Feeder-Level Deep Learning-based Photovoltaic Penetration Estimation Scheme

Xiao-Yu Zhang, Stefanie Kuenzel
Department of Electronic Engineering
Royal Holloway, University of London
Egham Hill, Egham TW20 0EX, UK
Xiaoyu.Zhang@rhul.ac.uk; Stefanie.Kuenzel@rhul.ac.uk

Chris Watkins
Department of Computer Science
Royal Holloway, University of London
Egham Hill, Egham TW20 0EX, UK
C.J.Watkins@rhul.ac.uk

***Abstract*—The increasing penetration of renewable energy to the distribution grids, especially photovoltaic (PV), helps smooth out supply and demand, and reduces greenhouse gas emissions. However, the PV generation is behind-the-meter, and cannot be detected by the smart meter. To address this problem, a hybrid regression multi-layer perceptron (MLP) deep neural network (DNN) model is designed to separate the PV generation from the overall grid measurements. The model utilizes grid measurements, weather-related measurements, satellite-driven irradiance measurements, and temporal information as inputs to evaluate the PV generation in real-time. We also examine the performance of the model with different levels of PV penetration. We show the proposed model reduces the mean square error by 49% compared to single variable input models.**

Keywords—deep neural network, photovoltaic penetration, online system, energy disaggregation, power system monitoring

I. INTRODUCTION

Simultaneous with the development of the smart grid and microgrid, the penetration of renewable and embedded energy generation in the grid increases, which provides great challenges to grid operators to manage and plan the distributed network. Above all other renewable energy generation, the capacity of photovoltaic (PV) increased 1.4 times from 2014 to 2019, from 5.4 GW to over 13 GW in June 2019 [1]. The large penetration of PV influences the steady-state stability and transient stability in the power system [2]. However, most PV generation is behind-the-meter. The measurements recorded by the smart meter are the aggregate of load and power generated by the PV, which means the generated PV power cannot be detected by the utility. The grid operator can neither make an accurate forecast of load nor PV generation. Moreover, lacking visibility prevents the grid operators from implementing essential management to minimize the stability problems in time. Hence, an algorithm to disaggregate the PV generation from the grid measurement is required.

The target of the project, PV penetration estimation (also named PV energy disaggregation), is to estimate the PV generation in a small geographic area (normally the feeder of the distribution line) from the grid measurements at the feeder-level by the smart meter. Both household-level and feeder-level methods to separate the PV generation are discussed in the literature [3-6]. Unsupervised learning approaches, including feature extraction based methods [5, 7], SunDance [4]; Supervised learning approaches, including Contextually Supervised Source Separation (CSSS) [8], linear regression [9]; and hybrid approach [3] are proposed. The unsupervised learning approaches do not require the access to historical records of PV generation, but more features (such as irradiance, weather) are required to build linear/ nonlinear correlation between the PV output and these features. In contrast, the supervised learning approaches require fewer features as input but need historical grid measurements to train their models.

Moreover, PV generation forecasting approaches are available in many works [10-16], especially these works, based on machine learning/ deep learning [13-17]. These works use historical PV generation, as well as satellite-driven measurement (e.g. Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), Diffuse Horizontal Irradiance (DHI)), and weather measurement (e.g. temperature, humidity) to forecast the PV generation both short-term (hours) and long-term (days to one month). The above works provide enriched theoretical foundations to construct a deep neural network (DNN) model to separate PV generation from the grid measurement.

Feeder-level energy disaggregation techniques are introduced in [18-21]. A shallow neural network is proposed by Xu *et al.* and Asres *et al.* respectively [19, 21]. The synthetic load components data is generated from ZIP/ exponential load models, so the overall load, as well as the portion of each component, are known to researchers. Then a neural network is trained to disaggregate the overall power consumption into individual components. This method is straight forward in the experiment, however, since the real power system is dynamic and grid components are complex; as such it is difficult to implement the algorithm with real-world measurements. Particularly relevant to our work, Ledva *et al.* [20] proposed an online learning method based on real-world smart meter measurements. Household-level smart meter measurements provided by the Pecan Street Dataport [22] are aggregated to build a feeder-level load. Then an online learning algorithm, Dynamic Fixed Share (DFS) is adopted to perform energy disaggregation.

