applied sciences

MDPI

Article

Photovoltaic Power-Stealing Identification Method Based on Similar-Day Clustering and QRLSTM Interval Prediction

Shurong Peng [1], Lijuan Guo [1], Bin Li [1,*], Shuang Lu [2], Huixia Chen [1] and Sheng Su [1]

[1] School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha 410114, China
[2] Jiangmen Power Supply Bureau, Guangdong Power Grid Co., Ltd., Guangzhou 529030, China
* Correspondence: libin407@stu.csust.edu.cn

Abstract: In order to defraud state subsidies, some unscrupulous users use improper means to steal photovoltaic (PV) power. This behavior brings potential safety hazards to photovoltaic grid-connected operations. In this paper, a photovoltaic power-stealing identification method based on similar-day clustering and interval prediction of the quantile regression model for long short-term memory neural network (QRLSTM) is proposed. First, photovoltaic data are clustered into three similar days by the similar-day clustering according to weather conditions. Second, compared with the quantile regression neural network (QRNN) prediction method, the good prediction performance of the QRLSTM method is illustrated. Third, using the prediction intervals with different confidence levels on three similar days, according to the time scale (short-term, medium-term and long-term) combined with different electricity-stealing judgment indicators, a three-layer photovoltaic power-stealing screening framework is constructed, and the degree of user power stealing is qualitatively analyzed. Last, the power generation data of eight photovoltaic users in a certain region of northwest China and the data of four groups of artificially constructed power-stealing users are used as an example for simulation. The simulation results prove the feasibility of the proposed method in this paper.

Keywords: photovoltaic power-stealing identification; interval prediction; similar-day clustering; quantile regression model for long short-term memory neural networks; three-layer photovoltaic power-stealing screening

Citation: Peng, S.; Guo, L.; Li, B.; Lu, S.; Chen, H.; Su, S. Photovoltaic Power-Stealing Identification Method Based on Similar-Day Clustering and QRLSTM Interval Prediction. *Appl. Sci.* **2023**, *13*, 3506. https://doi.org/10.3390/app13063506

Academic Editors: Tarek Gaber, Shu-Chuan Chu and Chin-Shiuh Shieh

Received: 20 January 2023
Revised: 1 March 2023
Accepted: 6 March 2023
Published: 9 March 2023

1. Introduction

At present, the problem of energy shortage is increasingly prominent, and the use of new energy instead of fossil energy has become an irreversible trend. Photovoltaic (PV) power generation has been promoted worldwide due to its advantages of low cost, high return and sustainability. However, the rapid development of photovoltaic power generation has also brought some problems. In order to obtain the high subsidies of the state in the field of photovoltaic power generation, some illegal users use certain technical means to make the measured power of photovoltaic meters falsely high. This behavior is called photovoltaic power stealing [1], which not only seriously damages the national interests but also brings potential security risks to the grid-connected operation of photovoltaic power.

The problem of power stealing is a significant problem of the grid [2,3]. For the identification of traditional power stealing, relevant technologies are relatively mature. Kong et al. [4] proposed a power-stealing detection method based on a similarity measure and the decision tree combined K-Nearest Neighbour and support vector machine (DT-KSVM). Huang et al. [5] introduced the stacked sparse denoising autoencoder (SSDAE) to detect the behavior of power stealing. The existing power-stealing identification methods have been able to relatively accurately identify traditional means of power stealing, but photovoltaic power stealing is different from traditional power stealing. Thus, the method applicable to traditional anti-power stealing may not be excellent for photovoltaic anti-power stealing.

There are four typical distributed photovoltaic power-stealing methods: increasing voltage method, increasing current method, photovoltaic simulator method and the grid reconnection method [6]. Both the increasing voltage method and the increasing current method need to change the external connection of the original line, which can be found through on-site inspection. The photovoltaic simulator method can control the output of the photovoltaic simulator according to the change of the external meteorological environment, which has high concealability and is not easy to identify. Power stealing by the grid reconnection method needs to ensure that the electricity consumed by the user is greater than the electricity emitted by the photovoltaic panels [7]. The existence of power stealing can be judged by analysing the electricity consumption of users.

The uncertainty of photovoltaic power generation, the diversity of photovoltaic power-stealing methods and the randomness of photovoltaic power-stealing behavior have brought great challenges to photovoltaic anti-power stealing. The existing photovoltaic power-stealing identification methods can be mainly divided into two categories: one needs to predict the photovoltaic power generation and compares the predicted value with the measured value of the photovoltaic meter [8,9]; the other does not need to predict the photovoltaic power generation and directly uses the data of the power station [10–12]. The first method requires the model to have high prediction accuracy. Shaaban et al. [8] proposed a power-stealing detection method based on the regression tree model, comparing the real value read from the photovoltaic meter with the predicted value obtained from the regression model to judge the power-stealing behavior. But the prediction accuracy of the proposed model is not high, leading to the identification accuracy of power-stealing behavior being not high.

The second method analyzes the data of the power station when there is behavior of power stealing and finds those characteristics of the data that can be used to identify photovoltaic power stealing. Xie et al. [13] put forward a new detection index for the virtual increase of photovoltaic power generation, that is, the changed slope between the data of the photovoltaic meter and the gate meter. The proposed detection index improves the accuracy of photovoltaic power-stealing identification, but this approach is only applicable to the photovoltaic simulator power-stealing method, not to the other three common power-stealing methods. Lu et al. [14] proposed an identification method for abnormal photovoltaic users based on MIV (mean impact value) and heuristic forward searching. However, this method is limited, has no timeliness, and the results have a certain randomness.

