scientific reports

OPEN

# Prediction of the corrosion rates of subsea pipelines via KPCA

Shiqi Xu[1,2]✉, Changhe Huang[1] & Hongda Zhao[3]

**The development of a precise model for predicting pipeline corrosion rates is essential for ensuring the safe operation of pipelines. To address the issues of inadequate stability and prolonged execution time associated with traditional models, the KPCA algorithm is used here to reduce the dimensionality of corrosion rate data for subsea pipelines, and the primary factors that influence the corrosion rate are identified. Based on the data characteristics, four algorithms (BP, LSSVM, SVM, and RF) were compared. Ultimately, the LSSVM algorithm was selected as the final prediction model. Then the LSSVM prediction model is subsequently developed, and the NGO algorithm is utilized to optimize the weights and thresholds of the LSSVM model, thereby increasing the accuracy of the prediction model and effectively reducing prediction instability. A combined KPCA-NGO-LSSVM model is developed to predict the corrosion rates of subsea pipelines and is compared with three other models: KPCA-PSO-LSSVM, PSO-LSSVM, and NGO-LSSVM. The mean absolute percentage error (MAPE), root mean square error (RMSE) and determination coefficient ($R^2$) of the integrated KPCA-NGO-LSSVM model are 1.791%, 0.06105 and 0.9922, respectively, these metrics are significantly lower than those of benchmark models, a finding consistently validated across multiple experimental datasets. This demonstrates the KPCA-NGO-LSSVM framework's enhanced prediction accuracy and stability. The model demonstrates effective performance in predicting the corrosion rates of subsea pipelines and offers novel methodologies and concepts for future research in this area.**

**Keywords** KPCA algorithm, NGO algorithm, LSSVM algorithm, Prediction accuracy, Prediction stability

Oil and natural gas occupy a significant position within China's energy strategy, which is of paramount importance to the country's national development. Pipeline transportation is one of the most prevalent modes of transportation for oil and gas and has evolved into the fifth largest transportation industry in China. However, when a pipeline is in operation, because of the environment and pipeline medium, pipeline corrosion occurs, which in turn leads to various accidents[1]. Corrosion is a significant factor contributing to pipeline failure, as evidenced by relevant statistical data. Increased pipeline operation time results in a gradual reduction in strength, ultimately leading to failure. Corrosion has been recognized as the principal cause of pipeline failure incidents. Consequently, the corrosion rate is commonly employed as an evaluation index for pipeline corrosion[2,3]. There are many factors affect the corrosion rates of pipelines, such as the pipeline medium components, temperature, flow rate, pH, and dissolved oxygen and $CO_2$ contents. These factors interact with each other and are interrelated, forming an intricate corrosion system[4]. Consequently, the development of a multifactor, high-dimensional model for accurately predicting the corrosion rate of subsea pipelines will be a focal point of future research.

Advances in computer science have led researchers worldwide to conduct extensive studies on predicting pipeline corrosion rates through machine learning[5–15]. Jin et al. proposed buffer operator theory to develop an enhanced DGM(1,1) model for forecasting pipeline corrosion rates over time, which significantly outperforms the conventional DGM(1,1) model in terms of predictive accuracy[16]. Biezma et al. proposed a fuzzy logic method to predict and analyse the corrosion rates of pipelines, considering six influencing factors. This approach improves both the accuracy and stability of the predictions[17]. Zhang et al. employed a distinctive BP neural network model for predicting pipeline corrosion rates, obtaining results that align more closely with measured values and effectively illustrating the correlation between various factors and the corrosion rate[18]. Nagoor et al. employed an ANN model to predict the service life of a crude oil pipeline, achieving a prediction accuracy of 99.97%[19]. Bo et al. predicted the corrosion rates of pipelines via the PSO-MGM(1,1) model, and its prediction accuracy was 16% higher than that of the MGM(1,1) model[20]. Xiao et al. used the WOA-BP algorithm to predict the corrosion rates of subsea pipelines, and the average absolute percentage error of their predictions was 3.689%, which was much lower than that of the comparison model[1]. Jia et al. used kernel principal component analysis

[1]College of Petroleum Engineering, Xi'an Shiyou University, Xi'an 710065, China. [2]Shaanxi Provincial Key Laboratory of Special Production Enhancement Technology for Oil and Gas Fields, Xi'an Shiyou University, Xi'an 710065, China. [3]PetroChina Changqing Oilfield Branch Company No. 3 Oil Extraction Plant, Dashuikeng Oil Extraction Operation Area, Wuzhong 751506, China. ✉email: 418902514@qq.com

to determine the corrosion rates of subsea pipelines and related factors and established a KPCA-SVM model[21]. This method reduces the interference of low-correlation data, improves the prediction accuracy and reduces the prediction difficulty. Nagoor et al.employed a Bayesian regularization-based neural network framework to predict dry airway lifespan with high accuracy, even when handling datasets containing missing parameters[22]. Xiao et al. predicted the corrosion rates of subsea pipelines via a combined PSO-TSO-BPNN model with an average absolute percent error of 1.8441%[4], which represents a significant improvement in both the accuracy and stability of the model. The modelling methods proposed by the above scholars all have unique advantages but are limited by the optimization algorithms and the neural network's own limitations, which may make them unable to obtain accurate predictions of pipeline corrosion rates for multifactorial and high-dimensional problems.

This paper presents a hybrid model, KPCA-NGO-LSSVM, for predicting the corrosion rates of subsea pipelines, utilizing kernel principal component analysis (KPCA) and Northern Goshawk optimization (NGO) to increase the performance of the least squares support vector machine (LSSVM). Kernel principal component analysis (KPCA) is employed to downscale the data and determine the principal factors influencing the corrosion rates of subsea pipelines, thus reducing the complexity of processing model data and increasing the efficiency of modelling operations. The penalty parameter γ and the kernel parameter $\sigma^2$ are optimized through the NGO algorithm to increase the precision of the prediction model and address the challenges of inconsistent predictions and insufficient generalization capability. Through experimental validation and a comparison of the error metrics, the KPCA-NGO-LSSVM model is shown to outperform existing methods. Specifically, the mean absolute percentage error (MAPE) is reduced to less than 2%, and the root mean square error (RMSE) is significantly lower than those of conventional models. The KPCA-NGO-LSSVM model provides reliable technical support for accurately predicting subsea pipeline corrosion rates. This model provides a scientific basis for optimizing corrosion protection strategies, guiding pipeline maintenance decisions, and ensuring flow safety. Furthermore, the model has significant potential in extending the service life of subsea pipelines and reducing operational and maintenance costs.

## Principles of the NGO algorithm and LSSVM modelling

### Principles of the NGO algorithm

The Northern Goshawk optimization (NGO) algorithm was introduced in 2022 by Mohammad Dehghani and colleagues. The algorithm replicates the Northern Goshawk's behaviour during hunting, focusing on prey recognition, attack, pursuit, and evasion. The Northern Goshawk optimization algorithm divides the hunting process into two phases: prey identification and attack (exploration phase) and chasing and escape (exploitation phase)[23].

*Initialization*

The Northern Goshawk algorithm can be represented by the following matrix for the Northern Goshawk population:

$$X=\begin{bmatrix} X_1 \\ \vdots \\ X_{\mathrm{i}} \\ \vdots \\ X_N \end{bmatrix}_{N\times m}=\begin{bmatrix} x_{1,1} & \cdots & x_{1,j} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{i,j} \\ \vdots & & \vdots \\ \mathrm{x}_{N,1} & \cdots & x_{N,j} \end{bmatrix} \tag{1}$$

.

$X$ is the population matrix of Northern Goshawks; $X_{\mathrm{i}}$ denotes the position of the *i*th Northern Goshawk; $x_{i,j}$ indicates the *j*th-dimensional position of the *i*th Northern Goshawk; *N* is the number of Northern Goshawk populations; and m refers to the number of dimensions in the solution problem.

In the Northern Goshawk optimization algorithm, the objective function of the problem is utilized to compute the objective function value of each Northern Goshawk; the objective function value of the Northern Goshawk population can be represented as a vector of objective function values:

$$F=\begin{bmatrix} F_1 \\ \vdots \\ F_2 \\ \vdots \\ F_N \end{bmatrix}_{N\times 1}=\begin{bmatrix} F\left(X_1\right) \\ \vdots \\ F\left(X_2\right) \\ \vdots \\ F\left(X_N\right) \end{bmatrix}_{N\times 1} \tag{2}$$

.

where *F* is the objective function vector of the Northern Goshawk population and $F_{\mathrm{i}}$ is the objective function value of the *i*th Northern Goshawk population.

*Prey identification and attack (Global search)*

During the initial phase of hunting, the Northern Goshawk selects a prey item at random and attacks it quickly. This phase improves the NGO algorithm's exploration capability by randomizing the selection of prey in the search space. In this phase, a global search of the search space is conducted to determine the optimal region.

