Data-Driven Day-Ahead PV Estimation Using Autoencoder-LSTM and Persistence Model

Yue Zhang, *Student Member, IEEE,* Chuan Qin, *Student Member, IEEE,*
Anurag K. Srivastava, *Senior Member, IEEE,* Chenrui Jin, *Member, IEEE,* and Ratnesh K. Sharma, *Member, IEEE*

***Abstract*—Inherent variability in photovoltaic (PV) and associated impacts on power systems is a challenging problem for both the PV owners and the grid operators. Existing statistical and machine learning algorithms typically work well for weather conditions similar to historical data. However, uncertain weather conditions pose a great challenge to the estimation accuracy of the estimation models. With the enhanced integration of intelligent electronic devices and the realization of associated automation in the power grid, renewable energy data is becoming more accessible, which can be utilized by deep learning models and improve the PV power generation estimation accuracy. In this paper, a hybrid deep learning model driven by external weather data is proposed to do day-ahead PV output forecasting at 15-minute-interval. The proposed model is motivated by the recent advancement of Long-Short-Term-Memory (LSTM) networks and AutoEncoder (AE), which estimates uncertainties in sequence while making the prediction for complex weather conditions. Meanwhile, the persistence model (PM) is used to predict continuous sunny weather conditions. The forecasting result is validated with data from multiple locations.**

***Index Terms*—PV power estimation, Deep learning, Data processing automation, Renewable energy integration, AE-LSTM, Hybrid model, Day-ahead forecasting.**

I. Introduction

THE integration of photovoltaic (PV) generation into the power grid has been growing rapidly over the past few years with increasing public awareness of using sustainable and reliable power. The global PV installations have exceeded 100GW in 2018 [1]. Some countries have also developed portfolios and policies to support the growth of solar energy. For example, the solar Investment Tax Credit (ITC) made by the U.S. federal government has helped the U.S. solar industry to grow by more than 10,000% since 2006, and the cumulative PV capacity has exceeded 76GW in 2019 [2].

From the perspective of smart grid development, short-term PV output prediction is one of the essential prerequisites to ensure the secure integration of PV generators. With the increasing PV penetration level in the grid, the uncertainties of PV generations bring many new challenges [3]. The uncontrollable uncertainty could increase the grid operating costs. NREL reported in [4] that each additional 100MW solar integration will result in an additional operating cost of 1$/MW per hour. To mitigate the impact raised by PV integration, accurate PV estimation methods are required by the system operator. The study in [5] shows that state-of-art PV power estimation methods can effectively reduce five billion dollars in operation cost each year. For PV owners, a better PV estimation method can also assist them to minimize the miss-bidding cost and increase energy trading revenue in the power market.

Our previous work in [6] introduced a hybrid forecasting model with a combination of Long-Short-Term-Memory (LSTM) and Persistence Model (PM) to provide day-ahead PV forecasting at a 15-minute time interval. We extende our past work in this paper by enhancing the LSTM model with the AutoEncoder (AE). Besides, locations with even complex weather conditions are used to validate the proposed model. The contributions of this paper are summarized as follows:

1) Proposed a new two-step approach for forecasting PV output driven by statistical (PM) and machine learning (LSTM) techniques. The robustness of the proposed deep learning model is boosted by AE-LSTM, which enhances the uncertainty estimation during the model training process.
2) In addition to the original datasets from Cupertino, CA, USA, datasets gained from Catania, Sicily, Italy are also used to validate this model. The estimation accuracy is constantly improved.
3) The new proposed model is eligible to cope with extreme adverse weather conditions compared with the previous model, i.e, reducing the uncertainties between the perceptron mapping, in response to the unstable forecasted consequences in some complex weather conditions.

II. Background and Related Work

The conventional PV energy estimation methods have been well reviewed in the literature. In [7], an Seasonal Auto Regressive Integrated Moving Average (SARIMA) is used to do 24 hours ahead estimation. Validated in a PV site in Greece, the normalized Root Mean Square Error (nRMSE) is 11.12%. A historical similar mining model is developed in [8] for 24 hours ahead power estimation, and the error is around 10.14%. A Support-Vector Machine (SVM) model is proposed in [9] with a Mean Relative Error (MRE) of 8.64%. Multiple adaptive regression models for day-ahead estimation, such as partial functional linear regression model, multivariate adaptive regression splines are respectively introduced in [10]–

Yue Zhang graduated from the Washington State University and now working at the GE Digital, Bothell, WA. Yue Zhang also worked as an intern with the NEC Labs America from May - August, 2018.

Chuan Qin and Anurag K Srivastava is with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA. (e-mail: anurag.k.srivastava@wsu.edu)

Chenrui Jin and Ratnesh Sharma worked with the NEC Labs America, Cupertino, CA, where some of the work reported here was conducted.

[12]. Each model is fed with the corresponding external weather datasets.

In recent years, enhanced automation provides more opportunities for researchers to develop a PV estimation model using deep learning architectures. A one-step-ahead Deep Believe Neural Network (DBNN) model is proposed in [13], which utilizes panel surface, ambient temperature, accumulated energy, and solar irradiation as input. The model is validated using two weeks measurements during summertime, and the Mean Square Error (MSE) of the testing and training results are 4.80% and 7.52% respectively. An additional DBNN model is introduced in [14]. This model uses solar radiation, temperature, and humidity as selected feature inputs to perform day-ahead estimation at 30-minute intervals. The Mean Absolute Percentage Error (MAPE) in the February testing dataset is 5.02%, and the MAPE in May is 8.92%. Furthermore, an LSTM day-ahead PV forecasting model with an nRMSE of 7.13% is proposed in [15]. In [16], a uni-variate input machine learning model is introduced. The historical PV power is fed into the model to predict one-step-ahead PV power generation. nRMSE is used to validate the accuracy of the model, which is around 2.7%. Another LSTM-based method has been designed for one-hour-ahead forecasting in [17]. This model uses irradiance, ambient temperature, and cloudiness index as input and validated the results using a 40kWp PV plant in Gumi. Authors in [18] study the hybrid model that combines conventional neural networks (CNNs) and LSTMs to do one day ahead prediction. Authors in [19] combine a variation mode decomposition (VMD) method with a CNN model to do short-term forecasting for a 100kWp plant in Nanjing. Authors in [20] combine wavelet packet decomposition (WPD) and LSTM to do one hour ahead forecasting with an average MAPE of 2.40%. Authors in [21] estimate temperature with stationary wavelet networks (SWN), extract historical powers with LSTM, and predict power with DNN.

