

Article

Data Augmentation-Based Photovoltaic Power Prediction

Xifeng Wang ^{1,†} , Yijun Shen ^{1,*,†}, Haiyu Song ² and Shichao Liu ^{3,†}
¹ Department of Automation, Zhejiang University of Technology, Hangzhou 310023, China; xifengwang0310@zjut.edu.cn

² School of Information Technology and Artificial Intelligence, Zhejiang University of Finance and Economics, Hangzhou 310018, China; haiyusong@gmail.com

³ Department of Electronics, Carleton University, Ottawa, ON K1S 5B6, Canada

* Correspondence: 2111903043@zjut.edu.cn

† These authors contributed equally to this work.

Abstract: In recent years, as the grid-connected installed capacity of photovoltaic (PV) power generation has increased by leaps and bounds, it has assumed considerable importance in predicting PV power output. However, the power prediction of newly built small-scale PV power plants often suffers from many unexpected problems, such as data redundancy, data noise, data sample imbalance, or even missing key data. Motivated by the above facts, this paper proposes a data augmentation-based prediction framework for PV power. Firstly, the daily power distance measurement is used to analyze feature correlation and filter erroneous data. Then, the autoencoder network trained based on the random discarding mechanism is used to restore the PV power generation data. At the same time, a specific data augmentation method for the PV power curve is designed to eliminate the influence of data sample imbalance. In the final experimental section, compared with the latest method, this method achieved the highest MAE accuracy of 8.26% and RMSE accuracy of 10.96%, which proves the effectiveness of this method.

Keywords: prediction; photovoltaic power; autoencoder; data augmentation



Academic Editor: Santiago Silvestre

Received: 6 January 2025

Revised: 20 January 2025

Accepted: 27 January 2025

Published: 6 February 2025

Citation: Wang, X.; Shen, Y.; Song, H.; Liu, S. Data Augmentation-Based Photovoltaic Power Prediction. *Energies* **2025**, *18*, 747. <https://doi.org/10.3390/en18030747>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As a virtually inexhaustible clean energy, solar photovoltaic (PV) power plays an important role in the world's energy field. Over the past decade, the global deployed on-grid PV capacity has grown by an astonishing 1860%, from 41.5 GW in 2010 to 773.2 GW in 2020 [1]. Furthermore, the global deployed on-grid PV annual new installed capacity is also increasing year by year. As shown in Figure 1, whether in terms of annual growth or overall growth trend, PV power generation has a promising development prospect.

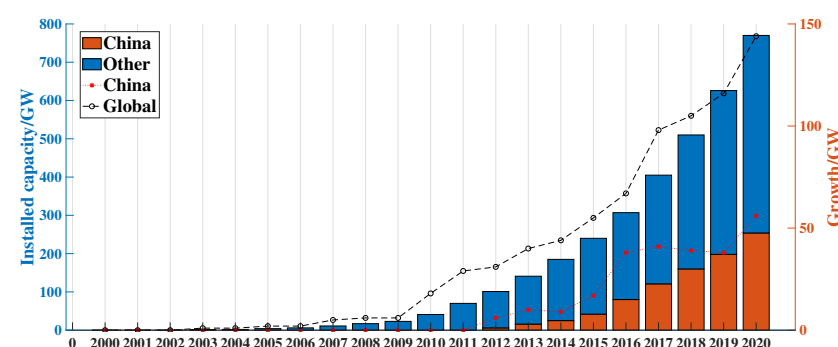


Figure 1. Photovoltaic installed capacity and its annual growth.

With huge strides made in PV technology and the popularization of PV applications, especially the extremely high market share of grid-connected PV power generation, the PV power prediction system has become particularly important. The accuracy of power generation prediction will have a direct impact on the grid load frequency, grid stability, system reliability, power quality, investment, and system operation efficiency [2–8]. The study on the short-term prediction of PV power generation is helpful to stabilize the operation of the power system, optimize grid scheduling, and provide effective protection for distribution load management.

Much research has been carried out on the prediction of photovoltaic power generation. According to the principle of prediction, the forecast methods can be divided into three categories: physical methods [9], statistical or time series analysis methods [10], and hybrid methods [11]. According to the predicted time scale classification, the forecast can be divided into four categories: very short forecasts [12], short forecasts [11], nowcasting [13], and long forecasts [10]. From the perspective of prediction methods, these technologies can be broadly divided into two categories: (1) indirect forecasting methods [9,13,14] and (2) direct forecasting methods [12,15,16]. For the indirect methods, solar irradiance is predicted based on historical solar irradiance and weather data, and then is converted to PV power output. The techniques used include the fuzzy logic method [9], wavelet analysis [14], and the artificial neural network (ANN) [13]. The direct methods predict PV power output according to their historical data and associated weather information. The techniques used consist of the support vector machine (SVM) [12], ANN [15], and hybrid ANN methods [16].

Solar power generation is, to a great extent, susceptible to weather, which is difficult to predict accurately [12,17–22]. Changes in weather conditions affect the stability and continuity of power output, which poses a serious challenge for the PV power prediction system. At present, a series of short-term PV power prediction methods have been proposed, such as the artificial neural network (ANN) [15], support vector machine (SVM) [12], Markov chain (MC) [11], and Long Short-Term Memory (LSTM) [10] network. In [23], an inverter-data-driven method to forecast the output power of PV plants under the time scale of a second was proposed, with a deduction accuracy of 94.8%, which could fully take the impact of moving cloud into consideration. In [24], historical PV power generation data and weather data were used to train Bayesian network-based data-driven reasoning models to predict ultra-short-term PV power. In [25], the correntropy criterion and ConvLSTM network were integrated to build a robust deep spatiotemporal framework. By leveraging the encoder–decoder ConvLSTM architecture, the resultant forecasts can cover multiple PV sites and look-ahead horizons simultaneously. In [26], the combination of nonlinear correlation measurement and weighted local outlier factor (WLOF) were utilized to build a Gaussian weighted regression algorithm to forecast short-term PV power generation with a normalized root mean square error (nRMSE) of 6.614%. In [11], the higher-order Markov chain was used to forecast the probability distribution function (PDF) of the generated power of PV systems, allowing the PV power point forecast as well as the PDF forecast to be carried out.