To address the problem discussed above, we propose a hybrid regression multi-layer perceptron (MLP) DNN model to estimate the PV penetration in real-time. This work learns from the hybrid model proposed in [12], and deep learning forecasting model proposed in [15], and energy disaggregation algorithm proposed in [20]. The model disaggregates the overall power consumption into PV generation and load power consumption. Rather than using grid measurements as input only, satellite-driven irradiance data, weather data, temporal information, and grid measurements are also adopted as inputs.

The contributions of this paper are as follows:

- We propose a deep neural network model to evaluate the PV penetration rate on a real-time basis.
- We analyze the variables (satellite-driven irradiance data, weather-related variables, temporal variables,

and grid measurements) that influence the accuracy of the model.

- We examine the accuracy of the model under different PV penetration rates.

II. Methodology

A. Problem Identification

The grid power measured by the feeder-level smart meter P_{Grid}, is the aggregation of the power load P_{Load} and the PV generation P_{PV}, see (1). To obtain the real value of the load consumption, the PV generation should be estimated at first. Normally, the PV generation is behind the meter, and cannot be detected by the smart meter.

$$P_{Grid}(t) = P_{Load}(t) - P_{PV}(t) \quad (1)$$

PV penetration is defined as follows:

$$PV\ peneration = \frac{Peak\ PV\ Load}{Peak\ Load\ Active\ Power} \quad (2)$$

Like PV energy forecasting, the PV penetration estimation focused on estimating the generation of PV in a targeted area. However, these two research problems have two differences:

- The input of the PV generation forecasting is the historical data of PV generation at the target geographic area. While the PV penetration estimation requires the measurements of the grid.
- The PV penetration estimation is evaluated in real-time for grid operation and management purposes, while the PV generation forecasting task is to estimate the future generation.

B. Input Features

The input variables are chosen from four datasets that are related to the generation of solar energy, which are satellite-driven irradiance data, weather-related measurement data, feeder-level power measurement data, and temporal data. The detailed description is shown as follows:

1) Grid Measurement: By accessing the grid measurements provided by the feeder-level smart meter, features such as active/ reactive power, voltage, current are collected. In this paper, active power P_{grid} is chosen as the grid measurement variable.

2) Satellite-driven Irradiance Features: Satellite-driven data includes GHI, DNI, DHI.

a) GHI: The total amount of shortwave radiation received from above by a surface horizontal to the ground.

$$GHI = DNIcos(\theta) + DHI \quad (3)$$

b) DNI: Amount of solar radiation received per unit area by a surface that is always held perpendicular (or normal) to the rays that come in a straight line from the direction of the sun at its current position in the sky.

c) DHI: The amount of radiation received per unit area by a surface (not subject to any shade or shadow) that does not arrive on a direct path from the sun, but has been scattered by molecules and particles in the atmosphere and comes equally from all directions.

$$P_s = kI_{incident}[cos(90-\theta)\, sin\beta\, cos(\varphi-\alpha) \quad (4\text{-a})$$

$$+ sin(90-\theta)\, cos\beta] \quad (4\text{-b})$$

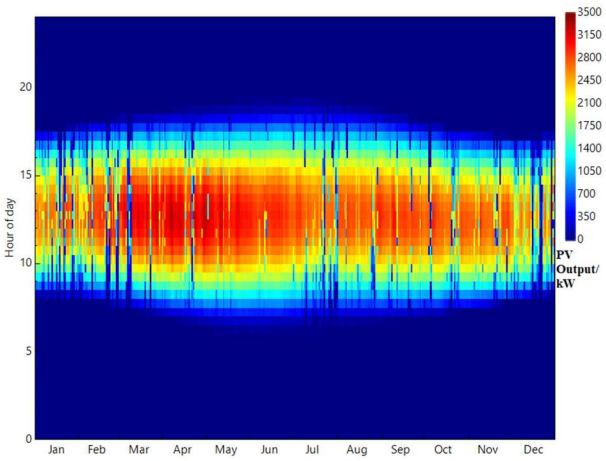

Fig. 1. Heatmap of PV generation thorugh the year.

