Hindawi
Complexity
Volume 2024, Article ID 9781610, 1 page
https://doi.org/10.1155/2024/9781610

Retraction

Retracted: Photovoltaic Generation Prediction of CCIPCA Combined with LSTM

Complexity

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] E. Zhu and D. Pi, "Photovoltaic Generation Prediction of CCIPCA Combined with LSTM," *Complexity*, vol. 2020, Article ID 1929372, 11 pages, 2020.

Hindawi
Complexity
Volume 2020, Article ID 1929372, 11 pages
https://doi.org/10.1155/2020/1929372

Research Article

Photovoltaic Generation Prediction of CCIPCA Combined with LSTM

E. Zhu[1,2] and D. Pi[2]

[1] *College of Internet of Things Engineering, Jiangsu Vocational College of Information Technology, Wuxi 214153, China*
[2] *College of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 211106, China*

Correspondence should be addressed to E. Zhu; erxi666@163.com

Received 4 May 2020; Revised 25 August 2020; Accepted 4 September 2020; Published 15 September 2020

Academic Editor: Zhihan Lv

In order to remedy problems encompassing large-scale data being collected by photovoltaic (PV) stations, multiple dimensions of power prediction mode input, noise, slow model convergence speed, and poor precision, a power prediction model that combines the Candid Covariance-free Incremental Principal Component Analysis (CCIPCA) with Long Short-Term Memory (LSTM) network was proposed in this study. The corresponding model uses factor correlation coefficient to evaluate the factors that affect PV generation and obtains the most critical factor of PV generation. Then, it uses CCIPCA to reduce the dimension of PV super large-scale data to the factor dimension, avoiding the complex calculation of covariance matrix of algorithms such as Principal Component Analysis (PCA) and to some extent eliminating the influence of noise made by PV generation data acquisition equipment and transmission equipment such as sensors. The training speed and convergence speed of LSTM are improved by the dimension-reduced data. The PV generation data of a certain power station over a period is collected from SolarGIS as sample data. The model is compared with Markov chain power generation prediction model and GA-BP power generation prediction model. The experimental results indicate that the generation prediction error of the model is less than 3%.

1. Introduction

With the gradual implementation of the "Internet + energy" policy, the PV generation industry is rapidly developing. The proposed incorporation of artificial intelligence has instilled new incentives for the PV generation industry. PV generation is susceptible to influence from extreme weather, distorting the predicted results and adding trouble in dispatching when the power system is connected to the grid [1], which may even lead to power resource wastage. Numerous methods in artificial intelligence exist for photovoltaic generation prediction such as support vector machines [2], Markov chains [3], and neural networks. Neural networks are widely used in PV generation prediction, and a number of power generation prediction models have surfaced to this effect, including neural networks based on time series [4], GA-BP neural network [5], deep belief network [6], and fuzzy neural network [7]. Such models find it difficult to overcome issues in convergence and local optimization, and they suffer from large prediction errors, mainly because many network input compositions are present, along with noises in the sample data. Various scholars have used principal component analysis (PCA) [8] and other algorithms to preprocess the collected data; however, PCA and other algorithms find it difficult to process large amounts of data collected in real time due to their own defects. In view of the above problem, this paper adopted CCIPCA, which is suitable in analysing large data sets to extract key data affecting power generation prediction, to reduce dimensions in data and eliminate noise. Simultaneously, combined with LSTM [9], a power generation prediction model was established to overcome problems such as local convergence and slow training speed, controlling the prediction error of PV generation within 3%.

2. Relevant Works

2.1. Analysis of Correlation Factors in Power Generation Prediction. The sample data adopted the data of PV station no. 11282, Zhangdian District, Zibo City, Shandong Province, China, which was provided by SolarGIS. Its geographical location is 118 degrees east longitude and 32 degrees north latitude. The data includes total solar radiation data such as horizontal radiation GHI and normal direct radiation DNI and meteorological parameters like temperature, humidity, and pressure, as well as environmental parameters such as elevation, surface inclination angle, and surface azimuth angle. There were a total of 37 dimensions and 30 minutes of resolution.

Considering that parameters like geographical location are constant and have little influence on power generation prediction, the prediction model will not be considered. Meteorological and historical power generation data were taken as the main influencing factors of the model, including power generation every 30 minutes, power generation at the moment in history, environmental temperature, environmental humidity, wind speed, wind direction, radiation amount, and other indicators. According to formula (1), the correlation analysis of the evaluation factors was carried out:

$$\zeta = \frac{\sum_{i=1}^{N} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{N} \left(x_i - \overline{x}\right)^2}\sqrt{\sum_{i=1}^{N} \left(y_i - \overline{y}\right)^2}}. \quad (1)$$

In Figure 1, the linear features denoted by (a) and (b) are consistent, which are either a line segment or a part of an object without any difference. However, given the overall information, the information represented by (c) is completely inconsistent. Thus, the question of how to extract this nonlinear feature effectively is raised. This question is determined by the micro network structure embedded in NIN network, that is, a full connection layer consisting of two layers of convolution. In the neural network, two-layer fully connected hidden neurons are capable of approximating arbitrary curves, where ζ is the correlation coefficient, x_i is the i-th value of factor x, $\overline{x}$ is the mean of factor x, y_i is the i-th value of factor y, and $\overline{y}$ is the mean of factor y. Table 1 shows the correlation coefficient analysis of PV generation with meteorological factors and historical power generation.

Table 1 demonstrates that electricity generation is highly correlated with solar radiation and historical electricity generation, valued at 0.92 and 0.86, respectively. Moreover, it had a moderate correlation with environmental temperature of 0.56 and had a weak correlation with wind direction and wind speed, though it was found to be negatively correlated with environmental humidity. Therefore, the prediction model adopted environmental temperature, environmental humidity, solar radiation, and historical power generation as its input components, and the correlation between various input components was observed to be strong; hence, carrying out a principal component analysis was required.

2.2. CCIPCA. Due to the advent of big data in PV generation, data preprocessing is particularly important. The PV generation system possesses a simple structure; however, it contains a large amount of equipment and has many failure-prone points. Additionally, the collected data has noise, which introduces obstacles in processing big data. The traditional principal component analysis (PCA) reduces dimensions by eliminating data in the dimension having small variance, which maximizes the information of the original data and removes noise to a certain extent. However, PCA must input all sample data before starting the analysis, which does not align with the objectives of big data; hence, the Candid Covariance-free Incremental Principal Component Analysis (CCIPCA) method was proposed. Unlike the batch method, which uses the covariance matrix to calculate eigenvalues and eigenvectors, CCIPCA initially eliminates the calculation of the covariance matrix using the asymptotic method to estimate the principal component values obtained by approximating the batch method. This data processing technique satisfies the requirements of photovoltaic data processing.