It is evident from [27–30] that the prediction accuracy of PV power is affected by the weather, majorly. In other words, when sudden changes in weather conditions occur, the accuracy of the power prediction system is reduced to varying degrees. For example, historical power output is classified as four different weather types, sunny, cloudy, foggy, and rainy, in [15], and the SVM is used for PV power prediction, with a mean relative error (MRE) ranging from 4.85 to 12.42%. In [31], the solar time, earth declination angle, and weather data were utilized to characterize the historical measurement power data using table look ups and weighted blending models, with an average nRMSE of 7.356%. In [30],

used by the power conversion model, the meteorological factors of the historical weather data of the target weather station and the surrounding weather stations were used for PV power prediction, with an average nRMSE of 6.72%.

Based on the above research situations, this paper discusses the characteristics of solar power generation time series and the challenges of solar power generation prediction (i.e., data redundancy, data noise, data sample imbalance, and missing data). On this basis, relying on data-driven methods, this paper proposes a PV power prediction framework under the condition of missing information. We demonstrate that, if the framework is executed correctly, the proposed method can effectively find relevant features and repair wrong data. The main contributions of this paper are:

1. A feature selection and data elimination (FS-DE) module is designed, which can effectively analyze the related external features and screen out abnormal samples. At the same time, a randomly discard mechanism autoencoder is designed to repair abnormal data.
2. A data augmentation module for PV-specific data is designed to help eliminate the imbalance of data distribution. The residual signal undergoes a series of transformations to generate multiple similar power curves, and then, in order to obtain external features synchronously, the reverse prediction network is used to generate sample pairs to achieve data enhancement.

2. Background

In this section, we provide a discussion on the challenges of solar power forecasting. The discussions are based on the real-life data that are later used for numerical simulations.

2.1. Data Source

The data we studied came from two different solar energy PV power generation stations in Zhejiang, China. The first station, Shaoxing station, is located in Shaoxing, Zhejiang (120.33° E, 30.12° N) and the second station, Jiaxing station, is located in Tongxiang, Zhejiang (120.61° E, 30.60° N). The detailed information is summarized in Table 1.

Table 1. The detailed parameters of the PV power stations.

	Shaoxing Station	Jiaxing Station
Location	Shaoxing	Jiaxing
Coordinates	120.33° E, 30.12° N	120.61° E, 30.60° N
Data	Cant radiation, Direct normal irradiance, Diffuse solar radiation, Total radiation, Atmospheric pressure, Relative humidity, Ambient temperature, The panel temperature, Wind speed, Wind direction, Precipitation	Cant radiation, Direct normal irradiance, Diffuse solar radiation, Total radiation, Ambient temperature, Humidity, Atmospheric pressure, The panel temperature
Sampling period	15 min	5 min/15 min
Data period	31 August 2019 to 29 August 2020	1 January 2021 to 26 March 2021

Shaoxing station contains seven units, and the rated power of each unit is 2.4 kW. The historical data collected include ambient temperature, wind speed, wind direction, humidity, direct radiation, diffuse radiation, and air pressure, with a sampling period of 15 min from 31 August 2019 to 29 August 2020.

Jiaxing station contains three units, and the rated power of each unit is 2 kW. For station 2, we used data that cover the period from 1 January 2021 to 26 March 2021, including power data with a sampling period of 15 min and other environmental parameters with a sampling period of 5 min.

2.2. Challenges of PV Output Forecasting

There is no doubt that PV power forecasting will have many challenges; without loss of generality, this paper considers several of the most common problems in PV power forecasting.

The low predictability of PV power generation is mainly due to the instability and discontinuity of power generation components affected by extremely unpredictable weather conditions. Figure 2 depicts the diurnal variation of PV power output under different weather conditions at Shaoxing Station. It can be seen that the PV power output is quite different under different weather conditions. On sunny days, the power curve presents a full bell-shaped curve. On rainy days, the output of power is at a low value, while on cloudy days it is in between. From the perspective of the time dimension, PV daily power generation is highly positively correlated with sunshine length. Figure 3 depicts the variation in sunshine duration with seasons and the variation in daily radiation values. It can be observed that there is a variation trend from the shortest sunshine duration in winter to the longest sunshine duration in summer. From the perspective of spatial dimension, places that are close to each other have a similar weather distribution. However, due to the instability and discontinuity of weather, daily power generation often has great differences. For example, the weather in Shaoxing and Jiaxing is completely different, although their annual weather distribution is similar. The weather uncertainty in time and space causes the instability and discontinuity of PV power generation, so the error of PV power prediction is relatively high.

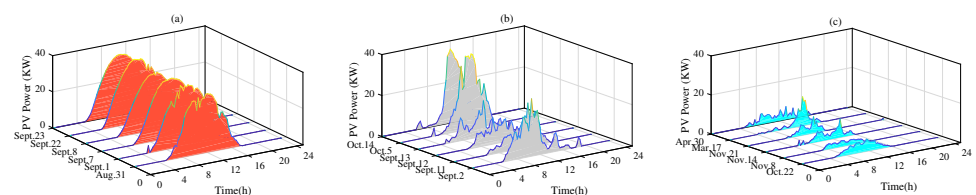


Figure 2. Power output curves under different weather conditions at Shaoxing station. (a) Sunny. (b) Cloudy. (c) Rainy.

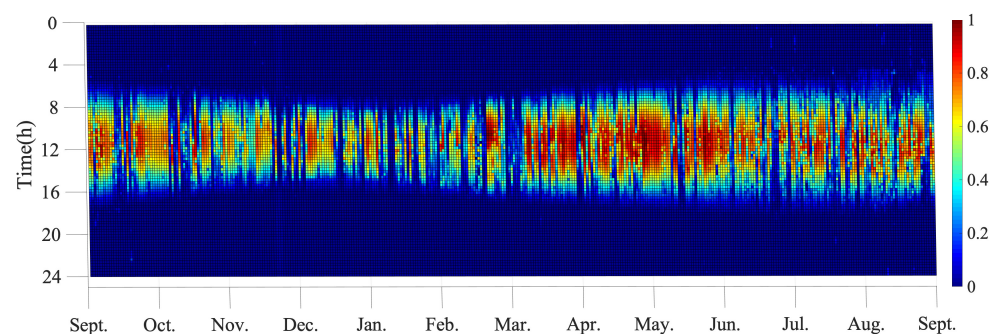


Figure 3. The variation trend in daylight hours with different seasons at Shaoxing station.

