EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Applying Artificial Intelligence in Forecasting the Output of Industrial Solar Power Plant in Vietnam

Ninh Nguyen Quang[1,*], Linh Bui Duy[1], Binh Doan Van[1], Quang Nguyen Dinh[1]

[1]Institute of Energy Science, Vietnam Academy of Science and Technology, Hanoi, Vietnam

Abstract

This paper uses recurrent neural network (Long Short – Term Memory - LSTM network) to build a model to forecast short-term generation capacity of Phong Dien solar power plant, (48 MWp – 35 MWAC) located in Thua Thien Hue Province, Viet Nam, with input factors including meteorological parameters. The authors conducted experiments to find the optimal structure of the model corresponding to the conditions of the plant and the data collection. Through this model, meteorological forecast data sets from commercial suppliers were used to forecast the plant's output power. The comments about the result as well as the further study direction are analysed and suggested.

Keywords: Long Short – Term Memory, Industrial PV power plant, Forecasting PV power, Artificial Intelligence.

Received on 10 November 2020, accepted on 24 March 2021, published on 29 March 2021

doi: 10.4108/eai.29-3-2021.169166

*Corresponding author. Email: nqninh@ies.vast.vn

1. Introduction

Vietnam is considered to have high solar potential with a lot of sunshine hours a year as mentioned in [1]. In recent years, solar power has had a strong boom in over the world. In Vietnam, since 2019, with strong incentives from the government, total installed capacity of PV power plants has increased rapidly and reached 4500 MWp [2]. Due to the uncertainty of solar source, in operation, both electricity system operators and the owner of industrial PV power plants need to know how many electric powers will be generated in next hour, next day. A forecasting method that could predict the output of the PV power plants based on influencing factors concerned as input data, will solve the problem. Recently, many forecasting techniques for generating capacity for solar PV systems have been developed and published. In [3], the authors have used two techniques to forecast the out-put power of a 6kWp PV system installed in a university in Malaysia. The results of Math processing machine learning SVR method (Support Vector Regression) and artificial neural network NAR method (Nonlinear Autoregressive) have been compared with the classical model. The results showed that the SVR method outperformed the NAR and the classical method in three typical weather conditions (clear, cloudy, and overcast). Jang in [4] has developed a new forecasting technique based on satellite images and SVM (Support Vector Machine). However, the results are not good enough due to the sporadic and random nature of the output power.

The statistical method is based on a set of observed values of one or more parameters measured at consecutive determined intervals [5]. This method includes many different types of prediction based on artificial intelligence algorithms (Artificial Neural Networks - ANN) such as Multilayer Perceptron (MLP) ([6], [7], [8]), Support Vector Machine (SVM) [9], Hybrid method [10], Markov string method, Fourier series, regression method… These methods rely only on data collected in the past to predict the generation capacity of solar power plants without requiring any information related to solar power plants such as panel capacity, number of panels... or location of construction area. Among the statistical-based forecasting models, ANNs forecasting models and regression models are currently the most widely used. According to many

studies, prediction based on ANN artificial intelligence model is one of the most effective methods. This is because ANN could adapt to drastic fluctuations in the input-output relationship due to varying environmental conditions [11]. However, the types of predictions using ANN model have the disadvantage of needing a large amount of data to serve the training process of the model, a set of values has been initially set up to start forecasting. This may reduce the reliability of the forecasting results, depending on the exact selection of the model's structure (the number of hidden layers, the number of neurons, etc.) [12].

The Auto-Regressive Moving Average (ARMA) model shows good performance when the data provided is static, while the Auto-Regressive Integrated Moving Average (ARIMA) performs forecasts for good results with uncertain time series data [13]. According to the calculations of the authors in [11], ANN models have better results than ARMA and ARIMA models and other statistical methods for accuracy and adaptability in unstable meteorological conditions. The grouping of weather conditions by day, [14] - [18], such as sunny days, rainy days, cloudy days will help improve the forecasting results of statistical models based on statistical methods [11], [8], [12], [19]. Statistical methods are mainly used in extremely short and short-term projections [12] and represent most forecasting techniques currently in use [20]. The Recurrent Neural Network (RNN) proposed by the authors in [7], [21], is a modified variant of the simple ANN model.

Finally, the authors in [13] developed the RNN model at a higher level, called the Long - short term memory model (LSTM). In the RNN model, the output of the previous step will be used as the input of the next step. RNN cannot forecast based on long-term historical data but may give accurate results from information that appears not too long ago. LSTM overcomes the disadvantages of RNN by retaining information for a long time. The authors in [22] compared forecasting models based on other neural network algorithms to the LSTM model, the results show that LSTM always give more accurate and stable results.

Most articles did not apply forecasts for large-scale solar power plants. The largest plant was considered is about 500kWp according to the authors in [14]. Moreover, articles do not consider the uncertainty, instability, and inaccurate forecast of weather factors.

