Hindawi
Geofluids
Volume 2021, Article ID 6641395, 13 pages
https://doi.org/10.1155/2021/6641395

Research Article

The Health Index Prediction Model and Application of PCP in CBM Wells Based on Deep Learning

Chaodong Tan,[1,2] Song Wang,[1] Hanwen Deng,[1] Guoqing Han,[1,2] Guanghao Du,[2] Wenrong Song,[3] and Xiongying Zhang[3]

[1]*State Key Laboratory of Petroleum Resource and Prospecting, China University of Petroleum, Changping, Beijing 102249, China*
[2]*College of Artificial Intelligence, China University of Petroleum, Changping, Beijing 102249, China*
[3]*Beijing Yadan Petroleum Technology Development Co., Ltd., Changping, 102200, China*

Correspondence should be addressed to Guoqing Han; hanguoqing@163.com

Received 25 November 2020; Revised 13 March 2021; Accepted 6 April 2021; Published 26 April 2021

Academic Editor: Guanglong Sheng

Aiming at the problems of the current production and operation status of the progressive cavity pump (PCP) in coalbed methane (CBM) wells which cannot be timely monitored, quantitatively evaluated, and accurately predicted, a five-step method for evaluating and predicting the health status of PCP wells is proposed: data preprocessing, principal parameter optimization, health index construction, health degree division, and health index prediction. Therein, a health index (HI) formulation was made based on deep learning, and a statistical method was used to define the health status of PCP wells as being healthy, subhealthy, or faulty. This allowed further research on the HI prediction model of PCP wells based on the long short-term memory (LSTM) network. As demonstrated in the study, they can reflect both the change trend and the contextual relevance of the health status of PCP wells with high accuracy to achieve real-time, quantitative, and accurate assessment and prediction. At the same time, the conclusion gives good guidance on the production performance analysis and failure warning of the PCP wells and suggests a new direction for the health status assessment and warning of other artificial lift equipment.

1. Introduction

Coalbed methane is a kind of clean energy; it is drained through depressor desorption; when the reservoir pressure is reduced to the desorption pressure of methane, the methane gas in the pores is desorbed, then diffuses and percolates into the wellbore [1–3]. The progressive cavity pump (PCP) is one of the lifting methods in CBM wells. The operation of PCP in CBM wells often fails, resulting in large production losses and short equipment life. Therefore, the monitoring, diagnosis, and early warning of the operation and health status of the PCP in CBM wells have attracted more and more attention from researchers and field engineers. Experience and statistical methods are not possible to evaluate the health status of the pump in the future and perform predictive maintenance. Some scholars thus have put forward some measures on PCP health management based on machine learning methods. For example, Saghir et al. discussed how to use data collected from a data acquisition system to apply data approximation and unsupervised machine learning methods to time series datasets to help analyze PCP performance and detect abnormal pump behavior [4]. Hoday et al. proposed a method based on abnormal monitoring to characterize PCP failures, maximize the information value of monitoring the operating conditions of each well, and minimize operating costs [5]. Saghir et al. proposed to convert the features extracted from time series data into images, which helps to detect abnormal behavior of PCP autonomously [6]. Prosper and West proposed the use of a machine learning framework that can be used to customize each workover configuration to optimize the service life of PCP while considering the heterogeneity and life of wells [7].

Due to the large number of parameters collected for CBM, quantitative evaluation of the health status of the PCP cannot be achieved and the evaluation results are not accurate. Some scholars also use some new technologies to

manage the health status of the PCP. For example, a tool called the Pressure Actuated Relief Valve (PAR Valve) is used above the PCP to eliminate solids settling during a shutdown [8]. Caballeroa et al. involved in supplying PCP technologies to the Orinoco Belt and have developed the exclusive and patented HR-PCP (hydraulically regulated PCP) technology in order to extend the run life of the conventional PCPs in these fields where the Mean Time Between Failure (MTBF) has shown a sharp decrease in the last few years [9]. In order to achieve continuous decision-making and control of the parameters of PCP wells, taking the maximum cumulative gas production as the optimization goal, a reinforcement learning model with the self-optimization ability and a model framework of the Q learning, Sarsa, Sarsa (lambda) algorithm were proposed [10]. Based on the above technical methods, although the service life of PCP can be prolonged and the output of CBM wells can be increased, real-time evaluation and prediction of the health status of the lifting equipment PCP cannot be carried out.

In fact, health status assessment has also been widely studied and applied in other equipment systems. Most of them use current detection data and historical operating data to evaluate the current health status of equipment systems or subsystems [11]. According to the different strategies of constructing the HI curve, it can be divided into two types: direct HI and indirect HI [12]. The former refers to the direct construction of health values with a certain physical significance based on the original monitoring data, guided by experts or empirical knowledge, through simple statistical analysis or feature extraction. Indirect HI is usually obtained by using machine learning methods to fuse or reduce the time domain features or frequency domain features of the sensor. It has no physical meaning and is often called virtual HI (VHI). Among the construction methods of VHI, the most popular is to use dimensionality reduction technology to construct VHI [13, 14]. Some scholars use the Mahalanobis distance to construct VHI [15–17] and use linear data transformation methods to construct VHI by fusing multiple features [18–20]. In the above method of constructing HI, the VHI constructed by dimensionality reduction calculation can best reflect the data change characteristics of the equipment collected and can better reflect the operating conditions of the equipment in real time. The methods of VHI provide a reference for the construction of the PCP health index.

Therefore, in view of the real-time evaluation and prediction of the health status of the PCP wells, this paper proposes a method based on deep learning to construct a health index calculation model and prediction model to reflect the before and after trends of the health status of the PCPs and realize the real-time, quantitative, and accurate evaluation and prediction of the health status of the wells.

2. Establishment of the HI Model

The health index calculation model is the basis for the analysis and prediction of the production performance of PCP wells. There are many parameters collected in CBM wells, and parameters that have an important impact on the health of PCP wells need to be selected as the principal parameters to form the input variables of the HI calculation model.

2.1. Principal Parameter Analysis. There are many parameters collected in CBM wells. However, some of these parameters have the same change trend, and these parameters show a strong correlation. There are also some parameters that cannot characterize whether the PCP fails or the influence of these parameters is small. Therefore, it is necessary to optimize the principal parameters before predicting the failure of the PCP. In this study, Pearson's correlation coefficient method was used for correlation analysis, and the principal component analysis method was used for principal parameter selection.

2.1.1. Pearson's Correlation Coefficient. Pearson's correlation coefficient is also called Pearson's product-moment correlation coefficient; it is a linear correlation coefficient, denoted as γ, used to reflect the degree of linear correlation between two variables X and Y. The value of γ is between -1 and 1; the larger the absolute value, the stronger the correlation. The calculation formula of γ is

$$\gamma = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(\left(X_i - \bar{X}\right)\right)^2}\sqrt{\sum_{i=1}^{n}\left(\left(Y_i - \bar{Y}\right)\right)^2}}, \quad (1)$$

where n is the number of samples. i is the serial number of the sample point.

The relationship between Pearson's correlation coefficient and the degree of correlation is shown in Table 1.

In this paper, it is stipulated that the correlation coefficient between the production parameters of PCP wells is extremely strong when the correlation coefficient γ is greater than 0.9.

