

Received May 25, 2021, accepted July 17, 2021, date of publication July 26, 2021, date of current version August 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3100105

Two-Stage Attention Over LSTM With Bayesian Optimization for Day-Ahead Solar Power Forecasting

MUHAMMAD ASLAM¹, SEUNG-JAE LEE¹, (Life Fellow, IEEE),
SANG-HEE KHANG¹, (Member, IEEE), AND SUGWON HONG²

¹Department of Electrical Engineering, Myongji University, Yongin, Gyeonggi-do 17058, South Korea

²Department of Computer Engineering, Myongji University, Yongin, Gyeonggi-do 17058, South Korea

Corresponding author: Sugwon Hong (swhong@mju.ac.kr)

This work was supported by Korea Electric Power Corporation under Grant R17XA05-2.

ABSTRACT The penetration of PVs into the power grid is increasing day by day, as they are more economical and environment-friendly. However, due to the intrinsic intermittency in solar radiation and other meteorological factors, the generated power from PVs is uncertain and unstable. Therefore, accurate forecasting of power generation is considered one of the fundamental challenges in power system. In this paper, a deep-learning model based on two-stage attention mechanism over LSTM is proposed to forecast a day-ahead PV power. In addition, the Bayesian optimization algorithm is applied to obtain the optimal combination of hyper-parameters for the proposed deep-learning model. Various input features that can affect the PV power generation such as solar radiation, temperature, humidity, snowfall, albedo etc. are considered and their impact with respect to the attention mechanisms on the forecasted PV power is analyzed. The input consists of data from 21 PVs installed at different geographical locations in Germany. The proposed model is compared with state-of-the-art models such as LSTM-Attention, CNN-LSTM, and Ensemble model for day-ahead forecasting. The model is also compared with various single attention mechanisms such as Input-attention, SNAIL, Raffel, and Hierarchical attention etc. The proposed model outperforms the traditional methods in terms of accuracy, hence proving its efficiency. Forecasting Skill (FS) score of the proposed model is 0.4813 whereas 0.4427 is for the Ensemble model, which is the best among other state-of-the-art models. Root Mean Square (RMSE) and Mean Absolute Error (MAE) of the proposed model is 0.0638 and 0.0324 respectively, whereas those of the Ensemble model are 0.0685 and 0.0369 respectively.

INDEX TERMS Attention, Bayesian optimization, day-ahead forecasting, deep-learning, LSTM, solar power forecasting, two-stage-attention.

I. INTRODUCTION

Due to increasing concerns about global environmental and economic challenges, the demand for clean energy such as photovoltaic and wind power is increasing [1]. According to International Energy Agency (IEA), photovoltaics (PV) is one of the fastest-growing and economical renewable energy resources [2]. PV power generation is mainly dependent on meteorological factors such as solar radiation, clouds, humidity, pressure, temperature etc. Due to the intrinsic intermittent nature of these factors, the nature of PV power is also variable and uncertain, resulting in unstable fluctuations [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Padmanabh Thakur¹.

Consequently, the inclusion of a large proportion of PV will cause grid oscillations [4]. Therefore, to reduce these uncertainties and fluctuations, accurate forecasting of PV output power is essential.

Numerous benefits can be obtained by hours to day-ahead solar power forecasting. Accurate forecasting can help companies to avoid penalties [5]. Forecasting helps in optimizing the scheduling of supply offers to the market, hence, increases revenues [6]. Operators can avoid problems in balancing generation and demand [7], which resultantly can improve the system stability, and reduce the costs of ancillary services [8], [9]. Furthermore, decisions on unit commitment, reserve requirement, and maintenance scheduling to obtain optimal operating cost can be performed with accurate

forecasting. Considering these reasons, accurate PV forecasting has been recognized as one of the fundamental challenges in power systems [10], [11].

Forecasting of solar power can be mainly categorized into physical, statistical and machine-learning models [12]–[14]. Physical models have well-established methods that rely on Numerical Weather Prediction (NWP) data [15] and satellite images [16]. Although these methods have good accuracies, they require extra information of images and cloud maps from satellites, resulting in a higher cost for operation and computation. In addition, these models should be designed for particular locations [12]. Statistical models are based on traditional regressive mathematical models such as linear regression and Automatic Regressive Integrated Moving Average (ARIMA) models. Since linear regression models build a linear mapping between inputs and the target power [17], they cannot efficiently capture the non-linear relationship between input features and outputs of solar power. Furthermore, as the forecasting horizon increases, the accuracy of these models decreases [12].

To overcome these issues, various machine-learning models have been proposed [18]–[20]. Support vector regressors are the machine-learning model that can generate non-linear relationship. Support Vector Regression (SVR), based on various weather information such as cloud, sun duration etc., has been used to forecast solar power in [18]. A hybrid model based on the genetic algorithm with Support Vector Machine (SVM) has been proposed for solar power forecasting in [19], which improved results as compared to simple SVM. An ensemble model is a hybrid machine-learning model based on the combination of various non-linear regression models. The ensemble model has been proposed as the best machine-learning model in [20]. However, these models depend on predefined parameters and predefined non-linear mapping. Therefore, it is difficult to capture the true underlying non-linear relationship between inputs and target values [13], [21], [22].

Recently, deep-learning models, which are advanced forms of traditional machine-learning techniques, are becoming very popular in various fields such as image processing [23], text translation [24] and time-series problems [25]. Solar power forecasting is a time-series problem where next time steps are sequentially dependent on past time steps in a non-linear relationship. Recurrent Neural Networks (RNNs) are deep-learning models specifically designed for time-series data [26]. Non-linear Autoregressive Recurrent Neural Network (NARX) has been successfully applied to solar power forecasting [27], [28]. However, conventional RNNs suffer from exploding and vanishing gradients [29]. Thus, they cannot capture long-term dependencies. The extension of RNN, Long Short-Term Memory (LSTM) [30], has been proposed to overcome these limitations of RNN. LSTM and combination of LSTM with Convolutional Neural Network (CNN) have been used with good accuracies in solar power forecasting [31]–[33].

Encoder-decoder networks based on LSTM are becoming popular deep-learning models in time-series forecasting, specifically in sequence-to-sequence mappings [34]–[38]. Therefore, these combinations of encoder-decoder with LSTM can be regarded as state-of-the-art. Although these models work well with small sequences, their performance degrades with the increasing length of sequences [37]. In time-series forecasting this is a big concern, as predictions usually require longer temporal sequences as well as many input features such as day-ahead solar power forecasting.

Attention mechanism is an extension of the encoder-decoder model specifically designed to improve the performance of longer sequences [39]. In [39], solar power has been forecasted using single self-attention over LSTM to capture important temporal states. However, single temporal attention mechanisms still lack in handling data containing many input features and long temporal sequences. Addressing these time-series forecasting challenges, a two-stage attention mechanism has been used for stock price forecasting in [40].