In this study, we build a new forecasting model based on the LSTM network considering the uncertainty and weather forecasts to predict the power output of the solar power plant with a large capacity in Vietnam. Specific steps are as follows:

(i) Develop a forecasting model based on the LSTM network with suitable layer chosen to short-term forecast the output power of the solar power plant with training data set in the past operational data of large-scale solar power plant in central Vietnam.
(ii) Apply the developed model for forecasting power output of large-scale solar power plant in central Vietnam with the input data being weather data provided by the commercial provider.
(iii) Compare and analysis the archived forecasting results between two cases: (1) the input data is the past operational data, and (2) the input data is commercial weather data provided by a third party.

2. Long Short-Term Memory model

LSTMs are a special kind of RNNs that can learn short term as well as long-term dependencies [23]. Classical RNN networks often use past data to train the model and find the correlation between that data and the forecasted one. However, with the arrangement of sequential data series, as the length of the series increases, effect of the information that is far from the forecasted position tends to decrease quickly, although such information is sometimes especially important. That phenomenon is so called vanishing gradient [24]. LSTM overcomes the vanishing gradient problem by introducing memory cell and gated units [25], [26]. Each gate does the tasks as follows:

- Forget gate f_t decides what information to remember or forget from the previous block.
- Input gate i_t decides which values from the input to update the memory state based on specific conditions.
- Output gate o_t decides what to output from the memory of the LSTM block and input and with specific conditions.

As shown in Fig. 1, after receiving input sequence, the LSTM block controls each gate activating its inputs to decide whether they are triggered or not. This operation makes the change of state and addition of information that flows through the block conditional. The gates have weights that can be learned during the training phase. Indeed, the gates make the LSTM blocks smarter than classical neurons and enable them to memorize recent sequences.

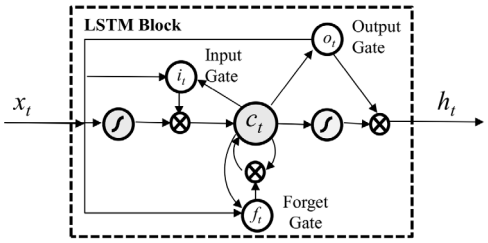

Figure 1. LSTM unit [23]

Each LSTM unit contains a cell which has a state c_t at time t. This cell can be considered as a memory unit. Reading/modifying this cell is controlled through the input gate it (a sigmoidal gate), forget gate f_t and output gate o_t. The LSTM unit receives inputs from two external sources at each of the four terminals (i.e., the three gates and the input) at each time step. The two external sources are:

- The current sample x_t.
- The previous hidden states of all LSTM units in the same layer h_{t-1}.

Each gate has an internal source, the cell state c_{t-1} of its cell block. The LSTM sums the inputs coming from different sources with a bias. The gates are activated by inputting their total input into the logistic function. The total input at the input terminal is passed through *tanh* nonlinearity. The LSTM multiplies the resulting activation by the activation of the input gate and then sums the result of the multiplication to the cell state after multiplying the cell state by the activation of the forget gate f_t. The LSTM passes the updated cell state through tanh nonlinearity and then multiplies it with the activations of the output gate o_t to determine the final output from the LSTM unit h_t. The previous steps and the updates of the LSTM unit can be formulated as follows [23]:

$$i_t = \sigma(W_i[X_t, h_{t-1}] + b_i) \quad (1)$$

$$f_t = \sigma(W_f[X_t, h_{t-1}] + b_f) \quad (2)$$

$$C_t = f_t C_{t-1} + i_t \tanh(W_C[X_t, h_{t-1}] + b_C) \quad (3)$$

$$o_t = \sigma(W_o[X_t, h_{t-1}] + b_o) \quad (4)$$

$$h_t = o_t \tanh(C_t) \quad (5)$$

3. Experiments and results

3.1. Input data

The historical data set was collected from Scada system of Phong Dien solar plant, during from December 2018 to January 2020 with the resolution of 5 minutes. Phong Dien solar plant is in Thua Thien Hue, Vietnam. Overall, the plant area is a rectangle whose length is around 1000 m and width is 485m. There is a station for measuring solar radiation and other meteorological indices located in the centre of plant. This plant has just started to operate by the end of 2018, and it is the longest available dataset of a solar farm all over Vietnam. Like other related papers, such as [14], one year data set of Phong Dien solar plant has been used for developing and checking the forecast model in this paper.

The collection of weather data includes global horizontal irradiance (GHI), panel temperature, air temperature, wind speed and humidity. The matching between input data for training and forecasting is considered carefully. With the resolution of 5 minutes, for Vietnamese regions, commercial suppliers of meteorological data can only provide GHI and air temperature. Due to this reason, in the training phase, two input meteorological parameters including GHI and air temperature were chosen to the model has ability to predict based on available forecasting data.