In regard to photovoltaic power generation forecasting, there has been much research to improve the accuracy of photovoltaic power generation forecasts. The higher the prediction accuracy, the more accurate the identification of electricity stealing. Talaat et al. [15] proposed two MFFNN-based optimization techniques to predict the DC output power of a PV plant. Zhang et al. [16] proposed an Ultra-short-term power forecasting method based on NWP similarity analysis. At this stage, for photovoltaic power generation forecasting, point forecasting is mostly used, but the information provided by point forecasting results is limited, which is not conducive to reflecting the randomness of forecasting. Therefore, the interval prediction method has been introduced. Interval prediction is an uncertainty prediction, which can reflect the volatility of photovoltaic power generation and can also obtain the probability distribution characteristics of photovoltaic output. Mei et al. [17] proposed a photovoltaic interval prediction method based on the adaptive-rolling matching-prediction correction mode. Wan et al. [18] established a linear programming-based prediction interval construction model for photovoltaic power generation.

In summary, the main method of photovoltaic power-stealing identification is to compare the predicted photovoltaic power generation with the photovoltaic metering meter. However, the main problem of this method is that the prediction accuracy of photovoltaic power generation in the existing research still has room for improvement, so it is necessary to constantly update the algorithm to improve the prediction accuracy. The other category of identification methods which do not rely on photovoltaic prediction also

has some problems, such as difficulty in data acquisition, insufficient accuracy and limited scope of application. These problems are considered comprehensively in carrying out the work of this paper. This work aims to study a photovoltaic power-stealing identification method based on similar-day clustering and quantile regression long short-term memory (QRLSTM) neural interval prediction. The following are the primary contributions of this work:

- The principles of four photovoltaic power-stealing methods and the data characteristics after power stealing were investigated, and four groups of power-stealing user data were artificially constructed to fully verify the effect of the proposed method.
- The method of combining similar-day clustering and the QRLSTM network was used to forecast photovoltaic power generation under different types of weather conditions to improve the prediction accuracy. It was considered that the efficiency of photovoltaic power generation in the same weather is similar.
- In order to make the identification result of the final power stealing more stable and accurate, three-layer screening criteria were set from the time scale, which were specifically manifested as short-term, medium-term and long-term.

2. Materials and Methods

2.1. Analysis of Photovoltaic Output Characteristics

2.1.1. Influencing Factors of Photovoltaic Output

The factors affecting photovoltaic output mainly include two aspects: one is internal factors, including the power generation performance and conversion efficiency of photovoltaic arrays, the installation angle of PV panels and its operation mode, the type of system components and geographical location, etc.; the other is external factors, which mainly refer to the climatic conditions of the location of the photovoltaic power generation grid [19,20]. Considering that the geographical locations of the users selected for the identification of electricity stealing in this paper are adjacent to each other, it is assumed that their power generation efficiency is similar, and only the influence of meteorological factors is considered. Meteorological factors that affect photovoltaic output mainly include weather type, actual irradiance, temperature, humidity, wind speed, pressure, etc. [21]. However, in addition to weather type and irradiance, which have obvious effects on PV output, the effect of other factors on PV output are often indirect and unclear. In order to improve the prediction effect of the model, further feature mining is carried out on the meteorological features, and some new features are constructed.

2.1.2. Feature Construction and Selection

Commonly used feature construction methods include statistical features and combined features [21]. Statistical feature construction refers to the construction of statistical indicators such as the mean and standard deviation of different features, which can reflect the fluctuation of meteorological data over a period of time. Combined feature construction is the use of four operations to "add, subtract, multiply, and divide" each feature to obtain new features, which allow new features of linear and nonlinear types to be obtained.

A large number of new features are obtained through feature construction, but the more features are not better. Too many features will make the model complex, resulting in increased training time; too few features will not improve the performance of the model. Therefore, all the constructed features should be selected according to the correlation between the feature and the actual photovoltaic power.

This paper uses the extreme gradient-boosting (XGBoost) algorithm to filter features. The feature importance score, a sub-module of the XGBoost algorithm, can calculate the importance of each attribute in the decision tree. Therefore, it can be used to better evaluate the degree of correlation between all structural features and the actual power of photovoltaics.

2.1.3. Similar-Day Clustering

Similar days refers to the days in a quarter that have the same weather characteristics as the predicted day. Compared with other dates, similar days can better reflect the photovoltaic output law of the predicted day [22]. Using clustered historical data for prediction can greatly improve the accuracy of model prediction. In this paper, a similar-day clustering method combining a fuzzy C-means (FCM) clustering algorithm and discrete Fréchet distance [23] is used. Similar days are clustered into three categories.

2.2. PV Output Interval Prediction

2.2.1. LSTM Model

The long- and short-term memory neural network (LSTM) is an improved version of the traditional recurrent neural network (RNN), and its network structure is shown in Figure 1.

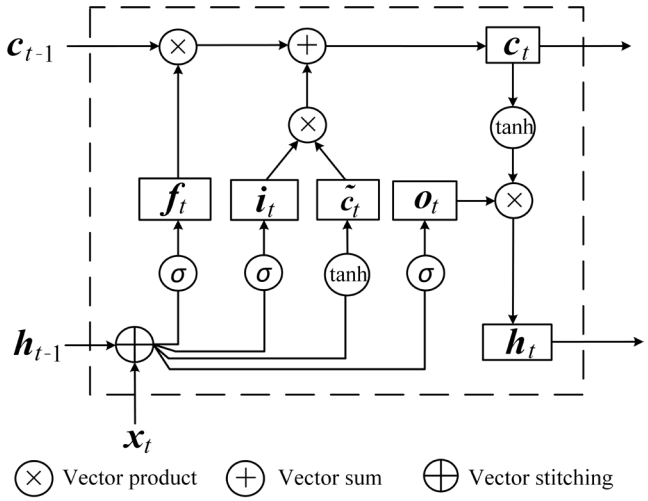


Figure 1. LSTM network structure diagram.

LSTM realizes the protection and control of information through forgetting gates, input gates, and output gates. The specific input–output relationship is shown in Equation (1).

$$\begin{cases} f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ \tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \\ c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t = o_t \odot \tanh(c_t) \end{cases} \tag{1}$$

where f_t is forgetting gates; i_t is input gates; o_t is output gates; $\sigma(\cdot)$ represents the sigmoid activation function; x_t is the current input vector; W_f, W_i, W_o, W_c, U_f, U_i, U_o, U_c are weight matrices; b_f, b_i, b_o, b_c are bias matrices; h_t is the current hidden layer vector; h_{t-1} is the hidden layer vector at last moment; $\tilde{c}_t$ is the candidate value; c_t is the updated status; $\odot$ represents dot product; and tanh is the hyperbolic tangent activation function.