In this paper inspired by [40], a two-stage attention mechanism-based deep-learning model is applied to day-ahead solar power forecasting using multiple input features. The model is optimized by using the Bayesian optimization algorithm to obtain the optimal combination of hyper-parameters. Following are the major contributions of this paper:

1. Two-stage attention-based encoder-decoder over LSTM is applied to day-ahead solar power forecasting. First, an attention layer is applied to the input, focusing on more relevant features at a particular time, which is followed by a temporal attention layer to focus on relevant temporal hidden states of LSTM units. Both attentions are applied over LSTM. 41 different input features from 21 different PV panels installed at different geographical locations in Germany are used as input data. The paper analyzes the performance of the attention mechanism with respect to some important input features such as solar radiation, temperature, snowfall, etc. The paper also analyzes the performance of the attention mechanism with respect to temporal values.
2. Deep-learning models have different hyper-parameters, which control their performance. On a particular problem, different combinations of these parameters produce optimal results. Therefore, the Bayesian optimization algorithm has been applied to the two-stage attention-based deep-learning model to obtain the optimal combination of hyper-parameters.
3. A comparative study of the proposed method with the persistence model [13], and the state-of-the-art methods such as LSTM [31], [32], CNN-LSTM [33], LSTM-Attention [39] and Ensemble model [20], has been carried out to show the effectiveness of the proposed method. Furthermore, single attention mechanisms can be carried out via different techniques such

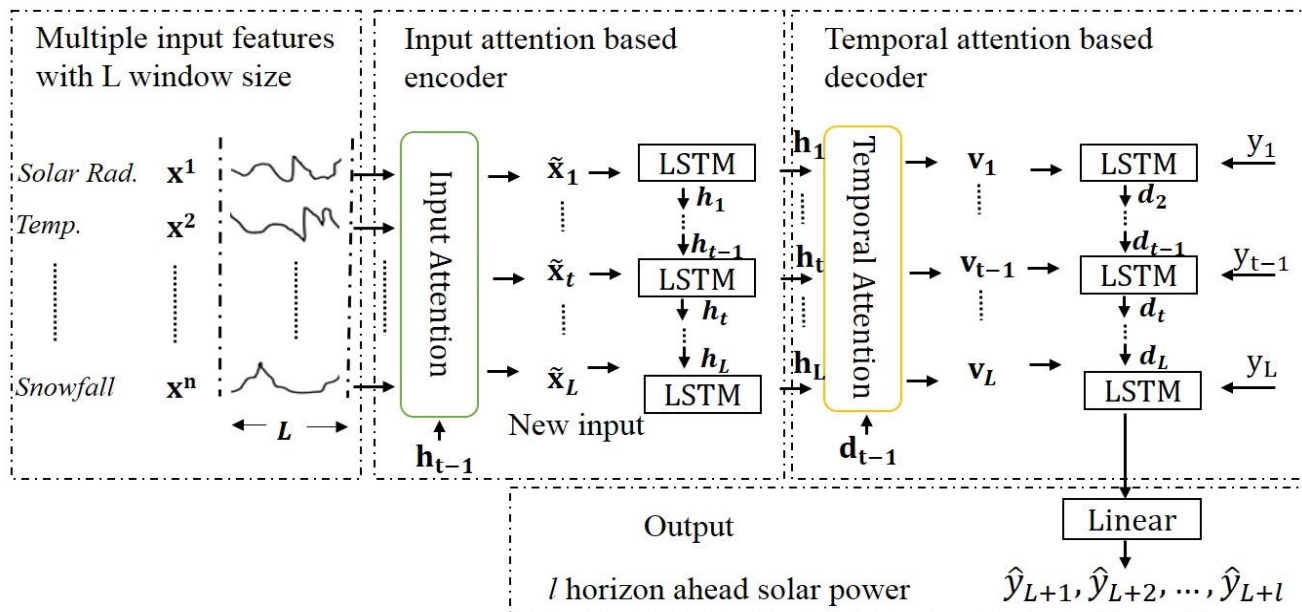


FIGURE 1. Complete illustration of two-stage attention-based encoder-decoder model for day-ahead solar power forecasting.

as Input-Attention, Raffel, Hierarchical, SNAIL attention, etc. [41]–[43]. Comparison of these single attention techniques over LSTM has also been performed with the proposed method to show the effectiveness of the two-stage attention mechanism.

The paper is organized as follows: in Section II, the two-stage attention mechanism over LSTM for day-ahead solar power forecasting is explained. Section III consists of the experiments. Results are given in Section IV. Section V gives the qualitative discussions and conclusion.

II. TWO-STAGE ATTENTION MODEL FOR SOLAR POWER FORECASTING

A. MODEL SUMMARY

Solar power is highly dependent on meteorological features like solar radiation, temperature, humidity, snowfall, etc. During normal conditions, it almost follows the trend of features like solar radiation. However, during extreme conditions like snowfall and albedo, the power production is almost zero. Therefore, an attention mechanism is required at the input to focus on the features that are more relevant at a particular time. Similarly, solar power forecasting is a time-series problem. The next time-step is correlated to past time-step outputs. Therefore, relevant time sequences must also be focused on with the attention mechanism. Considering the aforementioned objectives, in this paper the two-stage attention-based encoder-decoder model over LSTM has been applied to day-ahead solar power forecasting. First, encoder-based attention is applied to the input features to focus on important features at a particular time. Then, decoder-based temporal attention is applied to the hidden states of the encoder’s LSTM to extract important temporal states. Finally,

a linear layer is added to the output to predict a day-ahead solar power. The whole model is trained based on the standard backpropagation algorithm. The complete model is shown in Fig. 1.

B. INPUT FEATURES AND TARGET

41 different features have been used as input. Pearson correlation of these inputs with the target solar power is shown in Fig. 2. Although these features have different correlations with the output power, all of them have impacts on output power at particular instances. For the ease of mathematical representation, let $\mathbf{x}_t = (x_t^1, x_t^2, \dots, x_t^n) \in \mathbb{R}^n$ be a series of n input data features at time t . And let $\mathbf{x}^k = (x_1^k, x_2^k, \dots, x_L^k)^\top \in \mathbb{R}^L$ be a series of k^{th} input feature data over L time-steps window. Then, a series of n input data over L time-steps window can be expressed as $(\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n)^\top = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_L) \in \mathbb{R}^{n \times L}$.

Input data consists of historical solar power data and weather data. The target of the proposed paper is a day-ahead solar power. Given the past values of output as (y_1, y_2, \dots, y_L) together with the input $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_L)$, the complete model to predict a day-ahead solar power can be expressed by the following function F :

$$(y_{L+1}, y_{L+2}, \dots, y_{L+l}) = F(y_1, y_2, \dots, y_L, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_L) \quad (1)$$

where $y_{L+1}, y_{L+2}, \dots, y_{L+l}$ are l -time steps ahead solar power to be predicted. Using the proposed method, any horizon ahead forecasting can be carried out. In this paper eight-time steps ahead forecasting with a resolution of three hours is performed to obtain a day-ahead solar power.

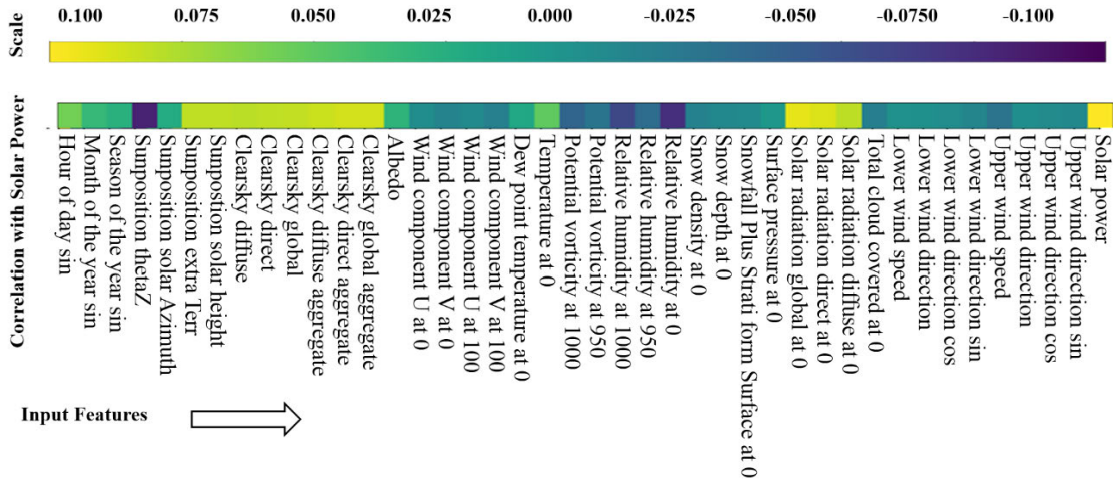


FIGURE 2. Pearson correlation between input features and target solar power.