2.1.2. Principal Component Analysis. The principal component analysis (PCA) is a statistical analysis method that reduces the original multiple variables to a few comprehensive indicators. From a mathematical point of view, this is a dimensionality reduction processing technology. There are many parameters automatically collected in PCP wells. Too many inputs will increase the difficulty and complexity of analyzing this problem. Therefore, this paper made use of the correlation between various factors to replace the original multiple influencing factors with the principal components after dimensionality reduction.

The input and output are as follows:

Input: n'-dimensional sample set $D' = (x^{(1)}, x^{(2)}, \cdots, x^{(n')})$, to be reduced to n-dimensional (where $x^{(i)}$ represents each parameter, $i = 1, 2, \cdots, n$).

Output: the sample set D after dimensionality reduction.

The process of dimensionality reduction algorithm is as follows:

(1) Centralize all samples: $x^{(i)} = x^{(i)} - (1/n)\sum_{j=1}^{n} x^{(j)}$

(2) Calculate the covariance matrix of the sample XX^T

Table 1: Correlation degree.

Pearson's correlation coefficient	Correlation
0.8–1.0	Extremely strongly correlated
0.6–0.8	Strongly correlated
0.4–0.6	Moderately correlated
0.2–0.4	Weakly correlated
0.0–0.2	Very weakly correlated or not correlated

(3) Perform eigenvalue decomposition on the covariance matrix XX^T, and the eigenvalue result is $W = (w_1, w_2, \cdots, w_n)$

(4) Calculate the weight of each parameter, and the calculation formula is $\omega_i = (w_i / \sum_{i=1}^{n} w_i) \times 100\%$. The weight result is $\Omega = (\omega_1, \omega_2, \cdots, \omega_n)$

(5) Set the threshold of the principal parameter. Add the weights of each parameter from large to small. When the weight sum is greater than 95%, it is considered that these parameters can characterize all the features, and the remaining parameters are removed

2.2. Health Index Calculation. The health index is a comprehensive indicator reflecting the health status of the PCP wells. Through data preprocessing and principal parameter optimization of the original data, n parameters are selected as the principal parameters for predicting the health status of the PCP wells. First, the principal parameters of all the failure wells are combined, and the PCA method is used to calculate the covariance matrix A of n principal parameters. Diagonalize the covariance matrix to obtain the eigenvalue of the covariance matrix, which is the weight of each principal parameter. Multiply the weights of the n principal parameters and add them together to obtain a comprehensive index that can reflect the health of the PCP, then normalize it to obtain the health index.

Table 2: Hypothetical dataset.

Time (t)	Principal parameter 1	Principal parameter 2	⋯	Principal parameter n
0	x_{10}	x_{20}	⋯	x_{n0}
1	x_{11}	x_{21}	⋯	x_{n1}
⋮	⋮	⋮	⋮	⋮
$m-1$	$x_{1,m-1}$	$x_{2,m-1}$	⋯	$x_{n,m-1}$

Assume that the hypothetical dataset is shown in Table 2.

The PCA method uses variance to measure the amount of information, and the sample set is $\{X = [X_{1j}, X_{2j}, \cdots, X_{nj}]^T\ 0 \le j \le m-1\}$, where n is the number of principal parameters, m is $t = m$ at a certain time, and $X_{nj} = [x_{n1}, x_{n2}, \cdots, x_{nj}]$. All samples are constructed into an $n \times m$ matrix, which is the covariance matrix. Let the covariance matrix be A; then,

$$A = \frac{1}{m-1}\begin{bmatrix} \sum_{j=0}^{m-1}(x_{1j}-\bar{x}_1)(x_{1j}-\bar{x}_1) & \sum_{j=0}^{m-1}(x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) & \cdots & \sum_{j=0}^{m-1}(x_{1j}-\bar{x}_1)(x_{nj}-\bar{x}_n) \\ \sum_{j=0}^{m-1}(x_{2j}-\bar{x}_2)(x_{1j}-\bar{x}_1) & \sum_{j=1}^{m}(x_{2j}-\bar{x}_2)(x_{2j}-\bar{x}_2) & \cdots & \sum_{j=0}^{m-1}(x_{2j}-\bar{x}_2)(x_{nj}-\bar{x}_n) \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=0}^{m-1}(x_{nj}-\bar{x}_n)(x_{1j}-\bar{x}_1) & \sum_{j=0}^{m-1}(x_{nj}-\bar{x}_n)(x_{2j}-\bar{x}_2) & \cdots & \sum_{j=0}^{m-1}(x_{nj}-\bar{x}_n)(x_{nj}-\bar{x}_n) \end{bmatrix}, \tag{2}$$

where x_{nj} is the sample attribute value corresponding to the n th principal parameter in the dataset at $t = j$. $\bar{x}_n$ is the average value of all attribute values of the principal parameter n, where $\bar{x}_n = (\sum_{j=0}^{m-1} x_{nj})/m$.

Let the set of eigenvectors of matrix A be υ, and the eigenvalue corresponding to υ is $\lambda_i (i = 1, 2, \cdots, n)$, so the relationship between the matrix, eigenvalue, and eigenvector can be obtained as

$$A\upsilon = \lambda_i \upsilon. \tag{3}$$

Construct the eigenvalue formula for solving the eigenmatrix:

$$|\lambda_i E - A| = 0, \tag{4}$$

where E is the identity matrix.

The principal parameter value input at a certain time t is $X_t = (x_{1t}, x_{2t}, \cdots, x_{nt})$; the calculation formula of the composite index (CI_t) is

$$\mathrm{CI}_t = \lambda {X_t}^T, \tag{5}$$

where λ is the eigenvalue vector composed of eigenvalues of matrix A, where $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_n)$.

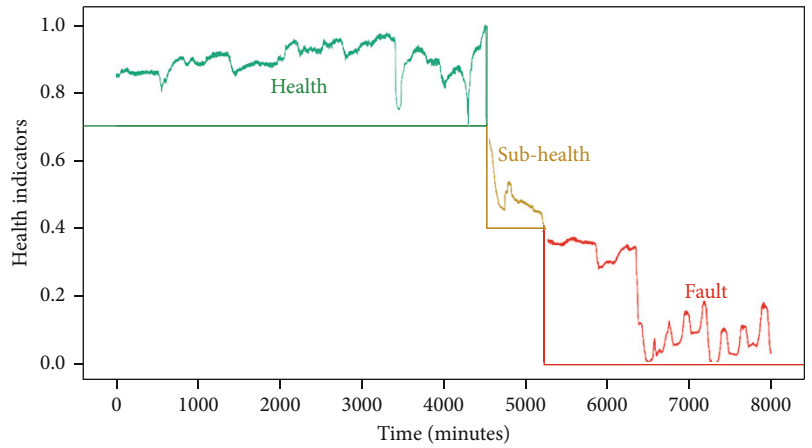

Figure 1: Change curve of the health status.

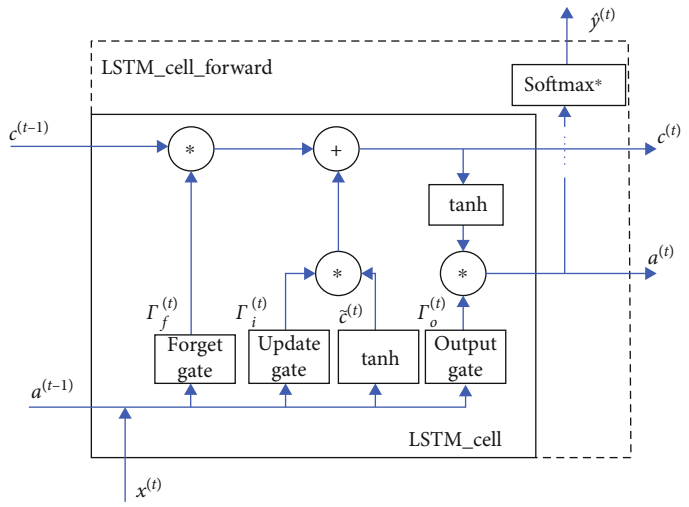

Figure 2: Cell structure of LSTM.