2.2.2. Quantile Regression

The quantile regression method can well analyze the quantile relationship between a set of explanatory variables and the explained variables. Compared with the least-squares method, the quantile regression method is more robust when dealing with outlier data and reflects more comprehensive data information [24,25]. Assuming a set of sample sequence

explanatory variables and explained variables, the parameters of the linear regression model are obtained by Equation (2).

$$\min \sum_{i=1}^{n} (Y_i - X_i\beta) \tag{2}$$

where β is the regression coefficient, which changes with the change of the quantile $\tau(0 < \tau < 1)$. The estimation of the β can be transformed into an optimization problem, which can be obtained by Equation (3).

$$\begin{aligned} &\min_{\beta} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i'\beta) \\ &= \min_{\beta} \sum_{i|Y_i \geq X_i'\beta} \tau\left|Y_i - X_i'\beta\right| + \sum_{i|Y_i < X_i'\beta} (1-\tau)\left|Y_i - X_i'\beta\right| \end{aligned} \tag{3}$$

where X_i' denotes the explanatory variable with quantile condition; $i|Y_i \geq X_i'\beta$ denotes that the actual value of the i-th explained variable is not less than the estimated value of linear regression; ρ is the optimization function—when ρ is the smallest, the optimal estimate of β is obtained. The calculation formula is shown in Equation (4), where $I(\cdot)$ is the indicative function.

$$\begin{cases} \rho_\tau(u) = u[\tau - I(u)] \\ I(u) = \begin{cases} 1 & u < 0 \\ 0 & u \geq 0 \end{cases} \end{cases} \tag{4}$$

It can be seen from Equation (3) that for each different quantile τ, the corresponding parameter estimate $\beta(\tau)$ can be calculated. When τ is continuously valued in the (0,1) interval, the complete range of the explained variable and its conditional distribution can be obtained, and the conditional density prediction can finally be obtained after the condition density is obtained.

2.2.3. QRLSTM PV Interval Prediction Model

The QRLSTM model can be obtained by combining the quantile regression method with the long short-term memory neural network. The expression of the loss function of the QRLSTM model is shown in Equation (5).

$$\begin{aligned} f_{\text{cost}} &= \sum_{i=1}^{N} \rho_\tau[Y_i - f(X_i, W, b)] \\ &= \sum_{i|Y_i \geq f(X_i,W,b)} \tau|Y_i - f(X_i, W, b)| + \sum_{i|Y_i < f(X_i,W,b)} (1-\tau)|Y_i - f(X_i, W, b)| \end{aligned} \tag{5}$$

where N represents the number of samples Y_i, and $Y_i \geq f(X_i, W, V)$ represents that the estimated value of the i-th regression model is less than or equal to the actual value of the explained variable. The objective function of the QRLSTM model is given by Equation (6).

$$\min_{W,b} f_{\cos t} + \frac{\lambda}{2}\left|\left(\hat{W}(\tau), \hat{b}(\tau)\right)\right|^2 \tag{6}$$

where $\hat{W}(\tau)$ is the weight under the quantile condition and $\hat{b}(\tau)$ is the bias item weight under the quantile condition. Using the Adam stochastic gradient descent algorithm, the parameters $\hat{W}(\tau)$ and $\hat{b}(\tau)$ can be solved, and then the results are put into Equation (7), where the condition quantile estimate of the explained variable can be obtained.

$$\hat{Q_Y}(\tau|X) = f\left(X, \hat{W}(\tau), \hat{b}(\tau)\right) \tag{7}$$

The quantile τ is continuously valued in the interval (0, 1), and the model learns the nonlinear relationship under different quantile conditions so as to obtain the corresponding

optimal estimated value under different quantile conditions. The conditional quantile curve is called the conditional distribution function, and the conditional density P can be derived from the distribution function, as shown in Equation (8).

$$P\left(\hat{Q}_Y(\tau|X)\right) = \frac{d\tau}{d\hat{Q}_Y(\tau|X)} \tag{8}$$

where $\hat{Q}_Y(\tau|X)$ is the conditional quantile curve. Then, after conditioning X and discretizing τ in Equation (8), the probability density function of the explained variable can be obtained through Gaussian kernel density estimation. The specific flow chart is shown in Figure 2.

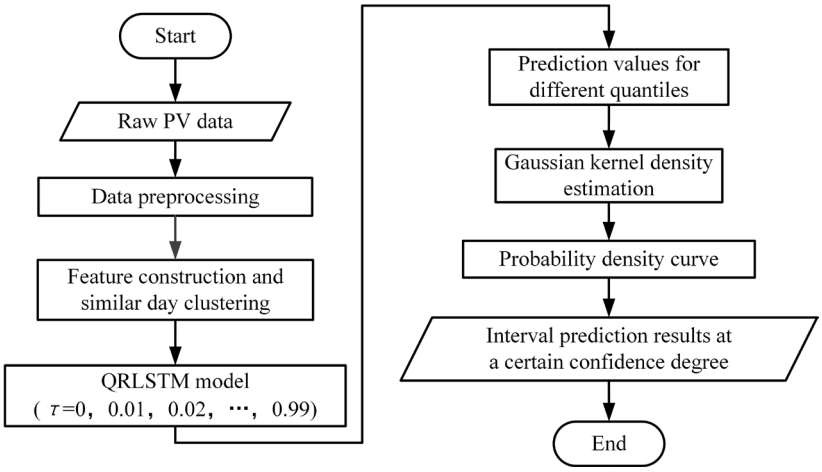


Figure 2. QRLSTM prediction flowchart.

2.3. Identification of PV Power Stealing Based on Interval Prediction

In order to identify PV power stealing more comprehensively and accurately, this paper adopts the short-, medium- and long-term time series scale as the hierarchical index of electricity stealing screening [26].