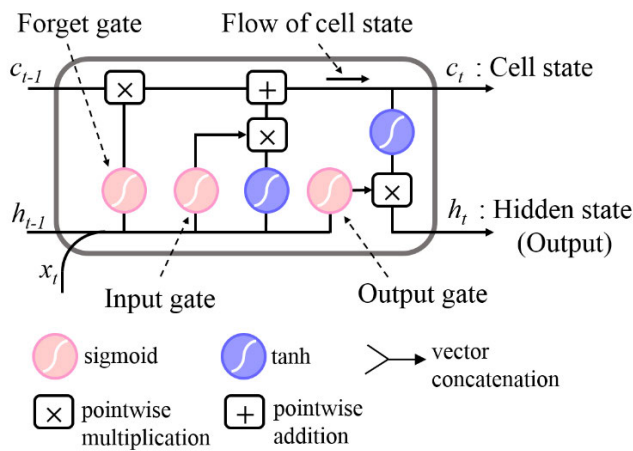


FIGURE 3. An illustration of an LSTM unit.

C. LSTM UNIT

Both input and temporal attention-based encoder-decoder are applied over LSTM units. An LSTM unit is shown in Fig. 3. LSTM unit consists of a hidden state h_t which is the output of an LSTM unit, and an internal state or cell state c_t which remembers the cell states. It also contains three gates: input i_t , forget f_t , and output gate o_t . An input gate controls the amount of current information to be passed. A forget gate controls the information to be processed or to be forgotten, and an output gate defines the internal state information that needs to be passed.

Provided that x_t is the input at t and h_{t-1} is the previous hidden state of LSTM, the following chain of equations can be used to obtain the current hidden state of LSTM unit h_t :

$$f_t = \text{sigmoid}(W_f [h_{t-1}; x_t] + b_f) \tag{2a}$$

$$i_t = \text{sigmoid}(W_i [h_{t-1}; x_t] + b_i) \tag{2b}$$

$$o_t = \text{sigmoid}(W_o [h_{t-1}; x_t] + b_o) \tag{2c}$$

$$c_t = f_t \otimes c_{t-1} \oplus i_t \otimes \tanh(W_c [h_{t-1}; x_t] + b_c) \tag{2d}$$

$$h_t = o_t \otimes \tanh(c_t) \tag{2e}$$

where $W_f, W_i, W_o, W_c, b_f, b_i, b_o, b_c$ are the weights and biases to be trained. Combining equations (2), an LSTM unit can be expressed by the following non-linear function f :

$$h_t = f(h_{t-1}, x_t) \tag{3}$$

D. INPUT ATTENTION BASED ENCODER

To extract relevant input features from the input series \mathbf{x}^k , an input attention is applied with the encoder as shown in Fig. 1. The input attention can be applied using \mathbf{x}^k , and the previous hidden and cell states of the encoder’s LSTM by using (4) and (5) as follows:

$$\varepsilon_t^k = z_\varepsilon \tanh(W_\varepsilon [h_{t-1}; c_{t-1}] + U_\varepsilon x^k) \tag{4}$$

$$\alpha_t^k = \frac{\exp(\varepsilon_t^k)}{\sum_{i=1}^n \exp(\varepsilon_t^i)} \tag{5}$$

where $z_\varepsilon, W_\varepsilon$ and U_ε are the parameters to be trained. α_t^k is an attention weight that shows the importance of the k^{th} input feature at time t . To keep the sum of all the attention weights to 1, a softmax activation is applied to ε_t^k . This attention mechanism gives important features more weights rather than treating all the inputs equally. A new input series can be extracted with these attention weights using (6). This new input is fed to update the encoder’s LSTM hidden state of (3) as shown by (7):

$$\tilde{x}_t = (\alpha_t^1 x_t^1, \alpha_t^2 x_t^2, \dots, \alpha_t^n x_t^n)^\top \tag{6}$$

$$h_t = f(h_{t-1}, \tilde{x}_t) \tag{7}$$

E. TEMPORAL ATTENTION BASED DECODER

The decoder model is designed to extract important temporal hidden states and to make the final output prediction. With the increasing length of input series, the results of the traditional encoder-decoder deteriorate. Therefore, after the attention encoder, a temporal attention-based decoder has been applied to select relevant hidden states of the encoder from all time-steps as shown in Fig. 1. The attention weights of each encoder’s hidden state can be calculated by using the

previous hidden state of the decoder's LSTM d_{t-1} and its cell state \hat{c}_{t-1} as in (8) and (9):

$$\rho_t^i = z_d \tanh(W_d [d_{t-1}; \hat{c}_{t-1}] + U_d h_i) \quad (8)$$

$$\beta_t^i = \frac{\exp(\rho_t^i)}{\sum_{j=1}^L \exp(\rho_t^j)} \quad (9)$$

where z_d , W_d and U_d are the weights to be trained. β_t^i represents the importance of the i^{th} encoder hidden state at time t . Since each encoder's hidden state h_i has been mapped to a temporal component of the input, the attention mechanism calculates the context vector v_t as a weighted sum of all the hidden states of the encoder (h_1, h_2, \dots, h_L) using (10). This vector is different at each time-step. The vector is then concatenated with the target values using (11). Then, using the decoder's hidden state d_{t-1} and the newly concatenated value \tilde{y}_{t-1} , the decoder's new hidden state d_t can be obtained using the decoder's LSTM non-linear function f based on (3) as shown in (12):

$$v_t = \sum_{i=1}^L \beta_t^i h_i \quad (10)$$

$$\tilde{y}_{t-1} = \tilde{w} [y_{t-1}; v_{t-1}] + \tilde{b} \quad (11)$$

$$d_t = f(d_{t-1}, \tilde{y}_{t-1}) \quad (12)$$

where \tilde{w} and \tilde{b} are weights and biases that are mapping the concatenation.

F. OUTPUT AND TRAINING MECHANISM

The final output follows the decoder, which consists of a linear layer to predict a day-ahead solar power. The final layer will predict l -time steps ahead. The complete model can be expressed by the following expression:

$$\begin{aligned} \hat{Y}_t &= (y_{L+1}, y_{L+2}, \dots, y_{L+l}) \\ &= F(y_1, y_2, \dots, y_L, x_1, x_2, \dots, x_L) \\ &= z_y^\top (W_y [d_L; v_L]) \end{aligned} \quad (13)$$

where \hat{Y}_t is the solar power to be predicted; W_y and z_y are the weights to be trained. In this paper, l is taken as eight to predict a day-ahead solar power with a resolution of three hours. The whole model has been trained using the standard backpropagation algorithm with the objective function defined by Mean Square Error (MSE):

$$O(\hat{Y}_t, Y_t) = \frac{1}{N} \sum_1^N (\hat{Y}_t^i, Y_t^i)^2 \quad (14)$$

where N is the number of training samples and Y_t is the actual values.