Where θ is the sun's zenith angle above, $(90-\theta)$ is the sun's elevation angle, α is the sun's orientation angle, β is the solar tilt angle, and φ is the solar module's azimuth angle. Solar tilt angle: 0° when lying flat on the ground, 90° when vertical; sun's zenith angle: 0° when the sun is directly overhead, 90° when sunrise or sunset; solar module's orientation angle: 0° when direct north, 90° when direct south.

$$T_{PV,t} = T_{a,t} + I_{PV,t}\frac{N_{OCT}-20}{800} \quad (5)$$

$$P_t \approx C\frac{I_{PV,t}}{1000}[1-\mu(T_{PV,t}-25)] \quad (6)$$

Where $T_{PV,t}$ is the temperature of the solar module, $T_{a,t}$ is the ambient air temperature, $I_{PV,t}$ is the solar irradiance that is striking the solar module, N_{OCT} is the nominal operating cell temperature, and 48℃ is selected as a typical value.

3) Temporal-related Features: the temporal variables include the number of the hour of the day *H*, the month of the year *M*. The heatmap shown in Fig.1 presents the PV outputs throughout the year. It is observed that both the hour of the day and the month of the year influence the PV generation. Normally, the maximum output powers are generated between 10 am and 15 pm, and between May to July.

4) Weather-related Features: the weather-related measurements include temperature *T*, humidity *U*, weather description (e.g. sunny, rainy, snowy, cloudy) *D*.

To summarize, all features can be divided into numerical variables $\boldsymbol{N}$ and categorical variables $\boldsymbol{C}$. The numerical variables $\boldsymbol{N}$ is:

$$\boldsymbol{N_t} = [P_{grid,t}, T_t, U_t] \quad (7)$$

A categorical variable is a category or type. Although some categorical variables are recorded as a number, they do not have numerical meaning, such as month and hour. Before feeding the data to the DNN model, all categorical variables should be converted to numerical forms via one-hot encoding. A new binary variable is used to represent the original variable [23]. In this paper, categorical variable matrix $\boldsymbol{C}$ contains:

$$\boldsymbol{C_t} = [D_t, H_t, M_t] \quad (8)$$

By implementing one-hot encoding, the variables are transferred to:

$$\boldsymbol{C_t^O} = f^O(\boldsymbol{C_t}) \quad (9)$$

where f^O is the one-hot encoding function, and $\boldsymbol{C_t^O}$ is the one-hot encoding matrix. Hence, the overall input matrix $\boldsymbol{X}$ is shown as follow:

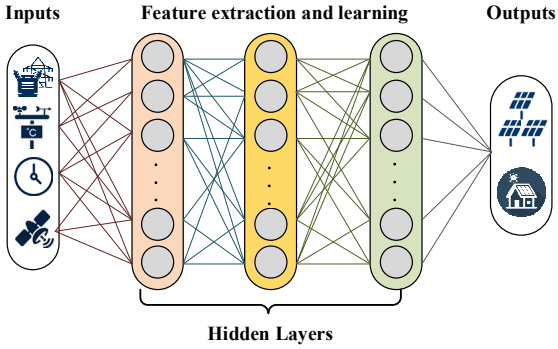

Fig. 2. MLP-DNN PV penetration estimation scheme .

$$\boldsymbol{X_t} = [\boldsymbol{C_t^0}, \boldsymbol{N_t}] \tag{10}$$

C. PV Penetration Estimation Framework

In this section, a regression MLP DNN model is designed to estimate the PV penetration. The DNN model contains one input layer, one output layer, and three hidden layers. Each layer contains a few neurons, and each neuron contains bias and an activation function [24]. The researchers would choose different activation functions for different tasks (e.g. ReLU, Tanh for nonlinear regression, SoftMax for classification task) The target of the model is to estimate the PV penetration and the real load, see Fig. 2. Hence, the output matrix Y can be expressed as:

$$\boldsymbol{Y} = [P_{PV}, P_{Load}] \tag{11}$$

The hidden layers, which are mathematical functions perform a nonlinear transformation to the inputs. An MLP-DNN has two steps: forward propagation and backpropagation [24].