Table 1 illustrates the structure of dataset for modelling and testing. The unit of Global Horizontal Irradiance is W/m2. Air temperature is measured in Celsius. Actual power output is recorded in MW. Day of Year column has the integer value in the range of 1 to 365, Hour of Day is from 1 to 24 and Minutes of Hour columns receives value for every five minutes. The meaning of each data column is as follows:

- GHI: (W/m^2) Global Horizontal Irradiance.
- Air Temperature: (^oC) Ambient temperature at power plant location.
- Actual Output: (MW) Output power of solar power plant.

Table 1. Input dataset for training and testing

Times	GHI (W/m^2)	Air Temperature (^oC)	Actual output (MW)
…	…	…	…
6/30/2019 8:10	330.15	33.91	13.09
6/30/2019 8:15	496.06	34.29	19.6
6/30/2019 8:20	435.19	34.61	17.31
6/30/2019 8:25	367.21	34.83	15.1
…	…	…	…

When it comes to radiation measuring, because the number of sensors for detect sun light is limited, historical GHI in each sample cannot reflect completely GHI for every single solar panel in this time. Normally, the higher value of GHI, the more power that plant can generate. But in some cases, recorded data showed the reverse trend. Dataset of June 30, 2019 show in Fig.2 is an example.

Historical data of Phong Dien Solar farm on June 30, 2019 from 10:50 to 11:20 is shown in Fig.3. In 11:00, recorded GHI and output generation is 719.77 W/m2 and 25.76 MW, respectively. In 11:05, the GHI decreased around 5% to 685.13 W/m2 but the generation output almost remained constant round 25.75 MW.

This phenomenon is not a rare in the collected dataset. It is not a fault in metering and the reason can be explained by the illustration shown in Fig.4 below. When the cloud is in the t1 position, the sensor cannot measure the considerable reduction of GHI. After moving to the t2 position, the cloud causes a significant decrease in the metering data. But in both cases, in general, the impact of the cloud to the total generation output of the plant is the same.

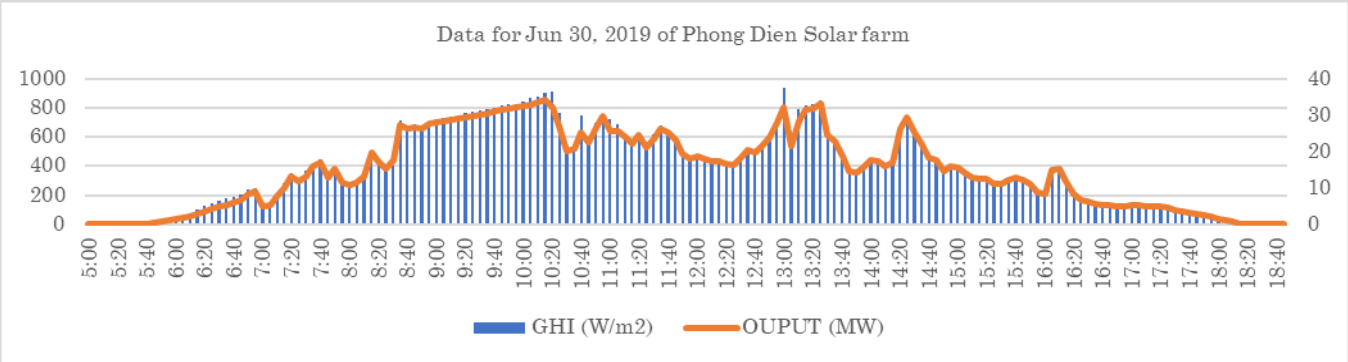

Figure 2. Five minutes historical data of Phong Dien Solar farm on June 30, 2019

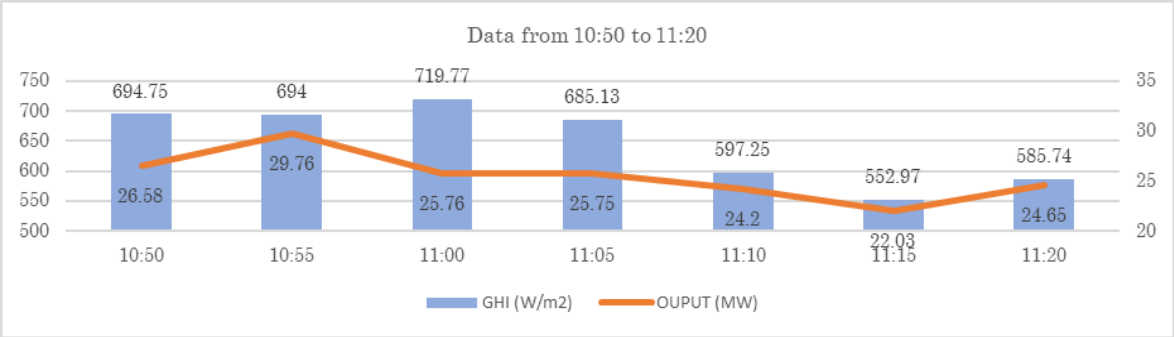

Figure 3. Historical data of Phong Dien Solar farm on June 30, 2019 from 10:50 to 11:20