2.3.1. Judgment of Suspected Short-Term Power Stealing

The judgement of short-term power-stealing suspicion is mainly on the daily scale. The judgement includes two layers. The first layer screening adopts the prediction results at the 95% confidence level. At this time, the coverage rate of the interval is highest, and the average width of the interval is the widest. So, it is suitable for screening the most obvious power-stealing behaviors of users, such as in the increasing voltage method and increasing current method. The first layer of power-stealing identification mainly uses two moments in the morning and evening (6:00–8:00 and 16:00–18:00). No matter what the weather type is, the solar radiation at sunrise and sunset is relatively weak, so theoretically, the power generation of the photovoltaic array is very small. However, for some users who steal electricity, the photovoltaic output in these two periods will be significantly higher than the actual output. Therefore, the photovoltaic power generation in these two periods can be used to determine whether the user has the behavior of stealing power. If the user is found to be three or more times outside the prediction interval in these two periods or fails to fall within the prediction interval for more than five times in a day, it is directly determined that the user is suspected of power stealing.

The second layer of screening is for users who cannot be determined in the first layer, and the key is to calculate the deviation of the points that fall outside the prediction interval.

To this end, a normalized average deviation (NAD) indicator is introduced. The expression is as shown in Equations (9) and (10).

$$e_{NAD} = \frac{1}{N_{interval}} \sum_{i=1}^{N_{interval}} \gamma_i, \tag{9}$$

$$\gamma_i = \begin{cases} \dfrac{y_i - U_i}{\frac{1}{N_{interval}} \sum_{i=1}^{N_{interval}} (U_i - L_i)} & y_i > U_i \\ 0 \quad y_i \in [L_i, U_i] \\ \dfrac{L_i - y_i}{\frac{1}{N_{interval}} \sum_{i=1}^{N_{interval}} (U_i - L_i)} & y_i < L_i \end{cases}, \tag{10}$$

where e represents the deviation and y_i represents the actual output value. The larger the e_{NAD} is, the farther the non-falling point is from the prediction interval, and a reasonable upper limit of deviation is set by referring to the identification of experts and comparing the specific samples of historical photovoltaic power-stealing users to screen out the suspected power-stealing users. If the value of the e_{NAD} is more than 20% on sunny and cloudy days or more than 30% on rainy days, the user is suspected of power stealing. The second layer of screening uses the prediction results at the 90% confidence level. Here, the average width of the interval is slightly smaller than that at the 95% confidence level, which is suitable for screening the less obvious power-stealing user behavior, such as in the photovoltaic simulator method.

2.3.2. Judgment of Suspected Medium- and Long-Term Power Stealing

The identification of medium- and long-term photovoltaic power stealing is mainly on a monthly scale. Due to the randomness and uncertainty of photovoltaic power generation, only using the daily scale to identify power stealing is not comprehensive enough, so the third layer of screening is carried out. The specific screening process is shown in Figure 3. By counting the number of days with suspicion of power stealing selected by users in the first two layers within a month, the identification result is given.

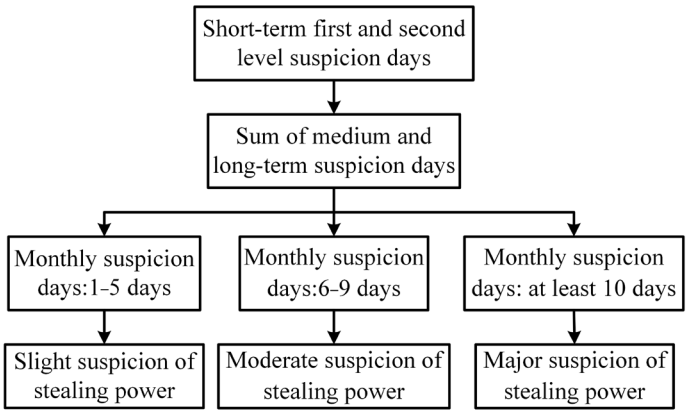


Figure 3. Flowchart of medium- and long-term power-stealing determination.

Through these three layers of screening, the user's degree of power stealing can be determined: when the number of days that a PV user is suspected of power stealing in a month is within the range of 1–5 days, it is determined that the user is suspected of mild power stealing; when the number of days is within the range of 6–9 days, it is determined that the user is suspected of moderate power stealing; when the number of days reaches 10 days, it is determined that the user is suspected of major power stealing. For users who are suspected of mild or moderate power stealing, appropriate offline inspections can be carried out to eliminate the deviation from the prediction interval caused by prediction errors and the photovoltaic power generation loss problem. If the suspicion remains

unchanged or aggravated in the next inspection cycle, inspectors can go directly to the site to rectify the users. For users suspected of major stealing of power, inspectors can go directly to the scene to conduct a comprehensive inspection to determine whether they had stolen power. The overall power-stealing identification process is shown in Figure 4.

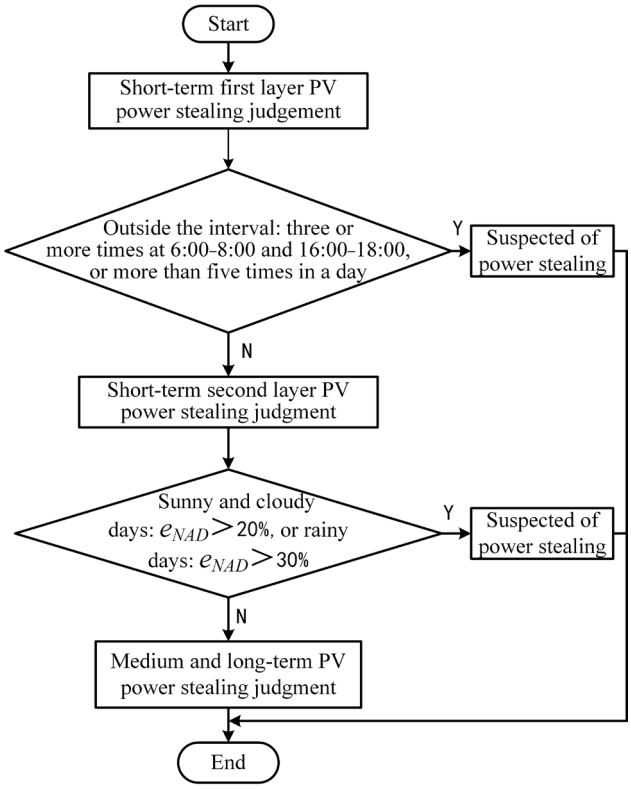


Figure 4. Flow chart of overall power-stealing identification.