Recently, research related to incremental principal component analysis has been ongoing. Oja and Karhunen et al. proposed the SGA algorithm [10], where, after receiving new sample data, the algorithm continuously refreshes the estimated value of feature vectors and normalizes the new estimated value with all of the previous feature vectors. Sanger put forward the GHA algorithm, which uses a single-layer linear neural network to extract principal elements [11]. Furthermore, Weng et al. described the CCIPCA algorithm [12], which is comprised of the following framework: gradually obtain data $x_1, x_2, \ldots, x_n, \ldots$ and calculate the previous k principal elements $v_{1n}, v_{2n}, \ldots, v_{kn}$. The calculation process of CCIPCA is given below.

According to derivation of PCA by the maximum variance theory [13], the total variance A of the data set X projected along a unit vector w is obtained as follows:

$$A = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(x_i - \overline{x}\right)^T = \frac{1}{n}\sum_{i=1}^{n} u_i u_i^T, \quad (2)$$

where $\overline{x}$ denotes the mean of data set X; that is $\overline{x} = ((\sum_{i=1}^{n} x_i)/n)$, $u_i = x_i - \overline{x}$, and A denotes the covariance matrix of data set X (the general covariance matrix is divided by $n - 1$, where n is used).

The calculation formula of the i-th eigenvalue and the i-th eigenvector is as follows:

$$\lambda_{\text{in}} w_{\text{in}} = A w_{\text{in}}, \quad (3)$$

where w_{in} denotes the n-th eigenvector to be calculated at the i-th input and λ_{in} denotes the corresponding eigenvalue. In order to speed up the iteration of CCIPCA algorithm, the whole iteration is based on the product $\lambda_{\text{in}} w_{\text{in}}$ of eigenvalues and eigenvectors. When the i-th data is input, the following formula is obtained:

$$v_{\text{in}} = \lambda_{\text{in}} w_{\text{in}} = A w_{\text{in}}. \quad (4)$$

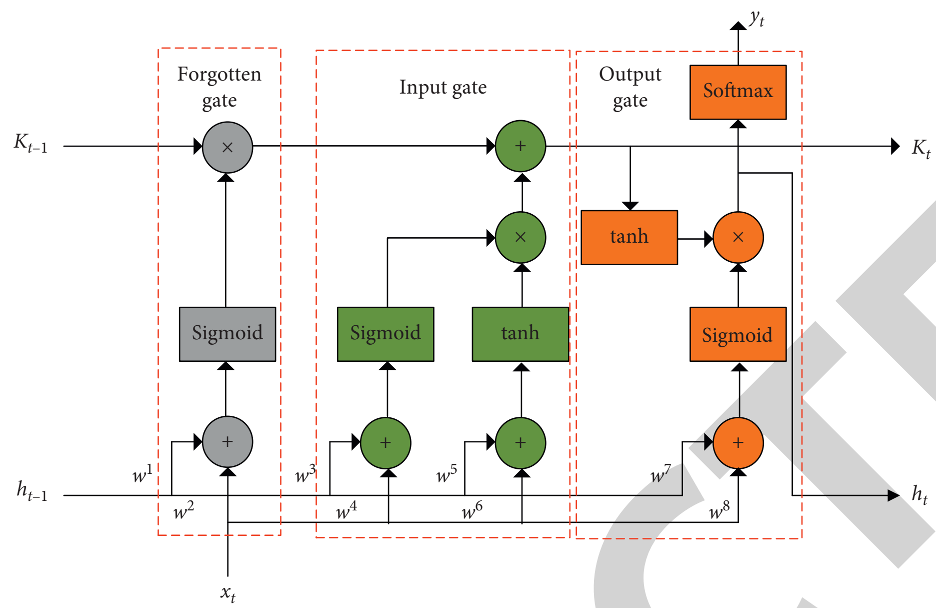


Figure 1: Cell structure of LSTM.

Table 1: Correlation coefficient between PV generation and other factors.

Factors	ET	EH	WS	WD	RQ	HQ	PG
ET	1.00	−0.42	0.12	0.02	0.48	0.55	0.56
EH	−0.42	1.00	−0.17	−0.33	−0.36	−0.24	−0.38
WS	0.12	−0.17	1.00	−0.21	0.22	0.12	0.22
WD	0.02	−0.33	−0.21	1.00	0.06	0.20	0.07
RQ	0.48	−0.36	0.22	0.06	1.00	0.96	0.92
HQ	0.55	−0.24	0.12	0.20	0.96	1.00	0.86
PG	0.56	−0.38	0.22	0.07	0.92	0.86	1.00

ET: environment temperature; EH: environment humidity; WS: wind speed; WD: wind direction; RQ: radiation quantity; HQ: historical quantity; PG: power generation.

Based on formulae (2) and (4),

$$v_{\mathrm{in}} = \frac{1}{n}\sum_{i=1}^{n} u_i u_i^T w_{\mathrm{in}}. \tag{5}$$

The product v_{in} of eigenvalue and eigenvector is obtained by iteration. As the eigenvector is normalized, λ_{in} and w_{in} can be separately obtained with $\|v_{\mathrm{in}}\|$ and $(v_{\mathrm{in}}/\|v_{\mathrm{in}}\|)$.

Formula (5) uses $(v_{i,n-1}/\|v_{i,n-1}\|)$ to approximate w_{in} and replace it. The basic iterative formula of CCIPCA can be obtained by transformation.

$$v_{\mathrm{in}} = \frac{n-1}{n} v_{i,n-1} + \frac{1}{n} u_n u_n^T \frac{v_{i,n-1}}{\|v_{i,n-1}\|}, \tag{6}$$

where $(n-1/n)$ denotes the weight of the previous iteration value $v_{i,n-1}$ and $(1/n)$ denotes the adjustment step size of the iteration.

With the adjustment of the iteration vector v_{in} by the new input data u_n, that is, $u_n = x_n - \overline{x}$, v_{in} gradually converges to the i-th eigenvector. For different v_i, formula (6) can be used for iteration with the different input vectors u_n.