The comprehensive index at each moment in the ΔT period is calculated as

$$\mathrm{CI} = (\mathrm{CI}_0, \mathrm{CI}_1, \cdots, \mathrm{CI}_{\Delta T}). \tag{6}$$

Normalize the obtained comprehensive index to obtain the health index (HI). The formula for calculating the health index at time t is

$$\mathrm{HI}_t = \frac{\mathrm{CI}_t - \mathrm{CI}_{\min}}{\mathrm{CI}_{\max} - \mathrm{CI}_{\min}}. \tag{7}$$

So the health index at each moment in the ΔT period is

$$\mathrm{HI} = (\mathrm{HI}_0, \mathrm{HI}_1, \cdots, \mathrm{HI}_{\Delta T}). \tag{8}$$

2.3. Health Degree Division. The health index will show different trends with the severity of the PCP failure. Before predicting the health status, the health status should be divided into different degrees according to the change trend of the health index, namely, health, subhealth, and fault, as shown in Figure 1. According to existing data, the range of the health index of health, subhealth, and failure of all sample wells is calculated. According to the statistics of the HI scope of all sample wells, the threshold of HI is obtained as the basis for the failure alarm.

The healthy state is the normal operation state of the pump unit, the HI value is close to 1 with little fluctuation, and the production of the CBM well is stable. The pump is running in a subhealthy state due to gas interference, stator swelling, wear, and leakage; the HI value gradually decreases with time; and the gas production continues to decrease. Owing to the gas locking, shaft broken, and serious leakage of the pump, the pump unit runs under the fault condition, the HI value is close to 0, and almost no gas is produced. When different faults occur, the drop rate of HI is different. For example, when the sucker rod is broken, HI will instantly fall to 0, and when the pump is running dry, HI will slowly decrease.

3. HI Prediction Model

Aiming at the characteristics of CBM well production data and the degree of PCP changes over time, a long short-term

Table 3: Evaluation index of the regression model.

Evaluation index	Calculation formula	Criteria
Mean absolute percentage error	$\mathrm{MAPE} = \sum_{i=1}^{n}\left\lvert\frac{y_t - y_{\mathrm{pre}}}{y_t}\right\rvert \times \frac{100}{n}$	The smaller the MAPE, the smaller the error
Mean absolute error	$\mathrm{MAD} = \frac{\sum_{i=1}^{n}\left\lvert y_t - y_{\mathrm{pre}}\right\rvert}{n}$	The smaller the MAD, the smaller the error
Root mean square error	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_t - y_{\mathrm{pre}}\right)^2}$	The smaller the RMSE, the smaller the error, and the larger the RMSE, the larger the error
Theil's inequality coefficient	$\mathrm{TIC} = \frac{\sqrt{(1/n)\sum_{i=1}^{n}\left(y_t - y_{\mathrm{pre}}\right)^2}}{\sqrt{(1/n)\sum_{i=1}^{n}y_t^2} + \sqrt{(1/n)\sum_{i=1}^{n}y_{\mathrm{pre}}^2}}$	The closer the TIC value is to zero, the higher the prediction accuracy will be; when it is equal to zero, it means 100% fitting
Decisive factor	$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_t - y_{\mathrm{pre}}\right)^2}{\sum_{i=1}^{n}\left(y_t - \bar{y}_t\right)^2}$	R^2 is between 0 and 1; the larger the value, the better the model fitting

memory (LSTM) neural network is selected to establish a HI deep learning model. LSTM is based on the general recurrent neural network (RNN) [21], adding memory units to the neural units of each hidden layer to achieve controllable memory information in time series. It is suitable for processing and predicting important events with relatively long intervals and delays in time series. LSTM is generally an artificial intelligence prediction algorithm based on deep learning.

3.1. Principle of LSTM. In order to solve the problem of vanishing gradient and maintain the long-term memory of the hidden layer, the long short-term memory (LSTM) network is improved on the basis of RNN [22]. LSTM uses three "gating" structures to control the state and output at different moments. The short-term memory and long-term memory are combined through the "gating" structure, which can alleviate the problem of gradient disappearance. The expansion of the LSTM structure is exactly the same as RNN in time; the difference lies in the difference of the cell. The cell calculation node of LSTM contains more structures, including update gates, forget gates, and output gates. As shown in Figure 2, the calculation formulas are as follows:

$$\begin{aligned}
\Gamma_f^{\langle t\rangle} &= \sigma\left(W_f\left[a^{\langle t\boxtimes 1\rangle}, X^{\langle t\rangle}\right] + b_f\right),\\
\tilde{c}_{\langle t\rangle} &= \tanh\left(W_c\left[a^{\langle t\boxtimes 1\rangle}, X^{\langle t\rangle}\right] + bc\right),\\
\Gamma_i^{\langle t\rangle} &= \sigma\left(W_i\left[a^{\langle t\boxtimes 1\rangle}, X^{\langle t\rangle}\right] + b_i\right),\\
c^{\langle t\rangle} &= \Gamma_f^{\langle t\rangle} * c^{\langle t-1\rangle} + \Gamma_i^{\langle t\rangle} * \tilde{c}_{\langle t\rangle},\\
\Gamma_o^{\langle t\rangle} &= \sigma\left(W_o\left[a^{\langle t\boxtimes 1\rangle}, X^{\langle t\rangle}\right] + b_o\right),\\
a^{\langle t\rangle} &= \Gamma_o^{\langle t\rangle} * \tanh\left(c^{\langle t\rangle}\right).
\end{aligned} \tag{9}$$

Among them, $\Gamma_f^{\langle t\rangle}$ represents the forget gate. If the value of a cell in the forget gate is close to 0, LSTM will "forget" the storage state in the corresponding cell of the previous cell state. If the value of a cell in the forget gate is close to 1, LSTM will mainly remember the corresponding value in the storage state; $\tilde{c}_{\langle t\rangle}$ represents a candidate value, which is a tensor containing information that may be stored in the cell state at the current time; $\Gamma_i^{\langle t\rangle}$ represents the update gate, which is used to determine which information of candidate value $\tilde{c}_{\langle t\rangle}$ is added to $c^{\langle t\rangle}$; $c^{\langle t\rangle}$ is the record of the current cell state information, and the information is used for transmission in subsequent time steps; the output gate $\Gamma_o^{\langle t\rangle}$ determines which information is used for the prediction of the current time step; $a^{\langle t\rangle}$ contains the current hidden node information, which is used to pass to the next time step to calculate the value of each gate and for label prediction calculation.

By introducing a gating mechanism into the computing nodes of the hidden layer, LSTM naturally overcomes the problem of gradient disappearance in the structure and has more parameters to control the model. By four times the parameter amount of RNN, time series variables can be predicted more finely. The prediction of the equipment health index is a long-term time series information processing process. Therefore, this paper chooses LSTM as the prediction model of HI.

