Power System Technology **Volume 48 · Number 1 · 2024**
ISSN: 1000-3673

The relationship between microstructure and hydrophobicity of pulverized composite insulator sheds

Deng T., Yang D., Tao W., Zhou K., Zhang X., Gu C.
School of Computer & Information Technology, Northeast Petroleum University, Daqing, China
Heilongjiang Provincial Key Laboratory of Oil Big data & Intelligent Analysis, Daqing, China

ABSTRACT
Short-term prediction of liquefied gas concentration is helpful to assisted analysis of storage tank operation status and trend, can discover the potential safety hazards timely inside the storage tank, and then take effective measures to prevent and control the risks, so as to ensure the safety and stability of the oil and gas gathering and transportation industry. Within the limited space, affected by factors such as complexity, high dimension, strong correlation and weak regularity of storage tank operation data, the existing short-term prediction method of liquefied gas concentration is difficult to ensure the real-time performance and accuracy of prediction results. Therefore, we propose a short-term prediction method of liquefied gas concentration based on mixed intelligence. Firstly, we bring in an Extreme Change Function, and calculate the weighted set kurtosis value of the feature curve to realize feature dimension reduction. Secondly, the Convolutional Neural Network is used to mine the correlation between features and extract effective feature vectors. Meanwhile, we use Long Short-Term Memory Network to learn the change law of the data, so as to obtain the predicted value of liquefied gas concentration. Finally, our method is applied to a real scenario to demonstrate that the short-term prediction method of liquefied gas concentration achieves superior results in prediction accuracy, running speed and stability compared with other methods.

1. INTRODUCTION

Short-term prediction of liquefied gas concentration is based on tank operation data, use intelligent analysis methods to predict potential hazards of tank operation, such as gas leakage, excessive concentration, abnormal pressure, etc. It can provide a reliable objective basis for the production decision-making department to formulate scientific and effective risk prevention and control measures. It is one of the auxiliary decision-making methods to ensure the safety and stability of oil and gas gathering and transportation industry [1-2]. Due to the complexity, high dimension, strong correlation and weak regularity of storage tank operation data, the existing prediction methods have low accuracy, so the short-term prediction method [3-4] of liquefied gas concentration is regarded as one of the key problems in the field of safe production in petroleum industry.

Traditional prediction methods of liquefied gas concentration include two types. One type is the statistical method represented by the auto regressive inter grated moving average model (ARIMA) [5]. However, such method has a large calculation amount and low prediction accuracy [6] when dealing with multi-dimensional features. The other type is the machine learning method represented by support vector machine (SVM) [7] and BP (back propagation) neural network [8]. Such method cannot learn the temporal feature of gas concentration data well, and requires manual setting of time characteristics, which still cannot improve the prediction accuracy [9-10].

With the rise of deep learning technology in the field of intelligent prediction, many scholars use the hybrid model of convolutional neural network (CNN) and long short-term memory (LSTM) for the short-term prediction of liquefied gas concentration [11-12]. This trend prediction model based on time series data can improve the prediction accuracy to some extent. However, this method takes the gas concentration and its multidimensional feature data as the model input directly, resulting in too high input dimension, which is not conducive to model training and accuracy improvement. In Reference [13], Pearson coefficient is used to select the features of gas concentration, and then the long short-term memory network is applied to predict the time series. In Reference [14], the principal component analysis (PCA) is used to screen effective features, and the long short-term memory network is employed to predict gas concentration. But Pearson coefficient and PCA method are more suitable for reducing dimension of stationary time series data, and the calculation time is long for large-scale data. However, the prediction of liquefied gas concentration involves a large amount of data, high feature dimension and non-stationary time series [15], so the traditional dimension reduction method leads to poor prediction results. As a common method of signal extraction, spectral kurtosis can detect and represent the non-stationarity of signals [16], and is widely used in the fields of audio processing [17], image processing [18] and mechanical equipment fault diagnosis [19]. In view of the advantages of spectral kurtosis for non-stationary signal processing, we design an extreme change function (ECF) based on spectral kurtosis to solve the problem of feature dimension reduction of non-stationary time series.

Based on the above research, by drawing on the design idea of hybrid intelligent algorithm of “divide-and-conquer, complementary advantages” [20-22], we propose a liquefied gas concentration short-term prediction method based on ECF, CNN and LSTM (ECL-LGSP). Firstly, the ECF is used to calculate the weighted set kurtosis value of the feature curve, so as to reduce the feature dimension; Secondly, we use CNN to mine the correlation between the features, and extract effective feature vectors; Then, the feature vectors are entered into the LSTM for training and learning the change rule of the data, so as to realize the prediction of liquefied gas concentration; Finally, it is proved by experiments that the method has obvious advantages in prediction accuracy and running speed. Our innovations include:

1. An extreme change function is proposed to solve the problem of feature dimension reduction of high dimensional and non-stationary time series.
2. CNN and LSTM are combined to realize feature extraction and short-term prediction of liquefied gas concentration.

The method we propose has been applied to real scenes and has good application effect. For example, when an oil extraction plant uses the anomaly monitoring and early warning system of CO2, this method is chosen as the basis for early warning. In the process of practical application, the accuracy and real-time performance of CO2 concentration prediction are better than the original method obviously, and it can prevent the occurrence of carbon leakage accidents effectively and contribute to the safe production of oil fields.

The article is organized as below: In Chapter 2, we arrange application scenarios, provide the workflow of ECL-LGSP method, and condense key issues. In Chapter 3, the definition of Extreme Change Function and the determination method of extreme change factor are expounded. In Chapter 4, the network model structure, training process and evaluation method of ECL-LGSP method are described. In Chapter 5, the effectiveness of the proposed method is demonstrated by experiments. In Chapter 6, we summarize the research results and offer the prospect of future research work.

2. BASIC WORK

2.1. Scenario definition and basic concepts

Defining scenario H: In a relatively closed environment composed of N groups of liquefied gas storage tanks, the number of physical parameters of each type of storage tank C is v, the number of detection instruments is d, and the number of regular inspection personnel is q. The monitoring center issues a liquefied gas concentration prediction task, and any storage tank $C_i \in C$ contains n features and can be solved by intelligent prediction method. The feature of C_i can be expressed as $C_i = (x, n)$, n represents the feature dimension, x represents the feature attribute set, including d dynamic attributes and q additional attributes.

Since the task is to predict the concentration of liquefied gas in a finite space, the feature attributes in the scenario H have integrity and satisfy the following three premises:

Precondition 1. Complete detection instruments are configured in the limited space for storing C.
Precondition 2. The number of inspection personnel, inspection time and gender of personnel in the additional attribute q are known.
Precondition 3. In the same prediction task, $\forall C_i \in C$ has the same features, and the feature attribute value is not empty.

The basic concepts and formulas are described as follows. Feature attribute: The operation data of any tank C_i can be expressed as $C = \{X_{ti} | 0 < t \leq m, 1 \leq i \leq a\}$, after data processing, the original feature set is $X = \{X_i | 1 \leq i \leq n,\ n < a\}$, and the feature attribute data within the time $0 < t \leq m$ is $D = \{x_{ti} | 0 < t \leq m, 1 \leq i \leq n\}$, wherein, x_{ti} is the observed value of the ith feature at time t, and D is expressed in matrix form as follows:

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \tag{1}$$

2.2. ECL-LGSP workflow

In the ECL-LGSP method, the model $\hat{p}_c$ is obtained by training based on a number of historical storage tank operation data, and the model $\hat{p}_c$ is used to predict the liquefied gas concentration value of a specific period in the future. Specifically, the ECL-LGSP model includes three parts: data preprocessing, feature dimension reduction and trend prediction, and the workflow is shown in Figure 1.

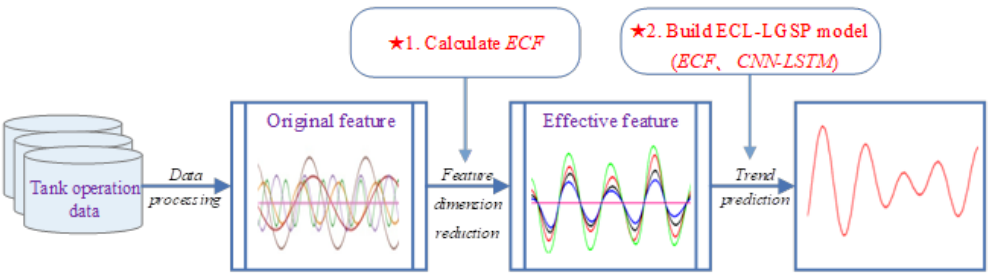


Figure 1. ECL-LGSP workflow

According to the ECL-LGSP workflow, the following two problems need to be addressed:

I. How to reduce the feature dimension for the high dimensional storage tank operation data.
II. How to build a trend prediction model of liquefied gas concentration based on CNN-LSTM network.

3. FEATURE DIMENSION REDUCTION

The essence of short-term prediction of liquefied gas concentration is regression prediction based on multi-dimensional time series data; Selecting a suitable feature dimension reduction method can reduce the computational complexity and overfitting risk of the model, and improve the prediction accuracy and running speed of ECL-LGSP model fundamentally.