III. EXPERIMENTS

A. DATA

The model is trained and tested on 21 different PV facilities installed at different geographical locations in Germany [44]. These facilities are installed on different spots ranging from rooftop to fully-fledged solar farms. Each dataset consists of NWP data and the historical power data in a resolution

TABLE 1. Parameters of the proposed model optimized by bayesian optimization algorithm.

Parameter	Range of Values	Optimal value
Batch size	[60, 100, 128, 150, 180]	60
Lookback	[2, 3, 4, 5]	2
Learning Rate (LR)	0.00001 ~ 0.0007	0.000451
Encoder number of units (Enc units)	[32, 64, 128]	64
Decoder number of units (Dec units)	[32, 64, 128]	64
Encoder activation function (Enc act)	[relu, leakyrelu, elu, tanh, none, sigmoid]	sigmoid
Decoder activation function (Dec act)	[relu, leakyrelu, elu, tanh, none, sigmoid]	sigmoid

of three hours for 990 days. The nominal power of the PVs ranges between 100kW and 8500kW. Out of the 990 days, 490 days are used for training, 250 days are used for validation, and 250 days are used for testing. After splitting the data, the data has been normalized. Except the output power, all input data are normalized between 0 and 1. The output power or the target value is normalized according to the power capacity of the respective PV facility.

B. BAYESIAN OPTIMIZATION

Hyper-parameter optimization is essential for the model's optimal performance. Traditionally, manual, or grid and random search techniques were used for tuning hyper-parameters [45]. Manual methods are time-consuming and depend on human expertise, while in grid search the efficiency decreases as the number of hyper-parameters increase. In random search, a combination of random parameters is sampled based on a statistical distribution given by the user, which may not spot optimal points in the search.

Bayesian optimization [45] considers past evaluations to select hyper-parameters to evaluate next. By selecting its parameters in an informed manner, it can more focus on areas of the parameter space that can validate more promising parameters. It has three main parts: search space from which parameters can be sampled out; objective function; and surrogate. It builds a probability model of the objective function and uses it to select the most promising hyper-parameters to evaluate the true objective function.

Table 1 shows the ranges of hyper-parameters to be optimized in the proposed model. Seven parameters of the proposed model have been optimized using the Bayesian optimization algorithm. The table also shows the optimal set of values after applying the Bayesian optimization. Fig. 4 shows the convergence points of the optimization algorithm where the optimal combination of the hyper-parameters is converged. These points show the optimal values of the loss functions.

C. MODELS FOR COMPARISON

The persistence model has been used as the benchmark model [13]. In this model, the forecasted PV power output

TABLE 2. Optimized parameters of models for comparison.

Models	Parameters	Optimal values
Ensemble method	Gradient Boosting	No. of Estimators = 200
	Random Forest Regressor	No. of Estimators = 250
	SVR	C = 1
	Neural Network	Hidden units = 200, activation = tanh
	K Neighbors Regressor	No. of neighbors = 100
LSTM attention	Batch size	16
	No. of LSTM layers	2
	LSTM units per layer	64
	Attention layer	Encoder units = 30, decoder units = 30, encoder activation = sigmoid, decoder activation = sigmoid
	Lookback	2
	Learning Rate	0.0002436
CNN-LSTM	Batch Size	50
	1 st Convolution layer	Filters = 64, kernel size = 2
	Followed by activation layer	Leakyrelu
	2 nd convolution layer	Filters = 32, kernel size = 2
	Followed by activation layer	Elu
	3 rd convolutional layer	Filters = 16, kernel size = 2
	Followed activation layer	Tanh
	Maxpool layer	Poolsize = 2
	1 st LSTM layer	Units = 64, activation = relu
	2 st LSTM layer	Units = 32, activation = relu
	Batch size Subsequences	64
	Lookback	15
Subsequences	3	
Learning Rate	0.00011	
Input attention	Attention layer	Encoder units = 64, decoder units = 128, encoder activation = sigmoid, decoder activation = relu
	Learning rate	0.0005650
	Batch size	128
	Lookback	2
Simple LSTM	No. of hidden layers	2
	LSTM unit per layer	128
	Look back	2
	Learning Rate	0.0000998
	Batch size	64

is assumed to remain the same at the same time of the previous or following day [13]. Various state-of-the-art models like LSTM, LSTM-Attention [39], CNN-LSTM [33], and Ensemble method [20] have been implemented on the dataset to compare with the proposed model for day-ahead forecasting. All the comparative models have been optimized to obtain their optimal set of parameters. Table 2 shows the optimized parameters of different models. The attention mechanism can be carried out via different techniques

such as Raffel [41], Hierarchical [42], and SNAIL [43] attention. These single temporal attention techniques have been applied over LSTM hidden states and are compared with the proposed method to check the performance of the two-stage attention mechanism. The Input-Attention model with attention only on the input features has also been compared with the proposed method to emphasize the importance of the combined effects of two stages of attention.

TABLE 3. RMSE of different models.

PV Panels	Models						
	Persistence	Simple LSTM	Input - Attention	LSTM - Attention	CNN-LSTM	Ensemble	Proposed Model
PV1	0.1077	0.059	0.0617	0.0583	0.06	0.0585	0.0544
PV2	0.1002	0.0558	0.0578	0.0553	0.0551	0.056	0.051
PV3	0.0782	0.041	0.0417	0.0449	0.0438	0.0418	0.04
PV4	0.0717	0.0388	0.0393	0.0389	0.04	0.0376	0.0394
PV5	0.0887	0.0723	0.0728	0.0723	0.0779	0.064	0.0546
PV6	0.1093	0.0701	0.0702	0.0706	0.0671	0.0651	0.064
PV7	0.1401	0.1123	0.1115	0.1171	0.1129	0.1114	0.0953
PV8	0.1314	0.0824	0.0892	0.0883	0.0844	0.0808	0.0735
PV9	0.1313	0.0674	0.0676	0.0674	0.0645	0.065	0.0596
PV10	0.1051	0.0519	0.0555	0.0526	0.0516	0.0496	0.0472
PV11	0.164	0.0991	0.1029	0.1004	0.0985	0.0947	0.0876
PV12	0.1653	0.1007	0.1017	0.1005	0.1015	0.0989	0.0902
PV13	0.1669	0.0947	0.0966	0.0948	0.0978	0.0951	0.0875
PV14	0.1296	0.0626	0.0654	0.0626	0.0658	0.0621	0.0575
PV15	0.1436	0.0684	0.0674	0.0661	0.0678	0.062	0.06
PV16	0.146	0.0762	0.0744	0.0756	0.0805	0.0745	0.0653
PV17	0.1253	0.0639	0.0668	0.0668	0.0828	0.0621	0.0573
PV18	0.1083	0.0561	0.0561	0.0582	0.0759	0.0564	0.058
PV19	0.1326	0.0653	0.0684	0.0646	0.0774	0.065	0.0585
PV20	0.1334	0.0807	0.0789	0.0814	0.0791	0.0765	0.0665
PV21	0.1029	0.0676	0.0726	0.0756	0.0665	0.0616	0.0716

D. EVALUATION

In this paper, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R2 score, and correlation coefficient are used to evaluate the performances of models. Furthermore, Forecast Skill (FS) score has been used to compare the models with the benchmark, i.e. the persistence model. Definition of FS score differs depending on literatures. This paper adopted the FS score from [46].

$$RMSE(y', y) = \sqrt{\frac{1}{N} \cdot \sum_{n=1}^N (x'_n - x_n)^2} \quad (15)$$

$$MAE(y', y) = \frac{1}{N} \cdot \sum_{n=1}^N |x'_n - x_n| \quad (16)$$

$$R2 \text{ score}(y', y) = 1 - \frac{\sum_{n=1}^N (x'_n - x_n)^2}{\sum_{n=1}^N (x_n - \bar{x})^2} \quad (17)$$

$$FS \text{ score} = 1 - \frac{RMSE(Model)}{RMSE(Persistence)} \quad (18)$$

where y' and y are predicted and the actual values respectively. Correlation-coefficient is the Pearson correlation-coefficient of the predicted and actual values.