The first eigenvector has been obtained by iteration. Firstly, $u_{1n} = u_n$ is projected onto the previous eigenvector (now the first eigenvector), and the residual data is obtained as follows:

$$u_{2n} = u_{1n} - u_{1n}^T \frac{v_{1n}}{\|v_{1n}\|} \frac{v_{1n}}{\|v_{1n}\|}. \tag{7}$$

u_{2n} denotes the output of the second eigenvector. The third and fourth eigenvector can be similarly obtained. Since the residual data is orthogonal to the data recovered from the previous eigenvector, all orthogonal eigenvectors can be obtained. In addition, the mean should be updated every time when a new data is input. The mean of the n-th data input adopts the following iterative formula:

$$\overline{x}_n = \frac{n-1}{n}\overline{x}_{n-1} + \frac{1}{n} u_n^T. \tag{8}$$

The CCIPCA solution process is summarized in Algorithm 1.

The CCIPCA algorithm has been widely used in the field of big data processing as well as the decomposition of large matrices due to its good convergence, which has achieved beneficial results. Meanwhile, as a benchmark algorithm, CCIPCA has been cited in various incremental algorithms. In this paper, it was applied in big data preprocessing of the PV station with a time dimension to make up for PCA preprocessing defects.

2.3. LSTM. In recent years, the improvement of LSTM has been continuously carried out. Yao et al. [14] proposed an improved NLP method based on Long Short-Term Memory (LSTM) structure, whose parameters are randomly discarded when they are passed backwards in the recursive projection layer, and it is used to overcome practical problems, such as high training complexity, computational

```
Input data:
  original data sequence $x_1, x_2, \ldots, x_n, \ldots,$
  % the algorithm can pause the output of projection matrix at any time.
  dimension of low-dimensional space $k$.
Initialization: $\overline{x}_0 = 0$
Output: projection matrix $W^* = \{w_1 = (v_1/\|v_1\|), w_2 = (v_2/\|v_2\|), \ldots, w_k = (v_k/\|v_k\|)\}$
Iteration steps:
  For $n = 1, 2, \ldots,$ % Run the following steps to update $k$ feature vectors
      $u_{1n} = u_n = x_n - \overline{x}_{n-1};$
    If $n == 1{:}\overline{x}_n = \overline{x}_{n-1} + u_n^T$
    Else: $\overline{x}_n = (n - 1/n)\overline{x}_{n-1} + (1/n)u_n^T$; % Update mean vector
    For $i = 1, 2, \ldots, \min\{k, n\}$ do:
      (a) If $i = n$: $v_{\text{in}} = u_{\text{in}}$; % Initialize the $i$-th eigenvector
      (b) Otherwise:
          $v_{\text{in}} = (n - 1/n)v_{i,n-1} + (1/n)u_{\text{in}}{u_{\text{in}}}^T (v_{i,n-1}/\|v_{i,n-1}\|)$
          $u_{i+1,n} = u_{\text{in}} - u_{\text{in}}^T (v_{\text{in}}/\|v_{\text{in}}\|)(v_{\text{in}}/\|v_{\text{in}}\|)$
```

Algorithm 1: CCIPCA solution process.

difficulties in large-scale content scenarios, high retrieval complexity, and lack of probabilistic significance. Tian et al. [15] proposed a hybrid prediction modelling strategy by combining the autocorrelation local characteristic-scale decomposition and the improved LSTM neural network. Zhang et al. [16] proposed a LSTM approach for bearing performance degradation evaluation. Numerical results show that the proposed LSTM method can effectively predict the remaining service life of the bearing. Li et al. [17] proposed a LSTM method for fault diagnosis and isolation of wind turbine, where stochastic forest algorithm is applied to make decision. Liu et al. [18] combined probabilistic decision-making methods and proposed a Bayesian LSTM algorithm for intelligent fault early-warning of nuclear power machinery. Mirza et al. [19] introduce efficient online learning algorithms based on the Long Short-Term Memory (LSTM) networks that employ the covariance information. They reduce the number of system parameters through the weight matrix factorization, where they convert the LSTM weight matrices into two smaller matrices in order to achieve high learning performance with low computational complexity.

LSTM is an evolutionary version of RNN, which effectively addresses the issue of long-term dependence of effective information in time series and has been broadly applied for different fields. In comparison to other models, the LSTM method is more sensitive to the trivial features in the historical data, easier to capture the details, suitable for big data processing, and more accurate in time series prediction. LSTM replaces the hidden layer of neurons in RNN with a memory unit to record the dependency relationship between time series data, which then rids itself of problems like gradient disappearance and gradient explosion occurring in RNN. LSTM utilizes a "gate" to control information selection. Compared with a dropout operation, this process is not random, though it implements Boolean selection based on sigmoid operation results with "0" signifying the forgetting of information and "1" meaning the remembering of information. The LSTM structure includes three gate structures that adjust the information flow, namely, the forgotten gate, input gate, and output gate, as shown in Figure 1.

The forgotten gate determines the degree to which the unit state K_{t-1} at the last time is retained to the current state K_t, which is calculated as

$$f_t = \text{Sigmoid}\big(W_f [h_{t-1}, x_t] + c_f\big), \tag{9}$$

where f_t is the output value and the bias of the forgotten gate.

The input gate controls the extent to which network input x_t is saved to the cell state K_t at the previous time, which is calculated as

$$\begin{aligned} K_t &= f_t c_t + i_t \widetilde{K}_t, \\ i_t &= \text{Sigmoid}\left(W_i [h_{t-1}, x_t] + c_i\right), \\ \widetilde{K}_t &= \tanh\left(W_k [h_{t-1}, x_t] + c_k\right). \end{aligned} \tag{10}$$

The output gate controls the current output value h_t of unit state K_t outputting to LSTM, which is calculated as

$$\begin{aligned} o_t &= \text{Sigmoid}\left(W_o [h_{t-1}, x_t] + c_o\right), \\ h_t &= o_t \tanh\left(K_t\right). \end{aligned} \tag{11}$$

If the output value has reached the threshold value required by the memory unit, the product of the output value with the calculated value of the current layer is taken as the output, and the calculation is carried out in the next layer. If the threshold is not reached, the memory unit will forget it.

Different from the traditional RNN, LSTM can solve the issues of gradient disappearance and gradient explosion in the training process, yielding more accurate prediction of a long-term time series. LSTM is developed to address data uncertainty while considering complicated situation of the operation.

3. PV Generation Prediction

3.1. Process of PV Generation Prediction. Power generation prediction conducted by CCIPCA combined with LSTM conforms to current trends in photovoltaic big data. The prediction process initially processes the CCIPCA algorithm for the collected big data samples, which then establishes the LSTM networks and implements the training of the sample data on the LSTM network. After completion, the sample data is input in order to acquire the prediction results output, as shown in Figure 2. When training big data samples, it is necessary to continuously tune the discarded neuron probability, attenuation rate, and learning rate.