As shown in Table 1, the data from Shaoxing station contain many external features, i.e., (1) cant radiation, (2) direct normal irradiance, (3) diffuse solar radiation, (4) total radiation, (5) atmospheric pressure, (6) relative humidity, (7) ambient temperature, (8) the panel temperature, (9) wind speed, (10) wind direction, and (11) precipitation. The proportion of features strongly correlated with the output power is very small, and most features have little correlation with the output power. Therefore, these redundant features will have a great impact on the prediction accuracy, and how to find the appropriate feature similarity metric is particularly important. In addition, due to the constraints of realistic conditions, data acquisition and data transmission anomalies often occur in sensor sampling, as shown in Figure 4 (the data collection interval is 15 min), which often leads to difficulties in improving the accuracy of data-based prediction methods. Therefore, a suitable data repair mechanism is required to integrate the original data. In this paper, we have built a data augmentation module to overcome this difficulty.

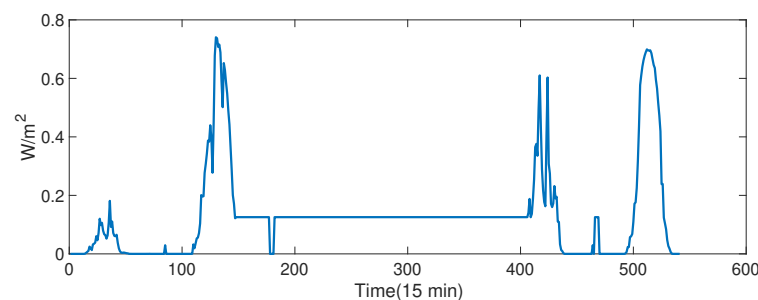


Figure 4. Abnormal data of Shaoxing station.

For many PV power generation systems, due to storage and other practical constraints, the sampled data often show a small amount of data accumulation and unbalanced data distribution, and so on. As shown in Figure 5, from the historical weather distribution, the weather conditions at Shaoxing station are similar to those at Jiaxing, but, looking at the history of sampling data, the sampling data of Shaoxing station are full, close to the historical weather distribution. At the same time, the historical sampling data of Jiaxing station are too small to reflect the historical distribution, thus causing a large prediction error.

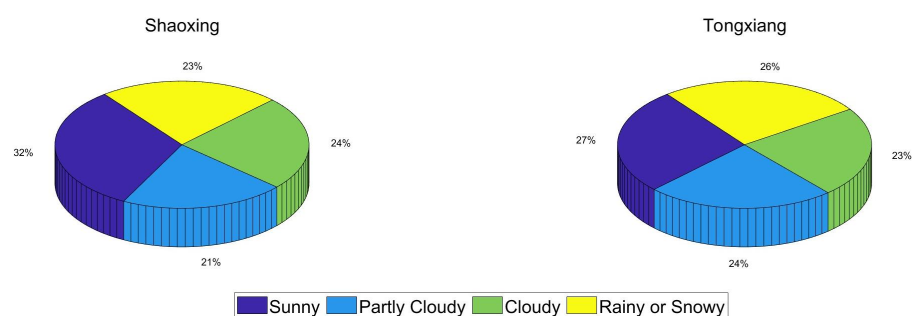


Figure 5. Weather distribution.

3. Methodology

3.1. The FS-DE Module

The feature selection and data elimination (FS-DE) module is designed for feature selection and abnormal data day extraction. Historical data were trained for feature selection by similarity metrics, while autoencoders (AEs) in [32] were used for data repair.

The power output of the PV system is directly or indirectly affected by a large number of external variables, such as direct normal irradiance, diffuse solar radiation, ambient temperature, relative humidity, cloud cover index, etc. In fact, there are only a few external

variables that have a great influence on power output, such as direct normal irradiance and diffuse solar radiation, and the rest of the external variables have little correlation. The FS-DE module is designed to pick out the most relevant features on the one hand and the days of data anomalies on the other. In Table 2, we have listed the symbols used in this paper.

Table 2. Summary of frequently used mathematical notations.

Notations	Mathematical Meanings
P	Historical power matrix
$P^{(d)}$	The power output vector of day d
τ	The sampling frequency
T	The number of intervals in a day
V	External variable matrix
N	The number of days for sampling data
M	The number of external features
PED	The Euclidean distance matrix of the historical power data
VED	The Euclidean distance matrix of the external features
PVD	The correlation coefficient between external feature and power
$PVED$	The external feature correlation metric
E_k	The loss function

In the FS-DE module, we are given the historical power output, P , and the historical external variables, V :

$$P = \{P^{(d)} \mid d = 1 \cdots N\} \quad (1)$$

$$P^{(d)} = \left(p_1^{(d)}, p_2^{(d)}, \dots, p_t^{(d)}, \dots, p_T^{(d)} \right)^T \quad (2)$$

$$V = \{V_k^{(d)} \mid d = 1 \cdots N; k = 1 \cdots M\} \quad (3)$$

$$V_k^{(d)} = \left(v_{k1}^{(d)}, v_{k2}^{(d)}, \dots, v_{kt}^{(d)}, \dots, v_{kT}^{(d)} \right)^T \quad (4)$$

where $P^{(d)}$ is the power output vector of day d with sampling frequency τ , t indicates the measurement interval, and T is the number of intervals in a day. $V^{(d)}$ is the k th external features of day d .

The concrete implementation of the FS-DE module is as follows:

Step 1: the Euclidean distance of the historical power data (PED) is calculated for each pair of historical days by the matrix norm, say PED, as follows.

$$PED = \begin{bmatrix} PED_{11} & \cdots & PED_{1N} \\ & \ddots & \\ \vdots & PED_{ij} & \vdots \\ & & \ddots \\ PED_{N1} & \cdots & PED_{NN} \end{bmatrix}_{N \times N} \quad (5)$$

where $PED_{ij} = \|P^{(i)} - P^{(j)}\|_2$ is the PED between day i and j according to the values of power. Note that days with similar power patterns will have very small PED values, while days with significantly different power patterns, such as sunny and cloudy days, will have high PED values.