In order to achieve high precision prediction of PV generation and inherit the advantages of different models, different models are0 tested in this work. In the final, the two models with the minimum error under different weather conditions are retained respectively to form a high-performance hybrid model. The architecture and validation of the proposed hybrid model will be described in this paper.

III. Methodology

The flowchart of the proposed PV power estimation method is demonstrated in Fig. 1. The forecasted weather data is passed through a classification module to determine whether the target date is a continuous sunny day. If the detection is true, the inputs data goes to a persistence sub-model for making the 24 hours ahead prediction. Otherwise, the data is fed into a trained AE-LSTM sub-model to create a new day-ahead estimation. Newly acquired weather data and recorded power data are used to update the dataset and retrain the model.

A. Persistence Model

PM is a simple and computationally effective method to execute time series forecasting, and the persistence can be interpreted as that the observed object exhibits periodic changes. In the future, the observed target will exhibit the same behavior as the current point in time. Once the PV panels are installed in the field, the generation efficiency factors regarding its physical properties, such as the tilt angle and affected surface area of the panel, will not likely to be changed in the near future. In addition, the distance between the earth and the sun does not change much over a few days, so the direct irradiation, diffuse radiation, and albedo radiation received by panels are approximately the same. Therefore, the PV generation will theoretically show the same characteristics in continuous sunny days.

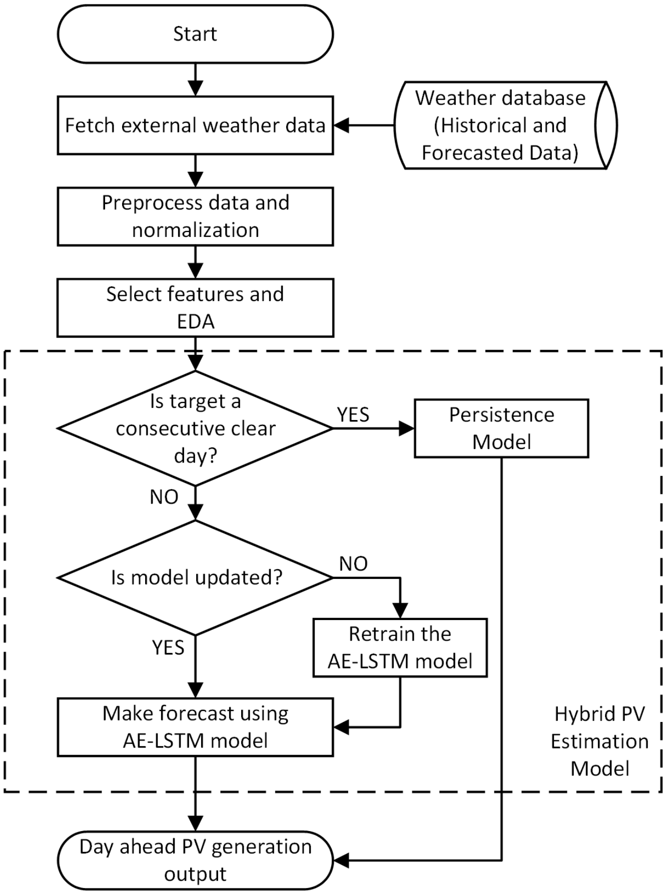
Start
Fetch external weather data
Weather database (Historical and Forecasted Data)
Preprocess data and normalization
Select features and EDA
Is target a consecutive clear day?
YES
Persistence Model
NO
Is model updated?
NO
Retrain the AE-LSTM model
YES
Make forecast using AE-LSTM model
Hybrid PV Estimation Model
Day ahead PV generation output

Fig. 1: The flowchart of the proposed method

Based on the properties of the stationary parts of PV, PM is proposed as a candidate prediction model in continuous sunny weather. The PM in this work can be simply defined as

$$\hat{P}(t+N) = P(t) \tag{1}$$

Where, $\hat{P}$ and P are forecast and measured PV power output at each timestamp t, respectively. N indicates the number of timestamps for each day.

B. AE-LSTM Model

The AE-LSTM model as shown in Fig. 2 combines AE and LSTM to generate PV prediction. In our work, the idea of adding AE to the LSTM network is to handle the training uncertainty given by the internal layer. The AE model has a bottleneck at the midpoint of the model to reconstruct the input data. The PV feature after AE processing is integrated with other weighted features as inputs and fed to the LSTM model for further training.