3.1. Feature trend analysis

The effective method of feature dimension reduction is to analyze the correlation between the predicted target and the features. In order to mine the relationship between the concentration of liquefied gas and its features fully, we take the operation data of storage tanks in different regions as samples and calculate the general change trend of the feature attributes such as temperature, pressure, humidity and the concentration of liquefied gas. It is found that the features of similar storage tanks in different regions have similar change trends. Figure 2 shows the changing trend of CO2 concentration and its features over time in a storage tank in a certain area. Experimental results show that the trend change of liquefied gas concentration and its features has temporal feature, non-stationarity and periodic correlation [23], which are manifested as follows:

I. The concentration of liquefied gas and its features change continuously with the advance of time series.

II. The change curves of liquefied gas concentration and its features are of impact in time domain and non-stationary in frequency domain.
III. The trend of liquefied gas concentration and its features presents periodic changes and has correlation.

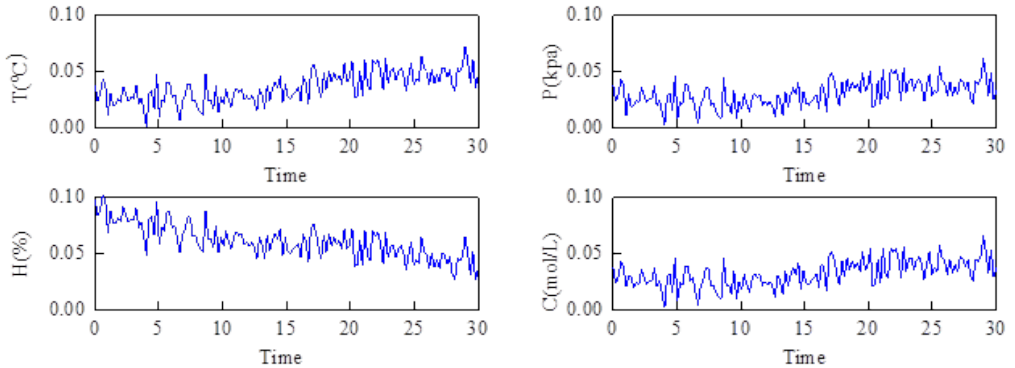


Figure 2. Trends of gas concentration and its features

3.2. Definition and formal representation of ECF

Aiming at the trend change characteristics of liquefied gas concentration and its features, we draw on the idea [24] of non-stationary signal extraction by spectral kurtosis and bring in an Extreme Change Function (ECF) to screen effective feature subsets from the perspective of features curves. The essence of ECF is to use the envelope kurtosis and envelope spectral kurtosis to reflect the impact and non-stationarity of the feature curves, calculate the correlation coefficient to reflect the correlation between the gas concentration and its features, and then establish the weighted set kurtosis f_{ECF} to judge the effectiveness of the features and achieve the feature dimension reduction. The ECF working principle is shown in Figure 3.

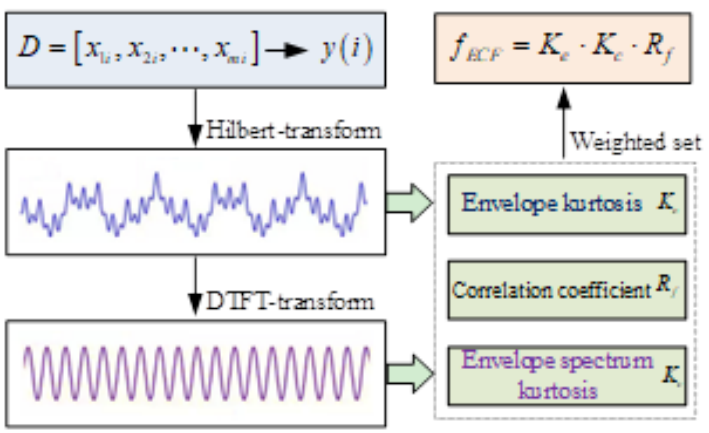


Figure 3. ECF working principle

The formal representation of ECF consists of three parts: envelope kurtosis, envelope spectral kurtosis and correlation coefficient. The formulas and the derivation process involved are described as follows:

(1) Envelope kurtosis: We define K_e to represent the envelope kurtosis of feature attribute, E_x is the envelope signal of $y(i)$ after the Hilbert transform ($y(i)$ represents the curve of feature X_i over time), $\mu_e$is the average of E_x, $\sigma_e$is the standard deviation of E_x. According to kurtosis formula, the representation form of envelope kurtosis of feature attributes is shown in formula (2):

$$K_e = \frac{E(E_x-\mu_e)^4}{\sigma_e^4} \tag{2}$$

(2) Envelope spectral kurtosis: We define K_c to represent the envelope spectral kurtosis of feature attribute, $E_c = DTFT[E_x]$ (Fourier transform of discrete aperiodic sequences), $\mu_c$is the average of E_c,$\sigma_c$is the standard deviation of E_c. According to spectral kurtosis formula, the representation form of envelope spectral kurtosis of feature attributes is shown in formula (3):

$$K_c = \frac{E(E_c-\mu_c)^4}{\sigma_c^4} \tag{3}$$

(3) Correlation coefficient: We define R_f to represent the correlation coefficient between the liquefied gas concentration and its features, $Cov(x,p)$ represents the covariance between feature X_i and the liquefied gas concentration p. The representation form of correlation coefficient between liquefied gas concentration and its features is shown in formula (4):

$$R_f(x,p) = \frac{Cov(x,p)}{\sqrt{D(x)}\sqrt{D(p)}} \tag{4}$$

In summary, by using the weighted set kurtosis, the representation form of ECF is shown in formula (5).

$$f_{ECF} = K_e \cdot K_c \cdot R_f \tag{5}$$

3.3. Judging criteria of effective features

ECF is the basis for screening effective features. We define $\gamma_i$to represent the ECF difference value between the liquefied gas concentration and its features, and δ to represent the fluctuation threshold, screen effective features by determining the relationship between γ_i and δ. The specific formulas of γ_i and δ are shown below:

$$\gamma_i = |f_{X_i} - f_o| \tag{6}$$

$$\delta = \frac{1}{n}\sum_{i=1}^{n}\gamma_i \tag{7}$$

wherein, f_0 and f_{X_i} represent the ECF value of liquefied gas concentration and its features.

Judging criteria of effective features:

1) If $\gamma_i > \delta$, it indicates that the curve difference between feature$X_i$and liquefied gas concentration is large and the correlation is weak, thereby judging that feature$X_i$is not an effective feature.
2) If $\gamma_i \leq \delta$, it indicates that the curve similarity between feature$X_i$and liquefied gas concentration is great and the correlation is strong, thereby judging that feature X_i is an effective feature.

According to the judging criteria of effective features, all features $X = \{X_i | 1 \leq i \leq n\}$ in the time range of 0~t are traversed to determine the effectiveness of the features one by one, thereby realizing the feature dimension reduction, and obtaining the effective feature matrix $X^* = \{X_k | 1 \leq k < n\}$.

4. ECL-LGSP MODEL

The ECL-LGSP model uses ECF to reduce feature dimension from the perspective of feature curve and integrates CNN network to extract features from the perspective of feature matrix and LSTM network to capture and learn time series data, which can further improve the prediction accuracy of ECL-LGSP model.

4.1. Design of CNN-LSTM network model

The CNN-LSTM hybrid network is composed of CNN layer, LSTM layer, fully connected layer and output layer. The network model structure is shown in Figure 4, and the detailed calculation process is shown below.

Firstly, the K-dimensional effective features after feature dimension reduction by ECF are entered into two CNN layers with Relu activation function and kernel size 2. In order to adapt to different prediction tasks, we select convolution kernels of 1×[k/2] and 1×[k/4] in series according to feature dimension k, which decreases the computation while ensuring the global feature, and reduces the training time while ensuring the model accuracy [25]. The convolution operation and pooling operation are respectively represented by formulas (8) and (9), and the output result of CNN network is an $m \times r$ matrix, as shown in formula (10),

$$P_j^i = f(\sum_{i=1}^{N} P_i^{l-1} * w_{ij}^l + b_j^i) \tag{8}$$

$$P_j^l = f(\alpha_j^l F_d(P_j^l - 1) + b_j^l) \tag{9}$$

in formula (8), P_j^i represents the *j*th convolution map at the convolution layer l, namely the liquefied gas concentration features extracted by the convolution layer; P_j^{l-1} represents the *i*th upper convolution map; w_{ij}^l represents the weight of the *j*th convolution kernel at the convolution layer l after the *i*th operation; b_j^i represents the bias of the *j*th convolution kernel at the convolution layer. In formula (9), P_j^l represents the *j*th feature map in the pooling layer l; α_j^l represents the multiplicative bias of the feature map; $F_d(x)$ represents the down sampling function.