Step 2: the values of VEDs are calculated for each pair of historical days based on each external variable, V_k , as follows:

$$VED_k = \begin{bmatrix} VED_{k11} & \cdots & VED_{k1N} \\ & \ddots & \\ \vdots & & VED_{kij} & \vdots \\ & & \ddots & \\ VED_{kN1} & \cdots & & VED_{kNN} \end{bmatrix}_{N \times N} \quad (6)$$

where $VED_{kij} = \|V_k^{(i)} - V_k^{(j)}\|_2$ is the VED between day i and j according to the values of the k th external variable. Similarly, days with similar external variables will have a very small value, while days with significantly different external variables, such as sunny and cloudy days, will have a very high value.

Step 3: we define the correlation coefficient, PVD , between the external feature, V_k , and power, P , as:

$$PVD(P, V_k) = \frac{\sqrt{2}\|PED - VED_k\|_2}{N \times (N - 1)} \quad (7)$$

A low value of the PVD indicates a higher similarity between V_k and P . We calculate PVD values for each external feature and power and then select M' external features that are most relevant to power.

Step 4: the value of $PVED_k$ s are calculated as follows:

$$PVED_k = [PVED_k^{(1)} \quad \cdots \quad PVED_k^{(i)} \quad \cdots \quad PVED_k^{(N)}]_{1 \times N} \quad (8)$$

where $PVED_k(i) = \|P^{(i)} - V^{(i)}\|_2$. If $PVED_k^{(i)} \gg \overline{PVED}$, this means there was an error in the data for that day. Pick out such data for later repair (i.e., $\varsigma = \{\varepsilon_i \mid i = 1, 2, \dots, N_e\}$, where N_e is the number of abnormal samples and $\varepsilon_i = [p \quad v_1 \quad v_2 \quad \cdots \quad v_L]$).

The abnormal data, ς , are selected, and the remaining data, $\kappa = \{\rho_i \mid i = 1, 2, \dots, N_m\}$ (where $\rho_i = [p \quad v_1 \quad v_2 \quad \cdots \quad v_L]$), are put into the autoencoder for training. The input vector is randomly set to zero for anti-noise training, and the specific formula is as follows:

$$\tilde{\kappa} = \{\tilde{\rho}_i \mid i = 1, 2, \dots, N_m\} \quad (9)$$

$$\tilde{\rho}_i = \rho_i * \Lambda \quad (10)$$

where $\tilde{\kappa}$ represents data generated by random dropout, and Λ is a Boolean diagonal matrix whose only diagonal elements are non-zero and whose diagonal elements are either zero or one. The concrete implementation is as follows:

$$\Lambda = \text{diag}(\theta_1, \theta_2, \dots, \theta_j, \dots, \theta_{L+1}) \quad (11)$$

$$\theta_j \in \{0, 1\} \quad (12)$$

The structure of the neural network (NN) used in this article to fix bad data is illustrated in Figure 6. After training, the designed model can predict the output power and

other vital external features with any given incomplete input, which is described by $\tilde{\rho}_i$. To evaluate the performance of NN, the loss function is defined as follows:

$$E_k = \frac{1}{n} \sum_{i=1}^n (\hat{\kappa}_k - \kappa_k)^2 \quad (13)$$

where κ_k is the real output in the k th training sample, and $\hat{\kappa}_k$ is the predicted output in the k th training sample.

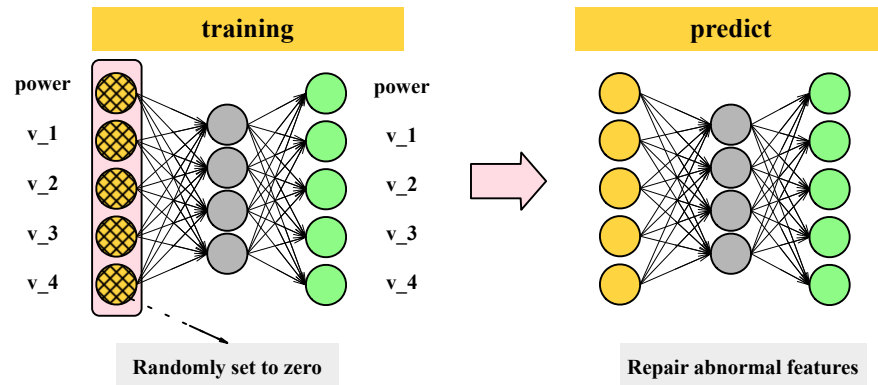


Figure 6. Feature repair network.

3.2. The Data Augmentation (DA) Module

To solve the problem of insufficient and unbalanced data samples, the feature enhancement module in Figure 7 is designed to improve data quality and solve the problem of insufficient and unbalanced data. According to the characteristics of the diurnal variation of the PV power curve, this paper has designed the inverse prediction data augmentation method based on the residual signal. The specific implementation method is as follows.

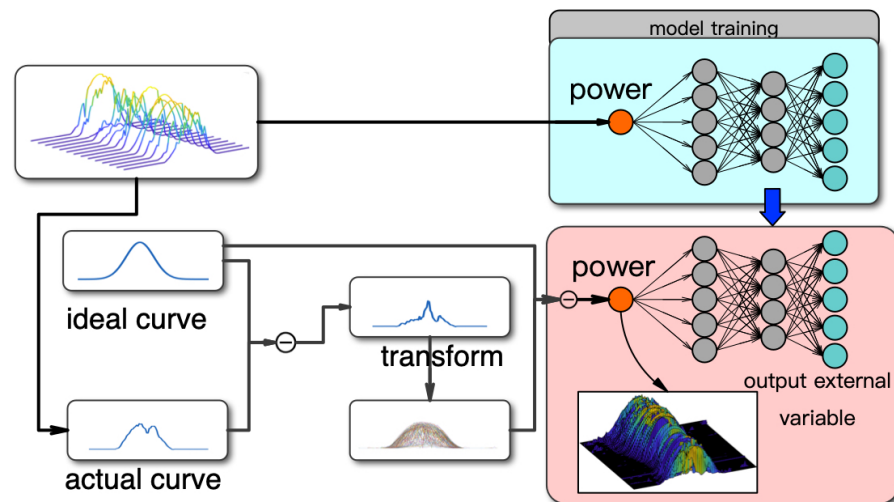


Figure 7. Schematic diagram of data enhancement module.