3. Results and Discussion

3.1. Data Preprocessing

This paper uses the historical photovoltaic data from February 2018 to January 2019 in a certain region in northwest China for simulation and selects the valid data for a total of 12 h during the day from 6:00 to 18:00. Eight users with high correlation of photovoltaic output in this area were selected, and then four power-stealing users were constructed according to the characteristics of common power-stealing methods to form the sample data of this paper.

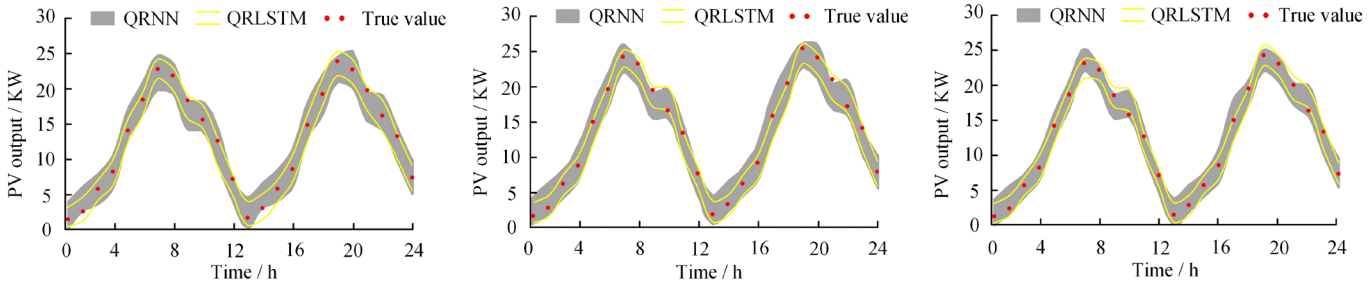
Before the simulation, the sample data was pre-processed, and the missing values in the data were filled using piecewise linear interpolation. At the same time, considering that the numerical ranges of different eigenvalues are different, the standard deviation standardization method was used to process the data, and the processed data presented a normal distribution with a mean value of 0 and a standard deviation of 1. The formula for standard deviation standardization is shown in Equation (11), where: x_{imean} is the mean value of the original data samples, and x_{ivar} is the standard deviation of the original data samples.

$$x'_{ij} = \frac{x_{ij} - x_{imean}}{x_{\mathrm{ivar}}}. \tag{11}$$

3.2. QRLSTM PV Interval Prediction Simulation

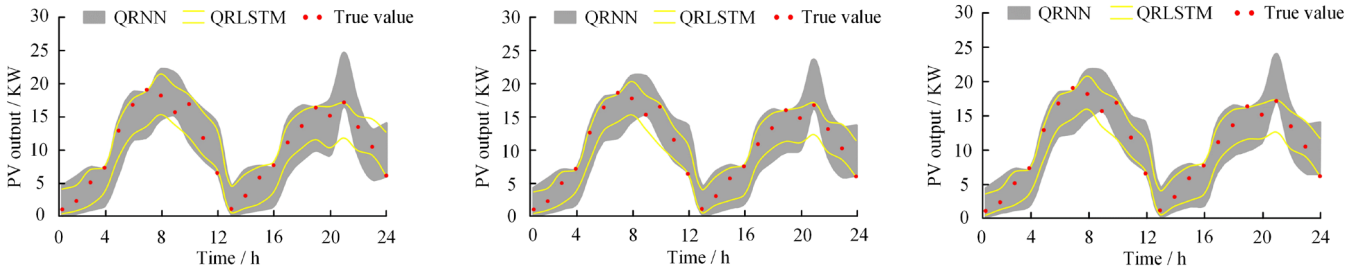
The pre-processed data was feature-constructed, and then divided into a training set and a test set in a ratio of 4:1. Firstly, to improve the prediction accuracy, the historical data were clustered into three types according to the weather conditions, using the FCM–Fréchet clustering method mentioned in Section 2.3. Then, the QRLSTM prediction model was

used to make interval predictions on three groups of similar-day data, the photovoltaic output in the next 24 moments (2 days) was predicted, and three confidence levels of 85%, 90%, and 95% were selected for analysis. In order to analyze the performance of the prediction method used in this paper, the prediction results were compared with those of the Quantile Regression Neural Network (QRNN) model. The prediction results are shown in Figures 5–7.



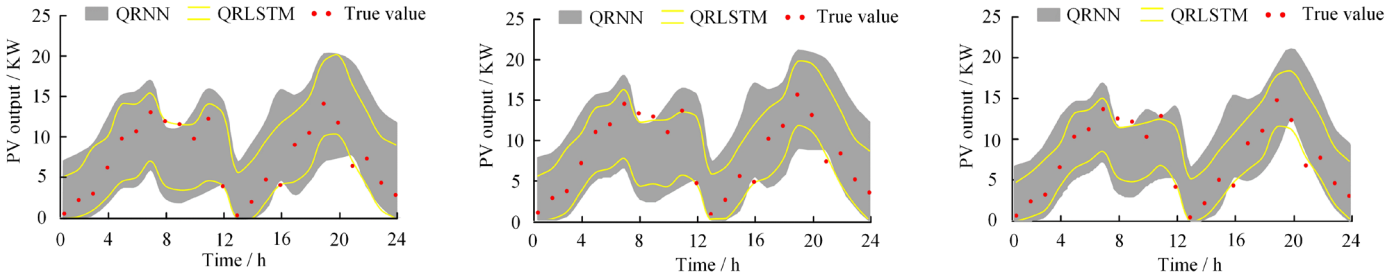
(a) 95% confidence level (b) 90% confidence level (c) 85% confidence level

Figure 5. Comparison of interval prediction results on similar-day 1.