3.2. Steps for HI Prediction. The methods of HI model establishment, training, and verification are as follows:

(1) Call the interface to create the model and set the initial parameters. Call the interface on TensorFlow to create an LSTM model; set the number of neural network layers, time series steps, number of neurons, number of training cycles (epochs), batch size, and other hyperparameters; and set the activation function and optimization function

(2) According to the model structure, establish a training set and a test set. According to the LSTM model

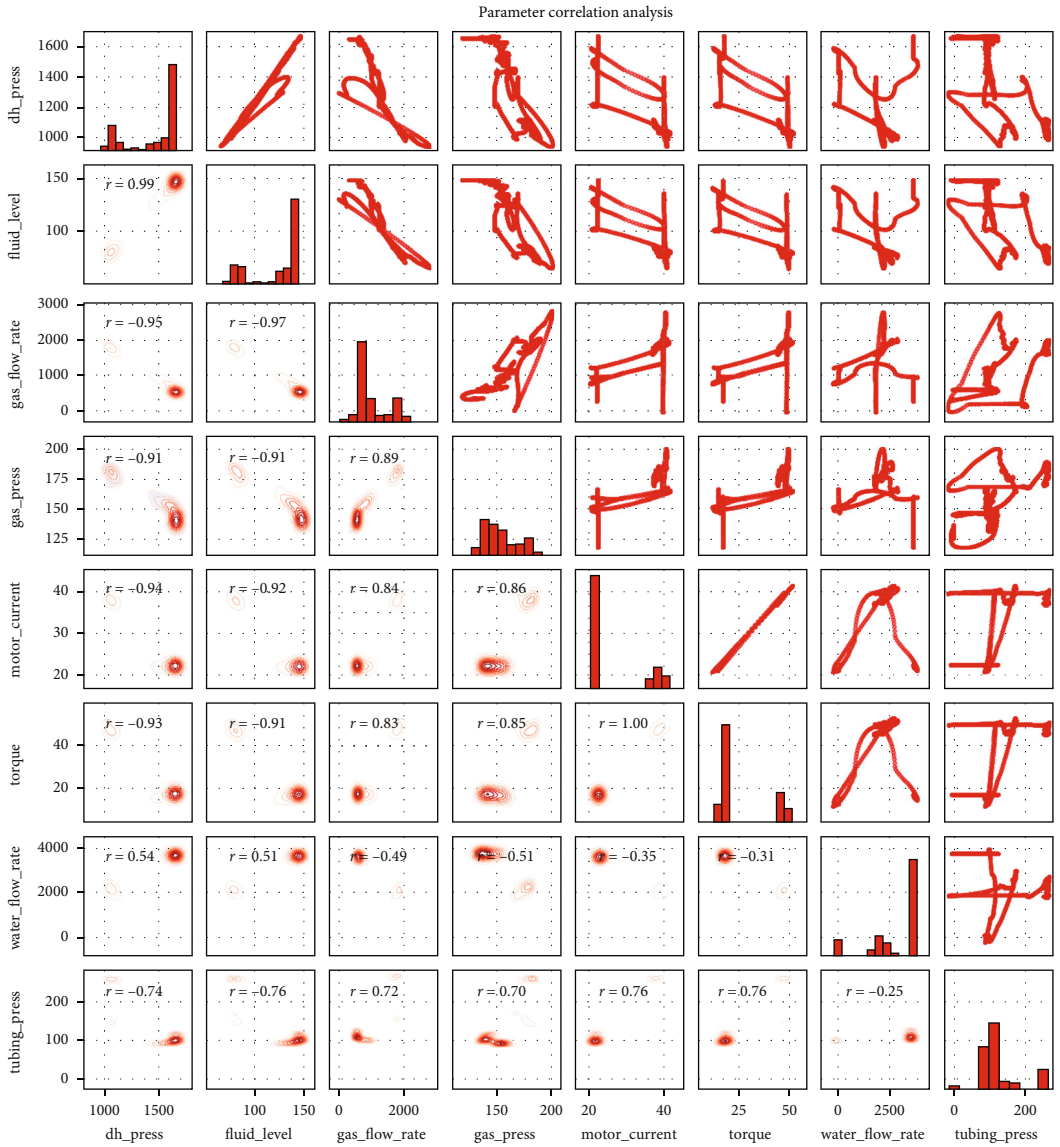

Figure 3: Parameter correlation distribution graph.

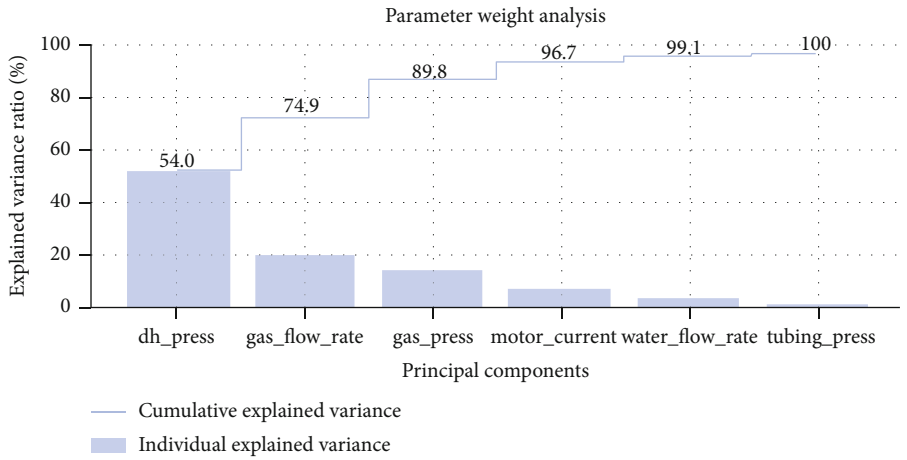

Figure 4: Parameter weight distribution graph.

Table 4: Principal parameters of 30 failure wells.

Well_ID	Principal parameter 1	Principal parameter 2	Principal parameter 3	Principal parameter 4
E001	dh_press	gas_flow_rate	gas_press	motor_current
E002	dh_press	gas_flow_rate	gas_press	tubing_press
E003	dh_press	gas_flow_rate	gas_press	tubing_press
E004	dh_press	tubing_press	—	—
E005	dh_press	gas_flow_rate	gas_press	tubing_press
E006	dh_press	gas_flow_rate	gas_press	motor_current
E007	gas_flow_rate	pump_speed	water_flow_rate	tubing_press
E008	dh_press	gas_press	motor_current	pump_speed
E009	dh_press	gas_flow_rate	gas_press	pump_speed
E010	dh_press	gas_flow_rate	gas_press	—
E011	dh_press	gas_flow_rate	gas_press	pump_speed
E012	dh_press	gas_flow_rate	gas_press	motor_current
E013	dh_press	gas_press	water_flow_rate	tubing_press
E014	gas_flow_rate	gas_press	pump_speed	tubing_press
E015	gas_flow_rate	gas_pres	tubing_press	—
E016	gas_flow_rate	gas_press	pump_speed	tubing_press
E017	gas_flow_rate	gas_press	motor_current	tubing_press
E018	gas_flow_rate	gas_press	motor_current	water_flow_rate
E019	dh_press	gas_flow_rate	gas_press	motor_current
E020	gas_flow_rate	gas_press	motor_current	water_flow_rate
E021	gas_press	tubing_press	dh_press	gas_flow_rate
E022	dh_press	gas_flow_rate	gas_press	pump_speed
E023	gas_flow_rate	tubing_press	pump_speed	—
E024	dh_press	gas_press	pump_speed	tubing_press
E025	gas_flow_rate	pump_speed	water_flow_rate	tubing_press
E026	gas_flow_rate	gas_press	motor_curren	tubing_press
E027	pump_speed	dh_press	gas_flow_rate	motor_curren
E028	gas_flow_rate	gas_press	torque	tubing_press
E029	dh_press	gas_flow_rate	gas_press	motor_current
E030	dh_press	pump_speed	water_flow_rate	tubing_press