3.2. Data Source and Preprocessing. SolarGIS is a solar resource assessment tool developed by SolarGIS S.R.O. in Europe, which uses satellite remote sensing data, GIS technology, and advanced scientific algorithms to obtain a high-resolution database of solar resources and climatic factors. In this paper, data from January 1, 2014, to December 31, 2019, were collected from the SolarGIS database as sample data, of which its volume reached 39.4 PB. Using formula (1), the data within a certain period were used for factor correlation evaluation, and factors with an absolute value coefficient above 0.2 were selected. In addition to selecting highly correlated factors, this paper also required two historical PV power generation horizontal and vertical mean data as the model's input sequences to control the disturbance brought about by extreme weather, as extreme weather is a gradual process that takes place in a short amount of time. The formulae are as follows:

$$\begin{aligned}\overline{Q}_k &= \frac{(Q_{k-1} + \cdots + Q_{k-m})}{m},\\ \overline{Q}'_k &= \frac{(Q_{k-1}{}' + \cdots + Q_{k-n}{}')}{n},\end{aligned} \qquad (12)$$

where Q_{k-i}, $i = 1, 2, 3, \ldots m$, represents the generating capacity of the previous continuous time units and $\overline{Q}_k$ represents the average of horizontal generating capacity. $\overline{Q}'_k$ is the average vertical generating capacity of the same time unit before the previous day and $Q_{k-i}{}'$, $i = 1, 2, 3, \ldots, n$, is the electrical generation of the previous n days.

Actual power generation always fluctuates around the mean curve of the horizontal and vertical power generation, which can be used as a stable factor in the prediction model to eliminate the interference of extreme weather in power generation prediction. Although weather conditions are very sporadic, the seasonal law always changes with Earth's revolution and rotation.

Here, the horizontal mean of the first 10 time units and the vertical mean of the same time unit of the first 5 days were selected as the input. This was done because if the number choices were more, the average value would change significantly, which was not conducive to the stability of the prediction model. If there were fewer number choices, the average value would not change significantly, which was not conducive to measuring the impact of extreme weather changes on the prediction model, resulting in inaccuracies of prediction.

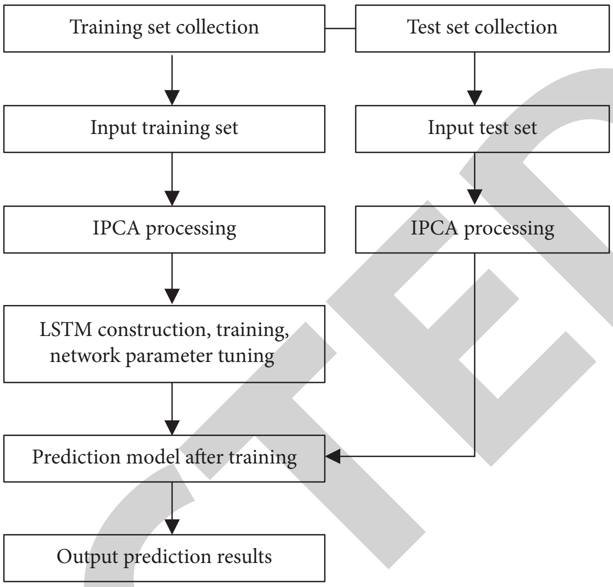

Training set collection
Test set collection
Input training set
Input test set
IPCA processing
IPCA processing
LSTM construction, training, network parameter tuning
Prediction model after training
Output prediction results

Figure 2: Flow chart of power generation prediction by IPCA combined with LSTM.

After correlation coefficient analysis, the dimension of sample data was reduced to 18. Some sample data are listed in Table 2. After the sample data were sorted, the data was preprocessed with *Z*-Score standardization to eliminate influences from different feature dimensions and improve the convergence speed of the model. Following completion of the prediction, a reverse operation would be carried out to achieve the real predicted value recovery. Subsequently, CCIPCA was preprocessed to map the sample data to dimensions set by the investigators, such as solar radiation amount and historical power generation. Finally, orthogonalization removed influences in the relationship between noise and dimensions.

Solar radiation adopts two aspects: global horizontal radiation (GHI) and direct normal radiation (DNI). Meteorological parameters are air temperature (TEMP) at 2 m, RH, average WS and WD at 10 m, precipitation RH, and atmospheric pressure.

CCIPCA preprocessing was implemented. 48 pieces of data on June 6, 2019, were used as data series input one by one. The low-dimensional spatial dimension was set to 18. The eigenvector and eigenvalue were calculated, and principal component analysis was conducted to remove the influence of noise and the relationship between dimensions. The results of principal component analysis are shown in Table 3.

It can be determined from Table 3 that the cumulative variance contribution rate of the first six eigenvalues is more than 80%. Therefore, six principal components were selected, and the corresponding eigenvectors of the six principal components were selected to construct the transformation matrix. The original data and the transformation matrix are calculated, and the input variables obtained were input into LSTM for prediction.

Table 2: Part of sample data after sorting.

Date	DNI (W. H/m^2)	GHI (W. H/m^2)	TEMP (°C)	RH (%)	WS (m/s)	WD (°)	AP (Pa)	$\overline{Q}_{10}$ (MWp)	$\overline{Q}'_5$ (MWp)
2020-06-06 05: 30: 00	11	11	21.4	73.2	0.3	294	1000.2	0	0.1372
2020-06-06 06: 30: 00	78	80	23.9	61	0.7	349	1000.2	0.0033	0.8562
2020-06-06 07: 30: 00	197	202	26.1	53.6	0.7	266	1000.5	0.0538	2.1322
2020-06-06 08: 30: 00	322	325	28.4	47.5	1.7	232	1000.9	0.1994	3.2794
2020-06-06 09: 30: 00	484	481	29.7	40.2	2.5	291	1000.8	0.4421	4.1634
2020-06-06 10: 30: 00	615	605	31.1	34.8	3.1	306	1000.4	0.8043	5.092

GHI: global horizontal irradiation; DNI: direct normal irradiance; RH: relative humidity; AP: atmospheric pressure; WS: wind speed; WD: wind direction; $\overline{Q}_{10}$: the mean of 10 consecutive time units; $\overline{Q}'_5$: the mean of 5 consecutive days at the same time; PVOUT: photovoltaic output per unit time.

Table 3: Eigenvalue and variance contribution of the input variables.