During this phase, the Northern Goshawks exhibit the prey selection and attack behaviours described in Eqs. (3)-(5):

$$P_{\mathrm{i}} = X_k,\ \mathrm{i} = 1, 2, 3 \cdots, N;\ \mathrm{j} = 1, 2, 3 \cdots, \mathrm{m};\ k = 1, 2 \cdots\ \text{i-1},\ \mathrm{i}, \mathrm{i}+1, \cdots N \tag{3}$$

$$x_{i,j}^{new,P1} = \{ \begin{matrix} x_{i,j} + r\left(p_{i,j} - I x_{i,j}\right), F_{P_i} < F_i \\ x_{i,j} + r\left(x_{i,j} - p_{i,j}\right), F_{P_i} \geqslant F_i \end{matrix} \tag{4}$$

$$X_i = \{ \begin{matrix} X_{\mathrm{i}}^{new,P1}, F_i^{new,P1} < F_i \\ X_i, F_i^{new,P1} \geqslant F_i \end{matrix} \tag{5}$$

.

Where $P_{\mathrm{i}}$ denotes the location of the *i*th Northern Goshawk's prey; $F_{P_{\mathrm{i}}}$ is the objective function value for the position of the *i*th Northern Goshawk's prey; *k* represents a random integer within a specified range [1,N]; $X_{\mathrm{i}}^{new,P1}$ represents the updated location of the *i*th Northern Goshawk; $x_{i,j}^{new,P1}$ represents the updated position in the *j*th dimension of the *i*th Northern Goshawk; $F_{\mathrm{i}}^{\mathrm{new},P1}$ is the value of the objective function pertaining to the *i*th Northern Goshawk following the update process in phase 1; *r* is a random number within the interval [0, 1]; and *I* denotes a randomly selected integer, either 1 or 2.

*Chase and escape (Localized search)*

After a Northern Goshawk attacks its prey, the prey will attempt to escape capture. Thus, in the concluding phases of hunting, the Northern Goshawk must sustain its chase. The Northern Goshawks' high pursuit speed enables them to chase and ultimately capture prey in nearly any circumstance. The simulation of this behaviour improves the algorithm's capacity for local search within the search space. This hunting activity is presumed to be in proximity to an attack position with a radius *R*. In the subsequent phase, it is described by Eqs. (6)-(8):

$$x_{i,j}^{new,P2} = x_{i,j} + R(2r-1)x_{i,j} \tag{6}$$

$$R = 0.02(1 - \frac{\mathrm{t}}{T}) \tag{7}$$

$$X_{\mathrm{i}} = \{ \begin{matrix} X_i^{new,P2}, F_i^{new,P2} < F_i \\ X_i, F_i^{new,P2} \geqslant F_i \end{matrix} \tag{8}$$

.

Where t represents the current iteration number and *T* denotes the maximum iteration limit; $X_i^{new,P2}$ represents the updated position of the *i*th Northern Goshawk during the second stage; $x_{i,j}^{new,P2}$ represents the updated position of the *j*th dimension of $X_{\mathrm{i}}^{new,P2}$; and $F_{\mathrm{i}}^{new,P2}$ is the value of the objective function pertaining to the *i*th Northern Goshawk following the update process in the second stage.

## LSSVM algorithm

Various machine learning algorithms, including backpropagation neural networks (BP), random forests (RF), and support vector machines (SVM), have been widely used to predict corrosion rates in subsea pipelines. While these methods have demonstrated varying degrees of success, they often face challenges in computational efficiency and model generalizability when dealing with small-to-medium scale datasets characterized by high dimensionality and strong nonlinearity. For instance, BP models typically require substantial computational resources and extensive hyperparameter tuning, while SVM and ensemble methods like RF may encounter overfitting risks in limited-data scenarios.

In contrast, Least Squares Support Vector Machines (LSSVM) show clear advantages in this particular application setting. In our preliminary study (see Figs. 2 and 3; Table 4), LSSVM consistently demonstrated superior performance metrics through systematic comparisons with three representative algorithms (BP, RF, and traditional SVM), with much higher prediction accuracy and stability than the other algorithms, and this improved performance stems from the unique mathematical formulation of the LSSVM, which converts the quadratic optimisation problem into a system of linear equations by means of equal constraints, thus ensuring a global optimisation solution while maintaining the simplicity of the model. ensuring a globally optimised solution while maintaining model simplicity. In addition, its structural risk minimisation principle enhances the generalisation capability, which is particularly important for offshore engineering applications where the cost of in situ data collection is high and the size of the dataset is limited.

As an advanced variant of support vector machines (SVM), LSSVM addresses the original algorithm's computational complexity through innovative problem reformation. Where conventional SVM solves convex quadratic programming problems, LSSVM transforms this into solving linear equations via kernel space mapping and regularization techniques. This fundamental improvement not only accelerates computation but also improves numerical stability, making it particularly suitable for handling the sparse, high-dimensional corrosion datasets typical of subsea pipeline monitoring systems.

The LSSVM is an advanced learning and predictive algorithm derived from the conventional support vector machine (SVM) algorithm. This algorithm streamlines the solution of quadratic optimization problems by converting them into linear Eq. [24].

The steps for using the LSSVM algorithm are as follows:

For a given value from the training dataset $(x_i, y_i)$, $x_i = (x_{i1}, x_{i2}, \cdots, x_{id})^T$ is the d-dimensional input vector, and the output data value is $y_i$; N is the total number of training data values.

(1) To transform the input space into the feature space, a nonlinear function is employed, $\phi(x_{\mathrm{i}})$. The process of estimating the nonlinear function is represented by Eq. (9)[25]:

$$\mathrm{f}(x) = b + \langle \phi(x), \omega \rangle \tag{9}$$

Where $\omega$ is the weight vector, $b$ is the bias term, and $\langle . \rangle$ denotes the inner product operation.

(2) The precise values of parameters $\omega$ and b are determined on the basis of the fundamental principle of risk mitigation:

$$\begin{cases} \min J(\overrightarrow{\omega}, \xi) = \frac{1}{2}\|\overrightarrow{\omega}\|^2 + c\sum\limits_{i=1}^{l}\xi_i^2 \\ s.t. y_i = \phi(x_i)\overrightarrow{\omega} + b + \xi_i \quad i = 1, \cdots, l \end{cases} \tag{10}$$

Where $c$ is the penalty factor and $\xi_i$ is the slack variable.

(3) Introducing the Lagrangian operator $\alpha$ yields the Lagrangian function:

$$L(\overrightarrow{\omega}, b, \xi, \alpha) = \frac{1}{2}\|\overrightarrow{\omega}\|^2 + c\sum_{i=1}^{l}\xi_i^2 - \sum_{i=1}^{l}\alpha_i\left[\overrightarrow{\omega}\phi(x_i) + b + \xi_i - y_i\right] \tag{11}$$

(4) Setting the derivatives of $\overrightarrow{\omega}, b, \ \xi_i, \alpha$ to zero provides the conditions for finding the optimal solution of the problem.

$$\begin{cases} \frac{\partial L}{\partial \omega} = 0 \rightarrow \sum\limits_{i=1}^{l}\alpha_i\phi(x_i) \\ \frac{\partial J}{\partial b} = 0 \rightarrow \sum\limits_{i=1}^{l}\alpha_i = 0 \\ \frac{\partial J}{\partial b} = 0 \rightarrow \alpha_i = c\xi_i \\ \frac{\partial J}{\partial \alpha} = 0 \rightarrow \overrightarrow{\omega}\phi(x_i) + b + \xi_i - y_i = 0 \end{cases} \tag{12}$$

(5) Eliminating the parameters $\overrightarrow{\omega}$ and $\xi_i$ in Eq. (11), we convert Eq. (12) into.

$$\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & K(x_i, y_i) + \frac{1}{c} & \cdots & K(x_i, x_j) \\ \vdots & \vdots & & \vdots \\ i & K(x_i, y_i) & \cdots & K(x_i, x_j) + \frac{1}{c} \end{bmatrix} \begin{bmatrix} b \\ \alpha_i \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} 0 \\ y_i \\ \vdots \\ y_n \end{bmatrix} \rightarrow f(x) = \sum_{i=1}^{l}\alpha_i K(x_i, x_j) + b \tag{13}$$

Where $K(x_i, x_j)$ is the kernel function, expressed as

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\delta^2}\right) \tag{14}$$

A is the parameter of the kernel function. Data prediction can be performed by resolving the unknown data in Eqs. (13)[26,27].

## KPCA-NGO-LSSVM model

### Kernel principal component analysis

High-dimensional feature values can lead to the curse of dimensionality; therefore, to extract the essential features and improve the predictive accuracy while minimizing the model complexity, kernel principal component analysis (KPCA) is applied to reduce the data dimensions[28].

Kernel principal component analysis (KPCA) is a nonlinear dimensionality reduction method that transforms raw data into a high-dimensional feature space by utilizing a kernel function, followed by the application of principal component analysis (PCA) within that feature space.