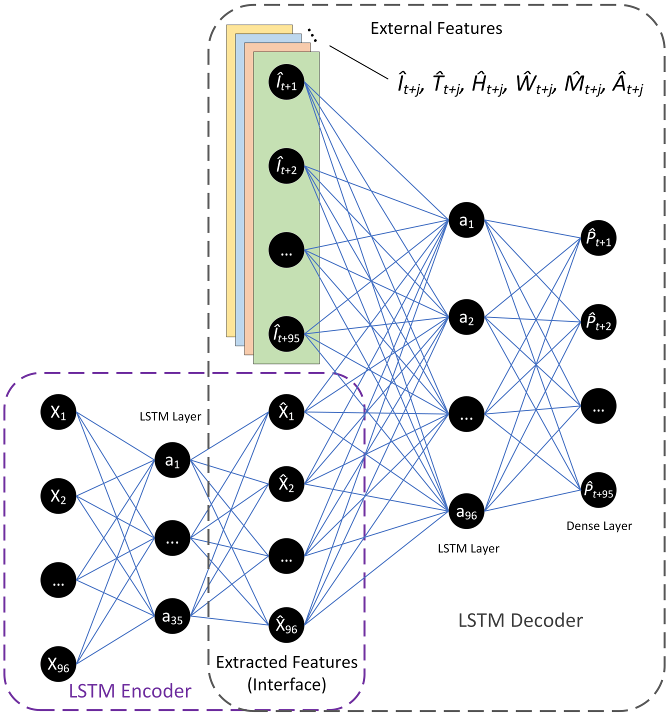

Fig. 2: The architecture of the proposed AE-LSTM network

LSTM has a recurrent architecture, which allows the system to recognize and predict sequences data. In addition to that, the LSTM can carry out tasks over long time series and discover long term features. Such long-term features may not be well captured due to the vanishing gradient problem [22]. With the help of input, forget and output gates, LSTM can hold only relevant data while 'forgetting' irrelevant information. For PV forecasting, LSTM can exploit the temporal and spatial dependence of data, which represents a major asset in utilizing contextual information.

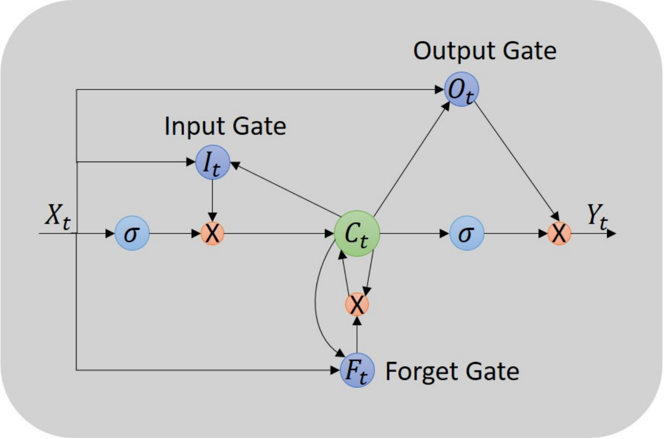

Fig. 3: The architecture of LSTM cell

The LSTM layer is composed of LSTM cells, whose structure is shown in Fig. 3. The main components of the LSTM unit are forget, input and, output gates. Each of those gates determines the portion of information to forget, to update, and to output through a σ function, where the output varies from 0 to 1. The inputs include current sample X_t and the output from the three gates $(Y_{t-1}, C_{t-1}, O_{t-1})$ in the previous iteration. The detailed output updating process is based on Equations (2) - (6), where $W_{k=1,\ldots,11}$ and $B_{l=1,2,3,4}$ are the weights and bias.

$$F_t = \sigma(W_1 X_t + W_2 Y_{t-1} + W_3 C_{t-1} + B_1) \quad (2)$$
$$I_t = \sigma(W_4 X_t + W_5 Y_{t-1} + W_6 C_{t-1} + B_2) \quad (3)$$
$$O_t = \sigma(W_7 X_t + W_8 Y_{t-1} + W_9 C_{t-1} + B_3) \quad (4)$$
$$C_t = F_t C_{t-1} + I_t \tanh(W_{10} X_t + W_{11} Y_{t-1} + B_4) \quad (5)$$
$$Y_t = O_t \tanh(C_t) \quad (6)$$

Extreme variability in the area of observed features is critical to the time-series PV power estimation. The estimation model based on deep learning network usually depends on the initialization of internal weights. The uncertainty from the neural network model, such as model specification errors and the inherent noises, will make the estimated results drift. The object with using AE-LSTM is to further quantify those uncertainties. In the preliminary test, no deterministic methods are used to select the optimal parameters for the model, and the estimation model is tuned by trial-and-error using the Keras python package [23]. The model compilation and training are optimized by Adam optimizer, and the MSE metric is used to reconcile the loss function. The sigmoid function is used as the activation function.

The training process of the AE-LSTM model includes two parts, feature selection, and model tuning. Even though one of the advantages of AE is extracting important features automatically, the application of feature selection in the PV estimation field is still critical. It can prevent models from overfitting and enable users to solve the prediction problem with limited data and computing resources [24]. For this paper, the Root Mean Squared Euclidean Distance Difference (RMSEDD) is utilized to help extract features for the training. RMSEDD for each feature v_i is defined as:

$$\text{RMSEDD}_i = \frac{\sqrt{\sum_{d'=d}^{G} \sum_{d=1}^{d'} \left(\text{ED}(P, d, d') - \text{ED}(v_i, d, d')\right)^2}}{\frac{1}{2} G(G-1)} \quad (7)$$

$$\text{ED}(x, d, d') = \sqrt{\sum_{t=1}^{N} \left(x_t^{(d)} - x_t^{(d')}\right)^2} \quad (8)$$

Where, $\text{ED}(x, d, d')$ measures the ED between day d and d' based on normalized variables x, which include normalized i_{th} feature v_i and normalized PV output P. N is the total number of time points in a day. G indicates the number of training days. Compared with the correlation coefficient, RMSEDD has the advantage of discovering the latent non-linear correlation between input features and outputs.