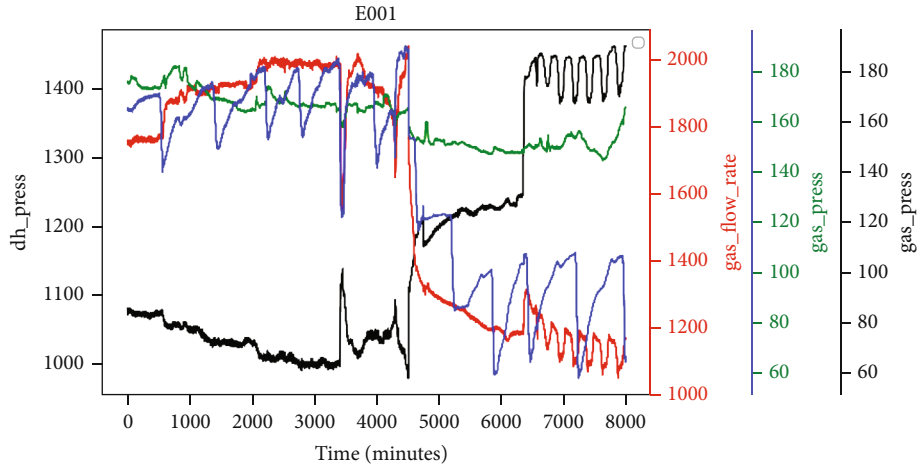

Figure 5: Four principal parameters of E001.

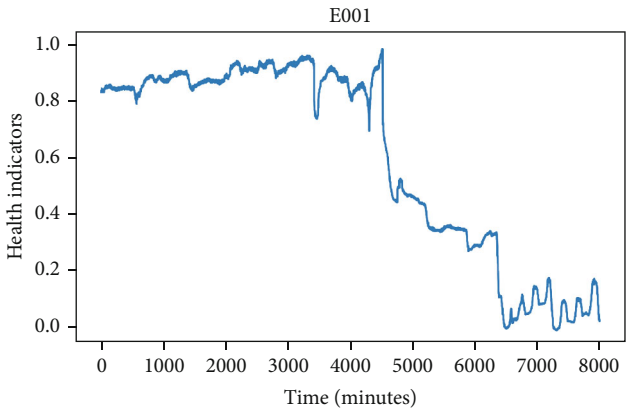

Figure 6: HI of E001.

structure, extract the PCP health index data, set up the time series, segment the sample data according to the set input time length, and create the HI training set and test set. And 80% of the data of each sample well is used for model training, and 20% of the data is used for model testing

(3) For model tuning, through the grid search method, the optimal hyperparameters, activation function, and optimization function of the model are selected

(4) For model training, use the tanh function as the activation function and the Adam function as the optimization function

(5) To verify the model, when evaluating the effect of the health index prediction model, set the prediction total of the model to n, the prediction value to y_{pre}, and the true value to y_t. The regression model evaluation indicators in Table 3 can be used to evaluate the accuracy of the model

(6) For model release, use the test set data to evaluate the prediction accuracy of the HI machine learning model. When the accuracy of the prediction result meets the requirements, the model training is completed and released as a formal prediction model

4. Application Results

The real-time production data of 30 PCP failure wells and 6 wells under normal conditions in a coalbed methane block in Australia's Surat Basin were collected. These data include downhole pressure, fluid level, gas production, water production, current, voltage, torque, tubing pressure, casing pressure, and pump speed. The data acquisition interval is one minute. The failure types of the collected failure wells include 6 types of failure, such as pump ran dry, tubing plugged, stator plugged, tubing broken, connection broken, and pump lost efficiency. The following mainly takes well E001 as an example to perform production characteristic analysis, health index calculation, early warning of failure.

4.1. Production Characteristic Analysis of PCP Wells. In the data preprocessing, the original data was deleted (removed noise points) and replaced (missing value processing), and the 10 parameters collected by the CBM well were processed into 8 items. Pearson's correlation coefficient analysis of these 8 parameters is shown in Figure 3.

The lower triangle r in the figure represents the correlation coefficient between the two parameters corresponding to the horizontal and vertical coordinates. A positive number indicates a positive correlation between the parameters, and the larger the positive number, the stronger the positive correlation. A negative number indicates a negative correlation between the parameters, and the smaller the negative number, the stronger the negative correlation. The upper triangle represents the corresponding correlation between the two parameters. The closer the slope of the line is to 1, the stronger the positive correlation between the two parameters; the closer the slope of the line is to -1, the stronger the negative correlation between the two parameters.

It is defined that the correlation between the two parameters is greater than 0.9, showing a strong correlation. It can be seen from Figure 3 that the correlation coefficient between downhole pressure (dh_press) and fluid level (fluid_level) is 0.99, and the correlation coefficient between current (motor_current) and torque (torque) is 1. Therefore, one of the downhole pressure and fluid level and current and torque can be deleted.

The principal component analysis is performed on the parameters screened by Pearson's correlation coefficient, and the weight analysis chart shown in Figure 4 is obtained. Each histogram in the figure represents the weight of each parameter, and the line graph represents the sum of the weight of each parameter. It is defined that when the sum of the weights of the parameters is greater than 95%, the parameters obtained can fully represent the characteristics of all parameters.

In this study, it can be seen from Figure 4 that when the first four parameters of downhole pressure (dh_press), gas production (gas_flow_rate), casing pressure (gas_press), and current (motor_current) are selected, the cumulative weight is greater than 95%. Therefore, these four parameters are selected as the principal parameters in well E001.

In order to make the obtained principal parameters adapt to the entire failure wells, the principal parameters of 30 failure wells are statistically analyzed, as shown in Table 4.

From the analysis of Table 4, downhole pressure (dh_press), gas production (gas_flow_rate), casing pressure (gas_press), and tubing pressure (tubing_press) are ranked in the top four for the most cumulative times in all cases. Therefore, these four parameters are selected as the principal parameters.