The principles of the KPCA algorithm are as follows:

(1) The sample set of the original running data $x_k$ is nonlinearly transformed on the basis of the nonlinear kernel function $\Phi$, which maps $x_k$ to a high-dimensional linear feature space. Then, its covariance matrix is computed for the new sample set; i.e.,

$$\overrightarrow{C} = \frac{1}{m}\sum_{j=1}^{m}\overrightarrow{\varphi}(x_j)\overrightarrow{\varphi}(x_j)^T \tag{15}$$

(2) The eigenvalues $\lambda$ and eigenvectors $\overrightarrow{v}$ of matrix $C$ are calculated. The following condition must be satisfied:

$$\lambda\overrightarrow{v}-C\overrightarrow{v}=0 \tag{16}$$

(3) A nonlinear function $\overrightarrow{\varphi}\left(x_i\right)$ is introduced on both sides, and the eigenvectors are represented linearly from $\overrightarrow{v}$ to $\overrightarrow{\varphi}\left(x_i\right)$; i.e.,

$$\overrightarrow{v}=\sum_{i=1}^{m}\alpha_i\overrightarrow{\varphi}\left(x_i\right) \tag{17}$$

(4) The kernel function matrix $\overrightarrow{K}\left(i,j\right)=\left\langle\overrightarrow{\varphi}\left(x_i\right),\overrightarrow{\varphi}\left(x_j\right)\right\rangle$ is defined and transformed:

$$mk\overrightarrow{\alpha}-K\overrightarrow{\alpha}=0 \tag{18}$$

where $\overrightarrow{\alpha}$ is the eigenvector of $K$, the eigenvalue is $mk$, and the subscript $i$ denotes an element in the input dataset.

For any sample, the projection to the principal element $\overrightarrow{\varphi}\left(x\right)$ in the feature space $F$ is[29]:

$$\overrightarrow{v}\overrightarrow{\varphi}\left(x\right)=\sum_{i=1}^{m}\alpha_i\overrightarrow{\varphi}\left(x_i\right)\overrightarrow{\varphi}\left(x_j\right)=\sum_{i=1}^{m}\alpha_i\overrightarrow{K}\left(x_i,x\right) \tag{19}$$

## Predictive modelling

Initially, the KPCA algorithm is employed to reduce the dimensionality of the data, and the NGO algorithm is then applied to optimize the penalty parameter γ and the kernel parameter A of the LSSVM algorithm, thereby yielding a composite corrosion rate prediction model for subsea pipelines, referred to as the KPCA-NGO-LSSVM model. The flowchart of this process is shown in Fig. 1. The NGO-LSSVM and NGO-LSSVM models are established for validation against the integrated KPCA-NGO-LSSVM model.

## Model evaluation indicators

To thoroughly assess the predictive accuracy of the KPCA-NGO-LSSVM corrosion rate model for subsea oil and gas pipelines, the mean absolute percentage error (MAPE), root mean squared error (RMSE) and coefficient of determination ($R^2$) were employed as evaluation metrics:

$$MAPE=\sum{}_{i=1}^{n}\frac{1}{n}\left|\frac{y_i-x_i}{x_i}\right|\times\ 100\% \tag{20}$$

$$RMSE=\sqrt{\frac{1}{n}\sum{}_{i=1}^{n}\left(y_i-x_i\right)^2} \tag{21}$$

$$R^2=\frac{\sum{}_{i=1}^{n}\left(y_i-\bar{x}\right)^2}{\sum{}_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \tag{22}$$

Where $x_i, y_i$ are the true and predicted values of the $i$th sample, respectively, for $i=1,2,\cdots,n$; $n$ is the total number of samples represented; MAPE indicates the model's overall error; and RMSE denotes the deviation of the predicted values from the actual values. A lower MAPE and RMSE indicate greater prediction accuracy and better predictive performance of the model.

# Example analysis

## Dataset segmentation

Three distinct types of pipeline corrosion rate data from the literature were selected for algorithmic prediction. Due to space constraints, the predictive research process is detailed only for data 1, while results for data 2 and data 3 are presented in result form.

The data 1 in this paper are 50 subsea pipelines corrosion data from reference [30]; some of the data values are shown in Tables 1 and  40 of which are chosen as the training set, and the remaining 10 of which are used as the test set for model prediction and error checking.

The data 2 in this paper are 100 overseas oil and gas pipelines corrosion data from reference[31]; some of the data values are shown in Tables 2,  80 of which are chosen as the training set, and the remaining 20 of which are used as the test set for model prediction and error checking.

The data 3 in this paper are 28 subsea multiphase flow pipelines corrosion data from reference[32]; some of the data values are shown in Table 3 and 22 of which are chosen as the training set, and the remaining 6 of which are used as the test set for model prediction and error checking.

## Data preprocessing

In kernel principal component analysis, the kernel function can be used to map the original data to a high-dimensional space, perform nonlinear dimensionality reduction, and mine the nonlinear information in the

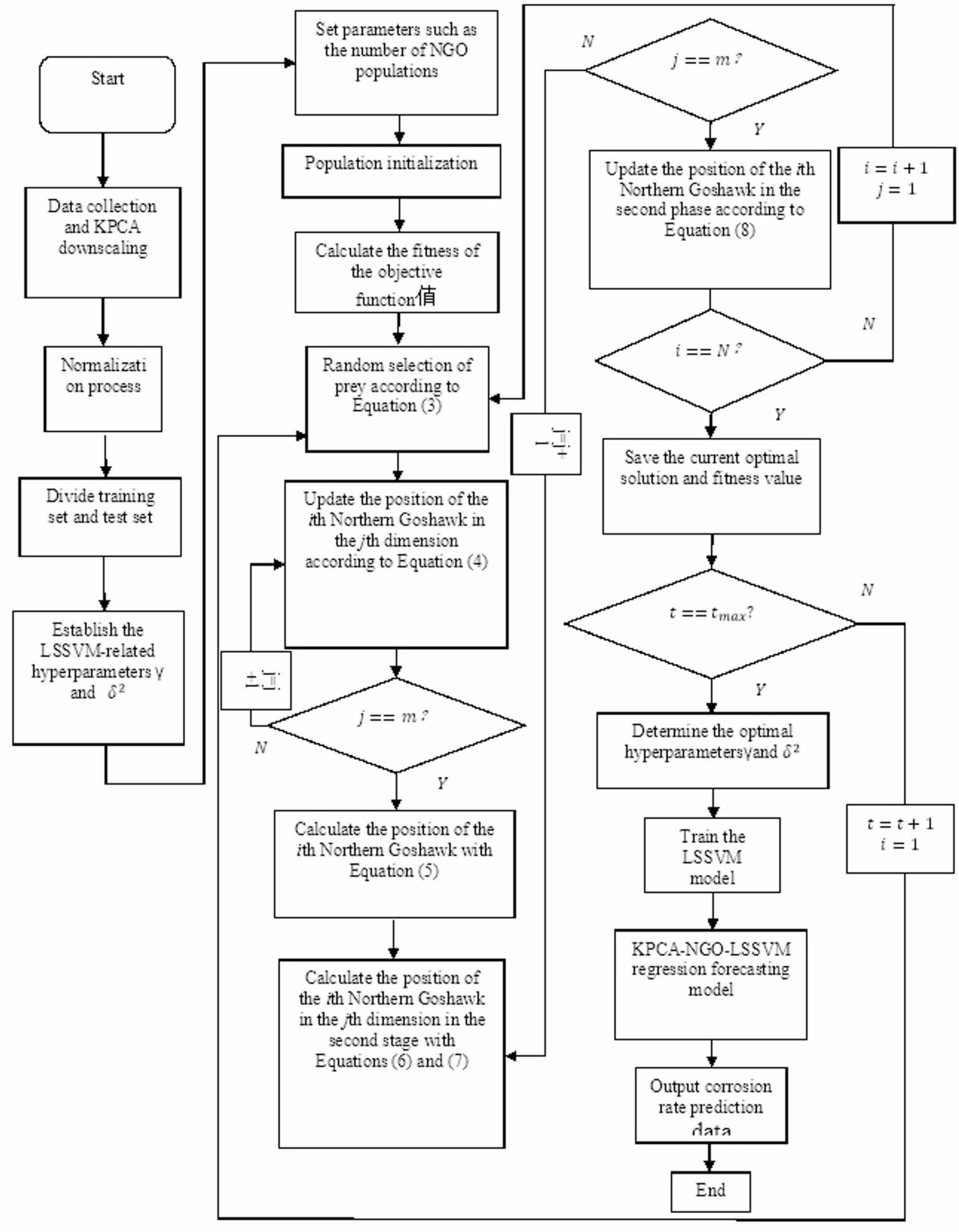
Start
Data collection and KPCA downscaling
Normalization process
Divide training set and test set
Establish the LSSVM-related hyperparameters γ and $\delta^2$
Set parameters such as the number of NGO populations
Population initialization
Calculate the fitness of the objective function值
Random selection of prey according to Equation (3)
Update the position of the ith Northern Goshawk in the jth dimension according to Equation (4)
$j == m$?
N
Y
Calculate the position of the ith Northern Goshawk with Equation (5)
Calculate the position of the ith Northern Goshawk in the jth dimension in the second stage with Equations (6) and (7)
$j == m$?
N
Y
Update the position of the ith Northern Goshawk in the second phase according to Equation (8)
$i = i + 1$
$j = 1$
$i == N$?
N
Y
Save the current optimal solution and fitness value
$t == t_{max}$?
N
Y
$t = t + 1$
$i = 1$
Determine the optimal hyperparametersγand $\delta^2$
Train the LSSVM model
KPCA-NGO-LSSVM regression forecasting model
Output corrosion rate prediction data
End

**Fig. 1**. Flowchart of the KPCA-NGO-LSSVM model.

data.[21] Therefore, nine influencing factors in subsea pipeline corrosion rate for data 1 were downscaled using KPCA. The magnitudes of the variance contributions of the nine principal components were obtained in MATLAB 2020a, as shown in Table 4.