C. Hybrid Model

The proposed hybrid model aims to combine the strength of each selected sub-model and develop the suitable criteria for sub-model selection. To select suitable sub-models from a variety of choices, the features of the forecasting model and PV data are both analyzed. For example, LSTM and RNN use similar forecasting techniques, there is no need to develop two sub-models for each of them. PM is the simplest estimation approach that duplicates the observed quantity of day d as the estimation for the day $d+1$, but provides good estimation results for a consecutive sunny day. Therefore, it is also selected as a sub-model. To take merit of PM under this circumstance, a mechanism driven by external forecasted weather is designed to determine the estimation model selection. For an instance, if the ED$(\hat{I}, d, d+1)$ is less than a certain threshold, day $d+1$ and d are most likely two clear days. Those continuous sunny days are retrieved and fed to train the value of the threshold using the method of grid search [25].

IV. Simulation Environment

In this work, three different datasets are used to evaluate the proposed hybrid model. The platform has an Intel(R) 2.2GHz I7-8750H CPU, 16GB RAM, and NVIDIA GeForce GTX 1070 with 8GB GDDR5 RAM.

A. Data Description

The first dataset is collected from the PV power station in Cupertino, (37.32N,122.01W). The capacity of the PV installation on this site is 6.41kW including twenty-one 305W SPR-305-WHT PV panels. The incline angle of the PV panels on the site is set to zero. The time interval for the sensor to measure the power is 15 minutes, and the data is saved from July 1, 2015, to December 31, 2016. The measurements are collected along with forecast weather data from the NAM numerical weather prediction model contributed by the National Oceanic and Atmospheric Administration (NOAA) [26]. NOAA publishes updated weather prediction values four times each day at the following times: 00 UTC, 06 UTC, 12 UTC, and 18 UTC. Referring to the literature [27], [28], five related weather features are chosen among the dozens published by NOAA: temperature $\hat{T}$, relative humidity $\hat{H}$, wind speed $\hat{W}$, total cloud cover $\hat{C}$, and solar radiation $\hat{I}$. In addition to NOAA features, one weather feature and two time-series characteristics are computed and added to the dataset for analyzing purposes. The weather feature is the solar zenith angle ($\hat{A}$) that is introduced in [28]. The time-series characteristic features are polarized minute and day index ($\hat{M}$ and $\hat{D}$), which are mentioned in [29].

Sunny days are the most common type of weather at this test site. According to the normalization of the average PV output of sunny days, the average power outputs under different weather types are summarized in Table I. The lowest average power outputs under rainy days as the table mentioned are 52.74%, which poses a critical challenge for making PV power estimation.

TABLE I: Average PV power output rate under different weather conditions

Weather Type	Mist	Clouds	Rain	Haze	Fog
Reduction Rate	78.58%	79.53%	52.74%	78.98%	54.46%

The second testing dataset is a 5.21kW PV site at Catania, Sicily, Italy (37°32′ 0″N, 15°5′ 25″W). Forecast data, including solar zenith angle, solar radiation, total cloud cover, and ambient temperature are provided by the regional atmospheric modeling system. For this site, the data covers the period from January 1, 2011 to December 31, 2011.

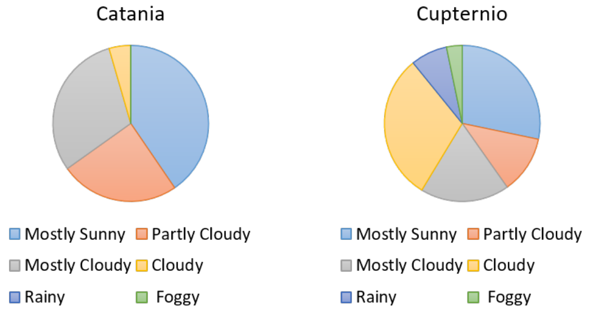

Fig. 4: Daily weather type distribution over three months testing period at Cupertino and Catania

Fig. 4 show the daily weather distribution at each site for testing periods from October to December. The parameters for weather-type classifications (e.g., sunny, cloudy, and foggy) are employed based on the definitions provided by Weather Underground [30]. The weather begins to get cloudier from the middle of September in Catania and get cloudier from late October in Cupternio [31].

B. Evaluation Metrics

In this work, the daily nRMSE and nMAE for tested days are used to assess the estimation model accuracy. The nRMSE and nMAE are defined as:

$$\text{nRMSE} = \frac{1}{P_c}\sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{P}_t - P_t)^2} \tag{9}$$

$$\text{nMAE} = \frac{1}{N \cdot P_c}\sum_{t=1}^{N}|\hat{P}_t - P_t| \tag{10}$$

Where, P_c is the maximum capacity of the PV power generation. P_t is PV power measurement at timestamp t, and $\hat{P}_t$ is its corresponding estimated value. For 24 hours ahead estimation with a resolution of 15 minutes, the sampling rate $N = 96$. According to [15], the forecast bias tends to be greater if nRMSE is much higher than nMAE.

V. Simulation Results and Analysis

A. Forecast Model Training

The AE-LSTM model is tuned to have an appropriate model size and structure to gain a balance between optimization and generation. Feature engineering is also utilized to enhance the performance of the model.

1) Feature Selection: Even though a modern deep learning approach can automatically extract useful features from raw data, a good feature selection is still needed in the field of PV forecasting. Practically, it is not economical to wait for several years to collect data and train a model. Selecting proper features can help to solve a problem with far fewer data [32]. In this paper, RMSEDD is utilized to help extract features for the training.