Number	Eigenvalue	Variance contribution rate (%)	Cumulative variance contribution rate (%)
1	6.40	23.16	23.16
2	5.02	20.32	43.48
3	4.26	17.55	61.03
4	3.24	10.12	71.15
5	2.60	7.61	78.76
6	1.07	4.23	82.99
. . .	. . .	. . .	. . .
18	0	0	100

3.3. Evaluation Indexes. After the completion of the model training, the generation capacity of the PV station in 1 day of continuous time was selected for prediction to judge the accuracy and efficacy of the entire model. Here, the mean square error was used as a loss function to test and measure the model's efficacy. The mean square error was calculated as

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_{ii}\right)^2, \tag{13}$$

where MSE is the mean square error, y_i is the actual power generation at time I, $\hat{y}_i$ is the prediction value of the model at time I, and m is the amount of sample data. Accordingly, the smaller the mean square error, the higher the model precision.

In addition, the maximum error was used to measure the prediction error range of the model. The larger the value of ME, the worse the accuracy of the model, which was computed as

$$\mathrm{ME} = \max\left\{\frac{\left|y_i - \hat{y}_i\right|}{y_i}, \quad i = 1, 2, 3, \ldots, m\right\}. \tag{14}$$

3.4. Model Building and Experimental Results Analysis. The experimental environment is Intel i9 processor, Linux + anaconda3 + tensorflow2.0 platform, Spyder software, and Python 3.7 programming language.

The LSTM model structure determines the optimization and prediction accuracy of the training process. In this paper, the LSTM network structure having 4 hidden layers of LSTM and 1 ordinary layer was adopted. The number of neurons in each layer was 512, 256, 128, and 64, respectively. Specifically, 64 neurons were used as the ordinary layer, and dropout operations were used between layers. The overall structure of LSTM is shown in Figure 3.

Relevant literature discusses the setting of training parameters of LSTM. When the learning rate and attenuation rate are different, the network performance is not the same, and the prediction effect is also different. If LSTM has dropout layers, the probability of dropping neurons is also the key to parameter optimization. According to the comprehensive literature [9, 11], the learning rate is generally small, or the learning rate is degraded dynamically; that is, when the accuracy rate in the network training process is no longer improved, the learning rate will be reduced to one-tenth of the original or lower. The probability of discarded neurons is 0.2 and its vicinity, and the decay rate is 0.9 and its vicinity.

The data taken between 2014 and 2018 were used as the training set, while the data taken from 2019 acted as the test set. After the sample data were standardized, PCA and CCIPCA pretreatment were performed, respectively, and training and testing were conducted on the LSTM. The probability of dropping neurons in the dropout layer was {0.1, 0.2, 0.3}; the attenuation rate was {0.8, 0.9} and the learning rate was 0.001. Moreover, the number of training iterations was 100. The experimental results are shown in Table 4.

Table 4 demonstrates that the prediction accuracy of power generation obtained after CCIPCA processing of the sample data was better than that of PCA + LSTM and LSTM. The LSTM model of $N = 0.3$ and $P = 0.9$ was selected to draw the loss function value curve of the training set after 100 iterations. As shown in Figure 4, CCIPCA + LSTM exhibited better convergence as well as a better convergence rate than that of LSTM and PCA + LSTM.

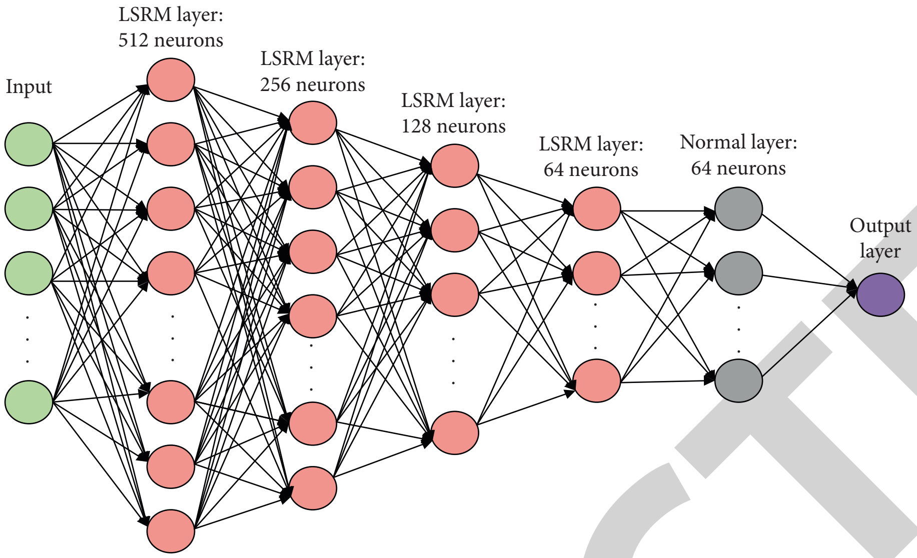


FIGURE 3: LSTM prediction model of the PV station's generation.

TABLE 4: Experimental results of the LSTM model under different parameters of optimization.

Model	$P=0.8$			$P=0.9$		
	$N=0.1$	$N=0.2$	$N=0.3$	$N=0.1$	$N=0.2$	$N=0.3$
LSTM [20]	0.081	0.072	0.064	0.079	0.078	0.074
PCA + LSTM	0.079	0.070	0.058	0.073	0.062	0.061
CCIPCA + LSTM	0.062	0.059	0.046	0.048	0.044	0.041