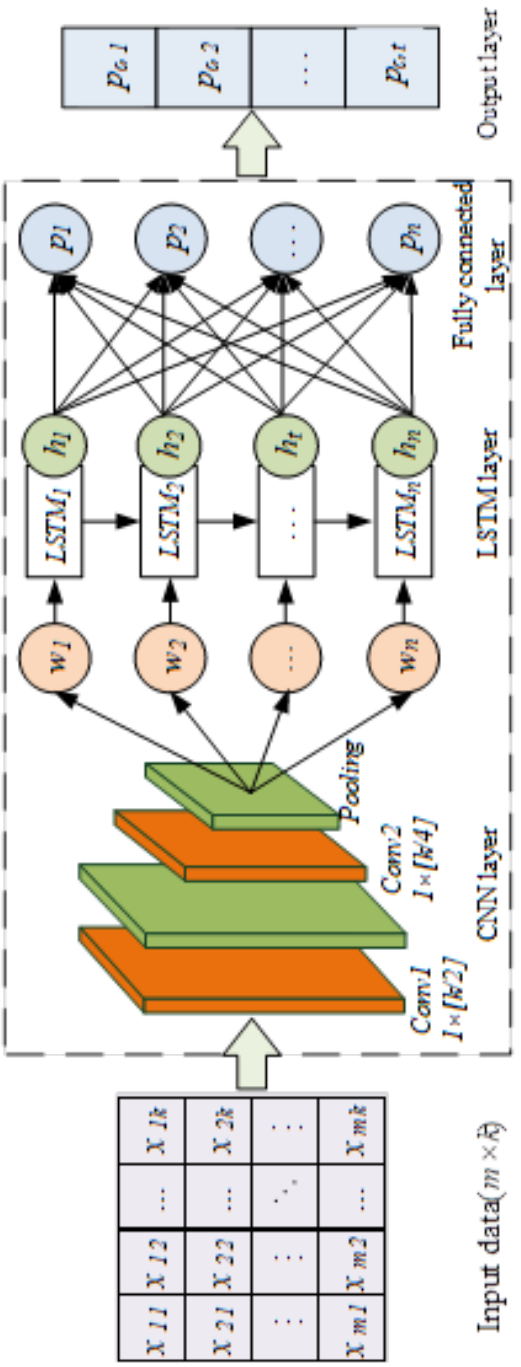


Figure 4. CNN-LSTM network model structure

$$W_t = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1r} \\ w_{21} & w_{22} & \cdots & w_{2r} \\ \vdots & \cdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mr} \end{bmatrix} = [w_1\ w_2 \ldots w_m]^T \tag{10}$$

Secondly, the $m \times r$ matrix of CNN output is divided into m row vectors $w_1, w_2, \cdots, w_m$ according to the time series order, and then input into the LSTM network. The memory unit of each LSTM is mainly composed of input gate, forget gate and output gate [26]. After calculating the input gate, forget gate and output gate, the hidden layer h_t of LSTM can be obtained as follows:

$$h_t = o_t \cdot tanh(C_t) \tag{11}$$

wherein, o_t is the output of the output gate, C_t is the state of the tth memory unit, and$tanh$is the hyperbolic tangent function.

Finally, h_t is taken as the input $h = [h_1\ h_2 \ldots h_m]$ of the fully connected layer, the input matrix$h$ is multiplied by the weight matrix trained by the hidden layer and the bias matrix is added to obtain the hidden layer p, and then the predicted value $\hat{p}_{c,t}$ of gas concentration is obtained by the same calculation for p,

$$p = \tau(h) \tag{12}$$

$$\hat{p}_{c,t} = w^{fc} p + b_{fc} \tag{13}$$

wherein, p is the fully connected hidden layer matrix, τ is the input matrix multiplied by the weight matrix plus the bias matrix function, $\hat{p}_{c,t}$ is the predicted value of liquefied gas concentration at time t, $\hat{p}_{c,t} = (\hat{p}_{c,1}, \hat{p}_{c,2}, \cdots, \hat{p}_{c,m})$, w^{fc} and b_{fc} are respectively the weight matrix and bias value matrix of the fully connected layer obtained by training set.

4.2. Training of CNN-LSTM network model

According to the working principle of ECF and the structure of CNN-LSTM network model, the training steps of ECL-LGSP model are given as follows:

Step 1. We enter and process the storage tank operation data to obtain the original feature attribute data $D = \{x_{ti} | 0 < t \leq m, 1 \leq i \leq n\}$.

Step 2. We calculate the ECF values of liquefied gas concentration and its features in the time range of $0 < t \leq m$.

Step 3. We calculate the ECF difference γ_i and fluctuation threshold δ according to formulas (6) and (7) and obtain the effective feature matrix X^* according to the judging criteria of effective features.

Step 4. We divide the corresponding attribute data of X^* into training set X'_{train} and test set X'_{test}.

Step 5. We input $X_{train}^{'}$ into the CNN network and realize feature extraction and dimension reduction through convolution layer and pooling layer, to obtain effective feature vector w. The operation process is shown in formulas (8) and (9).
Step 6. We input the data processed by Step 5 into the LSTM network, enter the data set of the hidden layer as row vectors $w_1, w_2, \ldots, w_n$, and then get the hidden layer output h_t.
Step 7. We take $h_t$as the input of the fully connected layer and get the predicted value of the liquefied gas concentration by calculation, that is, $\hat{p}_{c,t} = (\hat{p}_{c,1}, \hat{p}_{c,2}, \cdots, \hat{p}_{c,m})$. The objective function is the root mean square error (RMSE). When the MSE is minimum, the training stops to get the final ECL-LGSP model $y_{train} = f(w, b)x_{train}$.

The pseudo-code form of ECL-LGSP is shown in algorithm 1, wherein, "/**/" indicates the annotation.

Algorithm 1. ECL-LGSP
Input: $D = \{x_{ti} | 0 < t \leq m, 1 \leq i \leq n\}$: The original feature attribute data obtained by data preprocessing
Output:
Xpredict: Liquefied gas concentration value in the future continuous time.
Begin
/*①ECF feature selection */
01 for i=1 to n do
02 set $K_e = \frac{E(E_x - \mu_e)^4}{\sigma_e^4}$; /*Calculate envelope kurtosis according to formula (2). */
03 set $K_c = \frac{E(E_c - \mu_c)^4}{\sigma_c^4}$; /*Calculate the envelope spectral kurtosis according to formula (3). */
04 set $R_f(x, p) = \frac{Cov(x,p)}{\sqrt{D(x)}\sqrt{D(p)}}$; /*Calculate the correlation coefficient according to formula (4). */
05 set $f_{ECF} = K_e \cdot K_c \cdot R_f$; /*Calculate the weighted set kurtosis according to formula (5). */
06 set $\delta = \frac{1}{n}\sum_{i=1}^{n} |f_{X_i} - f_o|$; /* , Calculate the ECF difference value according to formula (6). */
07 set $\delta = \frac{1}{n}\sum_{i=1}^{n} \gamma_i$; /* Calculate the fluctuation threshold according to formula (7). */
08 if $\gamma_i \leq \delta$ $X^* += X_i$; /*Obtain X* according to judging criteria of effective feature. */
09 end for
/*②CNN-LSTM network model prediction*/
10 for i=0 to X* do
11 set $X_j = P_j^i(X_i)$; /*Carry out convolution calculation according to formula (8).*/
12 set $W_t = P_j^l(X_j)$; /*Carry out pooling calculation to get the feature vector according to formula (9).*/
13 end for
14 def Lstm(batch_size,time_step, W_t); /*Model training.*/
15 for i=0 to w_t do
16 set $f_{loss} = MSE()$;
17 end for
18 return $f(w, b)x_{train}$;
19 def Predict():
20 Lstm(W_t);
21 set Xpredict = Predict(); /*Get the predicted value of liquefied gas concentration.*/
22 return Xpredict
End

4.3. Evaluation function of ECL-LGSP model

The root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) are applied to evaluate the model prediction effect. The smaller the evaluation indexes values, the higher the model prediction accuracy. The specific formulas are shown as follows, wherein, $p_{c,t}$ and $\hat{p}_{c,t}$ represent the true and predicted values of the liquefied gas concentration respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{p}_{c,t} - p_{c,t})^2} \tag{14}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{p}_{c,t} - p_{c,t}\right| \tag{15}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{p}_{c,t} - p_{c,t}|}{p_{c,t}} \tag{16}$$

RMSE and MAE can reflect the error of the accurate value and the predicted value, and the smaller the value of both, the higher the model prediction accuracy. MAPE reflects the ratio of error to accurate value.

5. EXPERIMENT ANALYSIS

5.1. Experiment preparation

I. Experimental environment. The CO2 anomaly monitoring and early warning system of the overall control platform of instrument operation monitoring and early warning in an oil extraction plant is simulated, and the simulation environment structure diagram is shown in Figure 5. The station is a star network topology, and the instrument of CO2 storage tank is connected to PLC, and then to the terminal (host computer). The central control terminal is responsible for the intelligent control, and the terminal system monitors the real-time status of the operation of the CO2 storage tank.

II. Data preparation. We select the CO2 storage tank operation data of a key station of an oil extraction plant as the experimental data, and the time resolution is 5min. The data set includes 8640 storage tank operation data from April 1, 2022 to April 30, 2022. 10 storage tank operation parameters in Table 1 are selected as input features, and divided into training set, verification set and test set according to the ratio of 8:1:1. The features such as number of employees and gender of personnel are one-hot encoded, and other features are normalized.