Step 1: reverse prediction of existing historical data. By fine-tuning the encoder model trained by the FS-DE module, a reverse prediction model, $\zeta(\cdot)$, can be trained quickly.

Step 2: calculate the residual signal, and then transform:

$$E^{(d)} = P_{ideal}^{(d)} - P^{(d)} \quad (14)$$

$$\tilde{E}^{(d)} = \Theta(E^{(d)}) \quad (15)$$

where $P_{ideal}^{(d)}$ is the ideal power curve calculated by the mathematical model, and $\Theta(\cdot)$ is a random transformation, like randomly expanded, flipped, transformed, and other operations.

Step 3: calculate the new power output curve, $\tilde{P}^{(d)}$:

$$\tilde{P}^{(d)} = P_{ideal}^{(d)} - \tilde{E}^{(d)} \quad (16)$$

Calculate the external feature curve, $\tilde{V}_k^{(d)}$, by using the model trained in Step 1.

$$\tilde{V}_k^{(d)} = \zeta(\tilde{P}^{(d)}) \quad (17)$$

Step 4: repeat Step 2 and Step 3 every day to obtain the enhanced data, $\tilde{P} = \{\tilde{P}^{(d)} \mid d = 1 \cdots N\}$ and $\tilde{V}_k = \{\tilde{V}_k^{(d)} \mid d = 1 \cdots N\}$

To sum up, the flow chart of the PV power forecasting process is shown in Figure 8.

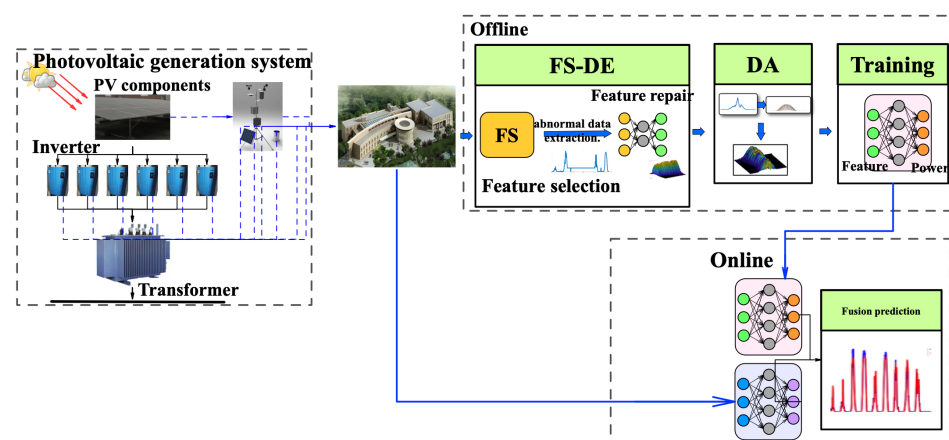


Figure 8. Schematic of the proposed PV power prediction framework.

4. Case Study

In this section, the data from the Shaoxing PV power station and Jiaying PV power station will be used for experiments to verify the effectiveness of the proposed method.

According to Algorithm 1, the correlation between the external characteristics and the output power is calculated and the error data are filtered out for subsequent repair. It can be seen from Table 3 that the characteristics of both stations that have a high correlation with power are cant radiation, direct normal irradiance, diffuse solar radiation, and total radiation. In feature correlation analysis, the FS-DE module can extract noise data at the same time, and then an autoencoder network is used for data repair.

Table 3. Correlation analysis of external features and power.

V_j	Feature	Shaoxing Station	Jiaying Station
V_1	Cant radiation	0.0066	0.0062
V_2	Direct normal irradiance	0.0056	0.0058
V_3	Diffuse solar radiation	0.0069	0.0072
V_4	Total radiation	0.0058	0.0055
V_5	Atmospheric pressure	0.0164	0.0148
V_6	Relative humidity	0.0176	0.0127
V_7	Ambient temperature	0.0127	0.0112
V_8	The panel temperature	0.0092	0.0116
V_9	Wind speed	0.0134	N/A
V_{10}	Wind direction	0.0163	N/A
V_{11}	Precipitation	0.0110	N/A

Algorithm 1 Feature selection and data elimination module.

Input: Power output values at each measuring interval t of each historical day d , denoted by $P = \{P^{(d)} \mid d = 1 \cdots N\}$. A set of external variables $V = \{V_k^{(d)} \mid d = 1 \cdots N; k = 1 \cdots M\}$

Output: $P' = \{P_t^{(d)} \mid d = 1, \dots, N'\}$, $V' = \{V_k^{(d)} \mid d = 1 \cdots N'; k = 1 \cdots M'\}$

- 1: The values of PEDs based on recorded power values are calculated for each and every pair of historical days, say day d and day d' , using (5).
- 2: The values of VEDs are calculated for each pair of historical days based on each external variable V_k , using (6).
- 3: Calculate the PVD value of each external variable using (7). Select M' variable with lower value PVD as the prime external variables (PEVs).
- 4: The value of PVED is calculated according to Formula (8), and then the samples with abnormal data were screened out.
- 5: Use an autoencoder neural network to repair abnormal data.
- 6: **return** P', V'

In this paper, two error measurement methods are used to measure the daily prediction accuracy: normalized mean absolute error (nMAE) and normalized root mean square error (nRMSE). They are defined as:

$$\text{nMAE} = \frac{1}{T} \sum_t \frac{|\hat{p}_t - p_t|}{P_r} * 100\% \quad (18)$$

$$\text{nRMSE} = \frac{\sqrt{\frac{1}{T} \sum_t (\hat{p}_t - p_t)^2}}{P_r} * 100\% \quad (19)$$

where P_t is the power output value at moment t , $\hat{p}_t$ is the predicted power value at moment t , and P_r is the PV installation capacity.

4.1. Results for Shaoxing Station

To verify the effect of the proposed framework, the online experiment is first carried out on Shaoxing station. As a control in the experiment, we take the original data without feature screening and the data augmentation as the baseline, the data that have not been enhanced with feature screening as the FS-DE, and the FS-DE+DA as the proposed method. The neural networks used for prediction all use LSTM.

The experimental results are shown in Table 4. It can be seen from the table that the accuracy of the baseline is much worse than that of the other three, indicating that the original data have fewer effective features and require feature selection to remove redundant features. Similarly, the accuracy of the model is further improved to 9.46% after data repair, which shows that abnormal data have a great impact on the accuracy of the model.