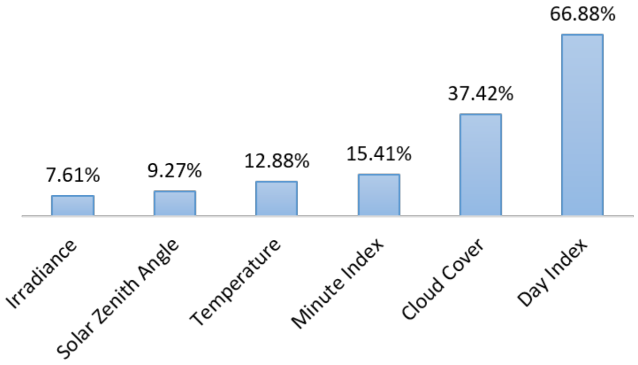

Fig. 5: The RMSEDD for all input features at Catania

The RMSEDD for all the available features is first calculated as shown in Fig. 5. The lower RMSEDD value indicates that the feature has a more similar trend as the PV output. As seen from the figure, highly related features like solar radiation and solar zenith angle have RMSEDD values that are less than 10%; while less relevant features such as polarized day index can have an RMSEDD value greater than 60%.

TABLE II: The impact of feature selection on estimation accuracy

Selected Input Features	nRMSE	nMAE
P, $\hat{I}$	9.93%	5.71%
P, $\hat{I}$, $\hat{A}$	8.99%	5.04%
P, $\hat{I}$, $\hat{A}$, $\hat{T}$	8.47%	4.60%
P, $\hat{I}$, $\hat{A}$, $\hat{T}$, $\hat{C}$	9.09%	5.15%
P, $\hat{I}$, $\hat{A}$, $\hat{T}$, $\hat{C}$, $\hat{M}$, $\hat{D}$	8.74%	4.91%

Based on the RMSEDD for each feature, different input feature combinations have been tested, ranging from one input feature with the lowest RMSEDD to a combination of all available features. As shown in Table II, utilizing only the most relevant feature or utilizing all the available features as inputs do not train the best estimation model. Since the values of previous days' PV power outputs (P) are utilized as a training input feature in many works in the literature, this paper also tests the effect of including them. As seen from the results, a model with feature P, $\hat{I}$, $\hat{A}$, $\hat{T}$ perform better than others.

2) Model Tuning: The AE-LSTM is tuned to achieve good forecasting accuracy as well as good regularization. The key parameters of the model are determined by trial-and-error. Reducing the size of the model is the simplest way to prevent over-fitting. The forecasting error is evaluated over different model size and the average and standard deviation (Std) of forecasting error for Catania in term of nRMSE is presented in Table III and Table IV. As seen from the table, the network with 5 layers and 400 hidden units in each layer performs better than other network setups. For 100 epochs, the training time for the selected network is 16 minutes. In comparison, the training time for the network with 10 layers and 1000 hidden units is about 55 minutes.

TABLE III: Average forecast error for models with different size in term of nRMSE

Neurons\Layers	2	3	5	6	10
200	11.44%	11.23%	10.52%	10.81%	19.31%
400	10.74%	10.98%	9.82%	10.58%	18.80%
500	12.26%	10.50%	10.37%	10.20%	19.10%
700	11.10%	10.44%	11.31%	11.03%	18.76%
1000	11.64%	9.93%	10.91%	10.93%	18.93%

TABLE IV: Std of forecast error for models with different size in term of nRMSE

Neurons\Layers	2	3	5	6	10
200	6.73%	6.62%	5.59%	6.30%	6.76%
400	6.47%	6.42%	5.68%	6.65%	6.50%
500	6.76%	6.07%	5.89%	5.95%	6.63%
700	6.25%	6.13%	6.77%	6.26%	6.53%
1000	6.48%	5.85%	6.37%	6.36%	6.62%

Once the key parameters of the model are determined, dropout regulation is implemented to overcome the ”over-fitting” issue. During dropout regulation, some randomly selected neurons are zeros out during training. Since some neurons are dropped out, other neurons have to adapt to handle the representation required to do predictions due to the missing neurons [33]. As shown in Fig. 8, the original network without applying dropout regulation starts over-fitting after 5 epochs. In contrast, the networks that applying the dropout technique has become more resistant to over-fitting than the original network. Through simulation, a 50% dropout rate yields better anti-over-fitting performance.

B. Evaluation of the AE-LSTM Model

To validate the forecasting accuracy of the AE-LSTM model, several benchmark models include PM, feed-forward neural network (FNN), deep neural network (DNN) have been built for comparison. FNN and DNN use the same input features as the AE-LSTM model. FNN has one hidden layer and DNN has five hidden layers. The forecast power output from different models over eight days are plotted in Fig. 6 and the daily weather type and average daily forecasting error for each testing day are presented in Fig. 7. The forecasting error from PM is very low on October 2 and 3 since October 1, 2, 3 are continuous sunny days. On the other hand, PM produces a very large forecasting error on October 9, since the weather condition on October 8 is very different from the