4.2. HI Analysis Results. Using the collected 4 principal parameters of 30 failure wells, a 4×30 sample matrix is

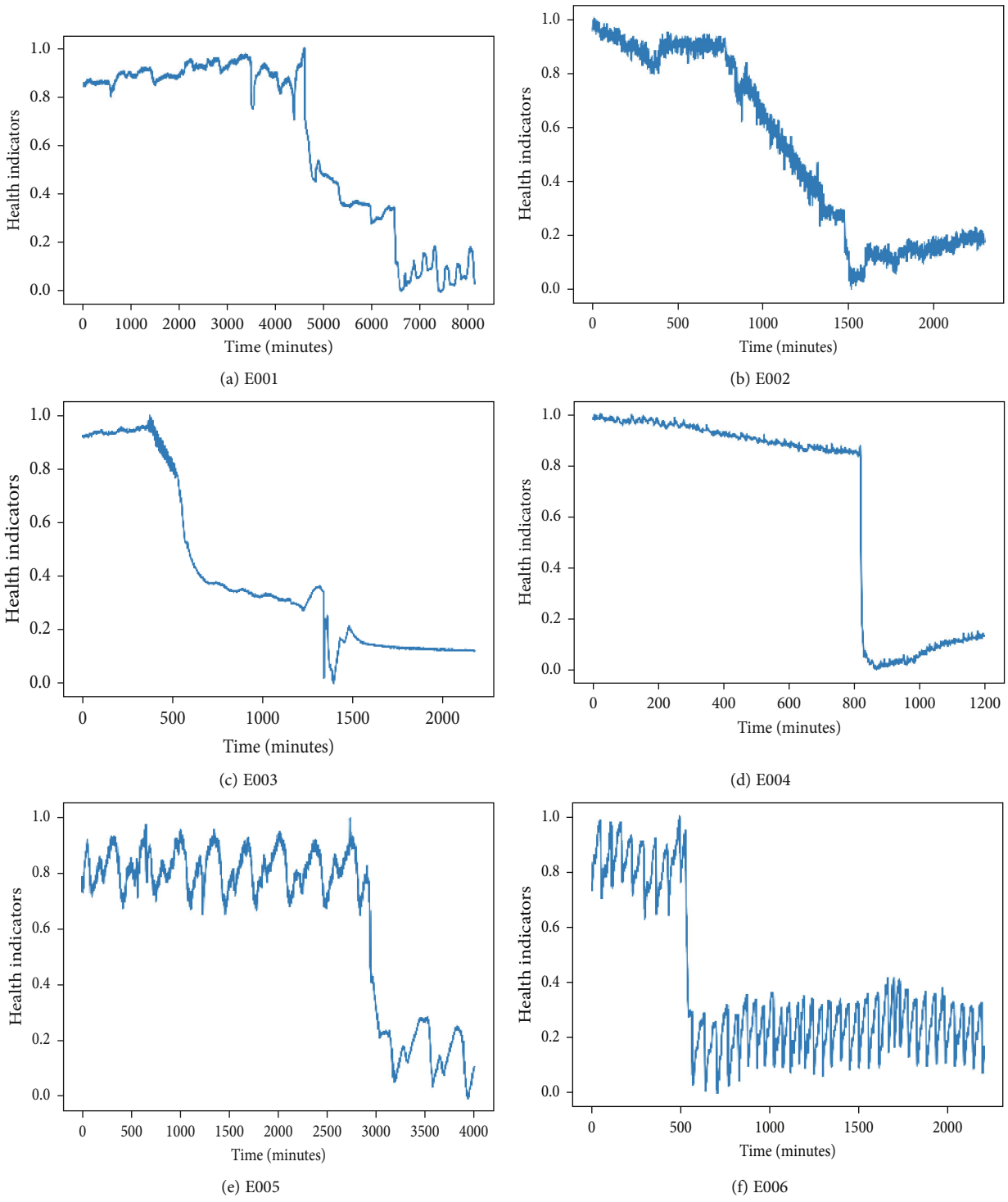

Figure 7: HI curves of 6 failure wells.

Table 5: HI range of 30 failure wells.

Well_ ID	HI range in normal condition	HI range in failure condition	Well_ ID	HI range in normal condition	HI range in failure condition
E001	0.8-1.0	0-0.2	E016	0.8-1.0	0-0.4
E002	0.7-1.0	0-0.3	E017	0.8-1	0-0.1
E003	0.8-0.1	0-0.2	E018	0.8-1	0-0.2
E004	0.8-1.0	0-0.2	E019	0.6-1.0	0-0.3
E005	0.7-1.0	0-0.3	E020	0.9-1.0	0.7-0.8
E006	0.7-1	0-0.2	E021	0.7-1	0-0.4
E007	0.8-1.0	0-0.2	E022	0.7-1	0-0.2
E008	0.7-1.0	0-0.3	E023	0.6-1.0	0-0.3
E009	0.7-1.0	0-0.3	E024	0.7-1.0	0-0.3
E010	0.8-1	0.5-0.7	E025	0.8-0.1	0-0.4
E011	0.8-1.0	0-0.2	E026	0.8-1	0-0.2
E012	0.8-1.0	0-0.2	E027	0.7-0.8	0.6-0.7
E013	0.7-1	0.4-0.6	E028	0.7-1	0-0.5
E014	0.8-1	0.1-0.3	E029	0.8-1.0	0-0.2
E015	0.7-1	0-0.2	E030	0.7-1	0.5-0.7

established, substituting equation (3), and the covariance matrix A applicable to the entire block is calculated as

$$A = \begin{bmatrix} 1.00 & -0.88 & -0.83 & -0.03 \\ -0.88 & 1.00 & 0.76 & 0.06 \\ -0.83 & 0.76 & 1.00 & -0.24 \\ -0.03 & 0.06 & -0.24 & 1.00 \end{bmatrix}. \quad (10)$$

From equations (4) and (6), the eigenvalue vector composed of the eigenvalues of matrix A can be obtained:

$$\lambda = (2.65, 1.06, 0.10, 0.18). \quad (11)$$

The data of 4000 points before and after the failure of well E001 is selected for health index analysis, and the principal parameters change with time, as shown in Figure 5. From formulas (5)~(8), the health index before and after the event of the Event_001 well can be calculated. As shown in Figure 6, the health index fluctuated between 0.7 and 1.0 before the failure, and the health index began to decline when the failure occurred, until the lowest value, fluctuating between 0 and 0.2.

The health index was calculated for 30 failure wells and 6 normal wells. Figure 7 shows the health index of 6 failure wells. It can be concluded that when the PCP is operating normally, the health index is between 0.7 and 1. When a failure occurs, the health index will gradually decrease. Therefore, it can accurately reflect the health status of the PCP operation.

Table 5 shows the range of health index variation of 30 failure wells. It shows that the health index of most wells under normal operating conditions is between 0.7 and 1, and the health index under failure conditions is between 0 and 0.4. When the health index is 0.7-1, the PCP is healthy; when the health index is 0.4-0.7, the PCP is subhealthy; when the health index is 0-0.4, the PCP is faulty. Therefore, when the health index is lower than 0.7, a failure warning will be sent and when it is lower than 0.4, a severe warning will be sent.

4.3. HI Analysis of Failure Types. The change trend of the health index curve under different types of working conditions is different. The following three working conditions of normal, tubing broken, and pump ran dry are taken as examples for analysis. The health index is shown in Figures 8–10.

It can be seen from Figure 8 that when the well is operating normally, the health index fluctuates between 0.8 and 1.0, which meets the scope of health index classification; Figures 9 and 10 show the HI curves of two different failures at the same time period.

Both mechanism analysis and data analysis have confirmed that tubing broken occurs in an instant, the process is fast, and the change in the health index appears to be a sudden drop; pump ran dry is a slow occurrence, the process is relatively longer, and the change in the health index appears to be a slow decline. This study counts the approximate time required for all wells from the beginning of the failure to the end of the failure according to different types of failures, as shown in Table 6.

Table 6 demonstrates that the health index not only can accurately represent the real-time health status of the PCP wells but also can be used for fault diagnosis.