IV. RESULTS

Table 3 and Table 4 show the comparison of the proposed model with different state-of-the-art models for day-ahead solar power forecasting. Table 3 shows the RMSE of all the models for each PV panel. From this table, it can be seen that the proposed model considerably outperforms all other models. Similarly, Table 4 shows the average values of RMSE, MAE, correlation coefficient, and R2 score of all models. RMSE and MAE errors indicate the losses. From Table 4 it can be seen that RMSE and MAE of the proposed model are the lowest among all the models. Similarly, R2 score and correlation coefficient refer to the accuracies. The R2 score and correlation coefficient of the proposed model is the highest as shown in Table 4.

In order to show the effectiveness of the proposed two-stage attention mechanism, the model has also been compared with various single attention mechanisms. SNAIL, Raffel, and Hierarchical attention are applied over LSTM hidden states to focus only on temporal sequences. An attention layer is applied to the input features in an Input-Attention model to focus only on important features. The comparison of the proposed model with various single attention mechanisms is

TABLE 4. Average values of RMSE, MAE, R2 Score, and correlation Co-efficient.

Error metrics	Models						
	Persistence	Simple LSTM	Input - Attention	LSTM Attention	CNN-LSTM	Ensemble Method	Proposed Model
RMSE	0.1229	0.0707	0.0723	0.072	0.0738	0.0685	0.0638
MAE	0.0605	0.0357	0.0358	0.0367	0.0378	0.0369	0.0324
R2 Score	0.5643	0.868	0.8648	0.8644	0.8525	0.874	0.8917
Correlation-Coefficient	0.7753	0.9313	0.9296	0.9294	0.923	0.9346	0.9441

TABLE 5. Average values Of RMSE, MAE, R2 score, and correlation Co-efficient of various attention models.

Error metrics	Attention Models				
	Input-Attention	LSTM - Raffel Attention	LSTM - Hierarchical Attention	LSTM - SNAIL Attention	Proposed Model
RMSE	0.0723	0.0721	0.0734	0.072	0.0638
MAE	0.0358	0.0364	0.037	0.0363	0.0324
R2 Score	0.8648	0.8654	0.858	0.8623	0.8917
Correlation-coefficient	0.9296	0.9299	0.9259	0.9295	0.9441

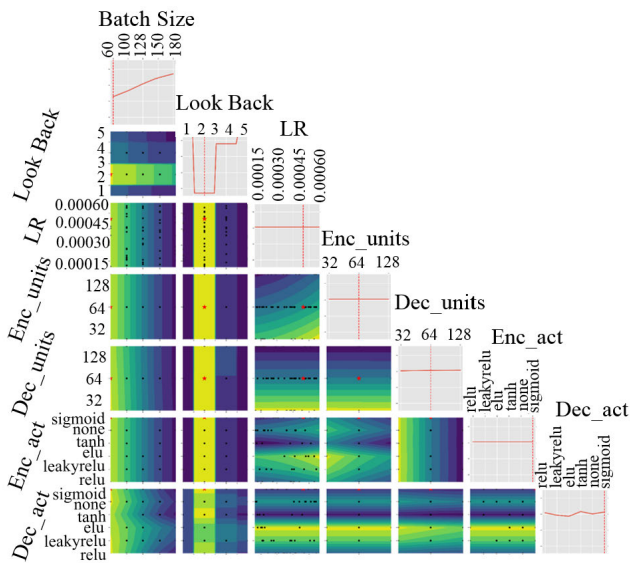


FIGURE 4. Convergence points of objective function of Bayesian optimization showing optimal combination of hyper-parameters.

given in Table 5. The combination of two-stage attention is highly efficient as compared to single attention mechanisms for day-ahead solar power forecasting, which can be seen from Table 5.

FS score is the criteria to check the performance of different forecasting models with respect to the persistence model. A higher value of FS score indicates the better performance of a model. The FS scores of all models are shown in Fig 5. The figure shows that the FS score of the proposed model is the highest among all the models. Among the other models, the ensemble model has the better FS score than the others.

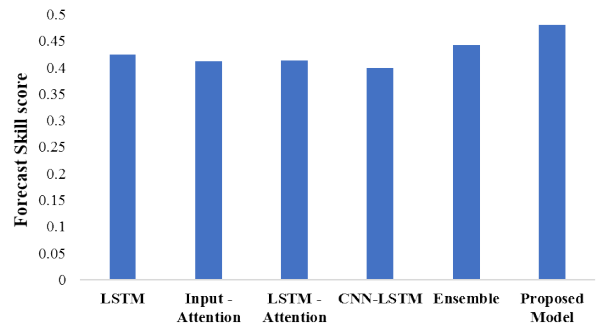


FIGURE 5. Forecasting Skill (FS) score of different models with respect to the persistence model.

The proposed paper considers 41 features. The performance of the model with respect to some important features such as hour of the day sine, month of the year sine, solar radiation direct, temperature, snowfall, and albedo is shown in Fig. 6. This figure shows that the input attention gives more weight to the input features which are more influencing on the target. It is obvious that the hour of the day has a very good correlation with the target power, which can be seen from Fig.6 (a). Similarly, Fig. 6 (b) shows the impact of the month of the year. Months in summer have a high impact on the output power, whereas winter months have the least impact. The temperature in summer is higher than that of winter, which also influences the power production as shown in Fig.6 (d). Solar radiation is the most correlated data among all inputs. The output power almost follows the trend of solar radiation as shown in Fig. 6 (c), unless some harsh weather conditions like snow or albedo are encountered, as shown in Fig. 6 (e) and (f). The model shows that the PV