Table 1. Storage tank operation parameters

Notation	Meaning	Attribute category
N	Tank life	*Static attribute (v)*
P	Pressure inside tank	*Dynamic attribute (d)*
T	Temperature inside tank	*Dynamic attribute (d)*
P_c	Pressure inside station	*Dynamic attribute (d)*
T_c	Temperature inside station	*Dynamic attribute (d)*
h	Humidness inside station	*Dynamic attribute (d)*
W_s	Wind speed inside station	*Dynamic attribute (d)*
W_d	Wind direction inside station	*Dynamic attribute (d)*
n	Number of employees	*Additional attribute (q)*
g	Gender of personnel	*Additional attribute (q)*

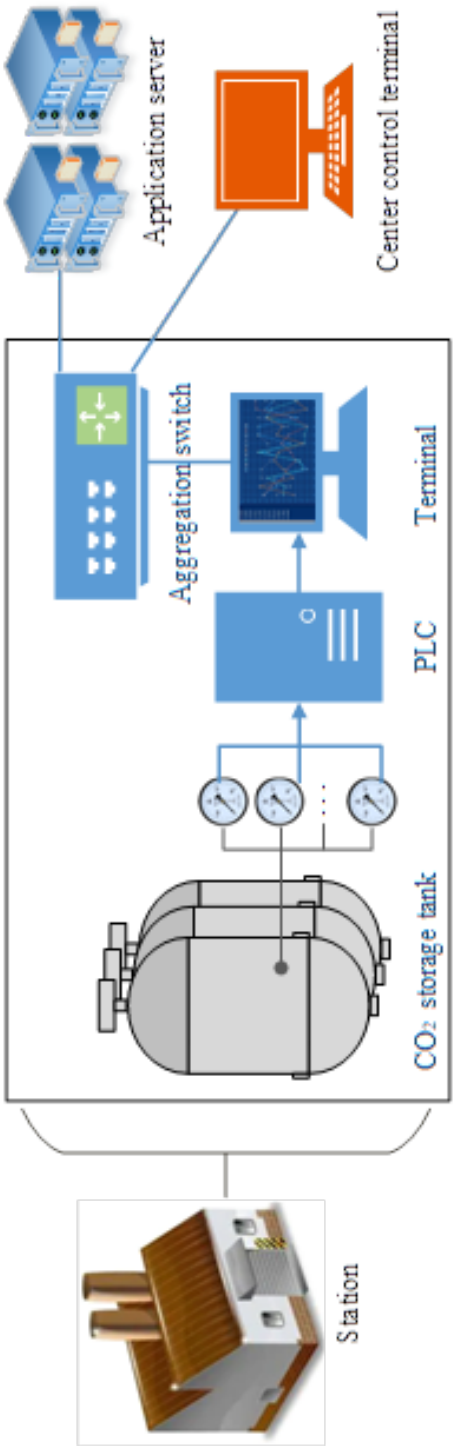
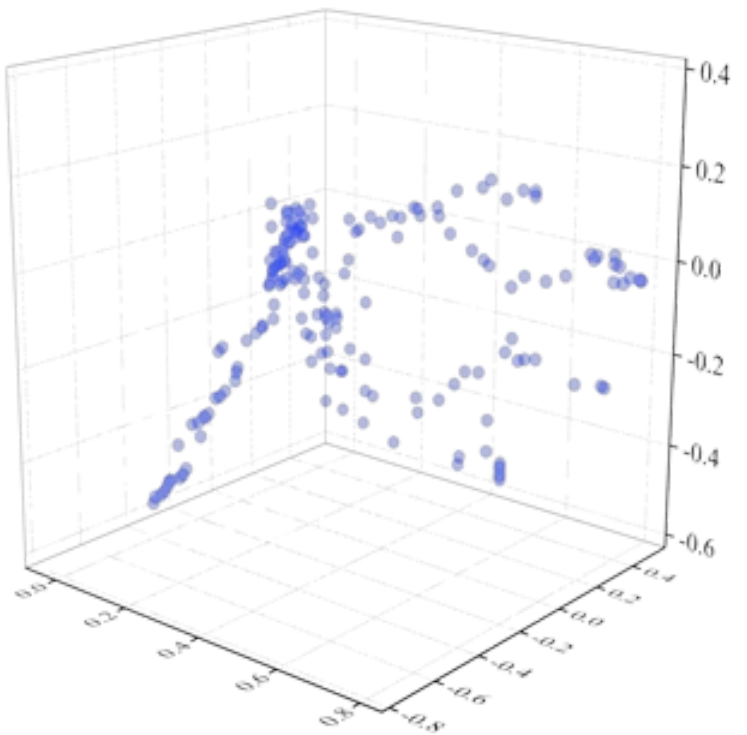


Figure 5. Experimental environment structure diagram

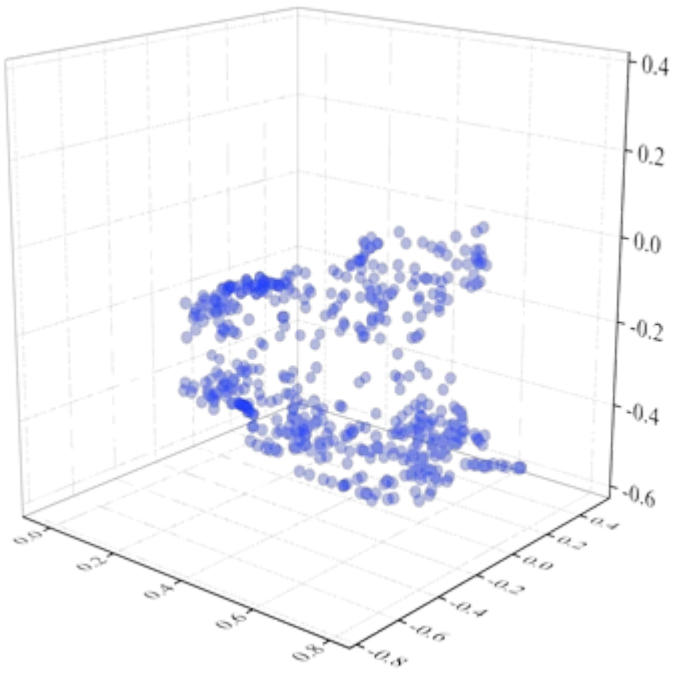
III. Contrast model. The experiments include different feature selection method experiments, performance evaluation experiments and tolerance experiments of ECL-LGSP model. We select CNN, LSTM and CNN-LSTM network models, and the hybrid model combining Pearson correlation coefficient, PCA, TSNE with CNN-LSTM respectively for comparison.

5.2. Feature dimension reduction experiment

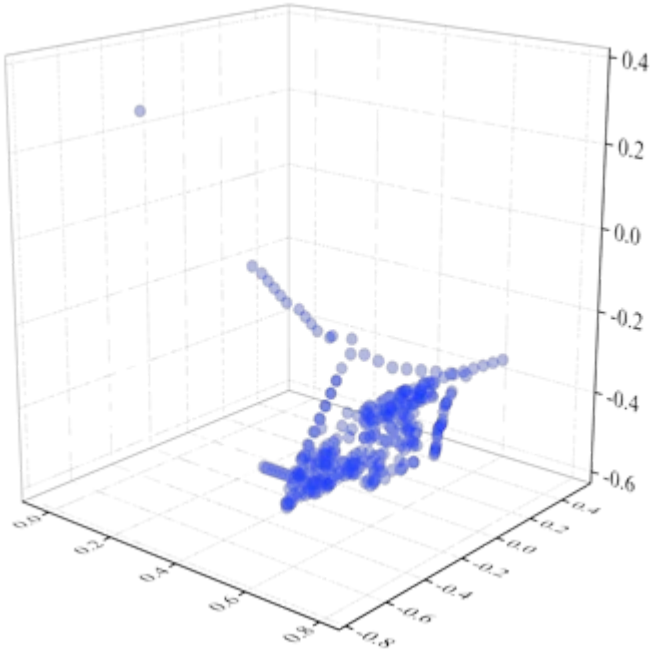
By using the Pearson correlation coefficient, PCA, TSNE and ECF method, the feature dimension reduction experiment is carried out after preprocessing of the 10-dimensional storage tank operation data, aiming to analyze the effectiveness and superiority of ECF method. The cumulative contribution rate threshold is set as 95% in the experiment, and the scatter diagrams of 3D feature data after feature dimension reduction are shown in Figure 6(a)-(d).



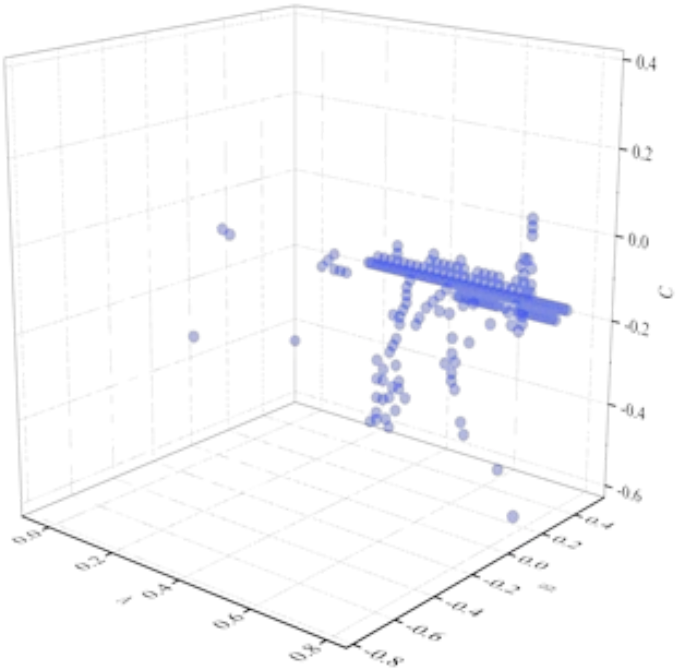
(a) ECF feature dimension reduction results



(b)TSNE feature dimension reduction results



(c) PCA feature dimension reduction results



(d) Pearson feature dimension reduction results

Figure 6. Dimension reduction results of different methods

It is obvious from Figure 6(a)-(d) that the ECF method achieves good feature dimension reduction results. However, after dimension reduction by TSNE, PCA and Pearson methods, the scatter diagram has an obvious aggregation degree, and some noise points are generated in the dimension reduction results of PCA and Pearson methods, which is not conducive to maintaining the trend of non-stationary time series.