Table 4. Prediction for Shaoxing station.

	nMAE		nRMSE	
	Average	Std	Average	Std
baseline	18.98%	10.64%	22.45%	14.43%
FS	9.82%	8.72%	12.32%	8.43%
FS-DE	9.46%	6.34%	10.24%	9.86%
FS-DE +DA	8.84%	5.43%	9.86%	7.96%

For the prediction analysis under different weather conditions, as shown in Table 5, the prediction accuracy of the FS method under different weather conditions is shown. It can be seen that the FS method achieves better results on sunny days. However, for severe weather conditions (partly cloudy, clouds, and rain), the accuracy of the FS method will be greatly lost and has a large variance. However, as shown in Table 6, compared with the FS method, the FS-DE method slightly improves the prediction accuracy, but the variance improvement is not obvious. This may be due to the fact that the impact of abnormal data in the data is mainly reflected in the misjudgment of the abnormal data themselves, while, for the imbalance of severe weather itself, the FS-DE method has not changed more in nature. Finally, the FS-DE+DA results verify the above statement, as shown in Table 7, that, after adding the data augmentation module, the prediction accuracy of the model is improved, and the prediction accuracy is stable in bad weather, indicating that the data augmentation module can effectively alleviate the data imbalance problem.

Table 5. FS prediction for Shaoxing station.

	nMAE		nRMSE	
	Average	Std	Average	Std
Sunny	10.24%	8.23%	13.43%	10.85%
Partly Cloudy	12.49%	11.34%	15.68%	16.23%
Cloudy	14.52%	12.54%	17.24%	17.37%
Rainy	9.33%	7.24%	11.24%	14.44%

Table 6. FS-DE prediction for Shaoxing station.

	nMAE		nRMSE	
	Average	Std	Average	Std
Sunny	8.99%	8.76%	12.36%	10.23%
Partly Cloudy	11.24%	10.40%	13.42%	14.54%
Cloudy	14.25%	12.42%	14.54%	15.26%
Rainy	9.13%	7.21%	10.24%	12.24%

Table 7. FS-DE+DA prediction for Shaoxing station.

	nMAE		nRMSE	
	Average	Std	Average	Std
Sunny	6.89%	4.46%	8.56%	6.85%
Partly Cloudy	8.94%	8.84%	11.68%	8.23%
Cloudy	8.35%	9.34%	12.64%	9.86%
Rainy	7.34%	5.25%	9.46%	7.34%

4.2. Comparison with Other Existing Methods

To verify the validity of our method, we compared the proposed method with the existing literature, and we used data from Jiaying station for the experiments. To manage the scenario that no historical data were available in the newly built PV power Jiaying station, we adopted a fine-tuning strategy in transfer learning.

As shown in Table 8, the final results indicate that the model accuracy of data enhancement is significantly higher than that of non-augmentation cases. The experiment

further shows that the data augmentation effectively alleviates the distribution imbalance. As shown in Figure 9, compared with other methods, the proposed method has a higher accuracy and lower std value.

Table 8. Comparison with existing methods.

	nMAE		nRMSE	
	Average	Std	Average	Std
baseline	25.43%	16.34%	29.35%	19.32%
FS	13.57%	11.27%	15.67%	13.56%
FS-DE	12.47%	9.98%	13.18%	10.98%
proposed	8.26%	7.28%	10.96%	9.26%
ConvLSTM [25]	15.12%	12.12%	15.98%	13.78%
MC [11]	14.81%	12.07%	15.69%	13.61%
WGPR [33]	12.86%	10.34%	13.26%	13.78%
SRPCNN [34]	9.90%	9.84%	12.89%	10.38%

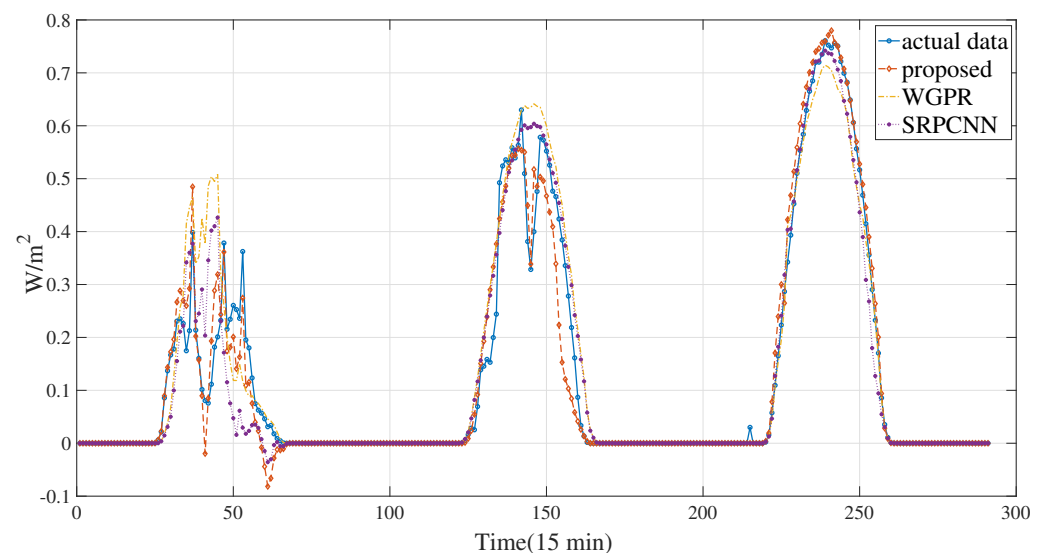


Figure 9. Final prediction effect.

4.3. Model Generalization Ability Experiment

To further demonstrate the effectiveness of the proposed method, data anomaly experiments and data imbalance experiments were designed manually to test the function of each module. As shown in Table 9, the first column is the proportion of abnormal data, and the normal data of Shaoxing station are processed manually into anomalies in a certain proportion. When normal data are treated as abnormal in a certain proportion, the prediction effect of the FS-DE method will be improved to some extent compared with that of the FS-DE method. The improvement effect is the most obvious on sunny days, and the worst on rainy days, because, when normal data are treated as abnormal, the changes on sunny days are the most obvious, and it is easy for them to be identified as abnormal by the FS-DE module for correction. However, the changes on rainy days are the least obvious, and the comparison between the FS-DE method and FS method is not obvious.