(a) 95% confidence level (b) 90% confidence level (c) 85% confidence level

Figure 6. Comparison of interval prediction results on similar-day 2.



(a) 95% confidence level (b) 90% confidence level (c) 85% confidence level

Figure 7. Comparison of interval prediction results on similar-day 3.

It can be seen from Figures 5–7 that the QRLSTM method can better predict the photovoltaic output. Compared with the prediction results of the QRNN method, the prediction interval of the QRLSTM method could contain more real values and the width of the interval was narrower, regardless of the type of the similar day.

In order to compare the two methods more intuitively, the evaluation indicators of interval prediction are introduced: prediction interval coverage probability (PICP) and prediction interval normalized average width (PINAW). Prediction interval coverage probability, which refers to the probability that the true value falls within the prediction interval, is used to assess the reliability of the prediction interval. Only when the width of the prediction interval is narrower, and the coverage rate of the interval is higher does it

mean that the prediction effect of the interval is better. The calculation formulas of the two evaluation indicators are shown in Equations (12) and (13).

$$\begin{cases} I_i^\alpha = \left[L_i^\alpha, U_i^\alpha\right] \\ \xi_i^\alpha = \begin{cases} 0 & y_i \notin I_i^\alpha \\ 1 & y_i \in I_i^\alpha \end{cases} \qquad (i = 1, 2, \cdots, N_{interval}), \\ PICP = \frac{1}{N_{interval}} \sum_{i=1}^{N_{interval}} \xi_i^\alpha \end{cases} \tag{12}$$

where I_i^α represents the i-th interval at a certain confidence level; L_i^α and U_i^α are, respectively, the lower and upper bounds of the corresponding interval; and ξ_i^α indicates whether the actual value y_i is within the prediction interval—1 represents that it is in the interval, 0 represents it is not. $N_{interval}$ is the number of intervals. The probability that the actual value y_i falls within the interval should not be less than or close to the set confidence level, which means that the prediction model is effective.

$$\begin{cases} \delta_i^\alpha = U_i^\alpha - L_i^\alpha \\ PINAW = \frac{1}{N_{interval}} \sum_{i=1}^{N_{interval}} \delta_i^\alpha \end{cases}, \tag{13}$$

where δ_i^α is the width of the i-th interval. The smaller the *PINAW* index is, the smaller the obtained prediction interval is.

Tables 1–3 show the comparison of the evaluation indicators of the two interval prediction models on three similar days.

Table 1. Comparison of evaluation indicators on similar-day 1.

Confidence Levels	Model	PICP (%)	PINAW (KW)
85%	QRLSTM	88.05	3.35
	QRNN	86.49	4.57
90%	QRLSTM	95.18	3.76
	QRNN	92.58	5.02
95%	QRLSTM	99.67	4.29
	QRNN	96.43	5.63

Table 2. Comparison of evaluation indicators on similar-day 2.

Confidence Levels	Model	PICP (%)	PINAW (KW)
85%	QRLSTM	86.05	4.71
	QRNN	82.11	6.37
90%	QRLSTM	92.18	5.16
	QRNN	89.58	6.94
95%	QRLSTM	96.67	5.78
	QRNN	95.00	7.53

Table 3. Comparison of evaluation indicators on similar-day 3.

Confidence Levels	Model	PICP (%)	PINAW (KW)
85%	QRLSTM	85.05	5.86
	QRNN	80.11	8.37
90%	QRLSTM	88.18	6.43
	QRNN	84.58	9.08
95%	QRLSTM	95.00	7.19
	QRNN	88.43	9.83

From Tables 1–3, it can be seen that on different similar days, with the increase of confidence level, the prediction interval coverage probability will increase, and the prediction interval normalized average width will also increase.

It can be seen from Table 1 that the interval coverage of the two methods exceeds the corresponding confidence level on similar-day 1. When the confidence level is 95%, the QRLSTM method can even achieve close to 100% coverage, and the average width of the interval is also smaller than that of the QRNN method. At other confidence levels, it can also be concluded that the prediction accuracy of the QRLSTM method is higher than that of the QRNN method on similar-day 1.

It can be seen from Table 2 that at the 95% confidence level, the interval coverage rate of QRLSTM is only about 95%, and full coverage of the interval is not achieved; at other confidence levels, it is only slightly larger than the corresponding confidence level. For the QRNN method, the interval coverage is even lower than the corresponding confidence level. At the same time, the average width of the interval of the QRLSTM method is smaller than that of the QRNN method at any confidence level. In summary, Table 2 shows that the prediction effect of the QRLSTM method is better than that of the QRNN method on similar-day 2. The QRLSTM method achieves a relatively high interval coverage in a relatively narrow interval.

It can be seen from Table 3 that the prediction effects of the QRLSTM and QRNN methods are not ideal on similar-day 3. The prediction effect of the QRNN method is significantly reduced, and the coverage of each prediction interval cannot reach the corresponding confidence level. However, the QRLSTM method can still achieve 95% coverage at the 95% confidence level, and the coverage under other confidence levels can also remain similar to the confidence level. Therefore, on similar-day 3, the prediction effect of the QRLSTM method is better than that of the QRNN method.

To sum up, Tables 1–3 all reflect the excellent prediction effect of the QRLSTM method numerically.

3.3. Simulation of PV Power-Stealing Identification Based on Interval Prediction

In Section 3.2, the good prediction performance of the QRLSTM interval prediction model was introduced. In this section, the trained QRLSTM model will be used for interval prediction, and then power-stealing identification will be completed based on the interval prediction results.