When the pump is running dry, the time period from the beginning to the complete failure is greater than 3000 minutes, and the time period from the beginning to the complete failure of other failures is less than 3000 minutes. Therefore, the severity of the fault can be judged by analyzing the slowness of the change of the health index curve. If the health index drops suddenly, it can be concluded that this type of fault is a serious fault; if the health index drops slowly, it is a slight failure.

4.4. Early Warning of Failure. First, initialize the LSTM neural network parameters randomly and set the number of neural network layers to 2, the time step to 200 minutes, the number of neurons to 8, the number of training cycles (epochs) to 8, and the batch size to 8.

Then, use the training data for model training. After the model training is completed, the grid search and the learning curve are drawn on the validation set to obtain the optimal network structure parameters of the LSTM model: epochs = 10, batch_size = 256, and time_step = 200.

The number of neurons in the first layer is 64; the number of neurons in the second layer is 16. The change process of the loss function with the training times during the training process is shown in Figure 11.

It can be seen from Figure 11 that the loss function of the model gradually decreases and tends to zero as the number of training increases. It shows that the LSTM prediction model has no overfitting or underfitting, and the model has good

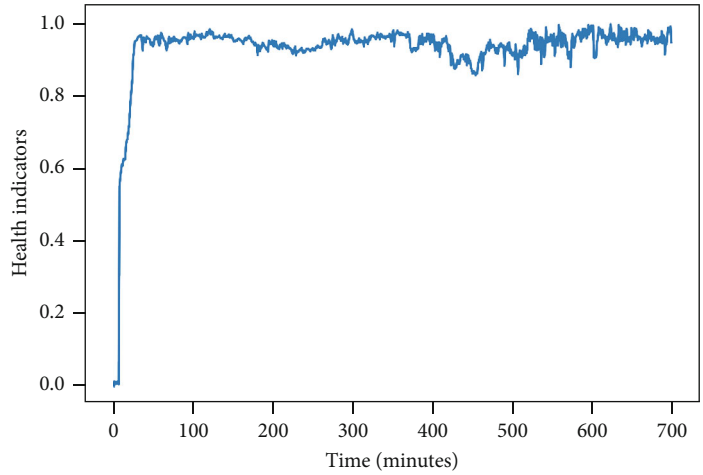

Figure 8: HI curve of the normal well.

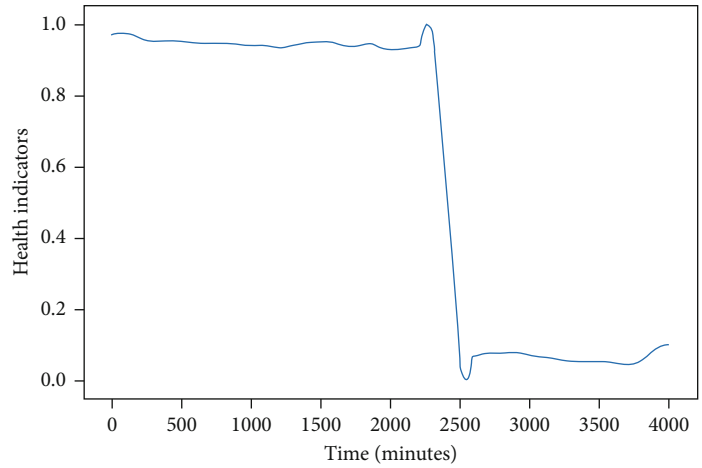

Figure 9: HI curve of tubing broken.

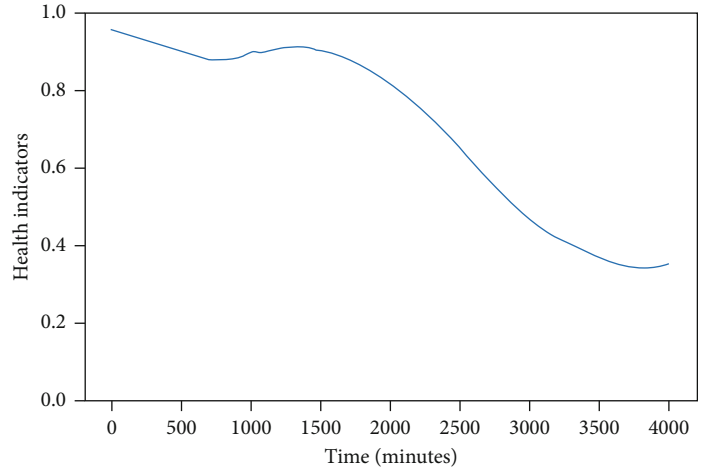

Figure 10: HI curve of pump ran dry.

Table 6: Statistics of the failure time range of different failure types.

Failure type	Well_ ID	Failure time range (minutes)	Well_ ID	Failure time range (minutes)	Well_ ID	Failure time range (minutes)
Tubing broken	E003	1500	E005	1000	E006	1000
	E009	1000	E011	1200	E012	100
	E013	1200	E016	1200	E023	1400
	E024	1200	E025	1400	—	—
Tubing plugged	E001	2200	E010	1300	E027	2200
	E002	3000	E015	2000	E018	1000
Stator plugged	E019	1200	E020	3000	E022	1200
	E026	1200	E028	3200	E029	2500
	E030	1400	—	—	—	—
Pump ran dry	E007	3200	E021	4000	E026	6000
Pump lost efficiency	E008	1700	E014	5000	E014	2200

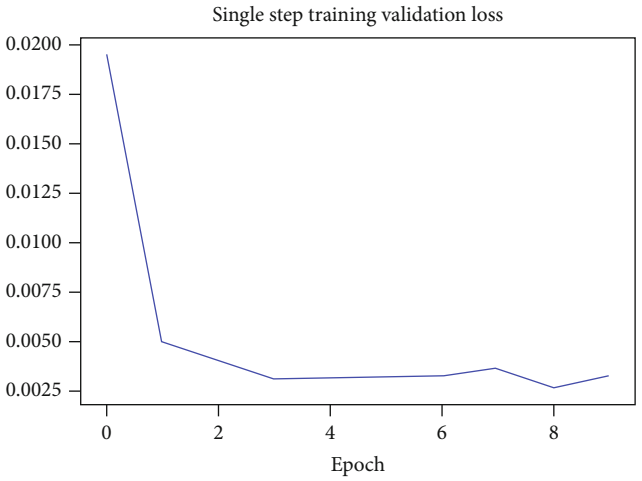

Figure 11: Change curve of training set loss with training times.

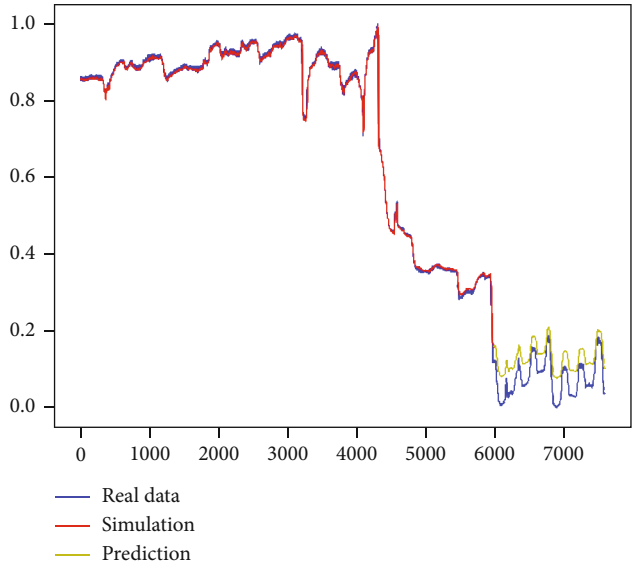

Figure 12: Curve of predicted data versus real data.