Similarly, we designed the data imbalance experiment manually. As shown in Table 10, the data in the first column are the proportion of data imbalance. For example, sunny 1:3 means that the number of days with sunny weather is reduced by 1/3. Under such

conditions, the DA module will play a certain role, and the prediction effect of the FS-DE+DA method will be improved more than that of the FS-DE method.

The improvement is most obvious on cloudy days, while the effect is not obvious on rainy and sunny days. We speculate that it is easier to enhance data on cloudy weather. On sunny days, due to the data's obvious characteristics, their features are more concentrated in the high-dimensional space, so the effect of data augmentation is not obvious, while, on rainy days, their features are more dispersed in the high-dimensional space, and the characteristics and effect of data augmentation are not obvious, and even produce side effects.

Table 9. Anomaly data model generalization experiment.

Error	Weather	nMAE		nRMSE	
		FS	FS-DE	FS	FS-DE
0%	sunny	10.24%	8.99%	13.43%	12.36%
	partly cloudy	12.49%	11.24%	15.68%	13.42%
	cloudy	14.52%	14.25%	17.24%	14.54%
	rainy	9.33%	9.13%	11.24%	10.24%
20%	sunny	12.53%	10.89%	15.70%	14.02%
	partly cloudy	14.67%	13.74%	18.34%	17.17%
	cloudy	16.25%	16.19%	19.12%	17.82%
	rainy	11.18%	10.74%	12.87%	13.08%
40%	sunny	14.86%	12.23%	16.74%	14.76%
	partly cloudy	16.13%	15.39%	18.62%	17.11%
	cloudy	17.36%	16.52%	19.69%	18.64%
	rainy	13.53%	12.79%	14.19%	13.82%

Table 10. Data imbalance experiment.

Rate	Weather	nMAE		nRMSE	
		FS-DE	FS-DE+DA	FS-DE	FS-DE+DA
1:1	sunny	8.99%	6.89%	12.36%	8.56%
	partly cloudy	11.24%	8.94%	13.42%	11.68%
	cloudy	14.25%	8.35%	14.54%	12.64%
	rainy	9.13%	7.34%	10.24%	9.46%
1:3	sunny	11.33%	10.21%	14.43%	12.85%
	partly cloudy	14.37%	10.74%	17.23%	14.57%
	cloudy	17.55%	13.03%	17.82%	14.73%
	rainy	11.67%	10.99%	14.47%	13.88%
1:9	sunny	14.36%	13.63%	16.83%	15.76%
	partly cloudy	18.56%	15.28%	19.62%	16.91%
	cloudy	18.89%	15.72%	19.69%	16.84%
	rainy	13.53%	13.49%	15.49%	15.32%

5. Conclusions

This paper discussed some unique challenges faced by the power generation output prediction of small grid-connected solar generators. Due to the influence of sensor equipment and environmental factors, the prediction accuracy of solar PV power generation is reduced, which leads to the failure of the prediction model in serious cases. Based on this, this paper proposes an online prediction framework for photovoltaic power in the absence of information, including the FS-DE module and DA module. The FS-DE module selects the high similarity features and filters out wrong data according to the similarity of historical daily power. At the same time, the autoencoder trained based on a random

discarding mechanism is used to repair photovoltaic power generation data. Then DA module is designed to eliminate the influence of data imbalance. The final experiments indicate that the proposed method is effective, whether it is sunny, cloudy, or rainy.

Author Contributions: Conceptualization, Y.S. and X.W.; methodology, Y.S., H.S. and X.W.; software, H.S. and S.L.; validation, H.S. and S.L.; formal analysis, Y.S.; investigation, X.W.; resources, X.W.; data curation, X.W.; writing—original draft preparation, X.W.; writing—review and editing, Y.S.; visualization, Y.S.; supervision, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this article are private and cannot be made public temporarily.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. SolarPower Europe (2021): Global Market Outlook for Solar Power 2021–2025. Available online: https://api.solarpowereurope.org/uploads/Solar_Power_Europe_Global_Market_Outlook_for_Solar_2021_2025_V3_7504499a2d.pdf (accessed on 15 October 2023).
2. Agalgaonkar, Y.P.; Pal, B.C.; Jabr, R.A. Stochastic distribution system operation considering voltage regulation risks in the presence of PV generation. *IEEE Trans. Sustain. Energy* **2015**, *6*, 1315–1324. [\[CrossRef\]](#)
3. Luna, A.C.; Diaz, N.L.; Graells, M.; Vasquez, J.C.; Guerrero, J.M. Mixed-integer-linear-programming-based energy management system for hybrid PV-wind-battery microgrids: Modeling, design, and experimental verification. *IEEE Trans. Power Electron.* **2016**, *32*, 2769–2783. [\[CrossRef\]](#)
4. Karimi, Y.; Oraee, H.; Golsorkhi, M.S.; Guerrero, J.M. Decentralized method for load sharing and power management in a PV/battery hybrid source islanded microgrid. *IEEE Trans. Power Electron.* **2016**, *32*, 3525–3535. [\[CrossRef\]](#)
5. Rose, A.; Stoner, R.; Pérez-Arriaga, I. Prospects for grid-connected solar PV in Kenya: A systems approach. *Appl. Energy* **2016**, *161*, 583–590. [\[CrossRef\]](#)
6. Kanchev, H.; Lu, D.; Colas, F.; Lazarov, V.; Francois, B. Energy management and operational planning of a microgrid with a PV-based active generator for smart grid applications. *IEEE Trans. Ind. Electron.* **2011**, *58*, 4583–4592. [\[CrossRef\]](#)
7. Tan, S.; Xu, J.X.; Panda, S.K. Optimization of distribution network incorporating distributed generators: An integrated approach. *IEEE Trans. Power Syst.* **2013**, *28*, 2421–2432. [\[CrossRef\]](#)
8. Wang, Y.; Huo, C.; Xu, F.; Zheng, L.; Hao, L. Ultra-Short-Term Distributed Photovoltaic Power Probabilistic Forecasting Method Based on Federated Learning and Joint Probability Distribution Modeling. *Energies* **2025**, *18*, 197. [\[CrossRef\]](#)
9. Tanaka, K.; Uchida, K.; Ogimi, K.; Goya, T.; Yona, A.; Senjyu, T.; Funabashi, T.; Kim, C.H. Optimal operation by controllable loads based on smart grid topology considering insolation forecasted error. *IEEE Trans. Smart Grid* **2011**, *2*, 438–444. [\[CrossRef\]](#)
10. Meng, X.; Gao, F.; Xu, T.; Zhou, K.; Li, W.; Wu, Q. Inverter-data-driven second-level power forecasting for photovoltaic power plant. *IEEE Trans. Ind. Electron.* **2020**, *68*, 7034–7044. [\[CrossRef\]](#)
11. Hossain, M.S.; Mahmood, H. Short-term photovoltaic power forecasting using an LSTM neural network and synthetic weather forecast. *IEEE Access* **2020**, *8*, 172524–172533. [\[CrossRef\]](#)
12. 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. [\[CrossRef\]](#)
13. Yona, A.; Senjyu, T.; Saber, A.Y.; Funabashi, T.; Sekine, H.; Kim, C.H. Application of neural network to 24-hour-ahead generating power forecasting for PV system. In Proceedings of the 2008 IEEE Power and Energy Society General Meeting—Conversion and Delivery of Electrical Energy in the 21st Century, Pittsburgh, PA, USA, 20–24 July 2008; pp. 1–6.
14. Zhang, P.; Takano, H.; Murata, J. Daily solar radiation prediction based on wavelet analysis. In Proceedings of the SICE Annual Conference 2011, Tokyo, Japan, 13–18 September 2011; pp. 712–717.
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. [\[CrossRef\]](#)
16. Sfetsos, A.; Coonick, A. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. *Sol. Energy* **2000**, *68*, 169–178. [\[CrossRef\]](#)
17. Voyant, C.; Muselli, M.; Paoli, C.; Nivet, M.L. Hybrid methodology for hourly global radiation forecasting in Mediterranean area. *Renew. Energy* **2013**, *53*, 1–11. [\[CrossRef\]](#)