At the same time, we combine the above feature dimension reduction methods with CNN-LSTM network respectively, to obtain the hybrid models of Pearson-CL, PCA-CL and TSNE-CL, and compare them with CNN, LSTM, CNN-LSTM and ECL-LGSP model. The prediction accuracy and time of each model are shown in Table 2.

Table 2. Predictive performance metrics comparison of different models

Models	MAE	MAPE (%)	RMSE	Prediction time (s)
CNN	24.98	5.25	26.07	249.54
LSTM	23.93	4.79	23.00	594.32
CNN-LSTM	18.90	3.68	20.04	277.62
Pearson-CL	22.64	4.61	25.14	238.59
TSNE-CL	16.25	3.15	18.11	771.35
PCA-CL	15.80	3.09	16.98	296.12
ECL-LGSP	14.82	2.87	16.73	189.36

By analyzing the experimental results, the following conclusions are given.

1) The prediction accuracy of the hybrid models combined with CNN-LSTM is better than that without CNN-LSTM, which proves the advantage of the CNN-LSTM network model.
2) In the hybrid models combined with CNN-LSTM, Pearson-CL has poor performance in the experiment. It cannot handle non-stationary time series well in feature selection, so the prediction accuracy is low.
3) By comparing ECL-LGSP with PCA-CL that has good overall performance, MAE, MAPE, RMSE and prediction time of ECL-LGSP are lower than those of PCA-CL, which proves the effectiveness and superiority of ECL-LGSP.

5.3. Model tolerance experiment

To analyze the influence of the "quality" and "quantity" of data on the prediction accuracy and real-time performance of the model, the prediction accuracy and prediction time of PCA-CL, CNN-LSTM and TSNE-CL models with better comprehensive performance and ECL-LGSP model are tested under different data volumes and different data loss rates, in order to analyze the effects of different data volumes and data loss rates on model tolerance.

5.3.1. Tolerance experiment under different data volumes

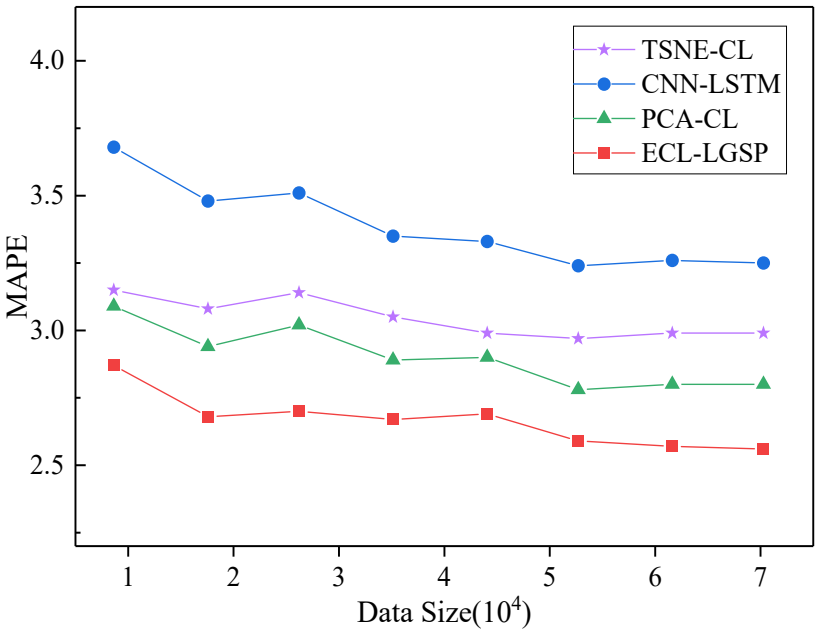
The original time series data set is expanded, and data sets of different time lengths are selected successively to study the influence of data volume on the prediction accuracy and prediction time of the model. The division of data volume is shown in Table 3.

Table 3. Division of data volume in different time lengths

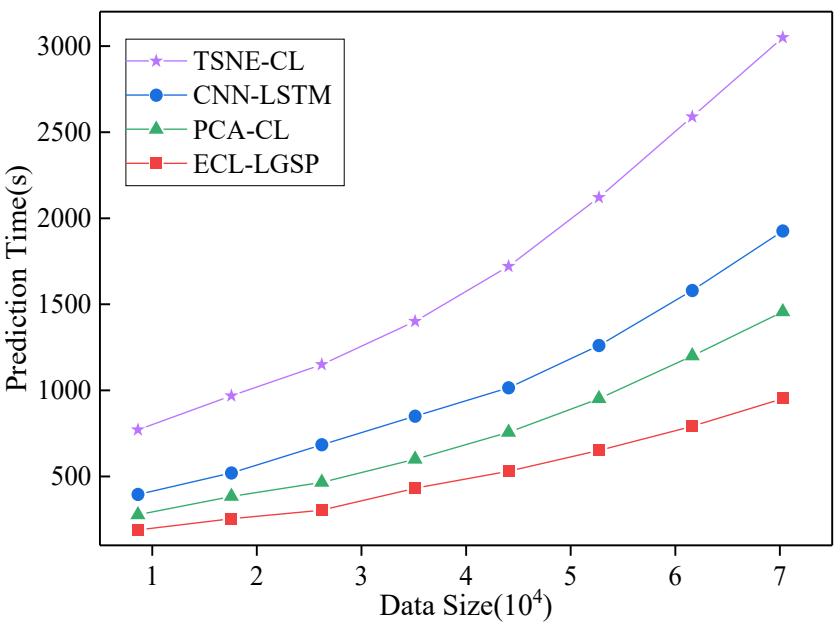
Date range	Data volume (piece)	Date range	Data volume (piece)
April 1, 2022-April 30, 2022	8640	April 1, 2022-August 31, 2022	44064
April 1, 2022-May 31, 2022	17568	April 1, 2022-Spetember 30, 2022	52704
April 1, 2022-June 30, 2022	26208	April 1, 2022-October 31, 2022	61632
April 1, 2022-July 31, 2022	35136	April 1, 2022-November 30, 2022	70272

By analyzing the experimental results, the following conclusions are given.

1) As shown in Figure 7(a)-(b), with the increase of data volume, the prediction errors of the four models show a trend of first decreasing and then stabilizing, and the prediction time shows an upward trend. Among them, the prediction accuracy of ECL-LGSP, PCA-CL and TSNE-CL models is obviously better than that of CNN-LSTM model, but the running time of TSNE-CL model is too long. When the data volume reaches 70,272 pieces, the prediction time exceeds 5min, which cannot guarantee the real-time performance of the model prediction.
2) Compared with the PCA-CL model that has better comprehensive performance, the ECL-LGSP has the lowest prediction error, the least prediction time and the most stable change, indicating that the ECL-LGSP model has a certain tolerance to the data size.



(a)



(b)

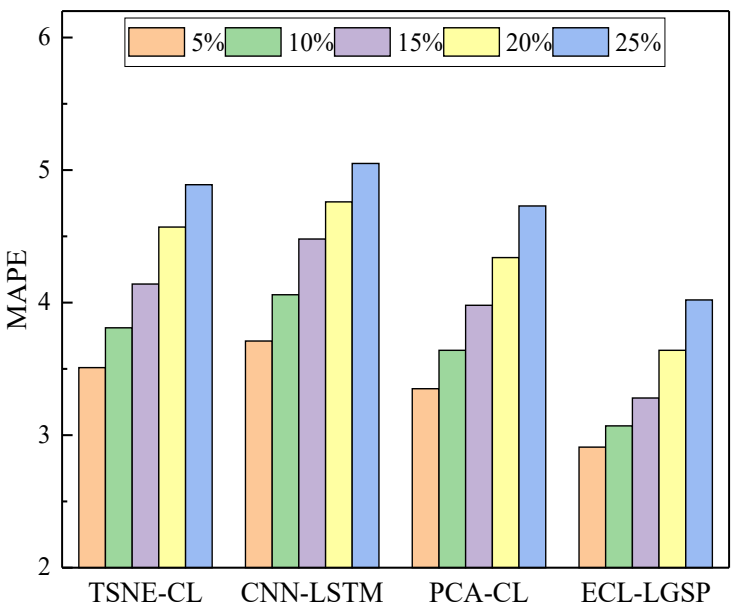
Figure 7. Tolerance experiment results under different data volumes