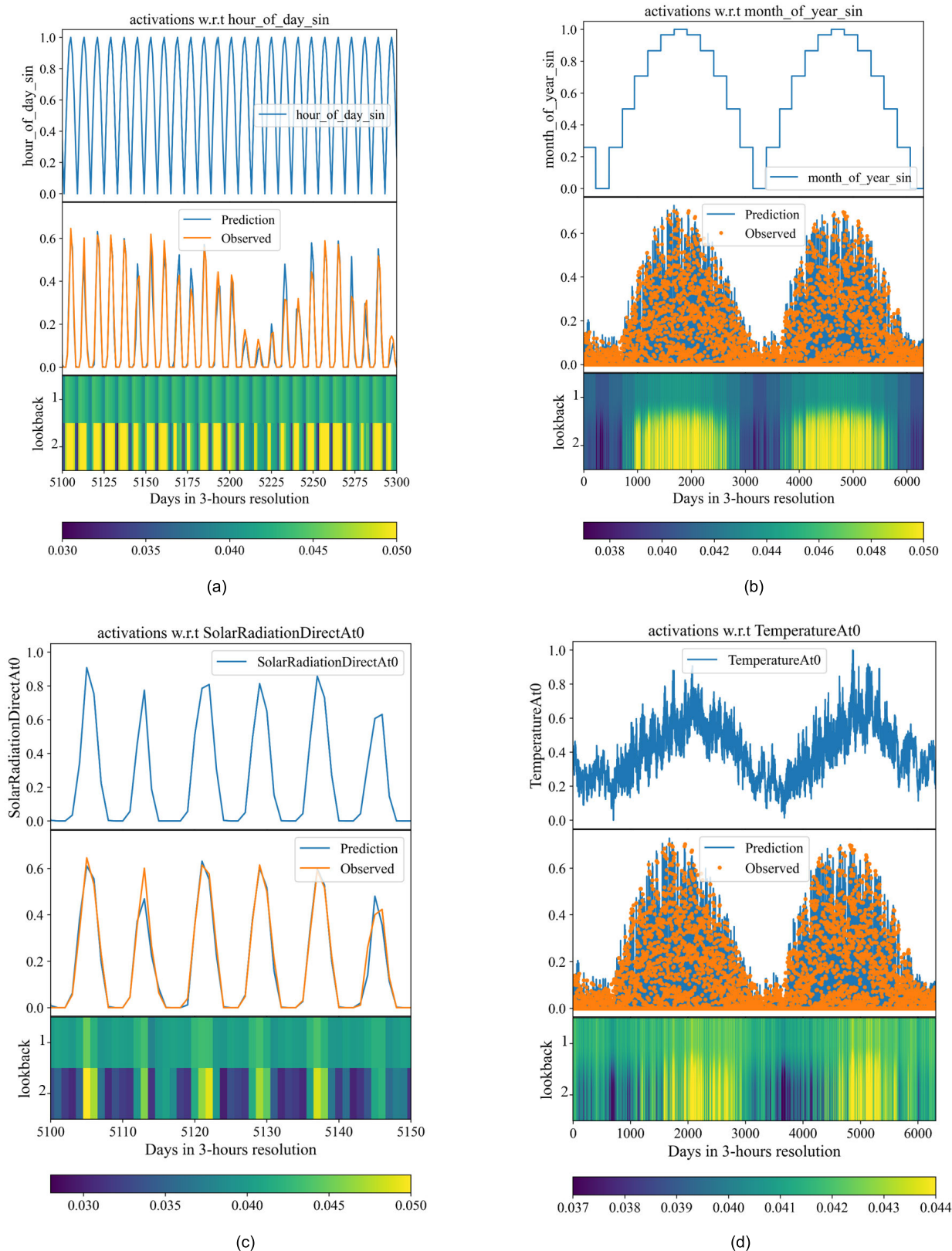


FIGURE 6. Model behavior with respect to some important input features and temporal lookbacks i.e. (a) Hour of the day sine, (b) month of the year sine, (c) Solar radiation direct (d) Temperature, (e) Snowfall, (f) Albedo.

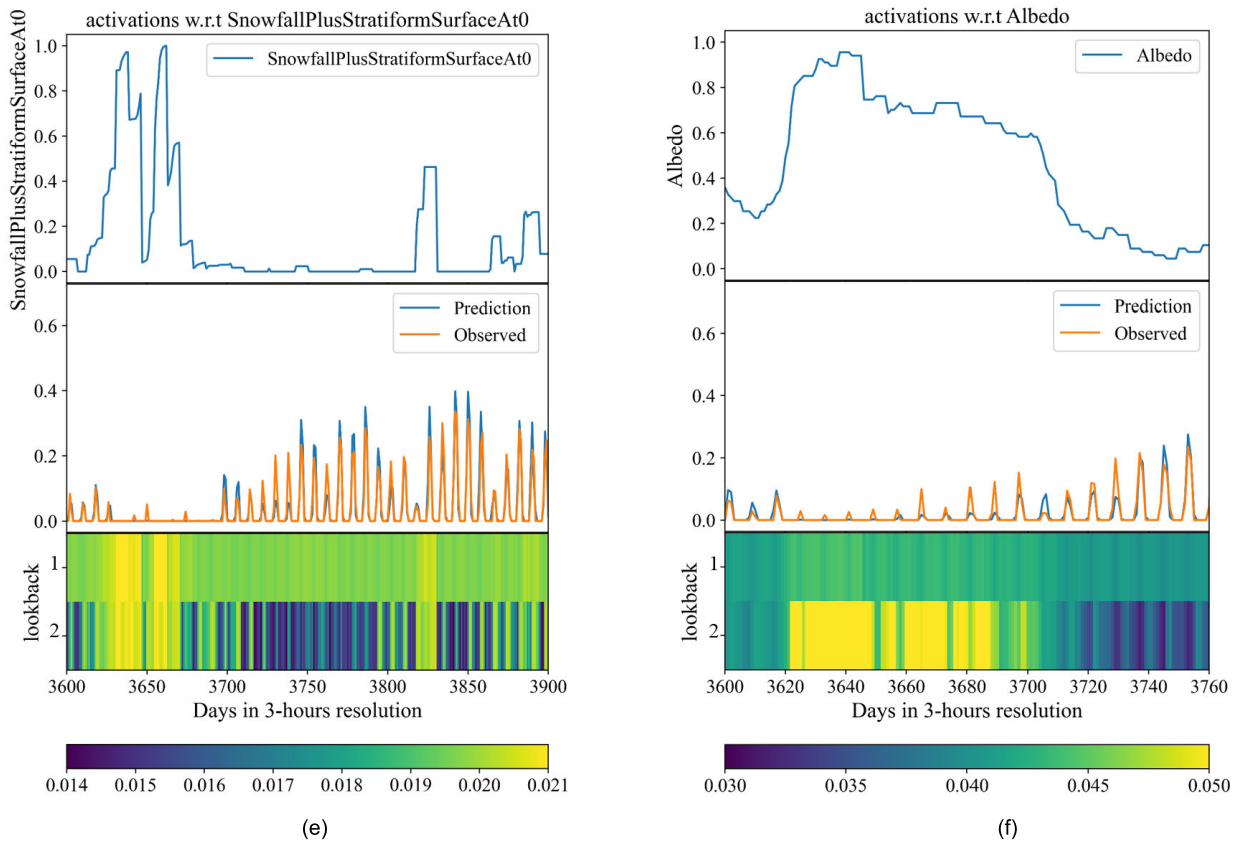


FIGURE 6. (Continued.) Model behavior with respect to some important input features and temporal lookbacks i.e. (a) Hour of the day sine, (b) month of the year sine. (c) Solar radiation direct (d) Temperature, (e) Snowfall, (f) Albedo.

TABLE 6. Time taken by each model for 200 epochs.

Models	Performance (seconds)
Simple LSTM	92.697
Input-Attention	49.8420
LSTM-Attention	111.592
CNN-LSTM	74.607
Ensemble method	28.555
Proposed model	155.928

performance decreases with the increase in the snow falling. Albedo, which accounts for the reflection from the panel, has also an effect on the output power. Fig.6 also shows the performance of the model with respect to temporal values. The model has given more weights when lookback is 2 as compared to when lookback is 1, which means the model is learning better when lookback is 2.

V. DISCUSSIONS AND CONCLUSION

Solar power forecasting is a time-series problem with non-linear relationship between inputs and targets. Traditional

methods either carry out linear mapping or lacks in handling long-term temporal dependencies. Although extensions of RNN such as LSTM with auto-encoders can handle long-term dependencies, the increase in the number of input features and long-sequences deteriorates their performance

In this paper, addressing aforementioned issues, day-ahead solar power forecasting has been carried out using a two-stage attention-based encoder-decoder model. The model applies two stages of attention over LSTM. At first, an encoder-based attention is applied to the input, which focuses on the important features at a particular time. At the second stage, a decoder-based temporal attention is applied to focus on important hidden states of the encoder. This combination of two stages of the attention mechanism with encoder-decoder model solves the time-series forecasting problems significantly, which can be seen from the results.