User 1 was selected to be the benchmark user. The benchmark user must ensure the accuracy of its power generation. If necessary, appropriate on-site inspection can be carried out on the benchmark user. Only after it is confirmed that it meets the requirements of power generation operation can it be established as the benchmark user. The meteorological data of the benchmark user is input into the trained QRLSTM model for prediction to obtain the prediction interval. The prediction interval and the power generation data of the users to be identified are input variables in the photovoltaic power-stealing identification process. According to the judgment standard and power-stealing identification process formulated in Section 2.3, four constructed power-stealing users in turn are identified, and the identification results are shown in Table 4. It can be seen from Table 4 that all the constructed users can be effectively identified. The constructed users 1 and 2 are identified in the first layer of screening, and the basis of identification is that 6 moments in one day and 3 moments at 6:00–8:00 and 16:00–18:00 exceed the prediction interval. The constructed users 3 and 4 have a more concealed way of power stealing and enter the second layer screening, but still can be identified, with the e_{NAD} of 22.6% and 27.1%. The judgment results are in line with the actual situation, which can confirm the effectiveness of the method proposed in this paper.

Table 4. Identification Results of Constructed Power-Stealing Users.

Users	Basis of Identification	Identification Result
Constructed user 1	The first layer, exceeds time in one day, H = 6	Suspected of power stealing
Constructed user 2	The first layer, exceeds time at 6:00–8:00 and 16:00–18:00, H = 3	Suspected of power stealing
Constructed user 3	The second layer, e_{NAD} = 22.6%	Suspected of power stealing
Constructed user 4	The second layer, e_{NAD} = 27.1%	Suspected of power stealing

By comparing the actual output value of other users with the falling situation of the predicted interval, and then checking one by one according to the judgment indicators of the three-layer screening, the final results are shown in Table 5. According to the monthly total suspicion days, the degree of the user's power-stealing can be identified. Users 2–7 are not suspected of power stealing, and user 8 is suspected of mild power stealing. Constructed users 1 and 2 are suspected of major power stealing, while constructed users 3 and 4 are suspected of moderate power stealing. This method effectively completes the identification and analysis of power stealing by photovoltaic users.

Table 5. Identification results of all users.

Users	Monthly Total Suspicion Days	Identification Result
2	0	not suspected of power stealing
3	0	not suspected of power stealing
4	0	not suspected of power stealing
5	0	not suspected of power stealing
6	0	not suspected of power stealing
7	0	not suspected of power stealing
8	3 days	suspected of mild power stealing
Constructed user 1	10 days	suspected of major power stealing
Constructed user 2	13 days	suspected of major power stealing
Constructed user 3	6 days	suspected of moderate power stealing
Constructed user 4	8 days	suspected of moderate power stealing

The recognition results of the proposed method are compared with those of the method proposed in Reference [14], as shown in Table 6. Both of them can effectively identify the power-stealing user, but the method proposed in this paper is more systematic and comprehensive, and it can qualitatively analyze the severity of power stealing. For PV User 7 and User 8, they are not screened in the method of Reference [14], but the final judgment result has not been determined, which needs to be combined with the results of the next cycle. As can be seen from Table 5, user 7 is not suspected of stealing power, which reflects that the MIV-heuristic forward search method has some errors. User 8 is indeed a mild power stealer, and the method in this paper can e effectively identified it within one cycle, which is more timely than the method in Reference [14]. Moreover, the three-layer screening structure can be used for more comprehensive identification, including short- and medium-term identification.

Although the method proposed in this paper can identify photovoltaic power-stealing users more accurately and comprehensively, there are still some limitations in this paper. In the subsequent research, the identification accuracy of power stealing can be improved by further optimization from the perspectives of identifying indicators of power stealing and improving the accuracy of the photovoltaic output prediction model.

Table 6. Comparison of identification results of two methods.

Users	MIV-Heuristic Forward Search Method in Reference [14]	Method Based on Similar-day Clustering and QRLSTM Interval Prediction
2	not suspected of power stealing	not suspected of power stealing
3	not suspected of power stealing	not suspected of power stealing
4	not suspected of power stealing	not suspected of power stealing
5	not suspected of power stealing	not suspected of power stealing
6	not suspected of power stealing	not suspected of power stealing
7	Suspected of power stealing or out of order	not suspected of power stealing
8	Suspected of power stealing or out of order	suspected of mild power stealing
Constructed user 1	suspected of power stealing	suspected of major power stealing
Constructed user 2	suspected of power stealing	suspected of major power stealing
Constructed user 3	suspected of power stealing	suspected of moderate power stealing
Constructed user 4	suspected of power stealing	suspected of moderate power stealing

4. Conclusions

Photovoltaic power stealing is a significant problem after photovoltaic grid-connected operations, which is related to the safe and stable operation of power systems and national interests. In this paper, a photovoltaic power generation interval prediction model based on similar-day clustering and QRLSTM is proposed. Experiments show that under three similar-day conditions, the prediction results of the QRLSTM model are better than that of the QRNN model. On this basis, a three-layer filter structure based on a time scale is used so as to identify the power-stealing users more comprehensively and accurately. In addition, the data used in this paper are historical meteorological data and photovoltaic user power data, which are easier to obtain than those used in other studies. Considering the lack of data of actual users of power stealing, this paper analyzes the principle of the typical means of photovoltaic power stealing and the data characteristics after power stealing, artificially constructs the data of power stealing, and fully verifies the effectiveness of the method proposed in this paper.

Author Contributions: Conceptualization, S.P. and S.L.; methodology, B.L. and S.L.; software, L.G. and H.C.; validation, L.G. and H.C.; writing—original draft preparation, L.G. and H.C.; writing—review and editing, L.G. and H.C.; supervision, S.P., B.L. and S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Key Project of Education Department of Hunan Province (grant number: No. 20A021) and National Natural Science Foundation of China (grant number: No. 52177069).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in this paper can be obtained by contacting the authors of this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Luo, Y.F.; Jiang, C.W.; Li, C.Z.; Xue, X.L. Integrated management method of anti-electricity stealing considering distributed generation. *Electr. Energy Manag. Technol.* **2015**, *59*, 49–55. [CrossRef]
2. Gong, X.J.; Tang, B.; Zhu, R.J.; Liao, W.; Song, L. Data augmentation for electricity theft detection using conditional variational auto-encoder. *Energies* **2020**, *13*, 4291. [CrossRef]
3. Liu, H.Q.; Li, Z.Q.; Li, Y.C. Noise Reduction Power Stealing Detection Model Based on Self-Balanced Data Set. *Energies* **2020**, *13*, 1763. [CrossRef]
4. Kong, X.Y.; Zhao, X.; Liu, C.; Li, Q.; Dong, D.; Li, Y. Electricity theft detection in low-voltage stations based on similarity measure and DT-KSVM. *Int. J. Electr. Power Energy Syst.* **2021**, *125*, 106544. [CrossRef]
5. Huang, Y.F.; Xu, Q.F. Electricity theft detection based on stacked sparse denoising autoencoder. *Int. J. Electr. Power Energy Syst.* **2021**, *125*, 106448. [CrossRef]