5.3.2. Tolerance experiments under different data loss rates

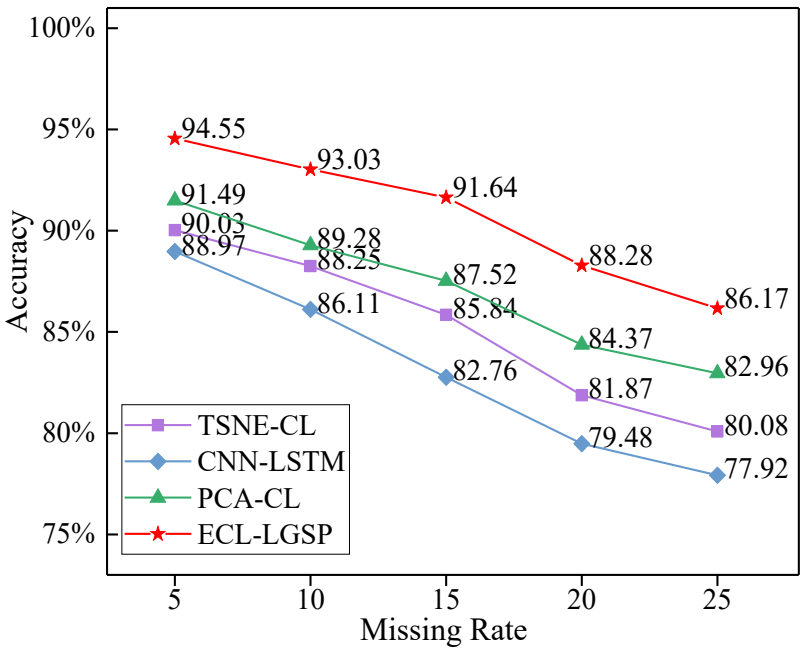
We select random data with a proportion of 5%, 10%, 15%, 20% and 25% as missing validation data to study the impact of data loss rate on the prediction accuracy of the model. Figure 8(a)-(b) shows the impact of the change of loss rate on the prediction accuracy of the four models.

By analyzing the experimental results, the following conclusions are given.

1) With the increase of the proportion of missing data, the prediction error of the four models shows an upward trend, and the prediction accuracy shows a downward trend, indicating that within a certain range, the higher the data loss rate, the lower the prediction accuracy of the model.
2) With the increase of the proportion of missing data, the prediction accuracy of PCA-CL, TSNE-CL and CNN-LSTM models declines rapidly, while the prediction accuracy of ECL-LGSP decreases steadily and is always the highest. This is attributed to that PCA, TSNE and CNN are linear combinations based on original data points and are more sensitive to data loss, while ECF is a curve fitting based on original data points, which can indirectly interpolate the data through approximate curve fitting where data loss occurs, so it has a certain tolerance to data loss.



(a)



(b)

Figure 8. Tolerance experiment results under different data loss rates

By analyzing the experimental results, the following conclusions are given.

1) With the increase of the proportion of missing data, the prediction error of the four models shows an upward trend, and the prediction accuracy shows a downward trend, indicating that within a certain range, the higher the data loss rate, the lower the prediction accuracy of the model.
2) With the increase of the proportion of missing data, the prediction accuracy of PCA-CL, TSNE-CL and CNN-LSTM models declines rapidly, while the prediction accuracy of ECL-LGSP decreases steadily and is always the highest. This is attributed to that PCA, TSNE and CNN are linear combinations based on original data points and are more sensitive to data loss, while ECF is a curve fitting based on original data points, which can indirectly interpolate the data through approximate curve fitting where data loss occurs, so it has a certain tolerance to data loss.

5.4. Model performance experiment

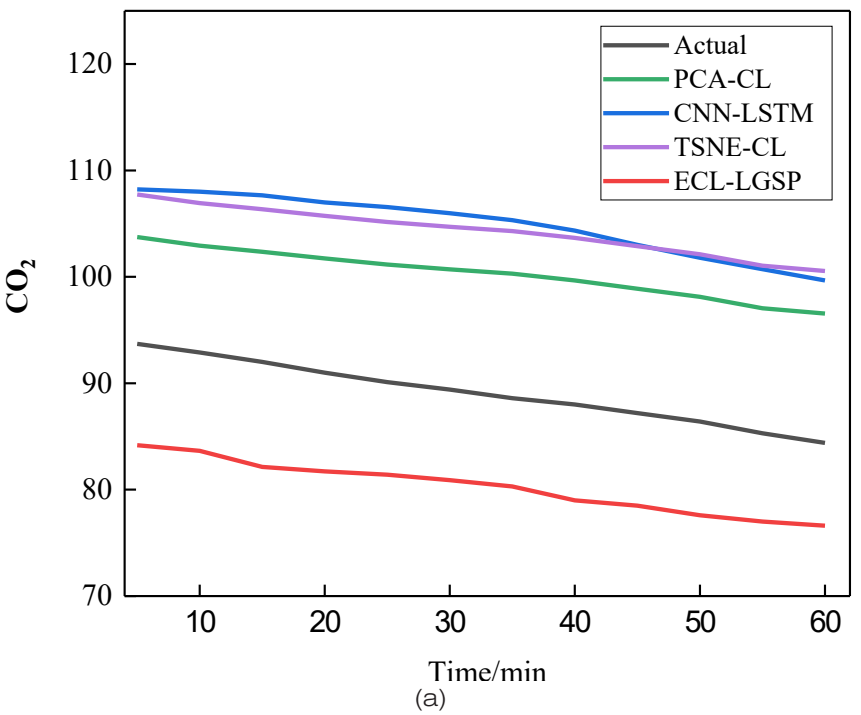
The average prediction time and prediction accuracy of ECL-LGSP with TSNE-CL, CNN-LSTM and PCA-CL are compared under different time scales and different prediction tasks. The purpose is to verify that ECL-LGSP has obvious advantages in prediction accuracy, running speed and stability. Table 4 shows the basic parameter settings of CNN-LSTM model.

Table 4. Model parameter setting

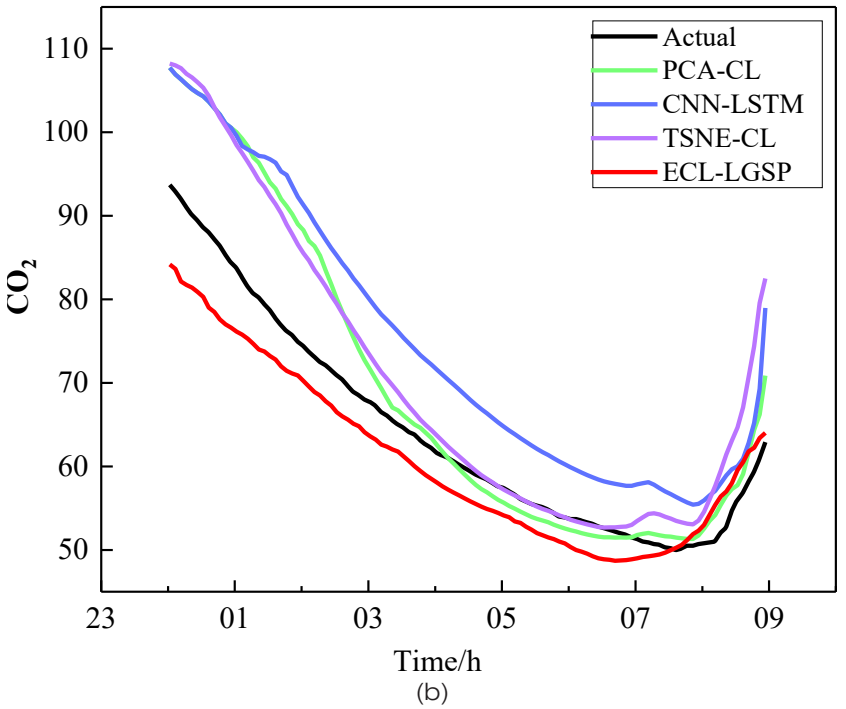
Parameters	Parameter values	Parameter meaning
α	0.001	Learning rate
hidden_layer	2	Hidden layer
hidden_layer_unit	10	Number of hidden layer neurons
batch_size	32	Batch size
Epoch	50	Number of iterations
Optimizer	Adam	Optimizer
activate_function	0.25	Activation function
dropout rate	0.1	Discard rate
objective_function	MSE	Loss function

5.4.1. Model performance experiment on different time scales

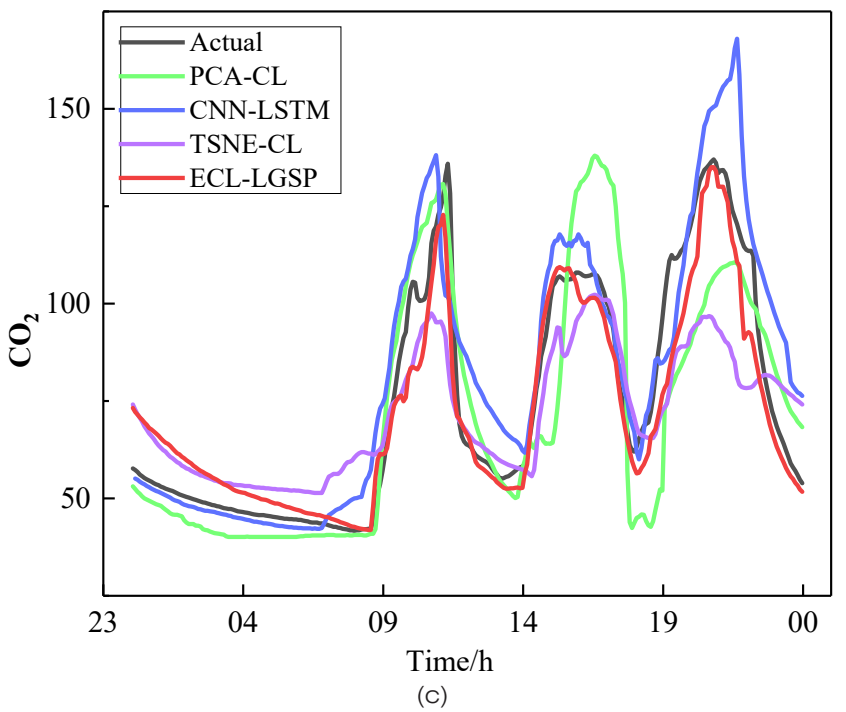
In order to verify the prediction performance of the ECL-LGSP model on different time scales, this method and the above three prediction models are applied to conduct short-term prediction experiments on CO2 concentration in the future time scales of 1 hour, 8 hours and 1 day. The prediction curves of CO2 concentration of the storage tank in the next 1 hour, 8 hours and 1 day are shown in Figure 9(a)-(c). The prediction error of each model is shown in Figure 9(d).