18. Muneer, T.; Gul, M.; Kambezidis, H. Evaluation of an all-sky meteorological radiation model against long-term measured hourly data. *Energy Convers. Manag.* **1998**, *39*, 303–317. [[CrossRef](#)]
19. Lin, Y. The model of the solar radiation energy's establishment and verification. *J. Anhui Inst. Archit. Ind. (Nat. Sci.)* **2007**, *5*, 44–46.
20. Heinemann, D.; Lorenz, E.; Girodo, M. Forecasting of solar radiation. In *Solar Energy Resource Management for Electricity Generation from Local Level to Global Scale*; Nova Science Publishers: New York, NY, USA, 2006; pp. 83–94.
21. Chang, G.W.; Lu, H.J. Integrating gray data preprocessor and deep belief network for day-ahead PV power output forecast. *IEEE Trans. Sustain. Energy* **2018**, *11*, 185–194. [[CrossRef](#)]
22. Wang, S.; Zhang, N.; Zhao, Y.; Zhan, J. Photovoltaic system power forecasting based on combined grey model and BP neural network. In Proceedings of the 2011 International Conference on Electrical and Control Engineering, Yichang, China, 16–18 September 2011; pp. 4623–4626.
23. Zhang, R.; Ma, H.; Hua, W.; Saha, T.K.; Zhou, X. Data-driven photovoltaic generation forecasting based on a Bayesian network with spatial–temporal correlation analysis. *IEEE Trans. Ind. Inform.* **2019**, *16*, 1635–1644. [[CrossRef](#)]
24. Chai, S.; Xu, Z.; Jia, Y.; Wong, W.K. A robust spatiotemporal forecasting framework for photovoltaic generation. *IEEE Trans. Smart Grid* **2020**, *11*, 5370–5382. [[CrossRef](#)]
25. Sheng, H.; Xiao, J.; Cheng, Y.; Ni, Q.; Wang, S. Short-term solar power forecasting based on weighted Gaussian process regression. *IEEE Trans. Ind. Electron.* **2017**, *65*, 300–308. [[CrossRef](#)]
26. Sanjari, M.J.; Gooi, H. Probabilistic forecast of PV power generation based on higher order Markov chain. *IEEE Trans. Power Syst.* **2016**, *32*, 2942–2952. [[CrossRef](#)]
27. Liaoliao, W.; Chenggang, C.; Ning, Y.; Hui, C. Evaluation of solar irradiance on inclined surfaces models in the short-term photovoltaic power forecasting. *J. Eng.* **2017**, *2017*, 2226–2230. [[CrossRef](#)]
28. Huang, C.J.; Kuo, P.H. Multiple-input deep convolutional neural network model for short-term photovoltaic power forecasting. *IEEE Access* **2019**, *7*, 74822–74834. [[CrossRef](#)]
29. Zhou, L.; Wu, H.; Xu, T.; Mei, F.; Li, Y.; Yuan, X.; Liu, H. Ultra-short term hybrid power forecasting model for photovoltaic power station with meteorological monitoring data. In Proceedings of the 2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC), Hefei, China, 19–21 May 2017; pp. 452–456.
30. Tascikaraoglu, A.; Sanandaji, B.M.; Chicco, G.; Cocina, V.; Catalao, J.P.S. Compressive spatio-temporal forecasting of meteorological quantities and photovoltaic power. In Proceedings of the 2017 IEEE Manchester PowerTech, Manchester, UK, 18–22 June 2017.
31. Zhang, X.; Li, Y.; Lu, S.; Hamann, H.F.; Hodge, B.M.; Lehman, B. A solar time based analog ensemble method for regional solar power forecasting. *IEEE Trans. Sustain. Energy* **2018**, *10*, 268–279. [[CrossRef](#)]
32. Wei, B.; Jun, Y.; Yulei, R.; Boris, P. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE* **2017**, *12*, e0180944.
33. Liu, W.; Ren, C.; Xu, Y. PV generation forecasting with missing input data: A super-resolution perception approach. *IEEE Trans. Sustain. Energy* **2020**, *12*, 1493–1496. [[CrossRef](#)]
34. Sheng, H.; Xiao, J.; Wang, P. Lithium iron phosphate battery electric vehicle state-of-charge estimation based on evolutionary Gaussian mixture regression. *IEEE Trans. Ind. Electron.* **2016**, *64*, 544–551. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.