1) Forward propagation: the input layer obtains the inputs and propagates the information through hidden layers and finally produce outputs. The mathematical expression of the MLP-DNN model is:

$$\hat{\boldsymbol{Y}} = \emptyset_N(\emptyset_{N-1}(\cdots\emptyset_1(\boldsymbol{X}, \boldsymbol{W_1}, c_1), \boldsymbol{W_{N-1}}, c_{N-1}), \boldsymbol{W_N}, c_N) \tag{12}$$

where $\emptyset_i$ is the activation function of the ith layer, and $\boldsymbol{W}_i$ and c_i is the weight matrix and bias of the ith layer. A loss function is adopted to measure the error between the ground truth output and the output generated by the model. Normally, a mean square error (MSE) is used as a loss function:

$$J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{Y}} - \boldsymbol{Y})^2 \tag{13}$$

where $\boldsymbol{\theta}$ is the matrix of all model parameters, including weights and bias. Referring to [9], a Taxicab norm (ℓ_1 norm) is added into the loss function to obtain a sparse model. The objectives of the ℓ_1 norm in this model are two aspects: (1) reduce the overfitting of the model; (2) extract features from the inputs. Hence, the final cost function becomes:

$$\tilde{J}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y}) = J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y}) + \alpha\|\boldsymbol{\theta}\|_1 \ (\alpha \in [0, +\infty)) \tag{14}$$

where α is the hyperparameter that reflects the regularization weight.

2) Backpropagation: Backpropagation is used to update the model parameters generated in forward propagation referring to the loss function, the direction is from outputs to inputs [24]. The gradients of the model parameters are calculated:

$$g \leftarrow \nabla_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y}) = \nabla_{\boldsymbol{W}}J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y}) + \alpha sign(\boldsymbol{\theta}) \tag{15}$$

Then the parameters are updated via the gradient direction with a learning rate ϵ:

TABLE I. HYPERPARAMETERS OF THE MODEL

Hyperparameters	Value	Description
Learning rate ϵ	0.05	The steps to adjust θ according to errors.
Hidden layers number	3	The total number of hidden layers.
Batch size B	128	The number of training examples utilized in one iteration.
Activation function for hidden layers	ReLU	$f_{ReLU} = \max[0, z]$.
Activation function for hidden layers	ReLU	Positive Output
Epoch number	100	One cycle through the full training dataset.
Loss function	MSE	Optimize the model
α	0.01	Reduce overfitting and extract features
Dropout	0.5	Reduce overfitting

$$\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta} - \epsilon\nabla_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y}) \tag{16}$$

Further hyperparameters in the DNN model are illustrated in Table I.

III. EXPERIMENTS AND RESULTS

A. Dataset

Pecan Street Dataport [22] is used as the dataset to train the proposed model. The dataset contains nearly one thousand household-level power consumption in the last five years with an interval resolution of 15 minutes. The household-level measurements are added together to construct a synthetic feeder model. In this paper, 75 houses are aggregated to build a feeder with a capacity of 100 kW during Jan 2018 and Dec 2018 in Austin, Texas, US. The PV penetration rate of the feeder is adjusted via adding or deleting houses with PV installed to match the requirement. The irradiance measurements and weather data at the same location are obtained from the National Climatic Data Center (NCDC) [25].

B. Software& Hardware

The simulation and computation are implemented on a Dell laptop equipped with Core i7-7700HQ CPU, NVIDIA GTX 1060 GPU, and 8GB RAM. The deep learning algorithm is run on Python 3.6, the TensorFlow framework is adopted to train the DNN model.

C. Error Metrics

In this paper, two error metrics are adopted to measure the errors between the ground truth and estimated values, which are Root Mean Squared Error (RMSE) and Normalized RMSE (nRMSE):

1) Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(\hat{P}_{PV,t} - P_{PV,t})}{T}} \tag{17}$$

2) Normalized RMSE（nRMSE）

$$nRMSE = \frac{RMSE}{\bar{P}_{Load}} \tag{18}$$

D. Case Study

Model 1 Grid Measurements-only Approach: in this model, only the feeder-level grid measurement P_{grid} is considered as the input of the DNN model.