The magnitudes of the eigenvalues and the cumulative contributions reflect the magnitudes of the influence of the principal components, as shown in Table 4. In this work, the first six principal components $F_1$, $F_2$, $F_3$, $F_4$, $F_5$, and $F_6$, with cumulative contributions greater than 85% were extracted.

| Serial Number | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_0$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 65.9 | 2.27 | 0.0317 | 5.1 | 0.64 | 7560 | 26.8 | 58.5 | 141 | 2.8385 |
| 2 | 67.7 | 2.41 | 0.0326 | 4.5 | 0.628 | 8020 | 19.8 | 62.3 | 150 | 2.5997 |
| 3 | 61.3 | 2.66 | 0.0331 | 6.1 | 0.639 | 6340 | 18.1 | 62.1 | 155 | 2.954 |
| 4 | 65.8 | 2.3 | 0.0329 | 6 | 0.443 | 6800 | 21.1 | 57.1 | 153 | 2.9615 |
| 5 | 41.7 | 2.39 | 0.0337 | 6.25 | 0.485 | 3560 | 18.3 | 63.9 | 161 | 2.5323 |
| 6 | 54.3 | 2.5 | 0.0331 | 6.1 | 0.625 | 3420 | 27.9 | 57.8 | 155 | 2.675 |
| 7 | 58.5 | 2.78 | 0.0329 | 6.01 | 0.531 | 6900 | 27 | 57.3 | 153 | 3.0823 |
| 8 | 60.7 | 2.41 | 0.0326 | 4.5 | 0.5 | 8020 | 25.6 | 63 | 150 | 2.6291 |
| 9 | 47.5 | 2.5 | 0.034 | 6.1 | 0.605 | 5700 | 18.6 | 55.3 | 154 | 2.2865 |
| 10 | 52 | 2.74 | 0.0344 | 5.95 | 0.534 | 6260 | 19.4 | 52 | 167 | 2.2647 |

**Table 1**. Corrosion rate data for selected subsea pipelines. Note: $X_1$ is the temperature (°C), $X_2$ is the system pressure (MPa), $X_3$ is the partial pressure of $CO_2$ (MPa), $X_4$ is the pH, $X_5$ is the medium flow rate ($m\cdot s^{-1}$), $X_6$ is the $Cl^-$ concentration ($mg\cdot L^{-1}$), $X_7$ is the $CO_2$ concentration ($mg\cdot L^{-1}$), $X_8$ is the $HCO_3^-$ concentration ($mg\cdot L^{-1}$), $X_9$ is the water content (%), and $X_0$ is the internal corrosion rate value ($mm.a^{-1}$).

| Serial Number | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | Y |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 109.03 | 6.5100 | 0.0717 | 1.1824 | 35.887 | 0.7636 | 0.9033 | 0.1050 | 0.1732 |
| 2 | 115.12 | 6.3578 | 0.0370 | 1.6696 | 37.932 | 0.7561 | 0.8283 | 0.1240 | 0.1425 |
| 3 | 115.05 | 6.2169 | 0.0838 | 3.2597 | 37.306 | 0.7701 | 0.8673 | 0.1102 | 0.1448 |
| 4 | 110.56 | 6.3860 | 0.1065 | 1.4988 | 36.520 | 0.7601 | 0.8581 | 0.1127 | 0.1620 |
| 5 | 110.66 | 6.5301 | 0.0535 | 2.5883 | 37.896 | 0.7622 | 0.8601 | 0.1232 | 0.1540 |
| 6 | 111.61 | 6.7053 | 0.0469 | 2.3399 | 37.767 | 0.7533 | 0.8595 | 0.1227 | 0.1518 |
| 7 | 114.09 | 6.5905 | 0.0565 | 4.8084 | 37.097 | 0.7712 | 0.8809 | 0.1149 | 0.1452 |
| 8 | 111.98 | 6.6371 | 0.0581 | 3.4667 | 37.389 | 0.7535 | 0.8785 | 0.1039 | 0.1488 |
| 9 | 109.94 | 6.6036 | 0.0153 | 0.3675 | 36.282 | 0.7551 | 0.8926 | 0.1253 | 0.1710 |
| 10 | 113.18 | 6.3849 | 0.0608 | 4.3186 | 37.838 | 0.7645 | 0.8497 | 0.1150 | 0.1418 |
| 11 | 110.97 | 6.3393 | 0.0507 | 4.0981 | 37.618 | 0.7733 | 0.8620 | 0.1144 | 0.1517 |
| 12 | 119.79 | 6.5631 | 0.0277 | 7.5143 | 39.429 | 0.7639 | 0.8574 | 0.1077 | 0.1113 |
| 13 | 120.69 | 6.7422 | 0.0089 | 3.8390 | 36.699 | 0.7526 | 0.8660 | 0.1222 | 0.1288 |
| 14 | 109.38 | 6.9032 | 0.1080 | 4.4791 | 37.205 | 0.7813 | 0.8182 | 0.1272 | 0.1573 |
| 15 | 117.57 | 6.8163 | 0.1083 | 1.8707 | 34.693 | 0.7629 | 0.8999 | 0.1046 | 0.1574 |
| 16 | 116.00 | 6.5560 | 0.0337 | 5.5114 | 36.578 | 0.7557 | 0.8513 | 0.1117 | 0.1337 |
| 17 | 120.04 | 6.8136 | 0.0352 | 1.1180 | 36.513 | 0.7538 | 0.8814 | 0.1158 | 0.1437 |
| 18 | 117.17 | 6.0568 | 0.0052 | 5.7569 | 36.708 | 0.7614 | 0.9235 | 0.1071 | 0.1339 |
| 19 | 115.69 | 6.7876 | 0.0302 | 3.3667 | 36.601 | 0.7789 | 0.9146 | 0.1153 | 0.1537 |
| 20 | 117.61 | 6.2195 | 0.0345 | 4.9774 | 36.867 | 0.7591 | 0.8788 | 0.1296 | 0.1322 |

**Table 2**. Corrosion rate data for selected overseas oil and gas pipelines. Note: $X_1$ is the operating temperature (°C), $X_2$ is the pH, $X_3$ is the $O_2$ concentration ($mg\cdot L^{-1}$), $X_4$ is the $CO_2$ concentration ($mg\cdot L^{-1}$), $X_5$ is the S concentration ($mg\cdot L^{-1}$), $X_6$ is the $N_2$ concentration ($mg\cdot L^{-1}$), $X_7$ is the operating pressure (MPa), and $Y$ is the process pipeline corrosion rate value ($mm.a^{-1}$).

The eigenvectors of the first six principal components selected in this paper are shown in Table 5. The eigenvector of each principal component indicates each factor's explanatory ability, and the closer the absolute value is to one, the stronger its explanatory ability is, implying that the factor has a greater influence on subsea pipeline corrosion. As shown in Table 5, $F_1$ has a greater correlation with system pressure, $F_2$ with a medium flow rate, $F_3$ with pH, $F_4$ with water content, $F_5$ with temperature, and $F_6$ with $CO_2$ partial pressure.

Finally, the system pressure, water content, medium flow rate, pH, temperature, and $CO_2$ partial pressure all have greater impacts on subsea pipeline corrosion rates for data 1 than the other factors. The above subsea pipeline corrosion factors are substituted into the combined model for the next prediction.

