on the meta-learner's prediction value using the training set in each fold and the actual target value.

B. ENSEMBLE MODEL

1) FEATURE SELECTION EFFECT ANALYSIS

In this section, we compare the models' predictive power depending on whether feature selection was applied. We used only the simple average methods, because MLP and CNN can produce good estimates even if a meta-learner's predictive power is poor. The control group for comparing the models' predictive power based on whether feature selection was applied has used the same four variables for fair comparison with the proposed ensemble model. Therefore, the control group used three variables excluded from feature selection (V_5, V_{12}, and V_{13}) and any one variable. Fig. 13 shows the ensemble model's prediction results based on whether feature selection was applied, and Table 12 shows the prediction error results. When the variables (V_1, V_9, V_{10}, and V_{14}) derived through feature selection were used, the proposed ensemble model showed higher predictive power than other models using the variable that arbitrary features were applied. The simple mean-based model that applied feature selection showed particularly high accuracy from the start of prediction to 4 months. Comparison model 2 and comparison model 4 showed poor predictive power than other models in all periods. However, we can see that the simple mean method of the meta-learner's predictions did not show strong enough performance overall. In the next section, we attempt ensemble predictions using MLP and CNN, as well as the simple mean method.

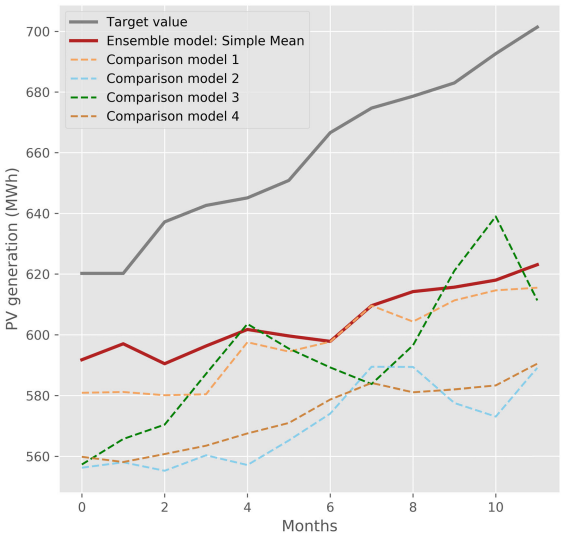


FIGURE 13. Comparison of predictive power with or without feature selection.

TABLE 12. Comparison of forecasting results (simple mean based).

	Methods	MAE	MSE	RMSE
Feature selection applied	Ensemble model (V_1, V_9, V_{10}, V_{14})	54.82	3294.64	57.39
Arbitrary feature applied	Comparison model 1 (V_1, V_5, V_{12}, V_{13})	62.11	4061.61	63.73
	Comparison model 2 (V_5, V_9, V_{12}, V_{13})	89.03	8196.06	90.53
	Comparison model 3 (V_5, V_{10}, V_{12}, V_{13})	66.06	4591.35	67.75
	Comparison model 4 (V_5, V_{14}, V_{12}, V_{13})	86.07	7659.86	87.52

2) ENSEMBLE MODEL PREDICTION RESULTS

After feature selection, we compared the ensemble models' predictive power using a test set. The meta-learners used here were trained through Fold 1 and Fold 2. Figs. 14 to 16 show the prediction error results of the calculated ensemble models. CNN showed the best predictive power in MAE, MSE, and RMSE. MLP and CNN outperformed the simple average of each meta-learner.

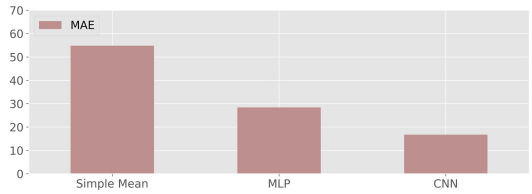


FIGURE 14. Prediction errors of ensemble models (MAE).

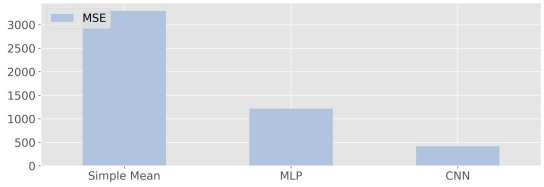


FIGURE 15. Prediction errors of ensemble models (MSE).

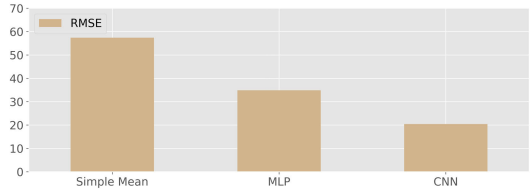


FIGURE 16. Prediction errors of ensemble models (RMSE).