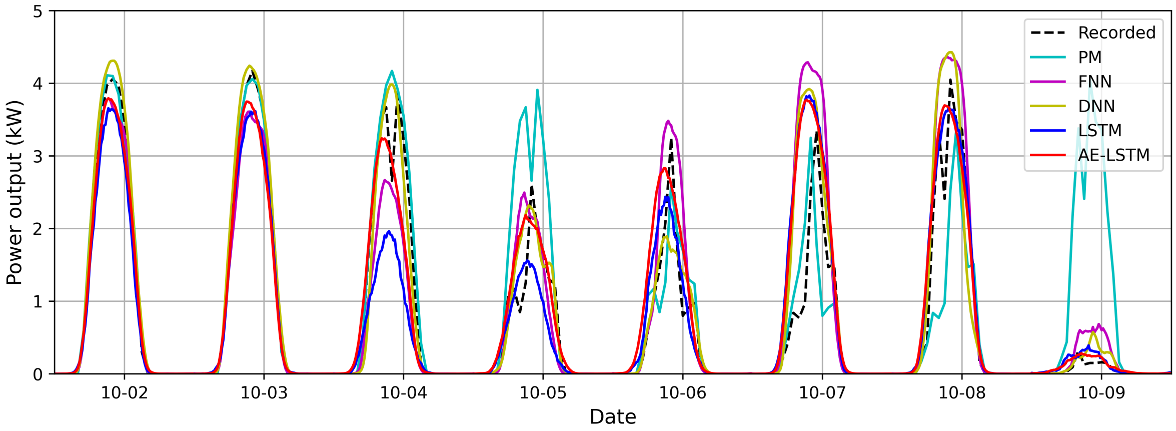

Fig. 6: Forecast output comparison at Catania from October 2 to October 9

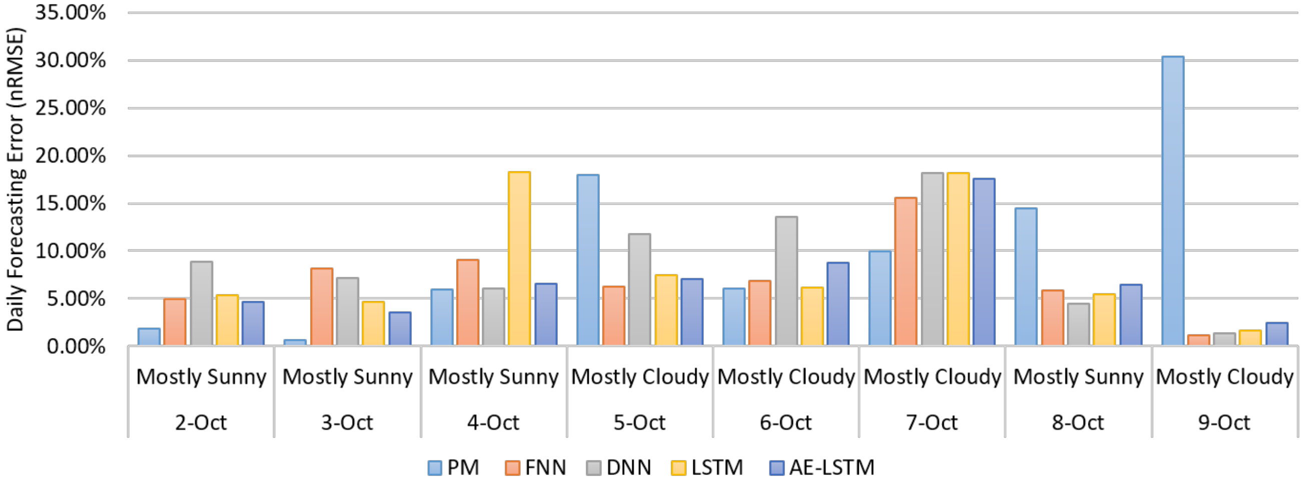

Fig. 7: Daily forecast error and weather type at Catania from October 2 to October 9

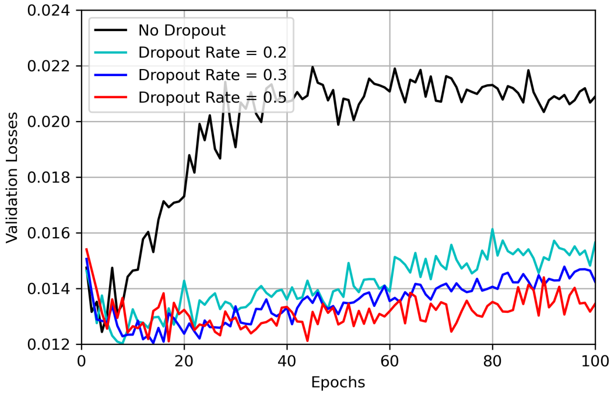

Fig. 8: Impact of dropout rate on the validation loss

weather on October 9. In comparison, other machine learning-based models have a relatively stable performance and the AE-LSTM model has an overall best performance. Noticing that the weather type is a general weather description for a certain day in a region, days with the same weather type could show different power output pattern. For example, October 8 is also recorded as a sunny day, but the power output pattern is significantly different from other sunny days' output. Similarly, the power output from some cloudy days is different as well. For example, October 8 and October 3 are both recorded as a mostly sunny day, but the output pattern is not the same. Similarly, power output under mostly cloudy shows different output patterns as well. Therefore, the overall forecast accuracy needs to be evaluated for a longer period to validate the performance. In Table. V, we calculated the forecasting error for three months from October 2 to December 31. The simulation result shows the proposed AE-LSTM model is more accurate than other benchmark models.

C. *Evaluation of the Hybrid Model*

Due to the nature of PV generation, the PM can perform very well for consecutive clear days: as an example shown in Fig. 9, the nRMSE is only 0.72%. While for the same day, the LSTM model generates an error of 3.16%. Fig. 10 provides estimation results for days with the highest and lowest PM

TABLE V: Forecast error from different models at Catania

Model	nRMSE		nMAE	
	Mean	Std	Mean	Std
PM	12.83%	6.59%	6.83%	3.74%
FNN	11.55%	5.70%	6.27%	3.29%
DNN	10.37%	5.52%	5.56%	3.10%
LSTM	8.87%	5.34%	4.78%	3.05%
AE-LSTM	8.39%	5.26%	4.56%	2.93%

error. November 19, 2016, is a rainy day, but the PM model considers it as a sunny day because the weather was sunny in the day before. So PM at the time produces the worst performance. However, LSTM, in this case, can capture the features from forecasted weather data to predict a much better estimation result from this transition. On the other hand, on the two continuous clear days before December 2, 2016, LSTM did not perform as well as persistence. The reason is that LSTM model optimization aims to reduce the overall error from supervised learning, so the performance on that day is slightly worse than PM.