Table 7: Evaluation result of the LSTM model.

	MAPE	MAD	RMSE	TIC	R^2
Training set	0.6	0.005	0.013	0.007	0.98
Testing set	32.8	0.009	0.012	0.024	0.98

generalization ability and can be used for pumping well power prediction.

Figure 12 shows the training and prediction effects of the LSTM model.

Table 7 shows the LSTM model evaluation results based on the model evaluation method.

It can be seen that the average percentage error MAPE of the model on the training set and test set is 0.6 and 32.8, respectively; the average absolute error MAD, root mean square error RMSE, and Theil's inequality coefficient TIC are all close to 0; and the evaluation coefficient R^2 is 0.98 on both the test set and the training set, which is close to 1.

Therefore, the LSTM prediction model accurately grasps the trend of the health index change and the correlation before and after and can accurately predict the health status of the PCP wells in real time.

5. Conclusions

This study proposed an artificial intelligence-based method for evaluating and predicting the health status of PCP in CBM wells and established a five-step method for failure prediction: data preprocessing, optimization of principal parameters, health index construction, health degree division, and health index prediction.

(1) Through data preprocessing and optimization of principal parameters for 10 production parameters of PCP wells, four principal parameters that are strongly related to the health status of the wells are determined, and a comprehensive index (health index) is constructed. According to the statistics of

the HI scope of all sample wells, the health status of PCP wells is characterized by degrees: health (0.7-1.0), subhealth (0.4-0.7), and failure (0-0.4).

(2) Use the long short-term memory (LSTM) neural network to train the sample set to obtain the machine learning model of the health index. This model can accurately predict the health status of PCP wells in real time and can realize early warning of well failures

(3) The health index model and LSTM prediction model in this study can reflect the health status of PCP wells timely and can realize early warning of failure, quantitative evaluation, and accurate prediction of the health status of the PCP in CBM wells

Data Availability

The CSV data used to support the findings of this study are currently under embargo while the research findings are commercialized. Requests for data (6/12 months) after the publication of this article will be considered by the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Y. Kang, Z. Li, J. Wang, L. Sun, and J. Gu, "Coalbed methane mobility and primary drainage rate control strategy in different coahrank blocks," *Acta Petrolei Sinica*, vol. 39, no. 10, pp. 1162–1174, 2018.

[2] D. Jia, Y. Qiu, C. Li, and Y. Cai, "Propagation of pressure drop in coalbed methane reservoir during drainage stage," *Advances in Geo-Energy Research*, vol. 3, no. 4, pp. 387–395, 2019.

[3] S. Liu, S. Tang, and S. Yin, "Coalbed methane recovery from multilateral horizontal wells in Southern Qinshui Basin," *Advances in Geo-Energy Research*, vol. 2, no. 1, pp. 34–42, 2018.

[4] F. Saghir, M. E. Gonzalez Perdomo, and P. Behrenbruch, "Machine learning for progressive cavity pump performance analysis: a coal seam gas case study," in *Proceedings of the SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference*, pp. 356–365, Brisbane, QLD, Australia, 2019, November 15.

[5] J. P. Hoday, M. Knafl, C. Prosper, and M. Braas, "Diagnosing PCP failure characteristics using exception based surveillance in CSG," in *Paper presented at the SPE Progressing Cavity Pumps Conference*, pp. 89–101, Calgary, Alberta, Canada, 2013, August 25.

[6] F. Saghir, M. E. Gonzalez Perdomo, and P. Behrenbruch, "Converting time series data into images: an innovative approach to detect abnormal behavior of progressive cavity pumps deployed in coal seam gas wells," in *Paper presented at the SPE Annual Technical Conference and Exhibition*, pp. 269–274, Calgary, Alberta, Canada, 2019, September 23.

[7] C. Prosper and D. West, "Case study applied machine learning to optimise PCP completion design in a CBM field," in *Paper presented at the SPE Asia Pacific Oil and Gas Conference and Exhibition*, pp. 177–187, Brisbane, Australia, 2018, October 19.

[8] R. Hicks, "New technology reduces Flushbys & extends PCP run life in Australia," in *Paper presented at the SPE Middle East Artificial Lift Conference and Exhibition*, pp. 436–439, Manama, Kingdom of Bahrain, 2016, November 30.

[9] D. Caballeroa, Y. Hurtado, A. Gomez, and L. Zimmer, "PCP run life improvement in Orinoco Belt with new PCP technology," in *Paper presented at the SPE Latin America and Caribbean Petroleum Engineering Conference*, pp. 637–647, Maracaibo, Venezuela, 2014, May 21.

[10] C. Tan, Z. Cai, H. Deng et al., "Intelligent decision making on PCP production parameters of CBM wells based on reinforcement learning," *Oil Drilling & Production*, vol. 42, no. 1, pp. 62–69, 2020.

[11] R. Gouriveau, K. Medjaher, and N. Zerhouni, *From Prognostics and Health Systems Management to Predictive Maintenance 1: Monitoring and Prognostics*, John Wiley and Sons Ltd, 2016.

[12] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: a systematic review from data acquisition to RUL prediction," *Mechanical Systems and Signal Processing*, vol. 104, no. MAY1, pp. 799–834, 2018.

[13] A. Widodo and B.-S. Yang, "Application of relevance vector machine and survival probability to machine degradation assessment," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2592–2599, 2011.

[14] T. Wang, "Bearing life prediction based on vibration signals: a case study and lessons learned," in *2012 IEEE Conference on Prognostics and Health Management*, pp. 1–7, Denver, CO, 2012.

[15] Y. Wang, Y. Peng, Y. Zi, X. Jin, and K. L. Tsui, "A two-stage data-driven-based prognostic approach for bearing degradation problem," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 924–932, 2016.

[16] X. Jin, Y. Sun, Z. Que, Y. Wang, and T. W. S. Chow, "Anomaly detection and fault prognosis for bearings," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 9, pp. 2046–2054, 2016.

[17] H. Kumar, S. P. Pai, N. Sriram, N. S. Sriram, and G. S. Vijay, "Rolling element bearing fault diagnostics: development of health index," *ARCHIVE Proceedings of the Institution of Mechanical Engineers Part C Journal of Mechanical Engineering Science 1989-1996*, vol. 203-210, 2017.

[18] A. Giantomassi, F. Ferracuti, A. Benini, G. Ippoliti, S. Longhi, and A. Petrucci, "Hidden Markov model for health estimation and prognosis of turbofan engines," in *Proceedings of the ASME 2011 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Volume 3: 2011 ASME/IEEE International Conference on Mechatronic and Embedded Systems and Applications, Parts A and B*, pp. 1–9, Washington, DC, USA, 2011.

[19] H. Ocak, K. A. Loparo, and F. M. Discenzo, "Online tracking of bearing wear using wavelet packet decomposition and probabilistic modeling: a method for bearing prognostics," *Journal of Sound and Vibration*, vol. 302, no. 4-5, pp. 951–961, 2007.

[20] L. Guo, N. Li, F. Jia, Y. Lei, and J. Lin, "A recurrent neural network based health indicator for remaining useful life prediction of bearings," *Neurocomputing*, vol. 240, pp. 98–109, 2017.

[21] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.