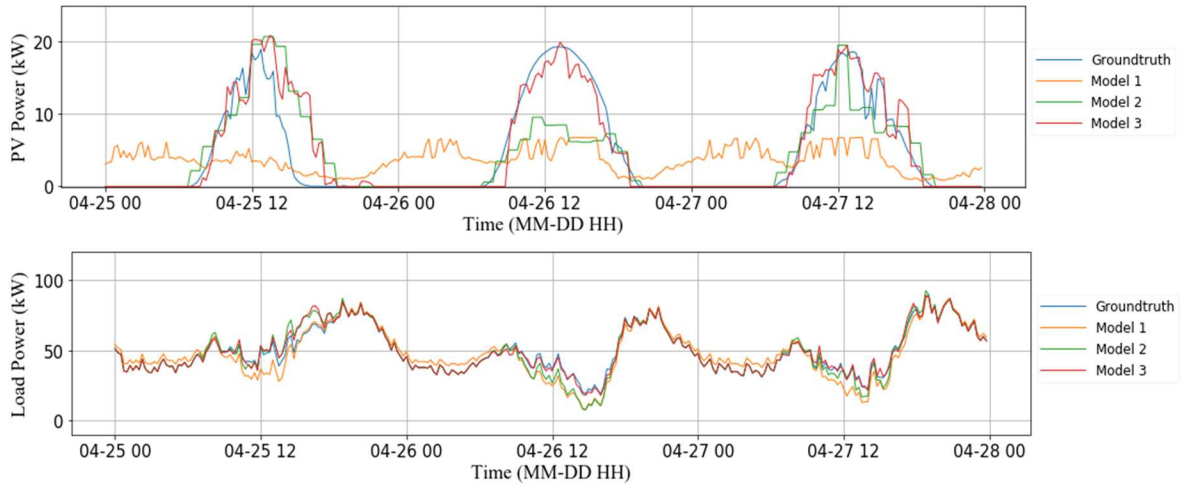

Fig. 4. Groundtruth and estimated value of PV output and load with 20% PV penetration.

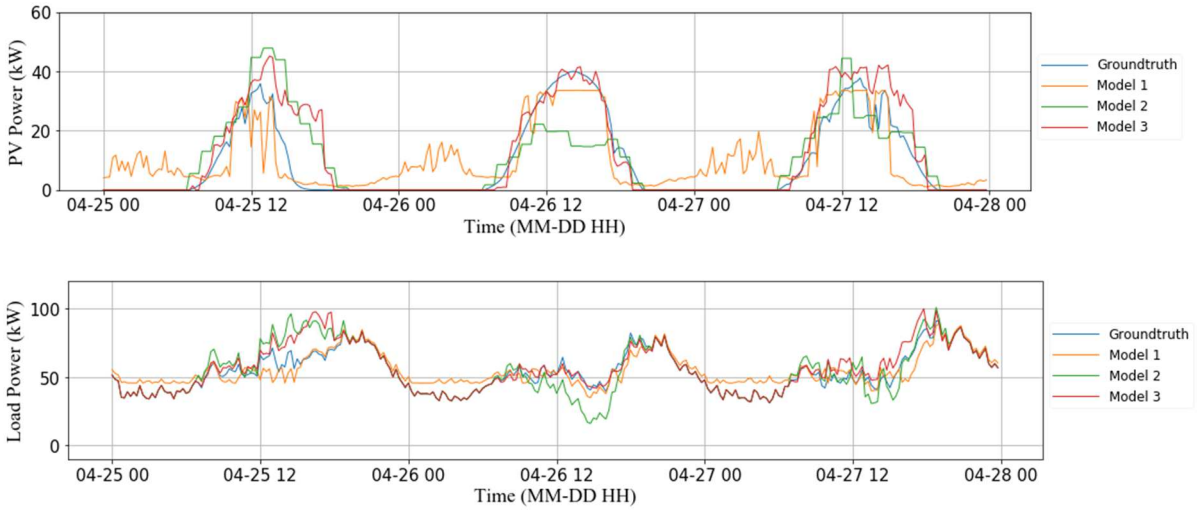

Fig. 5. Groundtruth and estimated value of PV output and load with 40% PV penetration.

Model 2 Satellite-driven Irradiance Data -only Approach: in this model, only the irradiance data DHI, DNI, GHI are considered as the input of the DNN model.

Model 3 Hybrid Data-Driven Approach: in this case, all features are described in Section II. C is adopted as inputs, including grid measurements, irradiance data, temporal data, as well as weather-related measurements.

We apply all three models to five PV penetration rate data (5%, 10%, 20%, 40%, 70%) respectively. Fig. 4 and Fig.5 show the disaggregated PV output as well as the load power in three days with 20% and 40% penetration rate. The results show that across all penetration rates, Model 3, which is the hybrid data-driven approach, performs the best. The estimation curve of Model 3 closely matches the real curve.