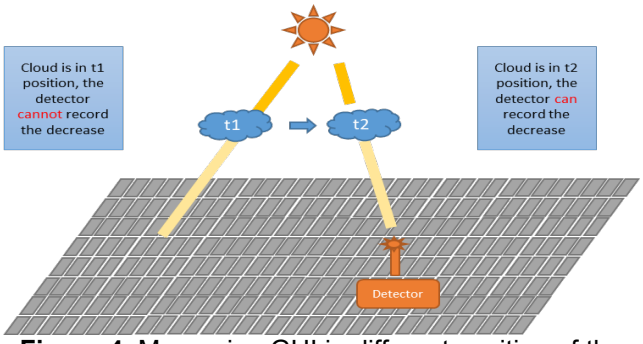

Figure 4. Measuring GHI in different position of the same cloud

With above analysis, we can see that GHI of previous observations has a hidden relationship with GHI of current observation and consequently indirectly affect to the output power. This shows the time-series characteristic of the dataset that we have.

3.2. Selecting the structure and hyper parameters of LSTM model

LSTM model has been proved to be one of the most suitable methodology to cope with the time-series problem. The next challenge for us is to find the best setting for the structure as well as the hyper parameters in LSTM model.

Regarding number of previous observations included in each input, we consider the wind speed data during a year of 2019 to support for our work. Fig.5 below shows the wind speed in the plant area.

The average speed of 2019 is 0.51 m/s. This relates to the moving speed of cloud around this area. If we assume that the cloud also moved with the same speed of wind, so it take 1960 second ~ 32.6 minutes to move 1000 m. This is corresponding to six previous intervals (each interval's duration is 5 minutes). Thus, the number of relative intervals from 1 to 6 will be tested in this experiment. In term of the structure, we will check the model with one, two, three and four layers of hidden unit to qualify the performance.

In the training process, we develop a model with hidden LSTM layers and an output layer. Each hidden layer contains 100 units. The number of units in the hidden layer is unrelated to the number of time steps in the input sequences. We will use the efficient Adam implementation of stochastic gradient descent and fit the model for 25

epochs. The mean squared error loss function is chosen for training.

In the testing phase, to evaluate the accuracy of the predict results, there are different commonly used metrics such as mean absolute error (MAE), absolute percentage error (APE), the mean absolute percentage error (MAPE), mean squared error (MSE), and root-mean-square error (RMSE). In this paper, the MAPE defined by equation (6) is used to evaluate the error in solar PV generation forecasting results because some periods in a day, the output generation is zero.

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\frac{|predict_i - actual_i|}{rate\ Capacity} \quad (6)$$

Where:

N: is the number of data point for error calculation.

$predict_i$: prediction of output power for data point i.

$actual_i$: Actual output power for data point i.

3.3. Results and analysis

Actual input data

To prove the effect and accuracy of the proposed model, the experiments were carried out with the input data is actual dataset, and using Python version 3.7.5, TensorFlow version 1.15 and Keras library version 2.2.4. Configuration of the computer is Intel Xeon 3.6GHz CPU, 8GB RAM, 64bit Window. In order to build the model and test the accurate of the model, similar to authors in [11], the actual dataset was divided into 2 parts. The first one was used to train the model while the remainder was considered as unknown observations to measure the accuracy of model. The length of the former set is 365 days to reflect full year context. The block diagram of the proposed procedure is shown in Fig.6.

After training and testing for all scenarios, the results are shown in Table 2 and Fig.7 as follows.