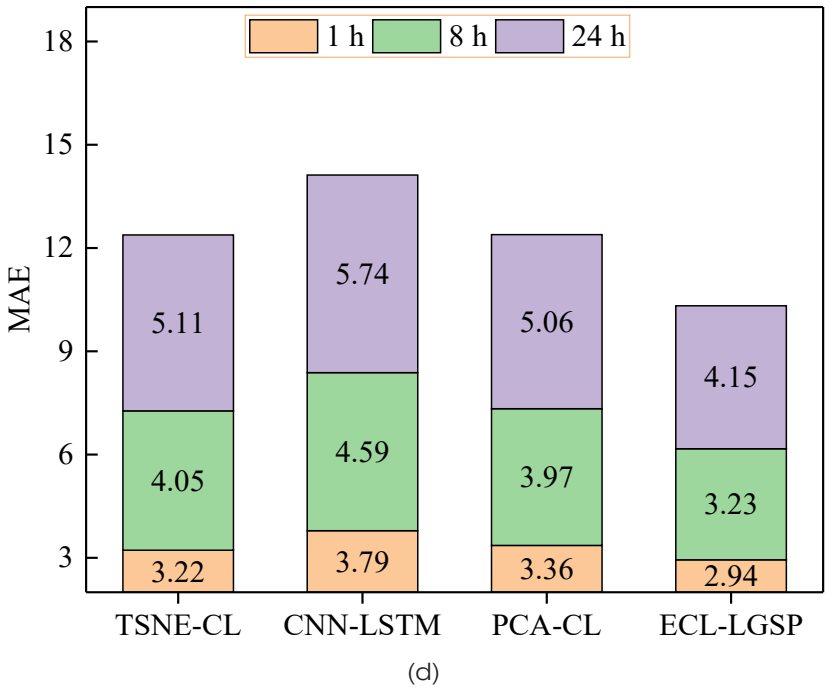
(a)


Actual
PCA-CL
CNN-LSTM
TSNE-CL
ECL-LGSP
110
100
90
80
70
60
50
CO_2
23
01
03
05
07
09
Time/h

(b)


Actual
PCA-CL
CNN-LSTM
TSNE-CL
ECL-LGSP
150
100
50
CO_2
23
04
09
14
19
00
Time/h

(c)



(d)

Figure 9. Prediction results on different time scales

By analyzing the experimental results, the following conclusions are given.

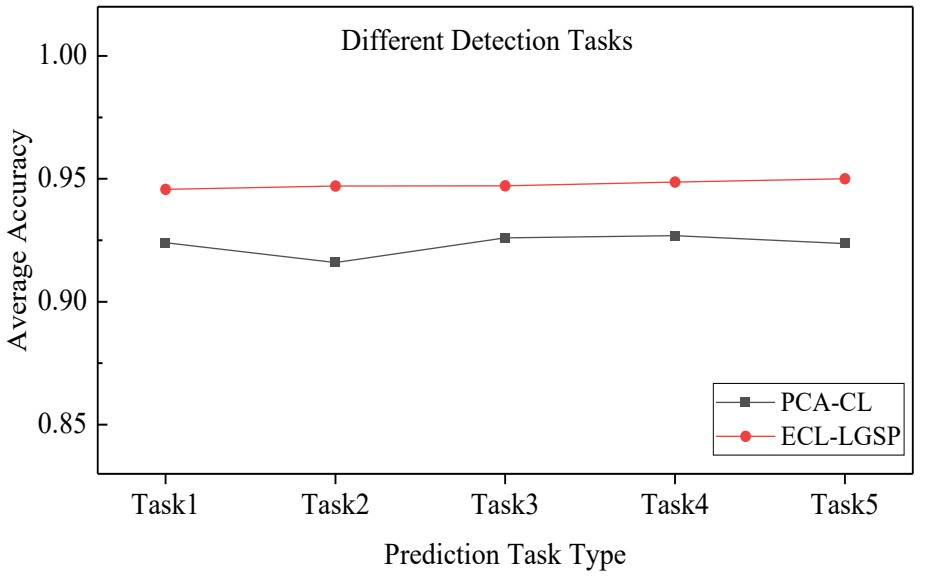
1) As shown in Figure 9(a)-(c), the curve fitted by ECL-LGSP method is highly consistent with the actual change, which proves that this method can predict CO2 concentration in the future, and the error is relatively stable.
2) As shown in Figure 9(d), on different time scales, compared to other prediction models, ECL-LGSP has the lowest prediction error, and the prediction accuracy is the highest under the condition in the next 1 hour, showing that ECL-LGSP has certain advantages for short-term prediction of liquefied gas concentration.

5.4.2. Model stability experiment

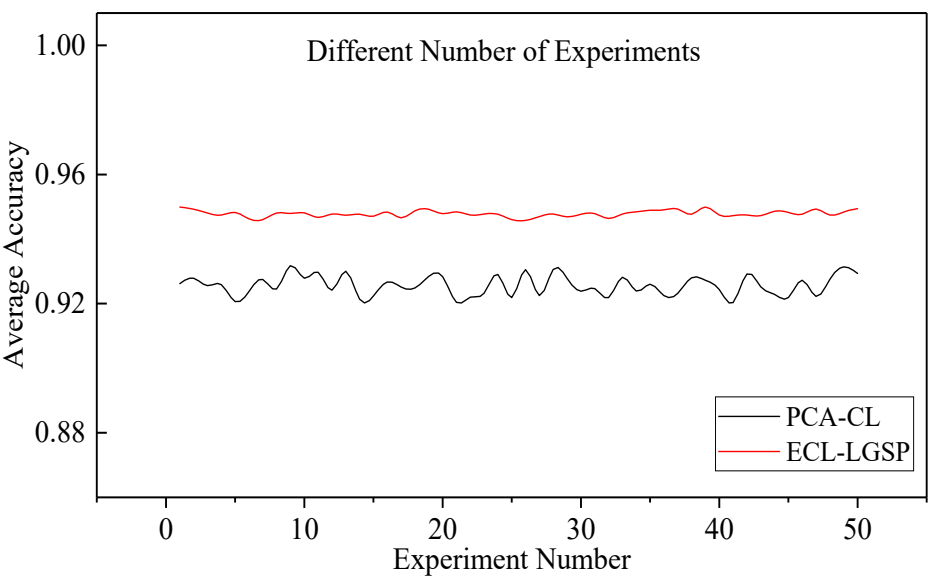
Under different prediction tasks and numbers of experiments, the PCA-CL model with better performance is selected as the comparison model to test the accuracy of the two models separately, in order to further analyze the stability of the model. In the experiment, the prediction task and the number of experiments are taken as variables, and the control variable method is used to test the change of the model accuracy separately.

We carry out 5 groups of experiments with different prediction tasks and select different gas storage tanks as prediction tasks respectively. 10 experiments are carried out in each group, and the experimental results are averaged. Figure 10(a) shows the influence of different prediction tasks on the stability of the two models.

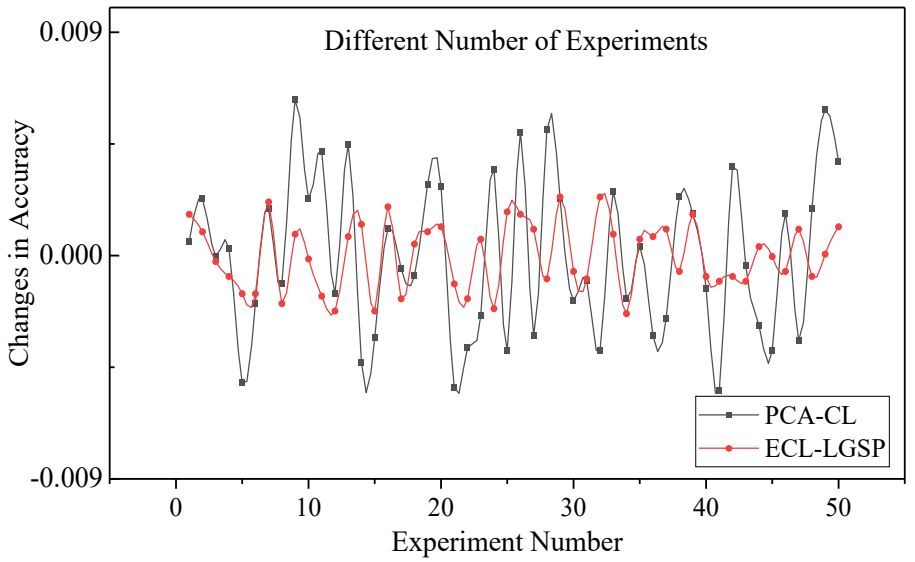
We conduct 50 groups of experiments with different numbers of experiments and take each additional experiment as one group (the first group is one experiment). The experimental results of each group are averaged. Figure 10(b)-(c) shows the influence of the number of experiments on the stability of the two models.