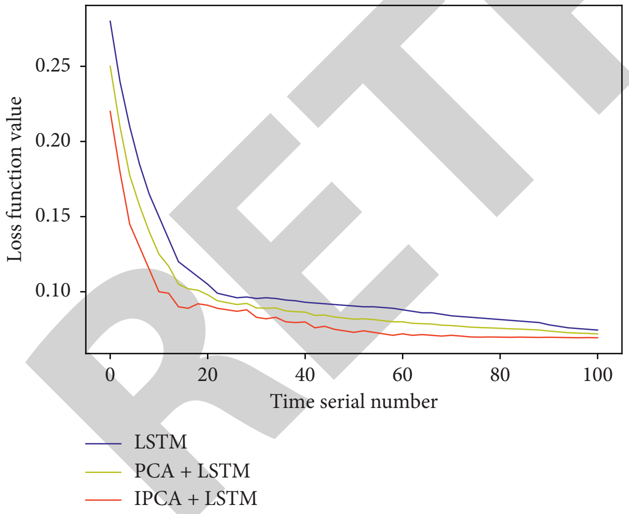


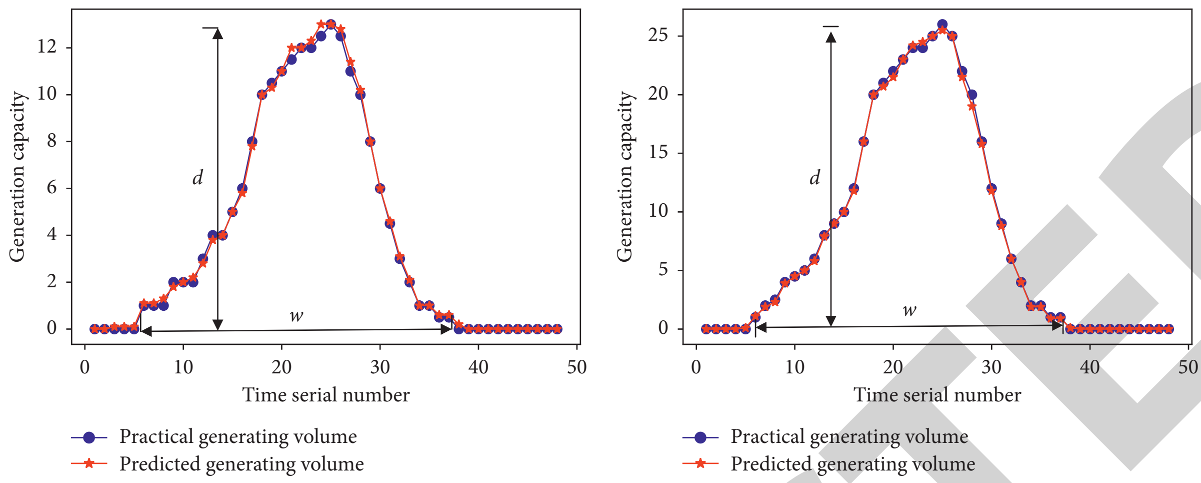
FIGURE 4: Loss function curve of training set and test set.

From the period of 2019, multidays data were selected as test sets in different seasons, and the prediction accuracy of the model was compared with the real power generation.

The data from January 3, 2019, was selected as the test set for comparison with the actual power generation, as shown in Figure 5(a). According to the record, it was light snow, but it was heavy snow three days ago and it turned to clear up after 14: 00. On that day, the temperature range was −2–4°C, with no wind, the wind speed range was 0–0.4 m/s, and the absolute humidity was 1.02–1.70 g/m^3. Then, the data from January 4, 2019, was selected to verify the model, as shown in Figure 5(b). It was sunny, accompanied by weak solar radiation and breeze, temperature range was −5–1°C, wind speed range was 0.6–1 m/s, and the humidity was the same as the previous day.

The meteorological conditions on January 3 and January 4 were similar, and the difference between two days was solar radiation, and solar radiation on January 4 was stronger than that on January 3. Compared with Figures 5(a) and 5(b), the peak of power generation on January 4 was much higher than that on January 3, which indicates that solar radiation is an important factor in power generation prediction. It can also be found from the power generation prediction curve that the photovoltaic power generation efficiency was reduced accompanied by the solar radiation gradually weakening, since the sunlight irradiated the snow surface and reflected, and the ambient temperature sharply dropped in relative to normal times after 14: 00 o'clock. When the night came, there was no sunlight and the ambient temperature dropped below zero, and the photovoltaic power generation was zero.

The data on April 3, 2019, were selected as the test set and compared with the real power generation, as shown in Figure 6. According to the data, April 3 was sunny, the temperature range was 9–14°C, the wind direction was strong, the wind speed ranged from 17.2 to 21 m/s, and the absolute humidity was 2.7–3.2 g/m^3.



(a)

(b)

Figure 5: Power generation forecast of two days in winter. (a) Comparison between predicted curve and real value of power generation on January 3. (b) Comparison between predicted curve and real value of power generation on January 4. Note: *w* represents the time zone of daily effective generation and *d* represents the peak value of daily generation.

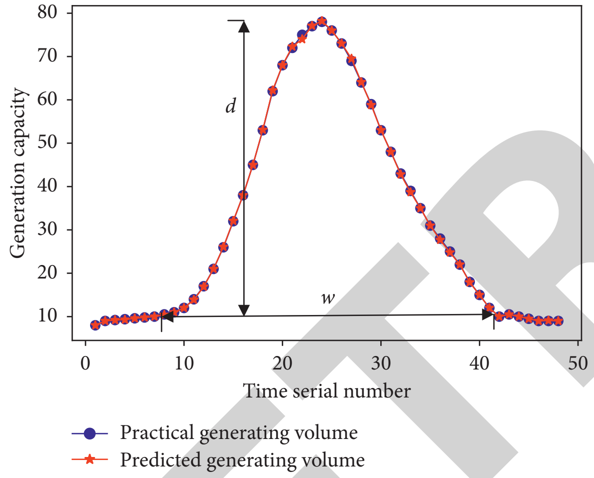


Figure 6: Forecast of one-day power generation in spring.

It can be found from Figure 6 that the predicted curve of the whole-day power generation was basically consistent with the real power generation curve, and the two curves were relatively smooth, which indicates that the power generation efficiency of photovoltaic equipment was not affected by wind direction and wind speed.

The data from July 7, 2019, were selected as the test set for comparison with the actual power generation. Figure 7 illustrates that the data recorded exhibited three thundershowers and severe convection weather on July 7, with periods of 3: 00–4: 00, 7: 00–7: 30, and 14: 00–14: 30, respectively. The meteorological conditions were temperature range of 20–32°C, gust, wind speed of 10.8–13.8 m/s, and absolute humidity of 22.22–25.76 g/m^3.

Figure 8 illustrates that the photovoltaic power generation efficiency is sensitive to the influence of solar radiation and ambient temperature. In the three periods of July 7, the solar radiation and environmental temperature changed rapidly, and the photovoltaic power generation also changed rapidly. The prediction results of the model are basically consistent with the real power generation under the condition of rapid changes of photovoltaic power generation factors.

The data on November 11, 2019, were selected as the test set and compared with the real power generation, as shown in Figure 9. According to the record, it was sunny, the temperature range was 9–15°C, the breeze level was 2-3, the wind speed reached 3.1–4.0 m/s, and the absolute humidity was 1.08–2.12 g/m^3.