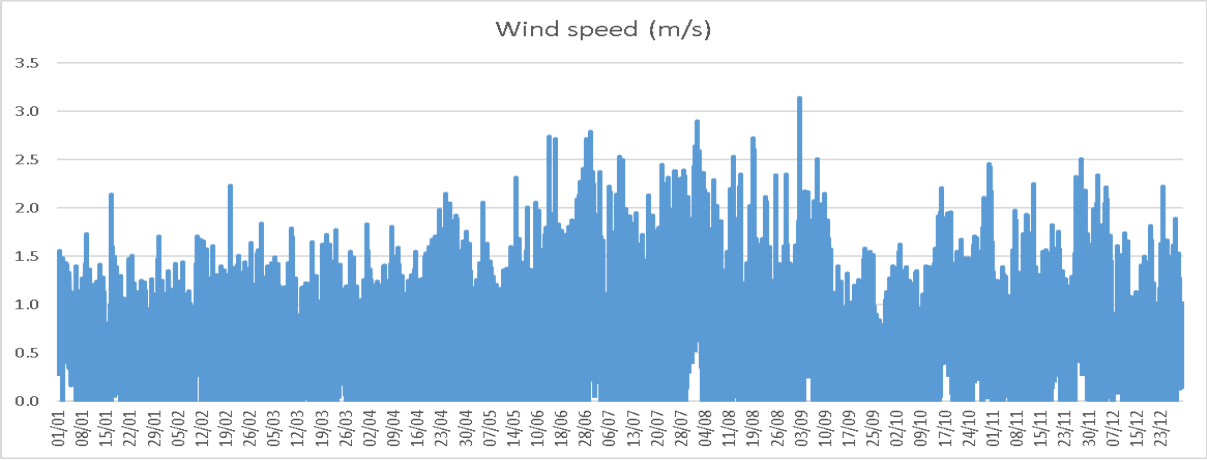

Figure 5. Historical wind speed at Phong Dien solar plant area in 2019

Table 2. Training time and testing result with different structure of LSTM model

Number of previous intervals for each input	1 hidden layer		2 hidden layers		3 hidden layers		4 hidden layers	
	Training time (second)	MAPE (%)	Training time (second)	MAPE (%)	Training time (second)	MAPE (%)	Training time (second)	MAPE (%)
1	152	1.41	275	1.19	500	1.13	667	1.18
2	188	**1.16**	348	**1.16**	650	**1.10**	843	**1.09**
3	222	1.21	413	1.22	763	1.12	1037	1.28
4	360	1.27	482	1.39	899	1.22	1168	1.30
5	288	1.24	536	1.30	1028	1.21	1366	1.23
6	444	1.38	600	1.35	1138	1.29	1517	1.18

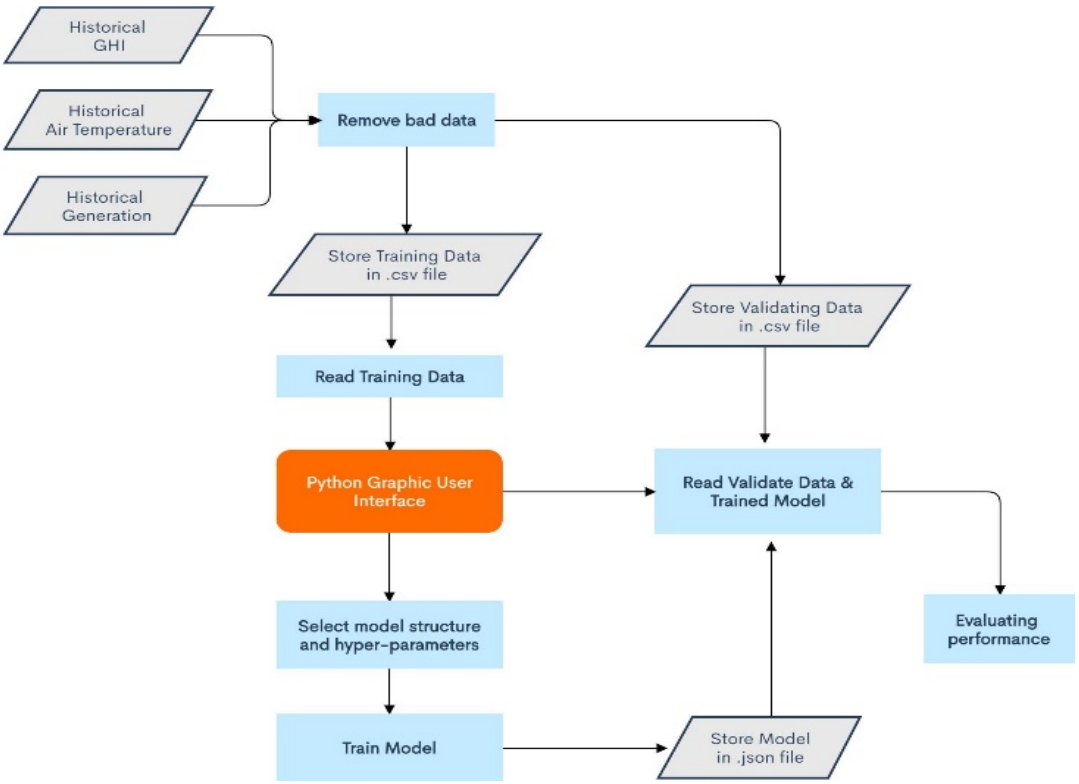

Figure 6. The block diagram of the proposed procedure

From Table 2, it can be clearly seen that with actual input dataset, the models show the accuracy and stable forecast results with the MAPE of all case lower than 1.5%. All the cases with two previous intervals of 5 minutes for each input give the results better than the others. And the model gives the best results (MAPE = 1.09%) in case we put 2 previous intervals of 5 minutes into each input and 4 hidden layers. This may be understood that previous meteorological factors within around 15 minutes has the greatest effect on current output power.