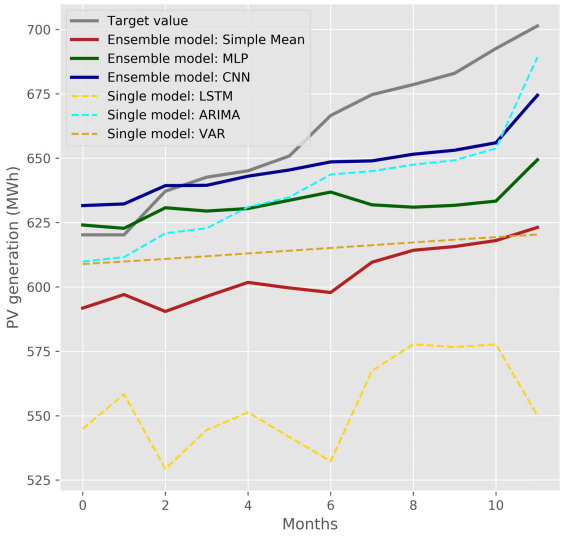


FIGURE 17. Forecasting results of PV generation in 2018.

To compare the ensemble models' predictive power, we compared them with single models using the same training data. Fig. 17 shows the results of forecasting PV power generation in 2018 based on training data up to 2017. The CNN-based ensemble model showed the best predictive power. However, the MLP-based ensemble model showed higher predictive power than the CNN-based ensemble model during the initial forecasting period.

TABLE 13. Comparison of forecasting results (simple mean based).

Methods		MAE	MSE	RMSE
Ensemble model	Simple mean	54.82	3294.64	57.39
	MLP	28.38	1215.70	34.86
	CNN	16.70	416.98	20.42
Single model	LSTM	93.31	9191.9	95.87
	ARIMA	26.32	890.53	29.84
	VAR	44.83	2517.91	50.17

The ensemble models all showed higher predictive power than the LSTM-based single model. The simple mean-based ensemble model, which simply averaged the meta-learner predictors, showed low predictive power because of LSTM's low predictive power. Finally, all models tended to predict slower PV power generation than actual PV generation trends.

Table 13 shows the prediction error for each model. The CNN-based ensemble model showed the highest predictive power, followed by ARIMA- and MLP-based ensemble models. In the case of MAE, the prediction error of the CNN-based ensemble model was 30% of the simple mean and 59% of the MLP results. Similarly, In the case of MSE, the prediction error of the CNN-based ensemble model was 13% of the simple mean and 34% of the MLP results. In the case of RMSE, the prediction error of the CNN-based ensemble model was 36% of the simple mean and 59% of the MLP results. In the single model, ARIMA showed relatively high predictive power, but it was 1.57 times higher than the predicted error of the CNN-based ensemble model (in the case of MAE).

VII. CONCLUSION

To estimate future regional PV power generation, we consider not only weather and equipment-related variables but also economic and PV government policy variables, because new PV installations can be strongly influenced by economic and policy factors. In addition, correlation analysis is performed to solve the problem of multicollinearity among candidate variables. As a result, 7 of the 14 variables are removed. Then, significant variables are selected through backward elimination for the first selected variables through correlation analysis. In this study, four variables are finally selected. When the variables derived through feature selection are used, the proposed ensemble models show higher predictive power than other models using the variable that arbitrary features are applied. To carry out long-term predictions of regional PV generation, we propose ensemble models based on the predictions of meta-learners (LSTM, ARIMA, and VAR). We use simple mean, MLP, and CNN methods as ensemble models. To compare the ensemble models' predictive power, we compare them with single models using the same training data. As a result of forecasting, the CNN-based ensemble model shows the best predictive power. However, the MLP-based ensemble model shows higher predictive

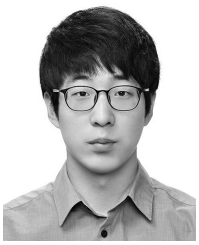
power than the CNN-based ensemble model during the initial forecasting period. MLP and CNN outperform the simple average of each meta-learner. Finally, ensemble models all show higher predictive power than the LSTM-based single model.

On the basis of feature-selective ensemble learning-based long-term regional PV power generation forecasting, we can drive further research using more diverse types of data, such as map images and distribution networks. PV forecasting using other types of data will increase actual utilization and increase predictive power.