6. Yuan, X.D.; Shi, M.M.; Sun, Z.Y. Research status of electricity-stealing identification technology for distributed PV. In Proceedings of the 2015 5th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), Changsha, China, 26–29 November 2015; pp. 2031–2034. [CrossRef]
7. Hu, C.; Sun, C.S.; Liu, J.; Feng, X.; Huaqin, H.; Binpeng, G. PV electricity stealing identification method considering weather type. *Distrib. Energy* **2017**, *2*, 13–19. [CrossRef]
8. Shaaban, M.; Tariq, U.; Ismail, M.; Almadani, N.A.; Ahmed, M. Data-Driven detection of electricity theft cyberattacks in PV generation. *IEEE Syst. J.* **2022**, *16*, 3349–3359. [CrossRef]
9. Zheng, G.D. Research of Short-Term Photovoltaic Power Prediction and Anti-Stealing Electricity Based on Machine Learning. Master's Thesis, Changsha University of Science and Technology, Changsha, China, 2020. [CrossRef]
10. Jing, J.; Li, Y. Research on the detection algorithm of HCM photovoltaic power theft in Qinghai. In Proceedings of the 2021 IEEE Sustainable Power and Energy Conference (iSPEC), Nanjing, China, 23–25 December 2021; pp. 55–60. [CrossRef]
11. Li, J.G.; Rong, N.; Chen, C.Q. A data augmentation method for distributed photovoltaic electricity theft using generative adversarial network. *J. Electr. Power Sci. Technol.* **2022**, *37*, 181–190. [CrossRef]
12. Xue, Y.; Yang, Y.N.; Liao, W.L.; Yang, D. Data augmentation method for distributed photovoltaic electricity theft based on non-linear independent components estimation. *Autom. Electr. Power Syst.* **2022**, *46*, 171–179.
13. Xie, Y.Y.; Yuan, X.D.; Sun, Z.Y.; Shi, M.; Yin, M.; Zou, Y. Feature analysis and detection method research of inflated electricity using PV generation simulator. *Power Syst. Technol.* **2016**, *40*, 1703–1708. [CrossRef]
14. Lu, S.; Peng, S.R.; Yang, Y.H.; Su, S.; Liu, D.; Zhang, H.; Wang, S. Identification method of abnormal photovoltaic users based on mean impact value and heuristic forward searching. *Electr. Power Autom. Equip.* **2022**, *42*, 106–111. [CrossRef]
15. Talaat, M.; Taghreed, S.; Mohamed, A.; Hatata, A. Integrated MFFNN-MVO approach for PV solar power forecasting considering thermal effects and environmental conditions. *Int. J. Electr. Power Energy Syst.* **2022**, *135*, 107570. [CrossRef]
16. Zhang, S.; Dong, L.; Ji, D.Y.; Hao, Y.; Zhang, X. Power forecasting of ultra-short-term photovoltaic station based on NWP similarity analysis. *Acta Energy Sol. Sin.* **2022**, *43*, 142–147. [CrossRef]
17. Mei, F.; Gu, J.Q.; Pei, X.; Zheng, J. Photovoltaic interval prediction based on adaptive rolling matching prediction correction mode. *Electr. Power Autom. Equip.* **2022**, *42*, 92–98. [CrossRef]
18. Wan, C.; Lin, J.; Song, Y.H.; Xu, Z.; Yang, G. Probabilistic Forecasting of Photovoltaic Generation: An Efficient Statistical Approach. *IEEE Trans. Power Syst.* **2017**, *32*, 2471–2472. [CrossRef]
19. Yang, Y. Study on Short-Term Probability Interval Prediction of Photovoltaic Power Based on Wavelet Packet Theory. Master's Thesis, Northeast Electric Power University, Jilin, China, 2020. [CrossRef]
20. Zhao, S.Q.; Wang, M.Y.; Hu, Y.Q.; Liu, C. Research on the Prediction of PV Output Based on Uncertainty Theory. *Trans. China Electrotech. Soc.* **2015**, *30*, 213–220. [CrossRef]
21. Zhang, J.; Liu, Y.; Peng, S.R. Photovoltaic power prediction based on feature mining and GRU-A. *Res. Explor. Lab.* **2020**, *39*, 25–30, 49.
22. Song, R.J.; Liu, F.S.; Ma, D.M.; Wang, L. A very short-term prediction model for photovoltaic power based on similar days and wavelet neural network. *Electr. Meas. Instrum.* **2017**, *54*, 75–80.
23. Chen, C.; Huang, G.Y.; Fan, Y.G.; Wu, J.; Wang, X. Short-term load forecasting based on discrete Fréchet distance and LS-SVM. *Power Syst. Prot. Control* **2014**, *42*, 142–147.
24. Rahman, A.S.; Khan, Z.; Rahman, A. Application of independent component analysis in regional flood frequency analysis: Comparison between quantile regression and parameter regression techniques. *J. Hydrol.* **2020**, *581*, 124372. [CrossRef]
25. Pei, Y. Quantile Regression and Its Application. Master's Thesis, Central China Normal University, Wuhan, China, 2014.
26. Zheng, Z.; Liu, G.; Zhang, L.J.; Dan, L. Photovoltaic generation three-layer electricity stealing recognition method based on LSSVM. *Power Electron.* **2017**, *51*, 30–32, 45.