FS score shows the effectiveness of a forecasting model with respect to the persistence model. The FS score of the proposed model is better as compared to other models as shown in Fig. 5. Table 3 shows the comparison of RMSE of the proposed model with the state-of-the-art models for each PV panel. It can be seen from this table that the proposed model outperforms the traditional methods considerably. In this result, the ensemble method, which is the combination of various machine-learning techniques, has shown some better results than the others, due to the combined effect

of various models together. However, the accuracy of the proposed method is much better due to its consideration of all challenges related to the time-series. This effectiveness of the proposed method can further be seen from Table 4, where the average values of RMSE, MAE, R2 score and correlation coefficient of the proposed method are considerably better than the traditional methods.

Different techniques can be used to apply attention mechanisms (such as Raffel, SNAIL, and Hierarchical techniques). These attention mechanisms have been applied over LSTM hidden states to emphasize temporal attention mechanism only. In addition, Input-Attention model with attention only on the input features has also been considered, emphasizing only input feature selection. All these attention mechanisms are compared with the proposed method in Table 5 to show the effectiveness of the two-stage attention mechanism. From this table, it can be seen that the combination of input attention and temporal attention has high accuracy as compared to considering temporal attention or input attention only.

The two-stage attention mechanism focuses on more relevant input features as well as temporal hidden states, which can be seen from Fig. 6. This figure shows that during normal weather conditions, the output power is following the trend of features like hour of the day, solar radiation, etc., and these features are obtaining more attention weights accordingly. However, under extreme weather conditions like snowfall or albedo, the power production is almost zero and these features are getting weights accordingly. Similarly, it can also be seen from Fig. 6 that the model is giving more weights depending on more relevant temporal values. For instance, the model has given more temporal weights when the lookback is taken as 2, as compared to when the lookback is taken as 1. This means the model is learning better considering 2 lookbacks as compared to 1 lookback.

The paper has some limitations to be addressed as future work. The proposed model consists of two layers of attention with encoder-decoder layers over LSTM. The proposed model requires more layers and parameters to be trained as compared to other models. Therefore, the performance of the proposed model in terms of speed is lower as compared to the other models, which can be seen in Table 6. Furthermore, similar to all day-ahead forecasting models, this model also relies on forecasted weather data. Therefore, the accuracy is dependent on the accuracy of weather forecasters. The aim of the paper is to show that under given same conditions, the proposed model performs better as compared to other state-of-the-art models.

Since the proposed model is very efficient in forecasting, this model can be applied in different multi-horizon forecasting applications such as microgrid demand response forecasting considering market participations, and in electric vehicles load and charge or discharge forecasting. Considering the ability of the model to focus on more relevant features, it can also be applied to various applications like event management, fault identification, faulty equipment identification,

intrusion detection, and power disturbance classification, and so forth.

REFERENCES

- [1] IRENA. (2018). *Global Energy Transformation: A roadmap to 2050*, International Renewable Energy Agency, Abu Dhabi. [Online]. Available: https://www.irena.org/~/media/Files/IRENA/Agency/Publication/2018/Apr/IRENA_Report_GET_2018.pdf
- [2] (2021). *Global Energy Review 2020—Analysis—IEA*. IEA. Accessed: Mar. 11, 2021. [Online]. Available: <https://www.iea.org/reports/global-energy-review-2020>
- [3] A. Şağlam, B. Oral, and S. Görgülü, "Measurements of meteorological parameter effects on photovoltaic energy production," *Int. J. Circuits, Syst. Signal Process.*, vol. 9, pp. 240–246, Jan. 2015.
- [4] S. You, G. Kou, Y. Liu, X. Zhang, Y. Cui, M. J. Till, W. Yao, and Y. Liu, "Impact of high PV penetration on the inter-area oscillations in the U.S. Eastern interconnection," *IEEE Access*, vol. 5, pp. 4361–4369, 2017.
- [5] A. Yona, T. Senjyu, T. Funabashi, and C.-H. Kim, "Determination method of insolation prediction with fuzzy and applying neural network for long-term ahead PV power output correction," *IEEE Trans. Sustain. Energy*, vol. 4, no. 2, pp. 527–533, Apr. 2013.
- [6] L. Bird, J. Cochran, and X. Wang, "Wind and solar energy curtailment: Experience and practices in the United States," Nat. Renew. Energy Lab., Golden, CO, USA, Tech. Rep. NREL/TP-6A20-60983, Mar. 2014.
- [7] S. Sobri, S. Koochi-Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," *Energy Convers. Manage.*, vol. 156, pp. 459–497, Jan. 2018.
- [8] M. G. De Giorgi, P. M. Congedo, and M. Malvoni, "Photovoltaic power forecasting using statistical methods: Impact of weather data," *IET Sci. Meas. Technol.*, vol. 8, no. 3, pp. 90–97, May 2014.
- [9] M. Q. Raza, M. Nadarajah, and C. Ekanayake, "On recent advances in PV output power forecast," *Sol. Energy*, vol. 136, pp. 125–144, Oct. 2016.
- [10] J. M. Morales, A. J. Conejo, H. Madsen, P. Pinson, and M. Zugno, *Integrating Renewables in Electricity Markets*. New York, NY, USA: Springer, 2014.
- [11] G. Kariniotakis, *Renewable Energy Forecasting, From Models to Applications*, 1st ed. Amsterdam, The Netherlands: Elsevier, 2017.
- [12] M. N. Akhter, S. Mekhilef, H. Mokhlis, and N. Mohamed Shah, "Review on forecasting of photovoltaic power generation based on machine learning and Metaheuristic techniques," *IET Renew. Power Gener.*, vol. 13, no. 7, pp. 1009–1023, May 2019, doi: [10.1049/iet-rpg.2018.5649](https://doi.org/10.1049/iet-rpg.2018.5649).
- [13] "Forecasting of photovoltaic power generation and model optimization: A review," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 912–928, 2018, doi: [10.1016/j.rser.2017.08.017](https://doi.org/10.1016/j.rser.2017.08.017).
- [14] R. Ahmed, V. Sreeram, Y. Mishra, and M. D. Arif, "A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization," *Renew. Sustain. Energy Rev.*, vol. 124, May 2020, Art. no. 109792, doi: [10.1016/j.rser.2020.109792](https://doi.org/10.1016/j.rser.2020.109792).
- [15] A. M. Pavan and V. Lughì, "Photovoltaics in Italy: Toward grid parity in the residential electricity market," in *Proc. 24th Int. Conf. Microelectron. (ICM)*, Dec. 2012, pp. 1–4.
- [16] S. Daliento, C. Aissa, P. Guerriero, A. M. Pavan, A. Mellit, and R. Tricoli, "Monitoring, diagnosis, and power forecasting for photovoltaics: A review," *Int. J. Photoenergy*, vol. 2017, pp. 1–13, Jan. 2017.
- [17] M. Abuella and B. Chowdhury, "Solar power probabilistic forecasting by using multiple linear regression analysis," in *Proc. SoutheastCon*, Apr. 2015, pp. 1–5.
- [18] A. Alfadda, R. Adhikari, M. Kuzlu, and S. Rahman, "Hour-ahead solar PV power forecasting using SVR based approach," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Apr. 2017, pp. 1–5, doi: [10.1109/ISGT.2017.8086020](https://doi.org/10.1109/ISGT.2017.8086020).
- [19] V. D. William, E. Jamei, G. S. Thirunavukkarasu, M. Seyedmahmoudian, T. K. Soon, B. Horan, S. Mekhilef, and A. Stojcevski, "Short-term PV power forecasting using hybrid GASVM technique," *Renew. Energy*, vol. 140, pp. 367–379, Feb. 2019, doi: [10.1016/j.renene.2019.02.087](https://doi.org/10.1016/j.renene.2019.02.087).
- [20] L. Gigoni, A. Betti, E. Crisostomi, A. Franco, M. Tucci, F. Bizzarri, and D. Mucci, "Day-ahead hourly forecasting of power generation from photovoltaic plants," *IEEE Trans. Sustain. Energy*, vol. 9, no. 2, pp. 831–842, Oct. 2018.
- [21] M. Lotfi, M. Javadi, G. J. Osório, C. Monteiro, and J. P. S. Catalão, "A novel ensemble algorithm for solar power forecasting based on kernel density estimation," *Energies*, vol. 13, no. 1, p. 216, Jan. 2020, doi: [10.3390/en13010216](https://doi.org/10.3390/en13010216).