## Analysis of the forecast results

Based on the characteristics of small sample size and high-dimensional features in the dataset, we initially selected four machine learning algorithms—BP, LSSVM, SVM and RF—for comparative analysis. The results (Figs. 2 and 3; Tables 6and 7) demonstrate that the LSSVM model outperforms the RF, BP and SVM models in

| Serial Number | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_0$ |
|---|---|---|---|---|---|---|---|
| 1 | 55.64 | 2515.34 | 0.92 | 251.53 | 0.53 | 4.88 | 2.79 |
| 2 | 55.09 | 2514.42 | 0.92 | 251.44 | 0.53 | 4.88 | 3.05 |
| 3 | 54.55 | 2514.42 | 0.92 | 251.35 | 0.53 | 4.88 | 3.05 |
| 4 | 54.02 | 2513.51 | 0.92 | 251.26 | 0.53 | 4.88 | 3.11 |
| 5 | 53.5 | 2512.6 | 0.92 | 251.17 | 0.53 | 4.89 | 2.6 |
| 6 | 52.99 | 2511.68 | 0.93 | 251.08 | 0.53 | 4.89 | 3.11 |

**Table 3**. Corrosion rate data for selected subsea multiphase flow pipelines. Note: $X_1$ is the temperature (°C), $X_2$ is the pressure (kPa), $X_3$ is the liquid-holding capacity, $X_4$ is the $CO_2$ partial pressure(MPa), $X_5$ is the liquid flow rate ($m\cdot s^{-1}$), $X_6$ is the pH, $X_7$ is the operating pressure (MPa), and $X_0$ is the internal corrosion rate value ($mm.y^{-1}$).

| Principal component | Variance contribution (%) | Cumulative contribution (%) |
|---|---|---|
| $F_1$ | 34.52 | 34.52 |
| $F_2$ | 14.83 | 49.35 |
| $F_3$ | 13.22 | 62.57 |
| $F_4$ | 10.30 | 72.87 |
| $F_5$ | 8.56 | 81.43 |
| $F_6$ | 5.29 | 86.72 |
| $F_7$ | 5.06 | 91.78 |
| $F_8$ | 4.71 | 96.49 |
| $F_9$ | 3.51 | 100.00 |

**Table 4**. Analysis of variance contribution ratios of nine principal components of the corrosion factors of a subsea pipeline.

| Serial Number | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ |
|---|---|---|---|---|---|---|---|---|---|
| $F_1$ | 0.18505 | 0.49258 | 0.17516 | −0.07907 | 0.07524 | −0.07028 | 0.19126 | 0.07929 | 0.08937 |
| $F_2$ | −0.05535 | 0.05948 | −0.03288 | −0.09702 | 0.39367 | 0.05285 | −0.23666 | −0.04639 | 0.11301 |
| $F_3$ | −0.11895 | −0.04651 | −0.03637 | 0.30246 | −0.13388 | −0.19042 | 0.02773 | 0.03341 | 0.00385 |
| $F_4$ | 0.08919 | 0.02214 | −0.01423 | 0.06840 | −0.14587 | −0.05581 | −0.06286 | 0.048021 | 0.38108 |
| $F_5$ | 0.49622 | 0.06190 | 0.14333 | 0.20628 | 0.03046 | 0.13325 | 0.16800 | 0.079734 | 0.10529 |
| $F_6$ | −0.08606 | −0.05097 | 0.63598 | 0.00302 | 0.03733 | −0.04974 | −0.02771 | 0.0135691 | −0.02778 |

**Table 5**. Eigenvectors of each factor of the first 7 principal components of the corrosion factors of a subsea pipeline.

terms of prediction accuracy and stability. Therefore, we adopted the LSSVM algorithm as the predictive model for estimating the corrosion rate of submarine pipelines.

Four optimized portfolio models based on the LSSVM algorithm were subsequently developed and compared. The results indicate that the predictive outcomes of the integrated KPCA-NGO-LSSVM model and three other combined models after training are shown in Figs. 4 and 5, and Table 8. Figure 4 shows that the KPCA-NGO-LSSVM model yields superior predictions and stability, followed by the NGO-LSSVM model, whereas the KPCA-PSO-LSSVM model and PSO-LSSVM model yield the worst predictions and least stability.

As shown in Figs. 4 and 5, and Table 8, the stability of the predicted values of the KPCA-PSO-LSSVM and PSO-LSSVM models is low overall, with maximum relative errors of 5.42% and 5.81% and average relative errors of 3.59% and 4.27%, respectively. The predicted values of the NGO-LSSVM model exhibit good stability, with a maximum relative error of 5.10% and an average relative error of 2.49%. The KPCA-NGO-LSSVM model exhibits superior performance, demonstrating optimal stability in terms of the predicted values, with a maximum relative error of 4.80% and an average relative error of 1.80%, both of which are lower than those of the other models, indicating the most effective predictions.

Similarly, prediction studies were also performed for Data 2 and Data 3 using these four algorithms. The comparative indicators of their prediction results are presented in Tables 10 and 11.

From Tables 9 and 10, and 11, along with the prediction results across multiple datasets, the KPCA-NGO-LSSVM algorithm exhibits optimal stability in predicted values and significantly outperforms the NGO-LSSVM, KPCA-PSO-LSSVM, and PSO-LSSVM algorithms in prediction accuracy.

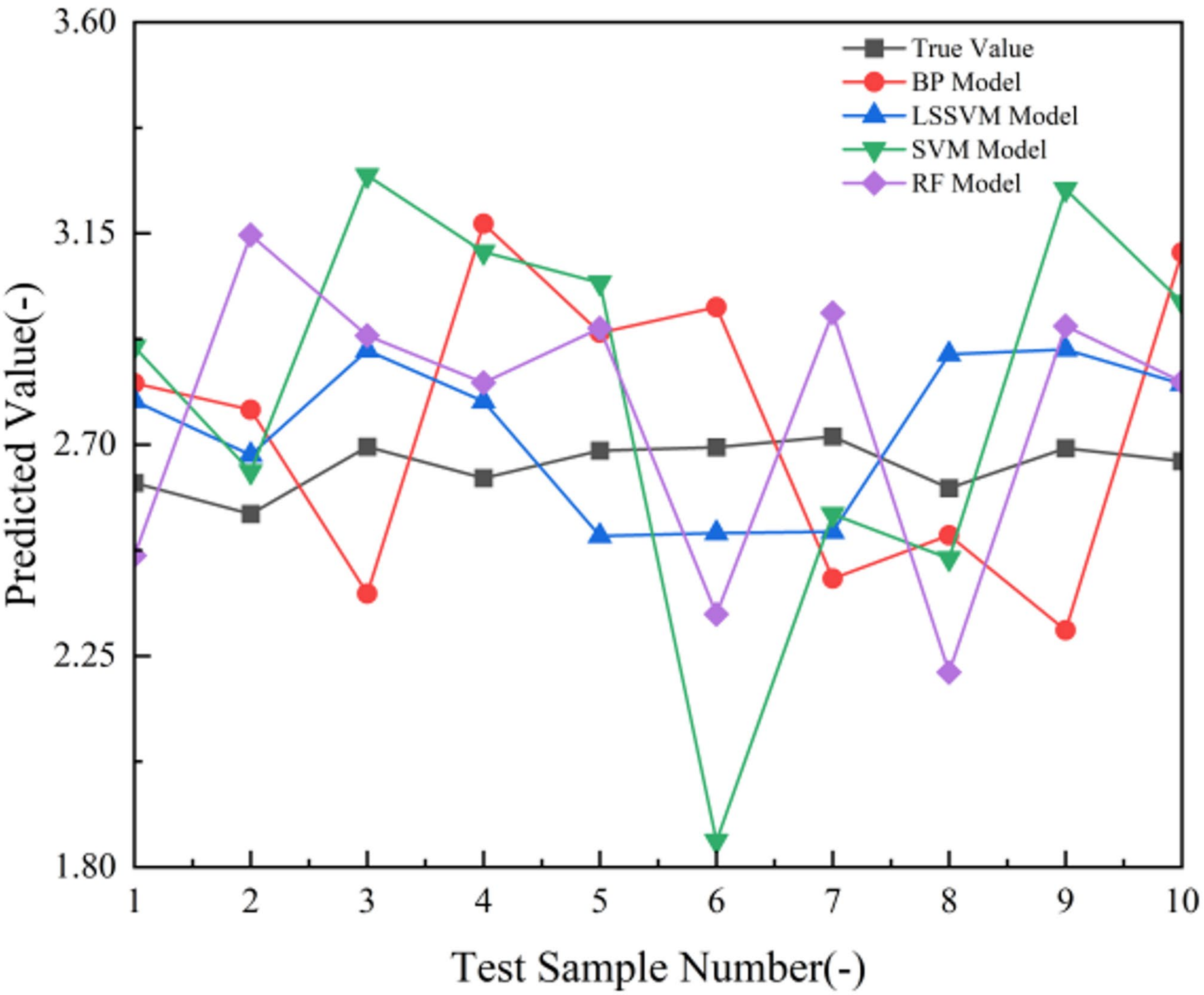

**Fig. 2**. Comparison of the predicted and real values for single models.

## Conclusion

This paper presents the fundamental principles of the KPCA, NGO, and LSSVM algorithms and establishes a composite model for predicting the corrosion rates of subsea pipelines utilizing the KPCA-NGO-LSSVM approach. The following conclusions are derived from the validation and error analysis of the corrosion rate data pertaining to subsea pipelines:

(1) The KPCA algorithm was utilized for data dimensionality reduction to obtain the six factors that have the greatest influence on the corrosion rates of subsea pipelines, i.e., system pressure, water content, pH, temperature, and $CO_2$ partial pressure. The multiple correlations between the influencing factors were eliminated, the complexity of the data was reduced, and the efficiency of the modelling operation was improved.