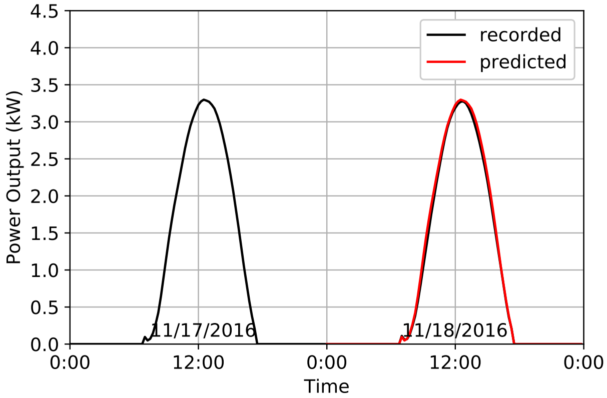

Fig. 9: Forecast power output from PM on two consecutive sunny days

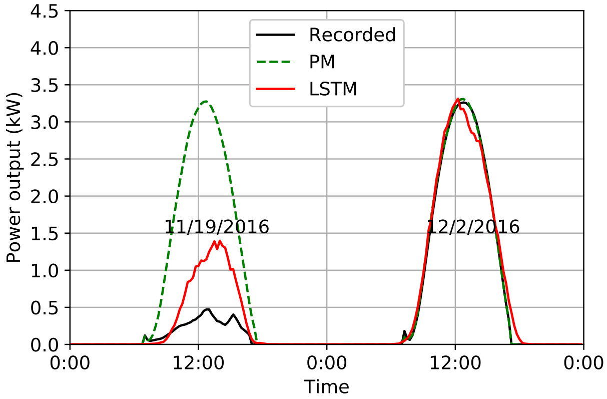

Fig. 10: Estimation results for days with highest and lowest PM error

The results clearly show that PM is not able to guarantee the estimation performance as the trans-day weather volatility increasing. The daily estimation errors are given by PM and LSTM model over a quarter testing period are presented in Fig. 11. The maximum daily estimated nRMSE from PM can reach nearly 20%. At the same time, the LSTM model is more effective in dealing with the high trans-day volatility condition.

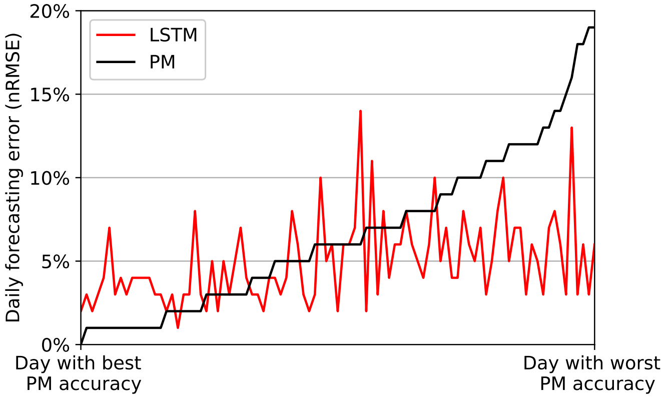

Fig. 11: Daily estimation error from PM and LSTM model over the testing period

The forecast solar radiation is utilized to determine whether this condition will happen or not. Since this irradiance forecast data is not perfect and has forecasting error as well, a suitable threshold needs to be determined. In this paper, the value of the threshold is tunned as 0.05%, and the value of $\mathrm{ED}(\hat{I}, d, d+1)$ is used to determine the switch between PM and optimized AE-LSTM model in the hybrid estimation model at the beginning of each day. Validated by the Cuptertnio data set, the hybrid model can improve the forecasting accuracy further by 14% (nRMSE) and 20% (nMAE).

VI. Conclusions

This paper presents an architecture for a hybrid day-ahead PV power estimated model, which utilizes an automated datafication approach to enhance the accuracy of PV power estimation. The hybrid model can effectively generate an estimate of PV output for the next 24 hours at 15-minute-interval. Through exploring the historical PV power generation data at different locations, PV power generation has different volatility in different weather conditions. In this paper, a hybrid model containing merits from both the models has been developed. According to the weather condition detecting algorithm with an overall accuracy of 80%, different sub-model will be used. PM is more adaptive to be used on consecutive sunny days, and AE-LSTM is more suitable to predict days with complex weather condition. LSTM integrated with AE forms a self-supervised learning structure that observes a compressed representation of time-series sequence data. This structure is the basis of complex sequential prediction problems, especially PV prediction in weather environments with unstable sunlight intensity. Using the encoder-decoder LSTM framework, we can manage and reduce uncertainty, misspecification, and noise. Compared with the existing methods in the literature, the proposed method shows a better forecasting accuracy using data from different PV sites.

VII. Acknowledgements

Authors would like to thank NEC Lab and the US Department of Energy AGGREGATE project for partial financial support to conduct this work.

References

[1] M. Osborne. (Jan 16, 2019) Global solar pv installations reach 109 GW in 2018 - BNEF. [Online]. Available: www.pv-tech.org

[2] SEIA. (2019) Solar industry research data. [Online]. Available: www.seia.org

[3] F. Wang, X. Ge, Z. Zhen, H. Ren, Y. Gao, D. Ma, M. Shafie-khah, and J. P. S. Catalão, "Neural network based irradiance mapping model of solar pv power forecasting using sky image," in 2018 IEEE Industry Applications Society Annual Meeting (IAS), Sep. 2018, pp. 1–7.

[4] K. Porter, S. Fink, M. Buckley, J. Rogers, and B. M. Hodge, "Review of Variable Generation Integration Charges," National Renewable Energy Lab. (NREL), Golden, CO (United States), Tech. Rep., 2013.