From Figures 6 and 7, it can be found that photovoltaic power generation has some relationship with environmental humidity. The meteorological conditions on November 11 and April 3 were basically the same; only the environmental humidity was different, but the peaks of the two pictures were inconsistent, which indicates that the environmental humidity has a negative effect on photovoltaic power generation. On November 11, the environmental humidity was lower, but the power generation effect was better. On April 3, the humidity was higher, and the power generation effect was poor.

It can also be found from Figures 5–7 that the efficiency of photovoltaic power generation is higher in summer, *d* in Figure 5 was much higher than that in other graphs, and the photovoltaic power generation efficiency in winter is the worst. Furthermore, *w* in Figure 5 was much longer than that in other graphs, which indicates that the photovoltaic equipment uses more time to generate electricity in summer.

The 62-day power generation data from July 1, 2019, to August 31, 2019, were selected as the test set and compared with the real power generation; the prediction accuracy of the model is investigated. As shown in Figure 9, the curve between the actual power generation and the predicted power generation in 62 days was shown.

Figure 9 showed that the two curves were relatively similar, indicating that the prediction accuracy of CCIPCA combined with LSTM was relatively high, and its prediction effect was satisfactory.

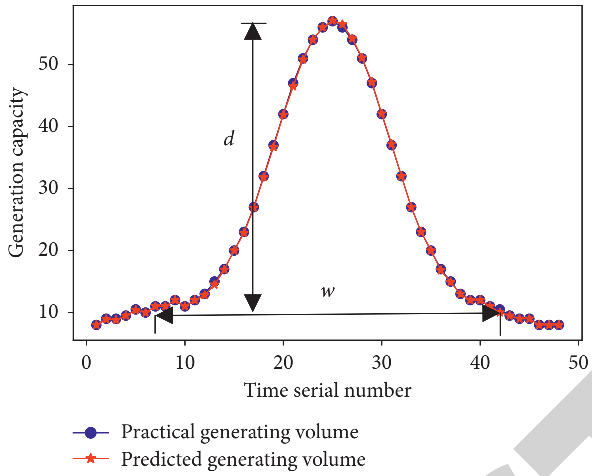


Figure 7: Forecast of one-day power generation in autumn.

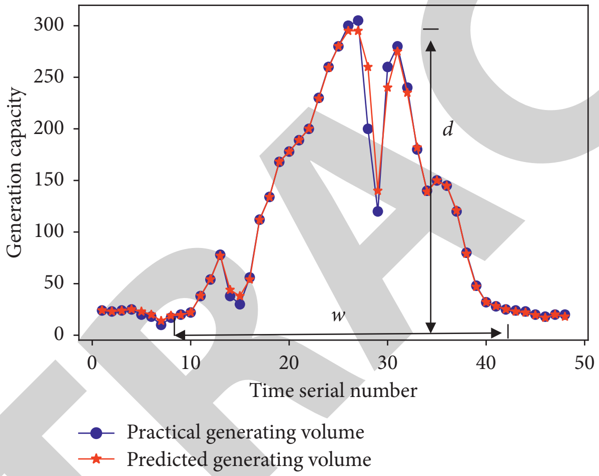


Figure 8: One-day power generation forecast in summer.

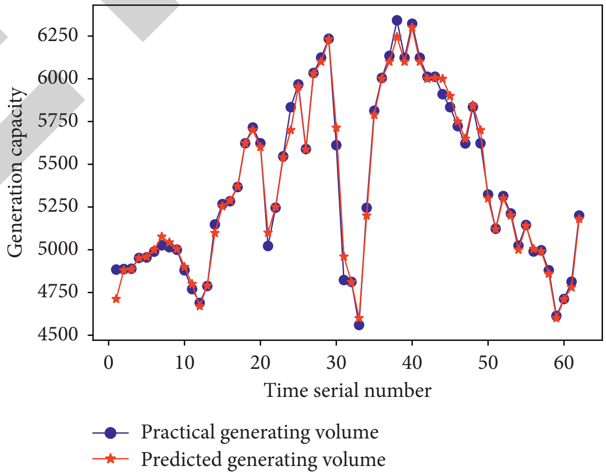


Figure 9: Comparison of photovoltaic power generation forecast from July 1 to August 31.

In view of the practicability of the prediction model, horizontal experiments were conducted to compare the prediction model with the GA-BP neural network [21] and Markov chain model (MC) [22], as shown in Table 5.

By comparison, prediction effect of MC is not as good as neural network, while prediction effect of LSTM is higher than GA-BP network due to the advantage of network layer number, and the maximum error is 6.4%; under the

Table 5: ME comparison between models.

Model	ME	Model	ME	Model	ME
MC	0.074	PCA + MC	0.052	CCIPCA + MC	0.048
GA-BP	0.068	PCA + GA-BP	0.044	CCIPCA + GA-BP	0.032
LSTM	0.064	PCA + LSTM	0.040	CCIPCA + LSTM	0.023

dimension reduction operation of PCA for photovoltaic data, the three prediction models are significantly improved, but, due to the limitations of PCA algorithm, the function mapping from high-dimensional space to low-dimensional space is linear. However, in many practical tasks, it may need nonlinear mapping to find the proper low-dimensional embedding, which leads to the poor dimensionality reduction effect of photovoltaic data. The prediction error of PCA combined with LSTM is 4%, which is higher than that of PCA combined with MC and GA-BP. However, the dimensionality reduction of PV data using IPCA avoids the complex covariance matrix calculation of PCA, and the regularization effect of IPCA is better than that of PCA, which is suitable for LSTM. The prediction error of IPCA combined with LSTM is 2.3%. In a word, the prediction model combining CCIPCA with LSTM displayed better prediction than other models, with an error range of prediction results within 3%.

4. Conclusions

In view of issues such as data being collected by PV stations in a matter of minutes, daily data volume reaching the GB or PB level, data scale being large, multiple instances of data dimensions [23], and existence of noise [24], this paper proposed a prediction model for PV generation by combining CCIPCA with LSTM. According to the simulation of the proposed prediction model, the following conclusions were drawn:

(1) CCIPCA has handled the super-large-scale data of PV station, realized the dimensionality reduction of data, made use of orthogonalization following dimensionality reduction of data, eliminated the influence of noise, and improved the convergence speed and training speed of the model.

(2) During model training, the historical horizontal and vertical mean values of PV generation were added to eliminate the disturbance of extreme weather conditions on the model, and 48 sets of data from a certain day were selected for testing. The obtained results aligned with the real values of power generation, demonstrating the model's stable performance.

(3) The model was compared to the other two PV generation prediction models horizontally, and the power generation prediction error was determined to be less than 3%, illustrating its practicality.