(2) Based on the data characteristics, four algorithms—BP, SVM, LSSVM, and RF—were selected and compared. The LSSVM model has a MAPE of 7.1398%, an RMSE of 0.1939 and an $R^2$ of 0.8047, which indicated that the LSSVM algorithm demonstrated significantly better predictive performance and stability than the other algorithms. Therefore, LSSVM was adopted as the predictive model for estimating the corrosion rate of subsea pipelines.

(3) Based on prediction results from three distinct datasets, the combined KPCA-NGO-LSSVM model demonstrates significantly superior prediction accuracy and stability compared to the other three models. These results demonstrate that the combined KPCA-NGO-LSSVM model achieves higher prediction accuracy and superior stability for subsea pipeline corrosion rate prediction. This model provides robust technical support for accurately predicting corrosion rates, offering significant potential for extending the service life of subsea pipelines and reducing operational and maintenance costs.

(4) The predictive accuracy and stability of the algorithmic model improve with larger data samples. Consequently, a comprehensive pipeline corrosion database could be developed to derive a corrosion rate prediction model with broader applicability and enhanced efficacy.

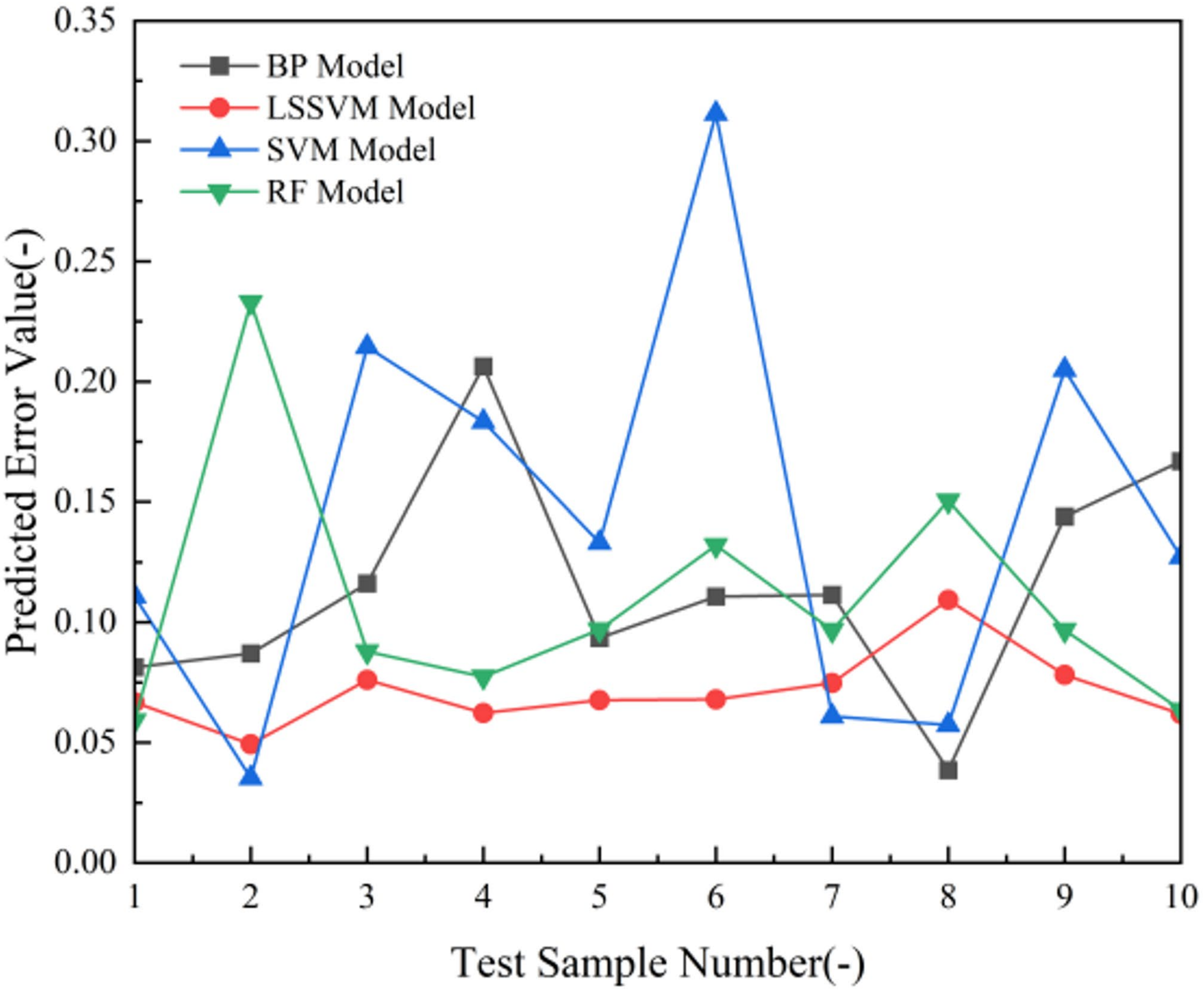

**Fig. 3**. Comparison of the relative errors for single models.

| | | BP | | LSSVM | | SVM | | RF | |
|---|---|---|---|---|---|---|---|---|---|
| Sample number | Real value | Projected value/ (mm.a$^{-1}$) | Relative error/ (%) | Projected value/ (mm.a$^{-1}$) | Relative error/ (%) | Projected value/ (mm.a$^{-1}$) | Relative error/ (%) | Projected value/ (mm.a$^{-1}$) | Relative error/ (%) |
| 1 | 2.6183 | 2.8311 | 8.13 | 2.7931 | 6.68 | 2.9084 | 11.08 | 2.4635 | 5.91 |
| 2 | 2.5523 | 2.7746 | 8.71 | 2.6783 | 4.94 | 2.6423 | 3.53 | 3.1467 | 23.29 |
| 3 | 2.6956 | 2.3825 | 11.62 | 2.9007 | 7.61 | 3.2736 | 21.44 | 2.9323 | 8.78 |
| 4 | 2.6285 | 3.1707 | 20.63 | 2.7922 | 6.23 | 3.1103 | 18.33 | 2.8319 | 7.74 |
| 5 | 2.6872 | 2.9384 | 9.35 | 2.5053 | 6.77 | 3.0450 | 13.32 | 2.9478 | 9.70 |
| 6 | 2.6945 | 2.9928 | 11.07 | 2.5115 | 6.79 | 1.8556 | 31.13 | 2.3388 | 13.20 |
| 7 | 2.7175 | 2.4148 | 11.14 | 2.5144 | 7.47 | 2.5520 | 6.09 | 2.9805 | 9.68 |
| 8 | 2.6074 | 2.5069 | 3.85 | 2.89251 | 10.93 | 2.4578 | 5.74 | 2.2151 | 15.05 |
| 9 | 2.6924 | 2.3049 | 14.39 | 2.9028 | 7.81 | 3.2441 | 20.49 | 2.9528 | 9.67 |
| 10 | 2.6647 | 3.1093 | 16.68 | 2.8296 | 6.19 | 3.0036 | 12.72 | 2.8341 | 6.36 |

**Table 6**. Statistics of the predicted values and relative errors for single models.

| Model | MAPE/(%) | RMSE | $R^2$ |
|---|---|---|---|
| BP | 11.5568 | 0.3299 | 0.6652 |
| LSSVM | 7.1398 | 0.1939 | 0.8047 |
| SVM | 14.3837 | 0.4426 | 0.5779 |
| RF | 10.9351 | 0.3144 | 0.6921 |