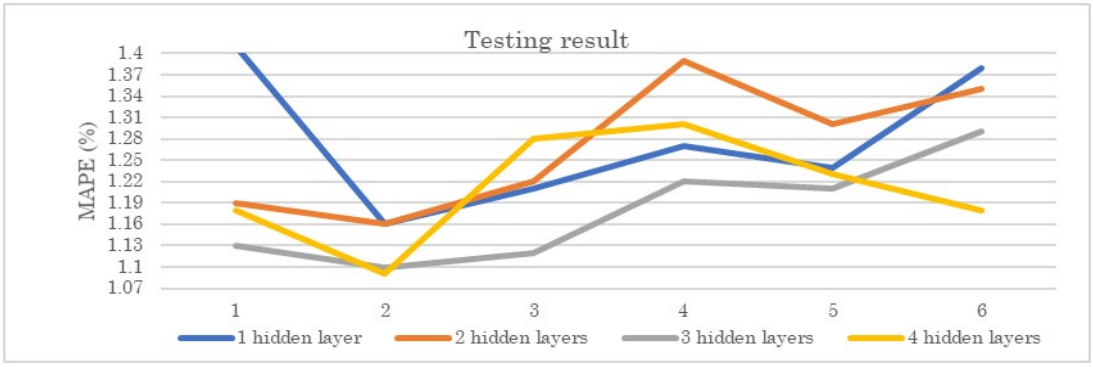

Figure 7. MAPE with different structure of LSTM model

As regards number of hidden layers, one-layer model and two-layer model have the same error percentage with 2 previous intervals input. For three-layer model, the performance improved dramatically in every single test compared with the two former ones. When the layers have been increased one more unit, the result was not improving considerably at only 0.01%. The experiment on 24 models, with different number of hidden layers, has shown a significant increase in training time as the model is more complex. But the rate of improvement of error in the best model is significantly reduced. The model with 2 previous interval input and four hidden layers has been chosen as the

final best one for Phong Dien Solar farm. The testing results of this model are shown in Fig.8. Through the graph, in case of actual input data, the predictive calculation data and the actual data are relatively close together, the number of large errors appears with a relatively small frequency. On some days, the error is barely noticeable.

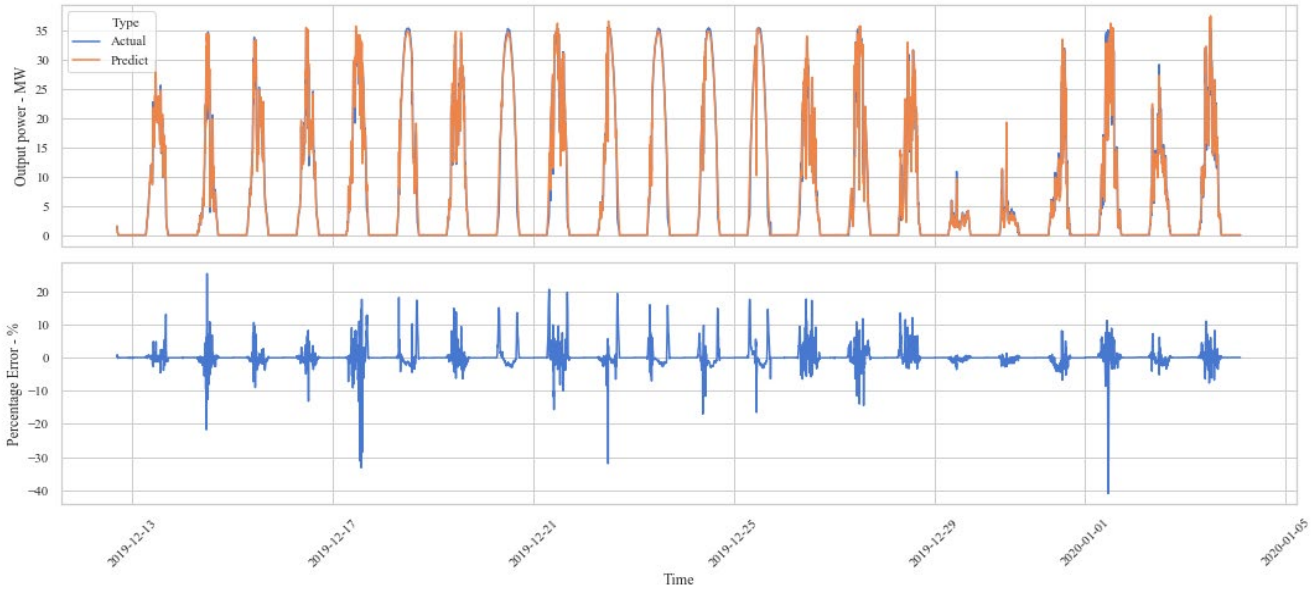

Figure 8. Detail 5-minute result with historical meteorological input

Prediction input data

In this part, the radiation and temperature prediction data set provided from an independent meteorological supplier has been used as prediction input data for the established model instead of using the actual input data in the past. The purpose of this experiment is to test whether the model works well on prediction input data or not. Due to in fact, most of the owners of solar plants must using this kind of input data for forecasting the output energy of plants. Experimental results obtained are presented in the Table 3 and Fig.9.