Different Detection Tasks
1.00
0.95
0.90
0.85
Average Accuracy
PCA-CL
ECL-LGSP
Task1
Task2
Task3
Task4
Task5
Prediction Task Type

(a)


Different Number of Experiments
1.00
0.96
0.92
0.88
Average Accuracy
PCA-CL
ECL-LGSP
0
10
20
30
40
50
Experiment Number

(b)



(c)

Figure 10. Model stability experimental results

As a supplementary note, Table 5 shows the values of each index of the two models in the model stability experiment.

Table 5. The index values of the two models in the experiment

Experiment condition	Index	PCA-CL	ECL-LGSP
Different prediction tasks	Accuracy interval	[0.9159,0.9268]	[0.9457,0.9500]
	Maximum difference (%)	1.12	0.44
	Minimum difference (%)	0.04	0.01
	Standard deviation	0.0044	0.0017
Different experiment times	Accuracy interval	[0.9203,0.9317]	[0.9458,0.9499]
	Maximum difference (%)	1.17	0.42
	Minimum difference (%)	0	0
	Standard deviation	0.0032	0.0010

By analyzing the experimental results, the following conclusions are given.

1) As shown in Figure 10(a), when the prediction task changes, the accuracy of both models fluctuates within a certain range, and the standard deviation of ECL-LGSP model is smaller than that of PCA-CL model. It shows that the stability of ECL-LGSP model is better than that of PCA-CL model under different prediction tasks.
2) As shown in Figure 10(b)-(c), with the increase in the number of experiments, the accuracy fluctuation of ECL-LGSP model is much smaller than that of PCA-CL model. The standard deviations of ECL-LGSP and PCA-CL are 0.0032 and 0.001, respectively. Therefore, the stability of ECL-LGSP model is better than that of PCA-CL model under different numbers of experiments.

6. CONCLUSIONS

We propose a short-term prediction method of liquefied gas concentration to solve the problem of safety hazards prediction of liquefied gas storage tank operation in oil and gas gathering and transportation industry. The method includes two stages: First, an Extreme Change Function is introduced to reduce the feature dimension and select effective features; Second, we use CNN network to achieve feature extraction, and capture the change law of data by LSTM to achieve short-term prediction of liquefied gas concentration. The experimental results are summarized as follows:

I. The ECL-LGSP method is suitable for solving the short-term prediction problem of liquefied gas concentration. Experimental results show that ECL-LGSP method has obvious advantages in prediction accuracy, prediction speed, tolerance and stability compared with similar methods.

II. ECF reduces the feature dimension by calculating the weighted set kurtosis value of the feature curve. Compared with other feature dimension reduction method, it has obvious advantages in prediction accuracy and running speed, and has a certain tolerance to high data volume and data loss cases.

III. The ECL-LGSP method integrates CNN-LSTM hybrid network, and combines the advantages of CNN feature extraction with LSTM time series processing, so as to realize the short-term prediction of liquefied gas concentration and improve the prediction accuracy.

IV. The short-term prediction method of liquefied gas concentration has large application scenarios in various fields. When our method is applied to real scenes, some more specific problems need to be solved. For example, how to add the influence of the nodes' own attribute value, how to obtain the best weight coefficients in the CNN-LSTM network.

REFERENCES

[1] Bożena K, Aneta K, Robert P, et al. Research on the safety and security distance of above-ground liquefied gas storage tanks and dispensers[J]. International Journal of Environmental Research and Public Health,2022,19(2).

[2] Guo X. Study on Near-source Release and Dispersion for Hazardous Liquefied Gas and Assessments of Accident Consequences [D]. Tianjin University, 2021.

[3] Geng T, Ju T, Li B, et al. Prediction of the tropospheric NO2 column concentration and distribution using the time sequence-based versus influencing factor-based random forest regression model[J]. Sustainability,2023,15(3).

[4] Wu C, He H, Song R, et al. A hybrid deep learning model for regional O3 and NO2 concentrations prediction based on spatiotemporal dependencies in air quality monitoring network[J]. Environmental Pollution (Barking, Essex :1987),2023,320.

[5] Sun Y. Analysis and prediction of CO2 emission calculation models in the steel industry[D]. Metallurgical Automation Research and Design Institute,2023.

[6] Lama A, Ran T, Bilal F, et al. Greenhouse gas emission prediction on road network using deep sequence learning[J]. Transportation Research Part D,2020,88.

[7] Djeziri, A. M, Djedidi, et al. A temporal-based SVM approach for the detection and identification of pollutant gases in a gas mixture[J]. Applied Intelligence,2021,52(6).

[8] Guoquan L,Zhichao J,Qi W. Analysis of Gas Leakage Early Warning System Based on Kalman Filter and Optimized BP Neural Network[J]. IEEE ACCESS,2020,8.

[9] Mohamad-Javad M,Faramarz B,Min Z, et al. Application of a hybrid mechanistic/machine learning model for prediction of nitrous oxide (N2O) production in a nitrifying sequencing batch reactor[J]. Process Safety and Environmental Protection,2022,162.

[10] Duan H, Meng X, Tang J, et al. Prediction of NOx concentration using modular long short-term memory neural network for municipal solid waste incineration[J]. Chinese Journal of Chemical Engineering,2023,56(04):46-57.

[11] Zhe Y,Chunlai Y,Xiaolei Y, et al. NOx concentration prediction in coal-fired power plant based on CNN-LSTM algorithm[J]. Frontiers in Energy Research,2023.

[12] Qin Y, Ouyang C, Fang P. Reservoir carbon dioxide flux prediction based on CNN-LSTM model and small sample data [J]. Journal of Chongqing Jiao tong University (Natural Science),2022,41(06):119-125.

[13] Chao L,Ailin Z,Junhua X, et al. LSTM-Pearson Gas Concentration Prediction Model Feature Selection and Its Applications[J]. Energies,2023,16(5).

[14] Chi D, Huang Q, Liu L, et al. Research on the Prediction Model of Dissolved Oxygen Content in Dished Lakes Based on PCA-MIC-LSTM [J]. Yangtze River,2022,53(06):54-60.

[15] HyungSub K,Florent N,NamJin N, et al. Future Projection of CO2 Absorption and N2O Emissions of the South Korean Forests under Climate Change Scenarios: Toward Net-Zero CO2 Emissions by 2050 and Beyond[J]. Forests,2022,13(7).

[16] A. L R, A. D E, J. D R, et al. Spectral Kurtosis Based Methodology for the Identification of Stationary Load Signatures in Electrical Signals from a Sustainable Building[J]. Energies,2022,15(7).

[17] Wang L. Research on the Detection Method of Centrifugal Pump Wear Ring Rubbing Sound Signal Based on Spectral Kurtosis[D]. Zhejiang University,2023.

[18] Bing D,Yingjie P,Ning L, et al. Bearing Fault Diagnosis Based on Prime Mean Spectral Segmentation Kurtogram[J]. Journal of Physics: Conference Series,2023,2419(1).

[19] Gao R, Hu D, Shi W, et al. Fault Feature Enhancement of Rolling Bearing Acoustic Signals based on Maximum Correlation Kurtosis Deconvolution and Spectral Kurtosis[J]. Noise and Vibration Control,2022,42(02):102-107.

[20] Liu J, Zhao X, Zhang W, et al. Prediction of NOx Concentration at SCR Inlet of Power Plant Boiler Based on CNN (1D)-LSTM Model[J]. Electronic Measurement Technology,2023,46(13):59-65.

[21] Jie J, Ke'nan L, Fang'ai L. Prediction of SO2 Concentration Based on AR-LSTM Neural Network [J]. Neural Processing Letters,2022.

[22] Ming F,Dan L,Siyan L. A deep learning-based direct forecasting of CO2 plume migration[J]. Geoenergy Science and Engineering,2023,221.

[23] Yuan Z, Chen W, Jiang Z, et al. Research Progress on Nonlinear Coupling Constitutive Relation of Rarefied Gas Flow [J]. Physics of Gases,2022,7(05):1-15.

[24] Anshul S,Pardeep K,Kumar H V, et al. Hilbert transform and spectral kurtosis based approach in identifying the health state of retrofitted old steel truss bridge[J]. World Journal of Engineering,2022,19(4).

[25] Li X, Bai C, Shi S. Prediction Method of Dissolved Gas Concentration in Locomotive Transformer Oil Based on CNN-BiLSTM Model [J]. Journal of the China Railway Society,2022,44(05):42-48.

[26] Surbhi K, Kumar S S. Machine learning-based time series models for effective CO2 emission prediction in India [J]. Environmental science and pollution research international,2022.