**Table 7**. Comparison of the evaluation indicators for single models prediction results.

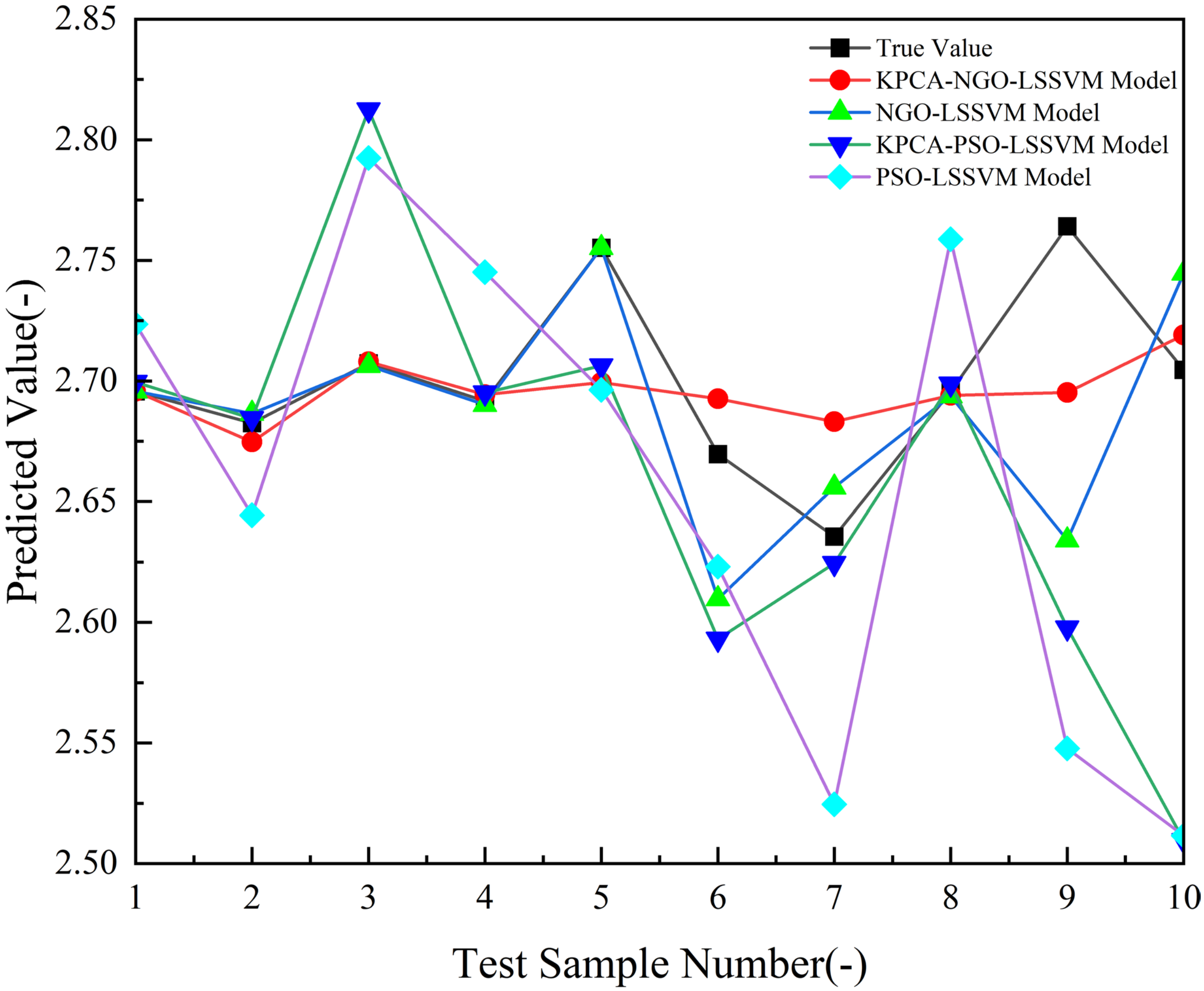

**Fig. 4**. Comparison of the predicted and real values for combined models.

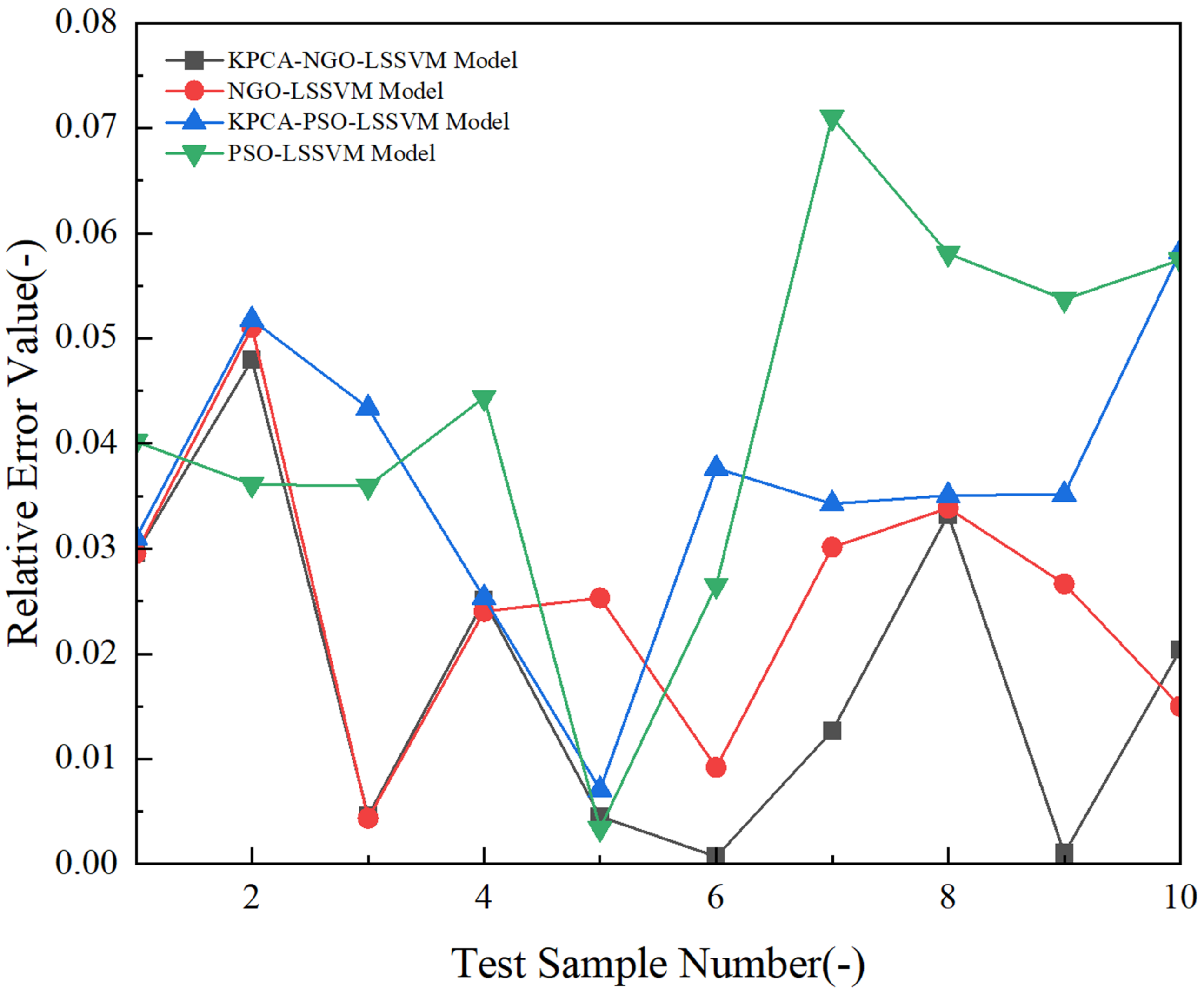

**Fig. 5**. Comparison of the relative errors for combined models.

| Sample Number | Real Value | KPCA-NGO-LSSVM | | NGO-LSSVM | | KPCA-PSO-LSSVM | | PSO-LSSVM | |
|---|---|---|---|---|---|---|---|---|---|
| | | Projected Value/ (mm.a$^{-1}$) | Relative Error/ (%) | Projected Value/ (mm.a$^{-1}$) | Relative Error/ (%) | Projected Value/ (mm.a$^{-1}$) | Relative Error/ (%) | Projected Value/ (mm.a$^{-1}$) | Relative Error/ (%) |
| 1 | 2.6183 | 2.6957 | 2.96 | 2.6957 | 2.95 | 2.6995 | 3.10 | 2.7235 | 4.02 |
| 2 | 2.5523 | 2.6747 | 4.80 | 2.6825 | 5.10 | 2.6844 | 5.18 | 2.6444 | 3.61 |
| 3 | 2.6956 | 2.7079 | 0.46 | 2.7073 | 0.43 | 2.8125 | 4.34 | 2.7925 | 3.59 |
| 4 | 2.6285 | 2.6944 | 2.51 | 2.6916 | 2.40 | 2.6951 | 2.53 | 2.7451 | 4.44 |
| 5 | 2.6872 | 2.6994 | 0.45 | 2.7552 | 2.53 | 2.7063 | 0.71 | 2.6963 | 0.34 |
| 6 | 2.6945 | 2.6926 | 0.07 | 2.6697 | 0.92 | 2.5931 | 3.76 | 2.6231 | 2.65 |
| 7 | 2.7175 | 2.6831 | 1.27 | 2.6356 | 3.01 | 2.6245 | 3.42 | 2.5245 | 7.10 |
| 8 | 2.6074 | 2.6941 | 3.32 | 2.6956 | 3.38 | 2.6988 | 3.51 | 2.7588 | 5.81 |
| 9 | 2.6924 | 2.6952 | 0.10 | 2.7641 | 2.66 | 2.5978 | 3.51 | 2.5478 | 5.37 |
| 10 | 2.6647 | 2.7191 | 2.04 | 2.7046 | 1.50 | 2.5203 | 5.42 | 2.5116 | 5.75 |

**Table 8**. Statistics of the predicted values and relative errors for combined models.

| Model | MAPE/(%) | RMSE | $R^2$ |
|---|---|---|---|
| KPCA-NGO-LSSVM | 1.7971 | 0.06105 | 0.9922 |
| NGO-LSSVM | 2.4903 | 0.07317 | 0.9734 |
| KPCA-PSO-LSSVM | 3.5887 | 0.10140 | 0.9381 |
| PSO-LSSVM | 4.2671 | 0.12339 | 0.9232 |