Data Availability

The SolarGIS data used to support the findings of this study are included within the article.

Additional Points

In the experiment, adopting SolarGIS Meteosat (EUMETSAT, DE) and GOES (NOAA, USA) radiation of satellite remote sensing data, combined with Meteosat (EUMETSAT, DE) and GOES (NOAA, USA) of cloud and snow index and Global Forecast System (GFS) database (NOAA, USA) of water vapor data, a series of meteorological elements including solar radiation and temperature value are calculated. Taking photovoltaic power station no. 11282 in Zhangdian district, Zibo city, Shandong province, China (118 east longitude, 32 north latitude), as an example, the generation data from 2014 to 2019 are selected/studied. All data can be downloaded from http://solargis.cn.

Disclosure

Some of the authors of this publication are also working on the following related projects: (1) higher vocational education teaching fusion production integration platform construction projects of Jiangsu province under Grant no. 2019 (26), (2) Natural Science Fund of Jiangsu Province under Grant no. BK20131097, (3) "Qin Lan project" teaching team in colleges and universities of Jiangsu province under Grant no. 2017 (15), and (4) high level of Jiangsu province key construction project funding under Grant no. 2017 (17).

Conflicts of Interest

The authors declare that there are no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] P. Wu, W. Huang, and N. Tai, "Novel grid connection interface for utility-scale PV power plants based on MMC," *The Journal of Engineering*, vol. 2019, no. 16, pp. 2683–2686, 2019.

[2] A. Ahmad, Y. Jin, C. Zhu, I. Javed, M. Waqar Akram, and N. A. Buttar, "Support vector machine based prediction of photovoltaic module and power station parameters," *International Journal of Green Energy*, vol. 17, no. 3, pp. 219–232, 2020.

[3] B. Chen and J. Li, "Combined probabilistic forecasting method for PV using an improved Markov chain," *IET Generation, Transmission & Distribution*, vol. 13, no. 19, pp. 4364–4373, 2020.

[4] P. Christonphe, V. Cyril, M. Marc, and M.-L. Nivet, “Forecasting of pre-processed daily solar radiation time series using neural networks,” *Solar Energy*, vol. 84, no. 12, pp. 2146–2160, 2010.

[5] X. Meng, A. Xu, W. Zhao, H. Wang, C. Li, and H. Wang, “A new PV generation power prediction model based on GA-BP neural network with artificial classification of history day,” in *Proceedings of the 2018 International Conference on Power System Technology (POWERCON)*, pp. 1012–1017, Guangzhou, China, November 2018.

[6] G. W. Chang and H.-J. Lu, “Integrating gray data preprocessor and deep belief network for day-ahead PV power output forecast,” *IEEE Transactions on Sustainable Energy*, vol. 11, no. 1, pp. 185–194, 2020.

[7] L. Kaiju, L. Xuefeng, M. Chaoxu, and W. Dan, “Short-term PV prediction based on T-S fuzzy neural network,” in *Proceedings of the 2018 33rd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 620–624, Nanjing, China, May 2018.

[8] S. Qijun, L. Fen, Q. Jialin, Z. Jinbin, and C. Zhenghong, “PV prediction based on principal component analysis and support vector machine,” in *Proceedings of the 2016 IEEE Innovative Smart Grid Technologies-Asia (ISGT-Asia)*, pp. 815–820, Melbourne, VIC, Australia, December 2016.

[9] C. Hua, E. X. Zhu, L. Kuang, and D. Pi, “Short-term power prediction of PV station based on long short-term memory-back-propagation,” *International Journal of Distributed Sensor Networks*, vol. 15, no. 10, Article ID 155014771988313, 2019.

[10] E. Oja and J. Karhunen, “On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix,” *Journal of Mathematical Analysis and Applications*, vol. 106, no. 1, pp. 69–84, 1985.

[11] T. D. Sanger, “Optimal unsupervised learning in a single-layer linear feedforward neural network,” *Neural Networks*, vol. 2, no. 6, pp. 459–473, 1989.

[12] J. Y. Weng, Y. L. Zhang, and W. S. Hwang, “Candid covariance-free incremental principal component analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 1034–1040, 2003.

[13] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[14] L. Yao and Y. Guan, “An improved LSTM structure for natural language processing,” in *Proceedings of the 2018, IEEE International Conference of Safety Produce Informatization (IICSPI)*, pp. 565–569, Chongqing, China, December 2018.

[15] H.-X. Tian, D.-X. Ren, and K. Li, “A hybrid vibration signal prediction model using autocorrelation local characteristic-scale decomposition and improved long short term memory,” *IEEE Access*, vol. 7, pp. 60995–61007, 2019.

[16] B. Zhang, S. H. Zhang, and W. H. Li, “Bearing performance degradation assessment using long short-term memory recurrent network,” *Computers in Industry*, vol. 106, pp. 14–29, 2018.

[17] M. Li, D. Yu, Z. Chen, K. Xiahou, T. Ji, and Q. H. Wu, “A data-driven residual-based method for fault diagnosis and isolation in wind turbines,” *IEEE Transactions on Sustainable Energy*, vol. 10, no. 2, pp. 895–904, 2019.

[18] G. J. Liu, H. X. Gu, X. C. Shen, and D. You, “Bayesian long short-term memory model for fault early warning of nuclear power turbine,” *IEEE Access*, vol. 8, pp. 50801–50813, 2020.

[19] A. H. Mirza, M. Kerpicci, and S. S. Kozat, “Efficient online learning with improved LSTM neural networks,” *Digital Signal Processing*, vol. 102, Article ID 102742, 2020.

[20] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, “Short-term residential load forecasting based on LSTM recurrent neural network,” *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2019.

[21] C. Dai, J. Wu, D. Pi et al., “Brain EEG time series clustering using maximum weight clique,” *IEEE Transactions on Cybernetics*, vol. 99, pp. 1–15.

[22] L. Cui, J. Wu, D. Pi, P. Zhang, and P. Kennedy, “Dual implicit mining-based latent friend recommendation,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 5, pp. 1663–1678, 2020.

[23] C. Dai, D. Pi, S. I. Becker, J. Wu, L. Cui, and B. Johnson, “CenEEGs,” *ACM Transactions on Knowledge Discovery from Data*, vol. 14, no. 2, pp. 1–25, 2020.

[24] W. Shao, D. Pi, and Z. Shao, “A pareto-based estimation of distribution algorithm for solving multiobjective distributed No-wait flow-shop scheduling problem with sequence-dependent setup time,” *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 3, pp. 1344–1360, 2019.