- [22] H. Sheng, J. Xiao, Y. Cheng, Q. Ni, and S. Wang, "Short-term solar power forecasting based on weighted Gaussian process regression," *IEEE Trans. Ind. Electron.*, vol. 65, no. 1, pp. 300–308, Jan. 2018, doi: [10.1109/TIE.2017.2714127](https://doi.org/10.1109/TIE.2017.2714127).
- [23] S. Masubuchi, E. Watanabe, Y. Seo, S. Okazaki, T. Sasagawa, K. Watanabe, T. Taniguchi, and T. Machida, "Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials," *NPJ 2D Mater. Appl.*, vol. 4, no. 1, pp. 1–9, Dec. 2020, doi: [10.1038/s41699-020-0137-z](https://doi.org/10.1038/s41699-020-0137-z).
- [24] M. Popel, M. Tomkova, J. Tomek, L. Kaiser, J. Uszkoreit, O. Bojar, and Z. Žabokrtský, "Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals," *Nature Commun.*, vol. 11, no. 1, pp. 1–15, Dec. 2020, doi: [10.1038/s41467-020-18073-9](https://doi.org/10.1038/s41467-020-18073-9).
- [25] C. Paoli, C. Voyant, M. Muselli, and M.-L. Nivet, "Multi-horizon irradiation forecasting for Mediterranean locations using time series models," *Energy Procedia*, vol. 57, pp. 1354–1363, Jan. 2014, doi: [10.1016/j.egypro.2014.10.126](https://doi.org/10.1016/j.egypro.2014.10.126).
- [26] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014, pp. 2104–2112.
- [27] M. Massaoudi, I. Chihi, L. Sidhom, M. Trabelsi, S. S. Refaat, H. Abu-Rub, and F. S. Oueslati, "An effective hybrid NARX-LSTM model for point and interval PV power forecasting," *IEEE Access*, vol. 9, pp. 36571–36588, 2021, doi: [10.1109/ACCESS.2021.3062776](https://doi.org/10.1109/ACCESS.2021.3062776).
- [28] E. Diaconescu, "The use of NARX neural networks to predict chaotic," *WSEA Trans. Comput. Res.*, vol. 3, no. 3, pp. 182–191, 2008.
- [29] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
- [30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [31] C.-H. Liu, J.-C. Gu, and M.-T. Yang, "A simplified LSTM neural networks for one day-ahead solar power forecasting," *IEEE Access*, vol. 9, pp. 17174–17195, 2021, doi: [10.1109/ACCESS.2021.3053638](https://doi.org/10.1109/ACCESS.2021.3053638).
- [32] M. Aslam, J.-M. Lee, H.-S. Kim, S.-J. Lee, and S. Hong, "Deep learning models for long-term solar radiation forecasting considering micro-grid installation: A comparative study," *Energies*, vol. 13, no. 1, p. 147, Dec. 2019.
- [33] Y.-Y. Hong, J. J. F. Martinez, and A. C. Fajardo, "Day-ahead solar irradiation forecasting utilizing Gramian angular field and convolutional long short-term memory," *IEEE Access*, vol. 8, pp. 18741–18753, 2020.
- [34] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: <http://arxiv.org/abs/1409.0473>
- [35] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.
- [36] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proc. EMNLP*, vol. 3, 2013, pp. 413–422.
- [37] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. SSST 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, 2014, pp. 103–111.
- [38] M. Aslam, J.-M. Lee, M. Altaha, S.-J. Lee, and S. Hong, "AE-LSTM based deep learning model for degradation rate influenced energy estimation of a PV system," *Energies*, vol. 13, no. 17, p. 4373, Aug. 2020.
- [39] 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.
- [40] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. W. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2627–2633.
- [41] C. Raffel and D. P. Ellis, "Feed-forward networks with attention can solve some long-term memory problem," in *Neural and Evolutionary Computing*, 2016.
- [42] Y. Tao, L. Ma, W. Zhang, J. Liu, and Q. Du, "Hierarchical attention-based recurrent highwat networks for time-series prediction," in *Proc. Stat.ML*, 2018, pp. 1–10.
- [43] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," in *Proc. Stat. ML*, 2018, pp. 1–8.
- [44] 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. (SMC)*, Oct. 2016, pp. 002858–002865, doi: [10.1109/SMC.2016.7844673](https://doi.org/10.1109/SMC.2016.7844673).
- [45] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. lei, and S. H. Deng, "Hyper-parameter optimization for machine-learning models based on Bayesian optimization," *J. Electron. Sci. Technol.*, vol. 17, no. 1, pp. 26–40, 2019.
- [46] N. Osorio, R. Escobar, R. Urraca, F. Martinez-de-Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Sol. Energy*, vol. 136, pp. 78–111, Oct. 2016, doi: [10.1016/j.solener.2016.06.069](https://doi.org/10.1016/j.solener.2016.06.069).



MUHAMMAD ASLAM received the B.S. degree in electrical engineering from the University of Engineering and Technology (UET), Peshawar, Pakistan, in 2013, and the M.S. and Ph.D. degrees in electrical engineering from Myongji University, Yongin, South Korea, in 2017 and 2021, respectively.

His research interests include application of AI in power systems, time-series analysis/forecasting, signal analysis, and image processing.



SEUNG-JAE LEE (Life Fellow, IEEE) received the B.E. and M.S. degrees in electrical engineering from Seoul National University, South Korea, in 1979 and 1981, respectively, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 1988. Since 2018, he has been serving as the Director for Next-Generation Power Technology Center (NPTC). He is currently working as a Professor with Myongji University. His major research interests

include protective relaying, distribution automation, and AI applications to power systems.



SANG-HEE KHANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Seoul National University, South Korea, in 1985, 1987, and 1993, respectively. He has been serving as the Director for Next-Generation Power Technology Center (NPTC), since 2019. He is currently serving as a Professor with Myongji University, South Korea. His research interest includes power system protection and control.



SUGWON HONG was born in Incheon, South Korea. He received the B.S. degree in physics from Seoul National University, Seoul, South Korea, in 1979, and the M.S. and Ph.D. degrees in computer science from North Carolina State University, Raleigh, USA, in 1988 and 1992, respectively.

His employment experience includes Software Development Center, Korea Institute of Science and Technology (KIST), Korea Energy Economics

Institute (KEEI), SK Innovation Company Ltd. (formerly Korea Oil Company), Electronic and Telecommunication Research Institute (ETRI). Since 1995, he has been with the Department of Computer Engineering, Myongji University, Yongin, South Korea, where he is currently a Professor. His current research interests include cyber security and smart grid.

• • •