Table II and Fig. 4 show the RMSE and nRMSE between the ground truth PV generation and estimated values by the three DNN models, respectively. As for Model 1, which performs the worst throughout the whole experiment, an increase of both RMSE and nRMSE is observed when the PV penetration rate is decreased. Model 2 has a better performance at a low penetration rate (e.g. 5% and 10%), RMSE, and nRMSE of Model 2 is approaching the values of Model 3 in these groups. However, with the penetration rate rising, the performance of the model becomes worse, nRMSE increase from 0.241% to 0.357% when the penetration rate increases from 20% to 40%. One reason for this result is that a high penetration rate would contain higher uncertainty caused by other conditions such as consumers' behaviors, weather conditions rather than irradiance variables. Finally, Model 3 has the best and stable performance throughout the experiment, although an increase of error is witnessed when we reduce the amount of PV generation into the grid.

To sum up, the hybrid model can have a stable estimation with high accuracy even the penetration rate is extremely small. Especially when we compare Model 3 with Model 1, the hybrid approach reduces the nRMSE up to 49%.

IV. Conclusion

In this paper, a hybrid regression MLP DNN model is proposed to separate the PV generation from the grid measurements. Multiple variables related to the PV generation (weather-related variables, irradiance variables, feeder-level grid measurement, etc.) are adopted as the inputs of the DNN model. The proposed model contains three hidden layers to extract and learn input features deeply. Then a synthetic feeder model used for the experiment is constructed by aggregating household-level data together. The performance of the proposed hybrid method is examined via a comparison

with the grid measurement-only approach and irradiance measurement-only approach. A conclusion is made that the hybrid model has a better performance than others under different PV penetration rates (from 5% to 70%).

TABLE II. PERFORMANCE METRICS OF DIFFERENT MODELS

Penetration Rate	Model	Performance Metrics	
		RMSE (kW)	*nRMSE (%)*
5%	1	1.982	0.440
	2	0.768	0.164
	3	1.251	0.262
10%	1	3.481	0.367
	2	1.705	0.185
	3	1.992	0.201
20%	1	6.815	0.343
	2	4.900	0.241
	3	4.322	0.216
40%	1	13.238	0.326
	2	8.256	0.357
	3	6.747	0.165
70%	1	19.553	0.254
	2	14.537	0.189
	3	12.511	0.162

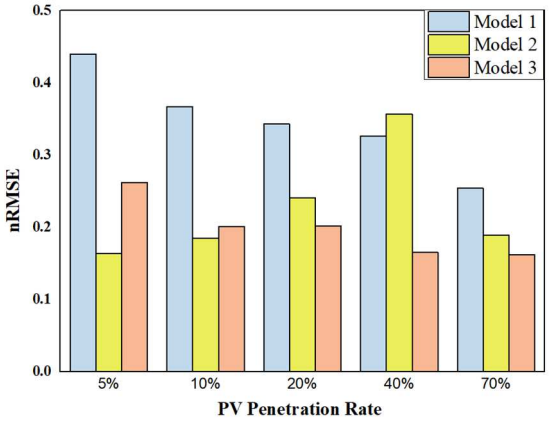

Fig. 6. NRMSE for different models.

REFERENCE

[1] M. Castaneda, S. Zapata, J. A. Herrera, A. J. Aristizábal, and I. Dyner, "Solar power business in the uk: A bright or cloudy future ahead?," 2020.

[2] S. Eftekharnejad, V. Vittal, G. T. Heydt, B. Keel, and J. Loehr, "Impact of increased penetration of photovoltaic generation on power systems," *IEEE transactions on power systems,* vol. 28, no. 2, pp. 893-901, 2012.

[3] J. M. Bright, S. Killinger, D. Lingfors, and N. A. Engerer, "Improved satellite-derived pv power nowcasting using real-time power data from reference pv systems," *Solar Energy,* vol. 168, pp. 118-139, 2018.

[4] D. Chen and D. Irwin, "Sundance: Black-box behind-the-meter solar disaggregation," in *Proceedings of the Eighth International Conference on Future Energy Systems*, 2017, pp. 45-55.