[5] D. Lew, N. Piwko, D. Miller, G. Jordan, K. Clark, and L. Freeman, "NREL: How do high levels of wind and solar impact the grid? the western wind and solar integration study," Tech. Rep., NREL, 2010.

[6] Y. Zhang, C. Jin, R. K. Sharma, and A. K. Srivastava, "Data-driven day-ahead pv estimation using hybrid deep learning," in 2019 IEEE Industry Applications Society Annual Meeting, 2019, pp. 1–6.

[7] E. G. Kardakos, M. C. Alexiadis, S. I. Vagropoulos, C. K. Simoglou, P. N. Biskas, and A. G. Bakirtzis, "Application of time series and artificial neural network models in short-term forecasting of PV power generation," in Power Engineering Conference (UPEC), 2013 48th International Universities', sep 2013, pp. 1–6.

[8] C. Monteiro, T. Santos, L. A. Fernandez-Jimenez, I. J. Ramirez-Rosado, and M. S. Terreros-Olarte, "Short-Term Power Forecasting Model for Photovoltaic Plants Based on Historical Similarity," Energies, vol. 6, no. 5, pp. 2624–2643, 2013.

[9] 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 Transactions on Industry Applications, vol. 48, no. 3, pp. 1064–1069, may 2012.

[10] G. Wang, Y. Su, and L. Shu, "One-day-ahead daily power forecasting of photovoltaic systems based on partial functional linear regression models," Renewable Energy, vol. 96, pp. 469–478, 2016.

[11] Y. Li, Y. He, Y. Su, and L. Shu, "Forecasting the daily power output of a grid-connected photovoltaic system based on multivariate adaptive regression splines," Applied Energy, vol. 180, pp. 392–401, 2016.

[12] L. Massidda and M. Marrocu, "Use of Multilinear Adaptive Regression Splines and numerical weather prediction to forecast the power output of a PV plant in Borkum, Germany," Solar Energy, vol. 146, pp. 141–149, 2017.

[13] Y. Q. Neo, T. T. Teo, W. L. Woo, T. Logenthiran, and A. Sharma, "Forecasting of photovoltaic power using deep belief network," in TENCON 2017 - 2017 IEEE Region 10 Conference, nov 2017, pp. 1189–1194.

[14] L.-L. Li, P. Cheng, H.-C. Lin, and H. Dong, "Short-term output power forecasting of photovoltaic systems based on the deep belief net," Advances in Mechanical Engineering, vol. 9, no. 9, p. 168781401771598, 2017.

[15] A. Gensler, J. Henze, B. Sick, and N. Raabe, "Deep Learning for solar power forecasting — An approach using AutoEncoder and LSTM Neural Networks," in 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), oct 2016, pp. 2858–2865.

[16] M. Abdel-Nasser and K. Mahmoud, "Accurate photovoltaic power forecasting models using deep LSTM-RNN," Neural Computing and Applications, oct 2017.

[17] D. Lee and K. Kim, "Recurrent neural network-based hourly prediction of photovoltaic power output using meteorological information," Energies, vol. 12, p. 215, 01 2019.

[18] K. Wang, X. Qi, and H. Liu, "A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network," Applied Energy, vol. 251, p. 113315, 2019.

[19] H. Zang, L. Cheng, T. Ding, K. W. Cheung, Z. Liang, Z. Wei, and G. Sun, "Hybrid method for short-term photovoltaic power forecasting based on deep convolutional neural network," IET Generation, Transmission & Distribution, vol. 12, pp. 4557–4567(10), November 2018.

[20] P. Li, K. Zhou, X. Lu, and S. Yang, "A hybrid deep learning model for short-term pv power forecasting," Applied Energy, vol. 259, p. 114216, 2020.

[21] J. Ospina, A. Newaz, and M. O. Faruque, "Forecasting of PV plant output using hybrid wavelet-based LSTM-DNN structure model," IET Renewable Power Generation, vol. 13, pp. 1087–1095(8), May 2019.

[22] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157–166, 1994.

[23] F. Chollet and Others, "Keras," https://keras.io, 2015.

[24] F. Chollet, Deep Learning with Python, 1st ed. Greenwich, CT, USA: Manning Publications Co., 2017.

[25] Y. Zhang, M. Beaudin, R. Taheri, H. Zareipour, and D. Wood, "Day-Ahead Power Output Forecasting for Small-Scale Solar Photovoltaic Electricity Generators," IEEE Transactions on Smart Grid, vol. 6, no. 5, pp. 2253–2262, 2015.

[26] Z. Janjic, R. Gall, and M. E. Pyle, "Scientific documentation for the NMM solver. NCAR Tech," Note NCAR/TN, vol. 477, 2010.

[27] B. Amrouche and X. L. Pivert, "Artificial neural network based daily local forecasting for global solar radiation," Applied Energy, vol. 130, pp. 333–341, 2014.

[28] L. M. Aguiar, B. Pereira, P. Lauret, F. Díaz, and M. David, "Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting," Renewable Energy, vol. 97, pp. 599–610, 2016.

[29] S. Sobri, S. Koohi-Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," Energy Conversion and Management, vol. 156, pp. 459–497, 2018.

[30] IBM. (2020) Weather calendar. [Online]. Available: www.wunderground.com

[31] Cedar Lake Ventures. (2020) Average weather. [Online]. Available: www.weatherspark.com

[32] B. Kraas, M. Schroedter-Homscheidt, and R. Madlener, "Economic merits of a state-of-the-art concentrating solar power forecasting system for participation in the Spanish electricity market," Solar Energy, vol. 93, pp. 244–255, 2013.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 56, pp. 1929–1958, 2014.