**Table 9**. Comparison of evaluation indicators for the prediction results of combined models based on data 1.

| Model | MAPE/(%) | RMSE | $R^2$ |
|---|---|---|---|
| KPCA-NGO-LSSVM | 2.5410 | 0.0621 | 0.9873 |
| NGO-LSSVM | 3.7629 | 0.0793 | 0.9434 |
| KPCA-PSO-LSSVM | 4.0974 | 0.0817 | 0.9217 |
| PSO-LSSVM | 5.2731 | 0.1003 | 0.9081 |

**Table 10**. Comparison of evaluation indicators for the predictive results of combined models based on data 2.

| Model | MAPE/(%) | RMSE | $R^2$ |
|---|---|---|---|
| KPCA-NGO-LSSVM | 3.1045 | 0.1035 | 0.9402 |
| NGO-LSSVM | 3.4588 | 0.1289 | 0.9310 |
| KPCA-PSO-LSSVM | 4.3479 | 0.1566 | 0.9152 |
| PSO-LSSVM | 4.7861 | 0.1640 | 0.9065 |

**Table 11**. Comparison of evaluation indicators for the predictive results of combined models based on data 3.

## Data availability

The authors declare that the data supporting the findings of this study are available within the paper (Data 1 from paper [30]: https://doi.org/10.27393/d.cnki.gxazu.2020.000585), (Data 2 from paper [31]: https://doi.org/10.27643/d.cnki.gsybu.2023.000192), (Data 3 from paper [32]: https://doi.org/10.26995/d.cnki.gdqsc.2020.000788).

Received: 12 November 2024; Accepted: 30 June 2025
Published online: 08 July 2025

## References

1. Xiao,R;Jin,S. Corrosion rate prediction of submarine pipelines based on WOA-BP algorithm. *J. Marine Science*,**46**(06):116–123(2022).
2. MAHMOODIAN M,LI CQ. Structural integrity of corrosion-affected cast iron water pipes using a reliability-based stochastic analysis method. *J. Structure and Infrastructure Engineering*,**12**(10):1356–1363(2016).
3. Cui,M. Research on corrosion and residual strength within $CO_2$ of multi-phase flow sea pipe. *D. China University of Petroleum(East China)*,2014.
4. Xiao,R;Liu,G;Liu,B;et al. Research on corrosion rate prediction in submarine pipelines based on PCA-TSO-BPNN model. *J/OL. Thermal Processing Technology*,1–7(2024).
5. Lu,P;Wang,X;Yang,W;et al. Improved reptile search algorithm to optimize ENN model to predict pipeline corrosion rate. *J. Science Technology and Engineering*, **23**(30):12942–12950(2023).
6. Xiao,S;Du,C;Wang,C. Improved sparrow search algorithm to optimize BP neural network pipeline corrosion rate prediction model. *J. Oil and Gas Storage and Transportation*, **43**(7):760–768, 795(2024).
7. Ma,M;Zhao,Z. Corrosion rate prediction of process pipelines based on KPCA-CSO-RVM model. *J. Safety and Environmental Engineering*, **28**(04):1–7 + 20(2021).
8. Lv,L;Wang,J;Qi,Q;et al. Corrosion rate prediction model for oil and gas mixed transport pipelines based on KPCA-IGOA-ELM. *J. Oil and Gas Storage and Transportation*, **42**(7):785–792(2023).
9. Xiao,B;Zhang,H;Liu,H. Application of improved PSO-BPNN algorithm in pipeline corrosion prediction. *J. Journal of Zhengzhou University (Engineering Edition)*, **43**(01):27–33(2022).
10. Jin,L;Zeng,D;Meng,K;et al. Research on corrosion prediction model of submarine pipeline based on GWO-LSSVM algorithm. *J. Oil and Gas Chemical Industry*, **51**(02):70–76(2022).
11. Li,S;Du,H;Cui,Q;Liu,P;Ma,X;Wang,H. Pipeline Corrosion Prediction Using the Grey Model and Artificial Bee Colony Algorithm. *J. Axioms* ,**11**,289(2022).
12. Zhou,Y; Peng,X;Geng,Y;et al. Internal corrosion rate prediction of shale gas gathering pipeline based on KPCA-GA-BP model. *J. Corrosion and Protection*, **45**(04):63–68(2024).
13. Huang,G;Zhou,Y;Yan S;et al. Prediction of external corrosion rate of buried pipelines based on KPCA-CS-SVM. *J. Thermal Processing Technology*,**51**(16):38–43(2022).
14. Chang.E. Research on the prediction of external corrosion rate of marine oil and gas pipelines based on GM(1,1) and ELM. *D. Xi'an University of Architecture and Technology*,2022.
15. Ling,X;Xu,L;Gao,J;et al. Prediction of external corrosion rate of long-distance pipeline based on IFA-BPNN. *J. Surface Technology*,**50**(04):285–293(2021).
16. Jin,W;Yao,S;,Jin,Y;et al. Improved DGM(1,1) modeling and validation of pipeline corrosion rate over time. *J. Journal of Safety and Environment*,**22**(01):77–83(2022).

17. Biezma M V, Agudo D, Barron G.A fuzzy logic method: Predicting pipeline external corrosion rate. *J .International Journal of Pressure Vessels and Piping*,**163**:55–62(2018).
18. Zhang,Y;Yang,J;. Prediction of pipeline corrosion rate based on BP neural network. *J. Total Corrosion Control*,**27**(9):67–71(2013).
19. Shaik, N.B., Pedapati, S.R., Othman, A.R. et al. An intelligent model to predict the life condition of crude oil pipelines using artificial neural networks. *Neural Comput & Applic* **33**, 14771–14792 (2021).
20. Bo,T;Yu,H;Song,W;. Construction of pipeline corrosion prediction model with improved MGM(1,1). *J. Mechanical Design and Manufacturing Engineering*,**51**(04):69–73(2022).
21. Jia,H;Hu,L;Li,X;et al. Corrosion risk prediction in submarine pipelines based on kernel principal component analysis algorithm. *J. Corrosion and Protection*,**44**(03):82–87(2023).
22. Shaik, N.B., Jongkittinarukorn, K., Benjapolakul, W. et al. A novel neural network-based framework to estimate oil and gas pipelines life with missing input parameters. *Sci Rep* **14**, 4511 (2024).
23. Fu,X;Zhu,L;Huang,J;et al. Multi-threshold image segmentation based on improved northern hawk optimization algorithm. *J. Computer Engineering*,**49**(07):232–241(2023).
24. Zhang,Jia;Li,L;Wang,H;et al. Research on the prediction of pipeline corrosion residual strength based on IWOA-LSSVM. *J. Mechanical Strength*,**46**(02):468–475(2024).
25. XUE X H,XIAO M. Deformation evaluation on surrounding rocks of underground caverns based on PSO-LSSVM. *J .Tunnelling and Underground Space Technology*,**69**:171–181(2017).
26. SHAYEGHIH,GHASEMI A,MORADZADEH M, et al. Day-ahead electricity price forecasting using WPT,GMI and modified LSSVM-based S-OLABC algorithm. *J .Soft Computing-A Fusion of Foundations Methodologies and Appications*,**21**(2):525–541(2017).
27. YARVEICY H,MOGHADDAM A K,GHIASI M, Practical use of statistical learning theory for modeling freezing point depression of electrolyte solutions: LSSVM model. *J .Journal of Natural Gas Science and Engineering*,**20**:414–421(2014).
28. Song,C;Zhang,X. A fast PCA-based algorithm for solving large systems of hyperdeterministic linear equations and applications. *J. Intelligent Computer and Applications*,**9**(4):91–95(2019).
29. GORJAEI R G,SONGOLZDEH R,TORKAMAN M, et al. A novel PSO-LSSVM model for predicting liquid rate of two phase flow through wellhead choke. *J .Journal of Natural Gas Science and Engineering*,**24**:228–237(2015).
30. Ying,S. Research on corrosion rate prediction in in-service submarine oil and gas pipelines. *D. Xi'an University of Architecture and Technology*,2020.
31. Ya.Z. Research on corrosion rate prediction model of overseas oil and gas pipelines *.D. China University of Petroleum (Beijing)*, 2023.
32. Bin.L. Research on corrosion rate prediction of a submarine multiphase flow pipeline based on artificial intelligence *.D. Northeast Petroleum University*, 2020.

## Acknowledgements

The authors wish to thank the Shaanxi Provincial Natural Science Foundation (2023-JC-YB-421) and the Xi'an Shiyou University Graduate Innovation Fund Program (YCX2413059) for financial support.

## Author contributions

X.S: Reviewed and edited the manuscript, Conceptualization, Methodology, Supervision. H.C: Wrote the main manuscript, Analysis and investigate data, Supervision. Z.H: Did experiment data, Projected administration, Supervision.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to S.X.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.