[5] C. M. Cheung, W. Zhong, C. Xiong, A. Srivastava, R. Kannan, and V. K. Prasanna, "Behind-the-meter solar generation disaggregation using consumer mixture models," in *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 2018: IEEE, pp. 1-6.

[6] E. C. Kara, C. M. Roberts, M. Tabone, L. Alvarez, D. S. Callaway, and E. M. Stewart, "Disaggregating solar generation from feeder-level measurements," *Sustainable Energy, Grids Networks,* vol. 13, pp. 112-121, 2018.

[7] K. Li, F. Wang, Z. Mi, M. Fotuhi-Firuzabad, N. Duić, and T. Wang, "Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation," *Applied Energy,* vol. 253, p. 113595, 2019.

[8] M. Wytock and J. Z. Kolter, "Contextually supervised source separation with application to energy disaggregation," in *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

[9] E. Vrettos, E. Kara, E. Stewart, and C. Roberts, "Estimating pv power from aggregate power measurements within the distribution grid," *Journal of Renewable Sustainable Energy, Grids and Networks,* vol. 11, no. 2, p. 023707, 2019.

[10] A. Yona, T. Senjyu, A. Y. Saber, T. Funabashi, H. Sekine, and C.-H. Kim, "Application of neural network to 24-hour-ahead generating power forecasting for pv system," in *2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century*, 2008: IEEE, pp. 1-6.

[11] H.-T. Yang, C.-M. Huang, Y.-C. Huang, and Y.-S. Pai, "A weather-based hybrid method for 1-day ahead hourly forecasting of pv power output," *IEEE transactions on sustainable energy,* vol. 5, no. 3, pp. 917-926, 2014.

[12] E. Ogliari, F. Grimaccia, S. Leva, and M. J. E. Mussetta, "Hybrid predictive models for accurate forecasting in pv systems," vol. 6, no. 4, pp. 1918-1929, 2013.

[13] 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, 2012.

[14] C. Chen, S. Duan, T. Cai, and B. Liu, "Online 24-h solar power forecasting based on weather type classification using artificial neural network," *Solar energy,* vol. 85, no. 11, pp. 2856-2870, 2011.

[15] Y. Wang, N. Zhang, Q. Chen, D. S. Kirschen, P. Li, and Q. Xia, "Data-driven probabilistic net load forecasting with high penetration of behind-the-meter pv," *IEEE Transactions on Power Systems,* vol. 33, no. 3, pp. 3255-3264, 2017.

[16] M. Abdel-Nasser and K. Mahmoud, "Accurate photovoltaic power forecasting models using deep lstm-rnn," *Neural Computing Applications,* vol. 31, no. 7, pp. 2727-2740, 2019.

[17] 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 *2013 48th International Universities' Power Engineering Conference (UPEC)*, 2013: IEEE, pp. 1-6.

[18] M. W. Asres, A. A. Girmay, C. Camarda, and G. T. Tesfamariam, "Non-intrusive load composition estimation from aggregate zip load models using machine learning," *International Journal of Electrical Power & Energy Systems,* vol. 105, pp. 191-200, 2019/02/01/ 2019, doi: https://doi.org/10.1016/j.ijepes.2018.08.016.

[19] Y. Xu and J. Milanovic, "Artificial-intelligence-based methodology for load disaggregation at bulk supply point," *IEEE Transactions on Power Systems,* vol. 30, pp. 1-9, 01/01 2014, doi: 10.1109/TPWRS.2014.2337872.

[20] G. S. Ledva, L. Balzano, and J. L. Mathieu, "Real-time energy disaggregation of a distribution feeder's demand using online learning," *IEEE Transactions on Power Systems,* vol. 33, no. 5, pp. 4730-4740, 2018, doi: 10.1109/TPWRS.2018.2800535.

[21] G. S. Ledva and J. L. Mathieu, "Separating feeder demand into components using substation, feeder, and smart meter measurements," *IEEE Transactions on Smart Grid,* pp. 1-1, 2020, doi: 10.1109/TSG.2020.2967220.

[22] P. J. P. S. I. Street, "Dataport: The world's largest energy data resource," 2015.

[23] C. Seger, "An investigation of categorical variable encoding techniques in machine learning: Binary versus one-hot and feature hashing," ed, 2018.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[25] I. S. J. A. Data, NC, "National climatic data center (ncdc)," 2001.