Table 3. Detail error result with actual weather data and forecast weather data

TYPE	UNIT	Predict with actual weather data	Predict with forecast weather data
MSE	MW^2	0.993	10.345
RMSE	MW	0.996	3.216
MAE	MW	0.380	1.551
MAPE	%	1.085	4.432
APE <5%	%	94.724	71.896
APE 5-10%	%	3.476	13.481
APE 10-20%	%	1.597	8.643
APE >20%	%	0.204	5.981

From the table 3, the error results in case of using actual input data are lower than the other such as: RMSE increased from 0.996 MW to 3.216 MW (increase 3.2 times), MAPE increased from 1.085% to 4.432% (increase 4.1 times), and the situations with APE being bigger than 20% also increased significantly from 0.204%, to 5.981%. The distribution of percentage errors of two cases has been shown in Fig.9. In case of using actual input data, the errors are distributed mainly around zero point and the number of big errors is a little. And in case of using prediction input data, big errors have been appeared.

The reason for the increase of such error metrics obviously may come from weather forecast errors, an uncontrollable factor for the owners of large-scale solar power plants. This problem has a big influence on the quality of the forecasting results. The raised problem in subsequent studies is to find ways to improve the forecast quality with prediction input data.

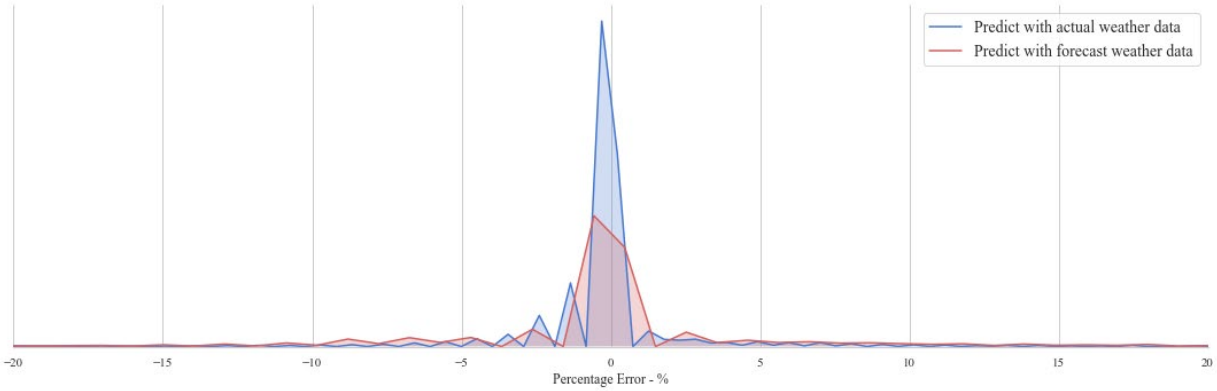

Figure 9. Distribution of Percentage Error

4. Conclusion

The paper presented method of applying LSTM algorithm to predict the output of large-scale solar power plants in Vietnam with consideration to the uncertainty of the weather. The two meteorological variables were considered as input data in this study are GHI and air temperature. The authors have been modelled, calculated, and compared to select the optimal number of hidden layers of LSTM model for large-scale solar power plants in Vietnam. The selected model has been tested in the actual operation of the large-scale solar power plant in Vietnam with the input meteorological data being supplied by commercial providers. The next steps should be concentrated to assess the root of errors and propose some potential approaches to solve errors affected by the weather forecast inputs.

Acknowledgements.

The authors wish to thank the Institute of Energy Science (IES), Graduate University of Science and Technology (GUST) and Vietnam Academy of Science and Technology (VAST) for their support to the research activity of Projects: “Research on methodology and develop software to short-term forecast the output of solar power plants based on Artificial Intelligence”, VAST07.01/21-22; “Building a tool to accurately forecast the energy output of a PV system in short-term in Vietnam”, QTIT01.03/20-21.

References

[1] Riva Sanseverino, E.; Le Thi Thuy, H.; Pham, M.-H.; Di Silvestre, M.L.; Nguyen Quang, N.; Favuzza, S. Review of Potential and Actual Penetration of Solar Power in Vietnam. *Energies* **2020**, *13*, 2529.

[2] EVNNLDC - Smart Grid Week Viet Nam 2019 – Day 3 Presentations. [Online] Available: http://smart-grid.vn/resources/.