REFERENCES

[1] *The 8th Basic Plan for Long-Term Electricity Supply and Demand (2017–2031)*, Minister of Trade, Industry and Energy, Sejong City, South Korea, 2017.
[2] N. Yorino, Y. Sasaki, S. Fujita, Y. Zoka, and Y. Okumoto, "Issues for power system operation for future renewable energy penetration: Robust power system security," *Electr. Eng. Jpn.*, vol. 182, no. 1, pp. 30–38, Jan. 2013.
[3] B. Kausika, W. Folkerts, W. van Sark, B. Siebenga, and P. Hermans, "A big data approach to the solar PV market design and results of a pilot in The Netherlands," in *Proc. 29th Eur. Photovolt. Sol. Energy Conf. Exhib.*, Apr. 2014, pp. 4030–4033.
[4] J. Shi, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting power output of photovoltaic systems based on weather classification and support vector machines," *IEEE Trans. Ind. Appl.*, vol. 48, no. 3, pp. 1064–1069, May 2012.
[5] C. Yang, A. A. Thatte, and L. Xie, "Multitime-scale data-driven spatio-temporal forecast of photovoltaic generation," *IEEE Trans. Sustain. Energy*, vol. 6, no. 1, pp. 104–112, Jan. 2015.
[6] G. TamizhMani, L. Ji, Y. Tang, L. Petacci, and C. Osterwald, "Photovoltaic module thermal/wind performance: Long-term monitoring and model development for energy rating," in *Proc. NCPV Sol. Program Rev. Meeting*, Mar. 2003, pp. 936–939.
[7] J. Zoellner, P. Schweizer-Ries, and C. Wemheuer, "Public acceptance of renewable energies: Results from case studies in Germany," *Energy Policy*, vol. 36, no. 11, pp. 4136–4141, Nov. 2008.
[8] R. Billinton and R. Karki, "Capacity expansion of small isolated power systems using PV and wind energy," *IEEE Trans. Power Syst.*, vol. 16, no. 4, pp. 892–897, Nov. 2001.
[9] T. Zhao and Y. Zhang, "Regional PV installed capacity forecasting considering generation costs and time lag of influential factors," *IEEJ Trans. Electr. Electron. Eng.*, vol. 13, no. 2, pp. 201–211, Feb. 2018.
[10] I. Colak, M. Yesilbudak, N. Genc, and R. Bayindir, "Multi-period prediction of solar radiation using ARMA and ARIMA models," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 1045–1049.
[11] A. J. Conejo, M. A. Plazas, R. Espinola, and A. B. Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 1035–1042, May 2005.
[12] E. Yatiyana, S. Rajakaruna, and A. Ghosh, "Wind speed and direction forecasting for wind power generation using ARIMA model," in *Proc. Australas. Universities Power Eng. Conf. (AUPEC)*, Nov. 2017, pp. 1–6.
[13] Y. Zhang, M. Beaudin, H. Zareipour, and D. Wood, "Forecasting solar photovoltaic power production at the aggregated system level," in *Proc. North Amer. Power Symp. (NAPS)*, Sep. 2014, pp. 1–6.
[14] L. Yujie, Y. Xiaoling, X. Jieyan, C. Zheng, M. Fei, and L. Haoming, "Medium-term forecasting of cold, electric and gas load in multi-energy system based on VAR model," in *Proc. 13th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, May 2018, pp. 1676–1680.
[15] M. He, V. Vittal, and J. Zhang, "A sparsified vector autoregressive model for short-term wind farm power forecasting," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Jul. 2015, pp. 1–5.
[16] K. Hua and D. A. Simovici, "Long-lead term precipitation forecasting by hierarchical clustering-based Bayesian structural vector autoregression," in *Proc. IEEE 13th Int. Conf. Netw., Sens., Control (ICNSC)*, Apr. 2016, pp. 1–6.
[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[18] Y. Yu, J. Cao, and J. Zhu, "An LSTM short-term solar irradiance forecasting under complicated weather conditions," *IEEE Access*, vol. 7, pp. 145651–145666, 2019.
[19] J. Zhang, Y. Chi, and L. Xiao, "Solar power generation forecast based on LSTM," in *Proc. IEEE 9th Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2018, pp. 869–872.
[20] D. Lee and R. Baldick, "Short-term wind power ensemble prediction based on Gaussian processes and neural networks," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 501–510, Jan. 2014.
[21] A. Gensler, J. Henze, B. Sick, and N. Raabe, "Deep Learning for solar power forecasting—An approach using AutoEncoder and LSTM Neural Networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2016, pp. 2858–2865.
[22] H. Zhou, Y. Zhang, L. Yang, Q. Liu, K. Yan, and Y. Du, "Short-term photovoltaic power forecasting based on long short term memory neural network and attention mechanism," *IEEE Access*, vol. 7, pp. 78063–78074, 2019.

HANEUL EOM received the B.E. degree in industrial engineering from Soongsil University, South Korea, in 2014, and the M.S. degree in management of technology from Sungkyunkwan University, South Korea, in 2018. He is currently pursuing the Ph.D. degree in science and technology studies (STS) with Korea University, Seoul, South Korea. He is also a Researcher with FnPricing, South Korea. His research interest includes distribution system planning.



YONGJU SON received the B.E. degree in electronic and electrical engineering from Chung-Ang University, Seoul, South Korea, in 2019. He is currently pursuing the M.S. degree in electrical engineering with Korea University. His research interests include microgrid operation, planning, renewable energy forecast, and data analytic.

SUNGYUN CHOI (Member, IEEE) received the B.E. degree in electrical engineering from Korea University, Seoul, South Korea, in 2002, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2009 and 2013, respectively. From 2002 to 2005, he was a Network and System Engineer, and from 2014 to 2018, he was a Senior Researcher with Smart Power Grid Research Center, Korea Electrotechnology Research Institute, Uiwang, South Korea. Since 2018, he has been an Assistant Professor with Electrical Engineering, Korea University, Seoul, South Korea. His research interests include smart grid technology, microgrid operation, control and protection, power system state estimation, phasor measurement units, and sub-synchronous.