[3] Kyairul Azmi Baharin; Hasimah Abdul Rahman; Mohammad Yusri Hassan; Chin Kim Gan, "Short-term forecasting of solar photovoltaic output power for tropical climate using ground-based measurement data," Journal of Renewable and Sustainable Energy 8, 053701 (2016); https://doi.org/10.1063/1.4962412.

[4] Jang HS, Bae KY, Park H-S, et al., "Solar power prediction based on satellite images and support vector machine," IEEE T Sustain Energy 2016; 7: 1255–1263.

[5] Montgomery, D.C.; Jennings, C.L.; Kulahci, M. Introduction to Time Series Analysis and Forecasting, 1st ed.; Wiley: Hoboken, NJ, USA, 2008.

[6] Mellit, A.; Massi Pavan, A. A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected {PV} plant at Trieste, Italy. Sol. Energy 2010, 84, 807–821.

[7] Mellit, A.; Shaari, S. Recurrent neural network-based forecasting of the daily electricity generation of a photovoltaic power system. Ecolog. Veh. Renew. Energy 2009, 2–7.

[8] Mellit, A.; Saglam, S.; Kalogirou, S.A. Artificial neural network-based model for estimating the produced power of a photovoltaic module. Renew. Energy 2013, 60, 71–78.

[9] Mellit, A.; Massi Pavan, A.; Benghanem, M. Least squares support vector machine for short-term prediction of meteorological time series. Theor. Appl. Climatol. 2013, 111, 297–307.

[10] Alfredo Nespoli, Emanuele Ogliari, Sonia Leva, Alessandro Massi Pavan, Adel Mellit, Vanni Lughi and Alberto Dolara, *Day-Ahead Photovoltaic Forecasting: A Comparison of the Most Effective Techniques,* Energies 2019, 12, 1621; doi:10.3390/en12091621.

[11] Raza, M.Q.; Nadarajah, M.; Ekanayake, C. On recent advances in PV output power forecast. Sol. Energy 2016,136, 125–144.

[12] Sobri, S.; Koohi-Kamali, S.; Rahim, N.A. Solar photovoltaic generation forecasting methods: A review. Energy Convers. Manag. 2018, 156, 459–497.

[13] Reikard, G. Predicting solar radiation at high resolutions: A comparison of time series forecasts. Sol. Energy 2009, 83, 342–349.
[14] Mellit, A.; Massi Pavan, A.; Lughi, V. Short-term forecasting of power production in a large-scale photovoltaic plant. Sol. Energy 2014, 105, 401–413.
[15] Shi, J.; Lee, W.J.; Liu, Y.; Yang, Y.; Wang, P. Forecasting power output of photovoltaic systems based on weather classification and support vector machines. IEEE Trans. Ind. Appl. 2012, 48, 1064–1069.
[16] Yang, C.; Thatte, A.A.; Xie, L. Multitime-scale data-driven spatio-temporal forecast of photovoltaic generation. IEEE Trans. Sustain. Energy 2015, 6, 104–112.
[17] Yang, H.-T.; Chao-Ming, H.; Huang, Y.-C.; Yi-Shiang, P. A Weather-Based Hybrid Method for one-day Ahead Hourly Forecasting of PV Power Output. IEEE Trans. Sustain. Energy 2014, 5, 917–926.
[18] Chen, C.; Duan, S.; Cai, T.; Liu, B. Online 24-h solar power forecasting based on weather type classification using artificial neural network. Sol. Energy 2011, 85, 2856–2870.
[19] Yona, A.; Senjyu, T.; Funabashi, T.; Kim, C.H. Determination method of insolation prediction with fuzzy and applying neural network for long-term ahead PV power output correction. IEEE Trans. Sustain. Energy 2013, 4, 527–533.
[20] Pelland, S.; Galanis, G.; Kallos, G. Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model. Prog. Photovoltaics Res. Appl. 2011, 9, 261–270.
[21] Mandal, P.; Madhira, S.T.S.; Ul haque, A.; Meng, J.; Pineda, R.L. Forecasting Power Output of Solar Photovoltaic System Using Wavelet Transform and Artificial Intelligence Techniques. Procedia Comput. Sci. 2012, 12, 332–337.
[22] Lee, D.; Kim, K. Recurrent Neural Network-Based Hourly Prediction of Photovoltaic Power Output Using Meteorological Information. Energies 2019, 12, 215.
[23] Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780.
[24] Razvan Pascanu, Tomas Mikolov, Y. Bengio, “On the difficulty of training Recurrent Neural Networks”, arXiv, November 2012
[25] Sepp Hochreiter, Jürgen Schmidhuber, “Long Short-Term Memory”, Neural Computation 9(8): 1735-1780, 1997.
[26] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke and J. Schmidhuber, "A Novel Connectionist System for Unconstrained Handwriting Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855-868, May 2009, doi: 10.1109/TPAMI.2008.137.